Київський національний лінгвістичний університет

Анотований науково-допоміжний бібліографічний покажчик статей журналу *Computational Linguistics*

(2000-2017 рр.)

Київ – 2017

Видавничий центр КНЛУ

УДК 81'32

ББК 91.9:81.1
А 69

Рекомендовано до поширення через мережу Інтернет вченою радою Київського національного лінгвістичного університету (протокол № 16 від 18 квітня 2017 року)

Рецензенти:
доктор філологічних наук, професор кафедри української мови та прикладної лінгвістики Інституту філології Київського національного університету імені Тараса Шевченка Н. П. Дарчук,
завідувач відділу лексикології, лексикографії та структурно-математичної лінгвістики Інституту української мови НАН України, доктор філологічних наук, професор Є. А. Карпіловська

Укладачі:
Коломієць, В. О. кандидат психологічних наук, доцент кафедри германської і фіно-угорської філології КНЛУ (керівництво колективом укладачів, загальна редакція покажчика, передмова, переклад анотацій); Драчова, М. О. магістрантка спеціальності "Прикладна лінгвістика" КНЛУ (переклад анотацій і редагування студентських перекладів); Дубок, М. Ю. магістрант спеціальності "Прикладна лінгвістика" КНЛУ (переклад анотацій); Мартинюк, О. М. магістрантка спеціальності "Прикладна лінгвістика" КНЛУ (переклад анотацій, укладання бази даних покажчика, бібліографічний опис статей); Павлущенко, Т. Р. магістрантка спеціальності "Прикладна лінгвістика" КНЛУ (переклад анотацій); Погорєлов, К. С. магістрант спеціальності "Прикладна лінгвістика" КНЛУ (переклад анотацій, бібліографічний опис статей); Погребна, М. В. магістрантка спеціальності "Прикладна лінгвістика" КНЛУ (переклад анотацій і редагування студентських перекладів); Попова, Д. М. магістрантка спеціальності "Прикладна лінгвістика" КНЛУ (переклад анотацій); Синящик, А. В. магістрантка спеціальності "Прикладна лінгвістика" КНЛУ (переклад анотацій); Снєгуров, І. О. магістрант спеціальності "Прикладна лінгвістика" КНЛУ (переклад анотацій, бібліографічний опис статей); Туз, В. А. магістрантка спеціальності "Прикладна лінгвістика" КНЛУ (переклад анотацій); Шульга, А. А. магістрантка спеціальності "Прикладна лінгвістика" КНЛУ (переклад анотацій); Беспальчук, Л. А. студентка 4-го курсу напряму підготовки "Філологія (Прикладна лінгвістика)" КНЛУ (укладання покажчика назв статей); Бобкова, А. І. студентка 4-го курсу напряму підготовки "Філологія (Прикладна лінгвістика)" КНЛУ (укладання покажчика авторів); Готовченко, К. Ю. студентка 4-го курсу напряму підготовки "Філологія (Прикладна лінгвістика)" КНЛУ (укладання покажчика назв статей); Кисіль, К. В. студентка 4-го курсу напряму підготовки "Філологія (Прикладна лінгвістика)" КНЛУ (укладання покажчика назв статей); Куліда, Ю. В. студентка 4-го курсу напряму підготовки "Філологія (Прикладна лінгвістика)" КНЛУ (укладання покажчика авторів)

А 69 Анотований науково-допоміжний бібліографічний покажчик статей журналу Computational Linguistics (2000-2017 рр.) [Електронний ресурс]: мережне електронне видання / ред. В.О.Коломієць, уклад. М.О.Драчова, М.Ю.Дубок, О.М.Мартинюк та ін. -- Електрон. текст. дані. -- К.: Видавничий центр КНЛУ, 2017. -- 17,79 др. арк. -- Режим доступу: http://cljai.weebly.com/, вільний. -- Назва з титул. екрана. -- Мови укр., англ.

У бібліографічному покажчику подано описи статей і коротких технічних звітів, опублікованих у журналі Computational Linguistics у 2000-2017 рр. Довідковий апарат видання містить покажчики назв і авторів статей.

Для науковців і практиків, викладачів вищих навчальних закладів і студентів, які цікавляться питаннями автоматичного опрацювання природної мови.

ПЕРЕДМОВА

Через стрімке зростання потреби в засобах автоматичного опрацювання природної мови, викликане значним посиленням ролі інформаційних технологій у забезпеченні конкурентоспроможності та інноваційного розвитку країни, набуває особливої актуальності питання організації ефективного доступу до зарубіжних публікацій на вказану тематику для ознайомлення з основними напрямами і результатами досліджень у цій галузі.

Журнал Computational Linguistics є одним із найстаріших рецензованих видань, присвячених створенню й аналізу систем автоматичного опрацювання природної мови. Журнал був заснований у 1974 році Асоціацією комп'ютерної лінгвістики, міжнародною організацією, яка об'єднує спеціалістів у галузі автоматичного опрацювання природної мови, під назвою American Journal of Computational Linguistics і спочатку публікувався тільки у вигляді мікрофіш. У 1980 році він став друкованим виданням, а в 1984 році змінив назву на Computational Linguistics. З 1988 року журнал випускає видавництво MIT Press (м. Кембрідж, США) від імені Асоціації комп'ютерної лінгвістики. З 2009 року Computational Linguistics є електронним журналом відкритого доступу. Журнал виходить чотири рази на рік.

Computational Linguistics належить до 10 найавторитетніших журналів у галузі комп'ютерної лінгвістики за версією Google Scholar Metrics. Імпакт-фактор журналу у 2015 році становив 2.017 (Journal Citation Report, Science Edition). Усім статтям журналу присвоєний ідентифікатор цифрового об'єкта DOI.

Крім наукових статей, які повідомляють про нові досягнення в царині автоматичного опрацювання природної мови, журнал публікує оглядові статті, технічні звіти, рецензії на книги, матеріали проблемного або дискусійного характеру, тексти виголошених на щорічній конференції Асоціації комп'ютерної лінгвістики промов учених, нагороджених премією за видатні професійні досягнення.

Мета анотованого науково-допоміжного бібліографічного покажчика статей журналу − полегшити пошук інформації про статті і технічні звіти, надруковані в журналі у 2000-2017 рр.

Матеріали в покажчику систематизовано за тематичними рубриками «Моделювання мови і мовної діяльності» і «Створення прикладних систем». Перша включає підрубрики «Автоматичний морфологічний аналіз», «Автоматичний семантичний аналіз», «Автоматичний синтаксичний аналіз», «Аналіз дискурсу», «Аналіз і синтез мовлення», «Аналіз тональності», «Встановлення референції», «Генерування тексту», «Зняття лексичної багатозначності», «Комп'ютерна лексикографія», «Корпусна лінгвістика», «Лінгвістичне анотування», «Проблеми машинного навчання», «Сегментація тексту», «Формальні моделі мови і їх застосування у комп'ютерній лінгвістиці». Друга рубрика складається з підрубрик «Автоматичне реферування», «Діалогові системи», «Інформаційний пошук», «Машинний переклад», «Мультимодальні системи», «Питально-відповідні системи». Бібліографічні записи всередині кожної підрубрики розташовано у хронологічному порядку і супроводжено анотаціями статей. Основою класифікації документів у покажчику є наукова класифікація, представлена на порталі знань з комп'ютерної лінгвістики.

Допоміжний апарат покажчика включає покажчик назв статей українською та англійською мовами і покажчик авторів статей, опублікованих у журналі за період 2000-2017 рр. У допоміжних покажчиках записи розташовано в алфавітному порядку.

Важливим елементом покажчика є система гіперпосилань, які пов'язують бібліографічні записи в різних розділах між собою та з відповідними публікаціями на офіційному веб-сайті журналу Computational Linguistics.

Анотований науково-допоміжний бібліографічний покажчик статей, опублікованих у журналі Computational Linguistics, стане у пригоді науковцям і практикам, викладачам і студентам вищих навчальних закладів, які цікавляться проблемами автоматичного опрацювання природної мови.

Укладачі висловлюють щиру подяку рецензентам доктору філологічних наук, професору кафедри української мови та прикладної лінгвістики Інституту філології Київського національного університету імені Тараса Шевченка Наталії Петрівні Дарчук і завідувачу відділу лексикології, лексикографії та структурно-математичної лінгвістики Інституту української мови НАН України,

доктору філологічних наук, професору Євгенії Анатоліївні Карпіловській за критичні зауваження і поради стосовно змісту покажчика. Укладачі також вдячні завідувачу сектору каталогізування відділу комплектування і обробки літератури бібліотеки Київського національного лінгвістичного університету Тетяні Євгеніївні Рябоконь за допомогу в підготовці покажчика до публікації.

Укладачі будуть щиро вдячні читачам за будь-які відгуки, критичні зауваження і побажання, які можна присилати на електронну адресу valentynak2004@yahoo.com.

Моделювання мови і мовленнєвої діяльності

Автоматичний морфологічний аналіз

Ofazer, K. Bootstrapping Morphological Analyzers by Combining Human Elicitation and Machine Learning [Створення морфологічних аналізаторів шляхом поєднання опитування інформантів і машинного навчання] / Kemal Ofazer, Sergei Nirenburg, Marjorie McShane // Computational linguistics. – 2001. – Vol. 27. – No. 1. – Pages 59–85. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120101300346804#.WH3oYn3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120101300346804

У статті описано напівавтоматичний спосіб створення скінченних морфологічних аналізаторів із широким охопленням для використання в системах опрацювання природної мови. Він складається з трьох компонентів – отримання лінгвістичної інформації від інформантів, алгоритму створення аналізатора за допомогою машинного навчання й середовища для тестування. Ці три компоненти застосовуються ітеративно, аж поки якість виведення досягне порогової величини. Вперше цей спосіб застосовано для аналізу морфології мов із обмеженими лінгвістичними ресурсами в рамках проекту-експедиції в лабораторії комп'ютерних досліджень університету штату Нью-Мексико. При цьому способі опитування-створення-тестування з отриманої від інформанта лексичної та флективної інформації укладається лексикон скінченного перетворювача, який поєднується із послідовністю морфографемних правил переписування, видобутою з отриманих прикладів за допомогою навчання на основі трансформацій. Потім за допомогою комплекту тестів здійснюється тестування створеного морфологічного аналізатора й усі виправлення вводяться до алгоритму навчання, після чого створюється удосконалений аналізатор.

Переклад В. Коломієць

Goldsmith, J. Unsupervised Learning of the Morphology of a Natural Language [Навчання морфології природної мови без учителя] / John Goldsmith // Computational linguistics. – 2001. – Vol. 27. – No. 2. – Pages 153–198. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120101750300490#.WH3xu33sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120101750300490

У статті повідомляються результати використання аналізу на основі мінімальної довжини опису (МДО) для моделювання навчання

морфологічної сегментації європейських мов без учителя за допомогою корпусів обсягом від 5000 до 500000 слів. Розроблено набір евристичних правил, які швидко створюють вірогіднісну морфологічну граматику, і в якості основного інструмента для визначення, чи будуть прийняті запропоновані евристичними правилами модифікації, використано МДО. Створена граматика добре узгоджується з аналізом, який здійснив би фахівець із морфології.

У заключному розділі обговорюється зв'язок цього типу граматичного аналізу на основі МДО з поняттям оціночної метрики у ранніх версіях породжувальної граматики.

Переклад К. Погорєлова

van Halteren, H. Improving Accuracy in Word Class Tagging through the Combination of Machine Learning Systems [Підвищення точності частиномовної розмітки шляхом об'єднання систем машинного навчання] / **Hans van Halteren, Jakub Zavrel, Walter Daelemans // Computational linguistics. – 2001. – Vol. 27. – No. 2. – Pages 199–229. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120101750300508#.WH3yEH3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120101750300508**

http://www.mitpressjournals.org/doi/pdf/10.1162/089120101750300508

Проаналізовано, як можна використати відмінності між мовними моделями, автоматично створеними різними керованими даними системами при виконанні однакових завдань опрацювання природної мови, для того щоб одержати вищу точність, ніж у найкращої окремої системи. Це зроблено за допомогою експериментів, які включали завдання морфосинтаксичної розмітки частин мови, на основі трьох різних розмічених корпусів. Чотири добре відомі генератори розмічувачів (прихована Марківська модель, на основі пам'яті, правила трансформації та максимальна ентропія) тренувались на однакових корпусних даних. Після порівняння їхні вихідні дані було об'єднано за допомогою кількох стратегій вибору й класифікаторів другого рівня. Всі комбіновані розмічувачі перевершили свої найкращі компоненти. Зменшення кількості помилок залежало від корпусу, але досягало 24,3% при використанні корпусу Ланкастер-Осло-Берген.

Переклад К. Погорєлова

Lee, G. G. Syllable-Pattern-Based Unknown-Morpheme Segmentation and Estimation for Hybrid Part-of-Speech Tagging of Korean [Сегментування і оцінювання нерозпізнаних морфем на основі моделі складу для гібридного частиномовного анотування корейської мови] / Gary Geunbae Lee, Jeongwon Cha, Jong-Hyeok Lee // Computational linguistics. – 2002. – Vol. 28. – No. 1. – Pages 53–70. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120102317341

774#.WH3yhX3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120102317341774

Нерозпізнані морфеми є причиною більшості помилок у морфологічному аналізі та частиномовній розмітці корейської мови. У статті представлено узагальнений метод оцінювання нерозпізнаних морфем на основі моделі складу за допомогою гібридної статистичної системи частиномовної розмітки на основі правил POSTAG (POStech TAGger)*. Цей метод угадування нерозпізнаних морфем базується на поєднанні словника моделей морфем, у якому представлено загальні лексичні моделі корейських морфем, з апостеріорною оцінкою складів триграм. Склади триграми допомагають вирахувати лексичні вірогідності нерозпізнаних морфем і вживаються для пошуку найкращого результату розмітки. За допомогою цього методу можна передбачити частиномовні теги нерозпізнаних морфем незалежно від їхньої кількості та/або позицій у eojeol (корейська мовна одиниця подібна до слова в англійській мові), чого не можна зробити за допомогою інших систем розмітки корейської мови. У низці експериментів із трьома різними корпусами розроблена система досягла точності розмітки 97%, хоча 10% морфем у тестових корпусах були нерозпізнаними. Система також показала дуже високу повноту охоплення і точність оцінювання усіх класів нерозпізнаних морфем.

*Бінарний код системи POSTAG знаходиться у вільному доступі для досліджень і оцінювання на веб-сторінці http://nlp.postech.ac.kr/. Перейдіть за посиланням OpenResources→DownLoad.

Переклад І. Снєгурова

Cohen-Sygal, Y. Finite-State Registered Automata for Non-Concatenative Morphology [Скінченні регістрові автомати для розпізнавання неконкатенативної морфології] **/ Yael Cohen-Sygal, Shuly Wintner // Computational linguistics. – 2006. – Vol. 32. – No. 1. – Pages 49–82. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2006.32.1.49#.WH3y433sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2006.32.1.49**

У статті розглядаються скінченні регістрові автомати для розпізнавання (CPAP), нові комп'ютерні засоби, які є різновидами скінченних автоматів для розпізнавання, спеціально пристосованими для реалізації неконкатенативних морфологічних процесів. Ця модель є розширенням наявних скінченних автоматів для розпізнавання, ще не оптимізованих для опису такого виду явищ. У статті спочатку подано означення моделі та описано її математичні й обчислювальні характеристики. Потім подано розширену регулярну мову, виразами якої позначені CPAP. Нарешті, наведено декілька прикладів складних морфологічних і фонологічних явищ, майстерно реалізованих за допомогою CPAP, для того щоб показати переваги моделі.

Переклад В. Коломієць

Daya, E. Identifying Semitic Roots: Machine Learning with Linguistic Constraints [Визначення коренів у семітських мовах: машинне навчання з використанням лінгвістичних правил] **/ Ezra Daya, Dan Roth, Shuly Wintner // Computational linguistics. – 2008. – Vol. 34. – No. 3. – Pages 429–448. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2008.07-002-R1-06-30#.WH3zUn3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2008.07-002-R1-06-30**

Слова в семітських мовах утворюються шляхом поєднання двох морфем – кореня й моделі. Корінь складається тільки з приголосних, як правило трьох, а модель є комбінацією голосних і приголосних, перемішаних із "пазами", в які вставляються кореневі приголосні. Визначення кореня заданого слова – важливе завдання, яке вважається обов'язковим компонентом морфологічного аналізу семітських мов; а інформація про корені потрібна як для лінгвістичних досліджень, так і для розв'язання практичних завдань. У статті описано застосування машинного навчання, вдосконаленого невеликим набором правил, у визначенні коренів слів у семітських мовах. Хоча існують прикладні програми, які можуть виокремлювати корені слів в арабській мові та ідишу, всі вони передбачають трудомісткий процес створення великих лексиконів, які є компонентами повномасштабних морфологічних аналізаторів. Перевага нашого методу полягає в автоматизації цього процесу, оминанні затримки, спричиненої необхідністю забарного укладання списків коренів і моделей всіх лексем у мові. Наскільки нам відомо, це перше застосування машинного навчання у розв'язанні цієї проблеми та одна з небагатьох спроб звернутися безпосередньо до неконкатенативної морфології, використовуючи машинне навчання. Загалом, отримані результати пролили світло на проблему об'єднання класифікаторів за наявності (лінгвістичних) правил.

Переклад В. Коломієць

Baldwin, T. Prepositions in Applications: A Survey and Introduction to the Special Issue [Прийменники у прикладних програмах: загальний огляд і вступ до спеціального випуску] **/ Timothy Baldwin, Valia Kordoni, Aline Villavicencio // Computational linguistics. – 2009. – Vol. 35. – No. 2. – Pages 119–149. – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2009.35.2.119**

Стаття містить загальний огляд досліджень прийменників і їхнього використання у прикладних програмах для опрацювання природної мови. Коротко описано синтаксис прийменників і його значимість для прикладних програм для опрацювання природної мови, при цьому особливу увагу приділено приєднанню прийменникових груп і прийменникам у

багатослівних виразах. Розглянуто формальні та лексико-семантичні характеристики прийменників і їхню значимість для прикладних програм для опрацювання природної мови, описано окремі прикладні дослідження, в яких прийменникам приділяється значна увага. Коротко викладено зміст статей, вміщених у спеціальному випуску журналу, й визначено напрями досліджень прийменників, для проведення яких настав час.

Переклад В. Коломієць

Hammarström, H. Unsupervised Learning of Morphology [Навчання морфології без учителя] / **Harald Hammarström, Lars Borin // Computational linguistics. – 2011. – Vol. 37. – No. 2. – Pages 309–350. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00050#.WH3z1n3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00050**

Стаття містить огляд досліджень навчання морфології без учителя. За визначенням авторів, навчання морфології без учителя є питанням породження опису (будь-якого, навіть якщо це тільки поділ на морфеми) будови орфографічних слів на основі лише необроблених текстових даних певною мовою. Коротко викладено історію й актуальність проблеми. Потім перераховано та стисло схарактеризовано більше 200 досліджень, критично проаналізовано найважливіші ідеї в цій галузі. Підсумовано наявні досягнення і вказано напрями подальших розвідок.

Переклад В. Коломієць

Ruokolainen T. A Comparative Study of Minimally Supervised Morphological Segmentation [Порівняльне дослідження морфологічного сегментування методом часткового навчання з учителем] / Teemu Ruokolainen , Oskar Kohonen , Kairit Sirts , Stig-Arne Grönroos , Mikko Kurimo, Sami Virpioja // Computational linguistics. – 2016. – Vol. 42. – No. 1. – Pages 91–120. – Режим доступу до анотації: https://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00243 – Режим доступу до повнотекстової статті: https://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00243

У статті представлено порівняльне дослідження однієї з галузей автоматизованого морфологічного аналізу – морфологічного сегментування методом часткового навчання з учителем. У морфологічному сегментуванні словоформи діляться на морфи, матеріальну реалізацію морфем. У керованому даними напівавтоматичному навчанні, система вчиться здійснювати сегментування за допомогою невеликої кількості словоформ, маркованих екпертами, та великого набору немаркованих словоформ. На додаток до огляду літератури, присвяченої опублікованим методам, у статті представлено докладне емпіричне порівняння трьох різних видів моделей, а

також детальний аналіз помилок. Спираючись на огляд літератури, було зроблено висновок про те, що існуючі методи значною мірою спираються на генеративні підходи на основі морфемного лексикону та методи на основі диференціального визначення меж. Що стосується більш успішного з двох підходів, як попередні дослідження, так і представлене у статті емпіричне оцінювання дають підстави вважати, що сучасні досягнення є результатом застосування методики диференціального визначення меж.

Переклад А. Шульги

Sun, W. Towards Accurate and Efficient Chinese Part-of-Speech Tagging [На шляху до ефективної частиномовної розмітки китайської мови] / Weiwei Sun, Xiaojun Wan // Computational linguistics. – 2016. – Vol. 42. – No. 3. – Pages 391–419. – Режим доступу до анотації: https://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00253 – Режим доступу до повнотекстової статті: https://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00253

За допомогою методів структурної лінгвістики досліджено парадигматичні та синтагматичні відношення між словами для автоматичної частиномовної розмітки китайської мови, важливого, але складного завдання автоматичного опрацювання китайської мови. Парадигматичні відношення між словами напряму визначаються шляхом кластеризації слів на базі великих нерозмічених корпусів і використовуються для створення нових правил для вдосконалення диференціального розмітника. Синтагматичні відношення між словами імпліцитно ідентифікуються шляхом автоматичного синтаксичного аналізу на основі граматики складників і використовуються шляхом об'єднання системи. Експерименти на базі корпусу Penn Chinese Treebank свідчать про важливість як парадигматичних, так і синтагматичних відношень. Завдяки запропонованим лінгвістично орієнтованим, гібридним підходам вдалося досягти відносного зменьшення помилок на 18% у порівнянні з сучасними базовими показниками. Незважаючи на ефективне підвищення точності, використання гібридних систем є недоцільним для багатьох практичних застосувань опрацювання природної мови через високу вартість обчислень. У статті також розглядається проблема підвищення ефективності маркування під час тестування. Зокрема, проаналізовано немарковані дані з метою передачі прогностичної здатності гібридних моделей моделям простих послідовностей. Точніше кажучи, гібридні системи використовуються для створення масштабних псевдотренувальних даних для дешевих моделей. Експериментальні результати свідчать, що створені заново моделі не тільки досягають вищої точності у класифікації окремих слововживань, але також слугують прекрасним зовнішнім інтерфейсом для аналізатора.

Переклад М. Дубка

Автоматичний семантичний аналіз

Merlo, P. Automatic Verb Classification Based on Statistical Distributions of Argument Structure [Автоматична класифікація дієслів на основі статистичного розподілу структури аргументів] / **Paola Merlo, Suzanne Stevenson // Computational linguistics. – 2001. – Vol. 27. – No. 3. – Pages 373–408. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120101317066122#.WH4VC33sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120101317066122**

У широкому колі завдань з обробки природної мови вирішальна роль належить автоматичному отриманню лексичних знань. Особливо важливою є інформація про дієслова, які є головним джерелом інформації про зв'язки у реченні, предикатно-аргументну структуру, яка пов'язує дію або стан з учасниками (тобто, хто що кому зробив). У статті описано експерименти з контрольованого навчання автоматичної класифікації трьох основних типів англійських дієслів на основі структури їх аргументів, а саме, тематичних ролей, які вони присвоюють учасникам. Для тренування класифікатора використовувались лінгвістично обґрунтовані статистичні показники, видобуті з великих за обсягом анотованих корпусів. Було досягнуто точність на рівні 69,8% для завдання, вихідна оцінка точності якого становила 34%, а вирахувана верхня експертна межа була на рівні 86,5%. Детальний аналіз продуктивності алгоритму та його помилок підтвердив, що запропоновані ознаки відображають характеристики, пов'язані з структурою аргументів дієслів. Отримані результати підтвердили гіпотези про те, що вирішальна роль у класифікації дієслів належить знанням про тематичні зв'язки і що їх можна автоматично видобути з корпусу. Таким чином, продемонстровано ефективне поєднання глибших лінгвістичних знань з надійністю та універсальністю статистичних методів.

Переклад І. Снєгурова

Clark, S. Class-Based Probability Estimation Using a Semantic Hierarchy [Оцінювання ймовірності на основі класу з семантичної ієрархії] / **Stephen Clark, David Weir // Computational linguistics. – 2002. – Vol. 28. – No. 2. – Pages 187–206. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120102760173643#.WH4Vyn3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120102760173643**

У статті розглядається оцінювання конкретного типу ймовірності, а саме, ймовірності появи іменника в певному значенні в ролі певного аргументу присудка. Для того, щоб вирішити додаткову проблему недостатньої

кількості даних, запропоновано визначати ймовірності відносно значень із семантичної ієрархії та скористатися тим фактом, що ці значення можна розбити на класи семантично схожих значень. Особлива увага приділяється питанню визначення класу, прийнятного для певного значення, або навпаки, визначенню рівня узагальнення, прийнятного для ієрархії. Для визначення прийнятного рівня узагальнення розроблено процедуру, яка використовує тест хі-квадрат. Ефективність цього методу оцінювання тестувалась шляхом імітування зняття омонімії і використання двох альтернативних методів оцінювання, в яких застосовано різні процедури узагальнення: в першому – принцип мінімальної довжини опису, а в другому – критерій сполучувальної переваги Резника. Окрім цього, ефективність запропонованого методу досліджувалась за допомогою як стандартного критерію хі-квадрат Пірсона, так і критерію хі-квадрат, що вираховується на основі логарифмів правдоподібності.

Переклад І. Снєгурова

Gildea, D. Automatic Labeling of Semantic Roles [Автоматична розмітка семантичних ролей] **/ Daniel Gildea, Daniel Jurafsky // Computational linguistics. – 2002. – Vol. 28. – No. 3. – Pages 245–288. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120102760275983#.WH4WLn3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120102760275983**

У статті описується система автоматичної ідентифікації семантичних відносин, або семантичних ролей, заповнених складовими речення у семантичному фреймі. За наявності вхідного речення, цільового слова і фрейму система автоматично присвоює складовим або абстрактні семантичні ролі, такі як Агенс або Паціенс, або більш предметно-орієнтовані семантичні ролі, такі як Мовець, Повідомлення і Тема.

Система базується на статистичних класифікаторах, навчених приблизно на 50 000 реченнях, які були вручну анотовані семантичними ролями учасниками проекту семантичної розмітки FrameNet. Потім було побудовано синтаксичні дерева всіх навчальних речень і видобуто різні лексичні та синтаксичні характеристики, зокрема тип словосполучення кожного складника, його граматичну функцію і місце в реченні. Ці характеристики були об'єднані з інформацією про дієслово-предикат, іменник чи прикметник, а також з інформацією про апріорну імовірність різних комбінацій семантичних ролей. Для того щоб зробити висновки щодо можливих заповнювачів ролей, використовувались різні алгоритми кластеризації лексики. Тестування включало синтаксичний аналіз речень, анотування їх виділеними характеристиками і пропускання через класифікатори.

Точність визначення системою семантичних ролей попередньо сегментованих складників досягає 82%. Точність виконання складнішого завдання одночасної сегментації складових і визначення їх семантичної ролі

досягла 65% при повноті 61%.

Здійснене дослідження також дозволило порівняти корисність різних характеристик та їх комбінацій для анотування семантичних ролей. Також досліджена інтеграція анотування ролей із статистичним синтаксичним аналізом і здійснена спроба зробити узагальнення для предикатів, які не зустрілися в навчальних даних.

Переклад К. Погорєлова

Lapata, M. A Probabilistic Account of Logical Metonymy [Вірогіднісне пояснення логічної метонімії] **/ Maria Lapata, Alex Lascarides // Computational linguistics. – 2003. – Vol. 29. – No. 2. – Pages 261–315. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120103322145324#.WIXKIH3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120103322145324**

У статті досліджується логічна метонімія, тобто конструкції, в яких аргумент слова, виражений синтаксичною одиницею, відрізняється від цього аргументу, вираженого одиницею логічною (наприклад, «отримати задоволення від книги» означає «*отримати задоволення від читання книги*», а «легка проблема» - це «проблема, яку легко вирішити»). Систематичне варіювання інтерпретації подібних конструкцій вимагає детального і складного пояснення утворень на основі зв'язку між синтаксисом і семантикою. Лінгвістичні пояснення логічної метонімії, як правило, не дають вичерпного опису всіх можливих інтерпретацій або не ранжують ці інтерпретації з точки зору їх вірогідності. Тому значення метонімічних дієслів і прикметників було видобуто з великого корпусу, також була запропонована вірогіднісна модель, яка дозволяє здійснити ранжування на основі набору можливих інтерпретацій. Інтерпретації визначаються автоматично на основі постійних відповідностей між поверховими синтаксичними ознаками і значенням. Отримані результати оцінювалися за допомогою перефразувань, отриманих від учасників експерименту. Показано, що здійснене моделлю ранжування значень надійно корелює з людською інтуіцією.

Переклад В. Коломієць

Mason, Z.J. CorMet: A Computational, Corpus-Based Conventional Metaphor Extraction System [Корпусно-базована система автоматичного видобування стертих метафор] **/ Zachary J. Mason // Computational linguistics. – 2004. – Vol. 30. – No. 1. – Pages 23–44. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120104773633376#.WH4W533sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120104773633376**

CorMet – це корпусно-базована система виявлення метафоричних відповідностей між концептами. Вона виконує це завдання, знаходячи систематичне варіювання в характерних для тематичної області преференціях вибору, отриманих із великих, динамічно досліджених Інтернет-корпусів.

Метафори переводять структуру з вихідної предметної області в цільову предметну область, роблячи деякі концепти в цільовій предметній області метафорично еквівалентними концептам у вихідній предметній області. Дієслова, які обирають концепт у вихідній предметній області, як правило, обирають його метафоричний еквівалент у цільовій предметній області. Ця закономірність, що виявляється за допомогою поверхневого лінгвістичного аналізу, використовується для знаходження метафоричних міжконцептуальних відповідностей, за допомогою яких можна потім зробити висновок про існування стертих метафор вищого рівня.

У більшості інших систем автоматичного виявлення метафор використовуються невеликі, запрограмовані вручну бази знань для семантичного аналізу й невелика кількість прикладів. Хоча єдиною базою знань системи CorMet є Word Net (C. Fellbaum, 1998), вона може виявити відповідності, які утворюють велику кількість стертих метафор, і в деяких випадках розпізнати речення, у яких ці відповідності реалізовані. Здійснена перевірка здатності CorMet виявити підгрупу списку основних метафор (G. Lakoff, J. Espenson, and A. Schwartz, 1991).

Переклад В. Коломієць

Girju, R. Automatic Discovery of Part-Whole Relations [Автоматичне виявлення відношень частина-ціле] **/ Roxana Girju, Adriana Badulescu, Dan Moldovan // Computational linguistics. – 2006. – Vol. 32. – No. 1. – Pages 83–135. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2006.32.1.83#.WH4Y4n3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2006.32.1.83**

Важливим компонентом видобування знань із текстів є автоматичне виявлення семантичних відношень. У статті представлено контрольований, семантично інтенсивний, незалежний від тематичної області підхід до автоматичного виявлення у тексті відношень частина–ціле. Спочатку описано алгоритм, який виявляє лексико-синтаксичні структури, які передають відношення частина–ціле. Складність полягає в тому, що ці структури також передають інші семантичні відношення і потрібен якийсь метод навчання, щоб з'ясувати, чи передає структура відношення частина–ціле, чи ні. Було проанотовано й уведено в спеціалізовану систему машинного навчання, яка навчається правилам класифікації, великий набір тренувальних прикладів. Правила генеруються за допомогою застосування ітеративного методу семантичної спеціалізації до складників іменних груп. Таким чином були згенеровані правила класифікації для різних структур,

таких як присвійний відмінок, складні іменники та іменні групи з прийменниковими словосполученнями, для того щоб виявляти у них відношення частина–ціле. Придатність цих правил була перевірена на тестовому корпусі, вони показали загальну середню точність 80,95% і повноту 75,91%. Наведені результати свідчать про необхідність зняття лексичної багатозначності для цього завдання. Вони також свідчать, що різні лексико-синтаксичні структури несуть різну семантичну інформацію й повинні оброблятись окремо, тобто до різних структур потрібно застосовувати різні правила тлумачення.

Переклад В. Коломієць

Schulte im Walde, S. Experiments on the Automatic Induction of German Semantic Verb Classes [Експерименти з автоматичною індукцією семантичних класів німецьких дієслів] **/ Sabine Schulte im Walde // Computational linguistics. – 2006. – Vol. 32. – No. 2. – Pages 159–194. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2006.32.2.159#.WH4Zq33sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2006.32.2.159**

У статті описано експерименти з класифікації німецьких дієслів. Джерелом дистрибутивного опису дієслів на стику лексичного синтаксису і лексичної семантики виступає статистична граматична модель німецкої мови, а алгоритм навчання ознак без учителя k-means використовує емпіричні характеристики дієслів для здійснення автоматичної індукції класів дієслів. Для порівняння за різними критеріями результатів класифікації з золотим стандартом семантичних класів німецьких дієслів використано різні мірки оцінювання. Основними цілями експериментів було (1) емпіричне застосування і дослідження добре усталеного зв'язку між значенням дієслова і його поведінкою у кластерному аналізі і (2) аналіз технічних параметрів кластерного аналізу, потрібних для виконання цього специфічного лінгвістичного завдання. Методика класифікації була розроблена на невеликому наборі дієслів, а потім застосована до великого набору, який складався з 883 намецьких дієслів.

Переклад В. Коломієць

Turney, P. Similarity of Semantic Relations [Схожість семантичних відносин] / Peter D. Turney // Computational linguistics. – 2006. – Vol. 32. – No. 3. – Pages 379–416. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2006.32.3.379#.WH4aQH3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2006.32.3.379

Існує принаймні два типи схожості. Реляційна схожість – це відповідність між відносинами, в той час як атрибутивна схожість – це відповідність між

характеристиками. Коли два слова мають високий ступінь атрибутивної схожості, вони називаються синонімами. Коли два слова мають високий ступінь реляційної схожості, кажуть, що їх відносини є аналогічними. Наприклад, пара слів каменяр:камінь є аналогічною парі тесляр:дерево. У статті описано латентний реляційний аналіз (Latent Relational Analysis, *скор.* LRA), метод визначення реляційної схожості. LRA може бути застосований у багатьох областях, зокрема видобуванні інформації, знятті лексичної багатозначності та інформаційному пошуці. Нещодавно для визначення релаційної схожості було адаптовано векторну модель (Vector Space Model, *скор.* VSM) видобування інформації й отримано результат 47% у тесті, який складався з 374 завдань вибору схожих слів із множин у межах університетської програми. У моделі VSM відносини між парою слів характеризуються вектором частоти попередньо заданих шаблонів у великому корпусі. LRA є розширенням моделі VSM у трьох напрямах. (1) Шаблони видобуваються з корпусу автоматично, (2) для згладжування даних частоти використовується сингулярне розкладання (Singular Value Decomposition, скор. SVD) і (3) варіанти пар слів аналізуються за допомогою автоматично згенерованих синонімів. LRA досягає результату 56% у тесті з 374 питань про схожіть слів, статистичного еквіваленту середнього результату виконання тесту людиною, що становить 57%. У спорідненому завданні класифікації семантичних відносин LRA має аналогічні переваги над VSM.

Переклад В. Коломієць

Padó, S. Dependency-Based Construction of Semantic Space Models [Розробка моделей семантичного простору на основі залежностей] / Sebastian Padó, Mirella Lapata // Computational linguistics. – 2007. – Vol. 33. – No. 2. – Pages 161–199. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2007.33.2.161#.WH4anH3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2007.33.2.161

Як правило, для представлення лексичного значення векторні моделі семантичного простору використовують статистику сумісного вживання слів у великих за обсягом корпусах текстів. У статті описано новий підхід до побудови семантичних просторів, який враховує синтаксичні зв’язки. Запропоновано алгоритм для цього класу моделей, завдяки якому процес розробки керується лінгвістичними знаннями. Запропонований підхід оцінено за допомогою низки завдань, пов’язаних із когнітивною наукою і обробкою природної мови: семантичного праймінгу, встановлення синонімії і зняття лексичної багатозначності. В усіх випадках, запропонований підхід не поступається за ефективністю існуючим методам або перевершує їх.

Переклад М. Погребної

Màrquez, L. Semantic Role Labeling: An Introduction to the Special Issue [Анотування семантичних ролей: вступ до спеціального випуску] / Lluís Màrquez, Xavier Carreras, Kenneth C. Litkowski, Suzanne Stevenson // Computational linguistics. – 2008. – Vol. 34. – No. 2. – Pages 145–159. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2008.34.2.145#.WJza-_LsSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2008.34.2.145

Анотування семантичних ролей, автоматична ідентифікація і маркування аргументів у тексті, стало сьогодні провідним завданням у комп'ютерній лінгвістиці. Хоча проблеми, пов'язані з цим завданням, вивчалися протягом багатьох десятиліть, наявність потужних ресурсів і розробка методів статистичного машинного навчання збільшили кількість досліджень у цій царині. В цьому спеціальному випуску журналу представлені вибрані й показові праці в цій царині. Цей огляд містить опис лінгвістичного підгрунтя проблеми, переходу від лінгвістичних теорій до їх комп'ютерної реалізації, основних використовуваних ресурсів, опис етапів роботи обчислювальних систем, а також перелік основних проблем і результатів анотування семантичних ролей (представлених у кількох міжнародних аналітичних звітах). В огляді проаналізовані недоліки в анотуванні семантичних ролей і вказані важливі проблеми в цій царині, які потребують розв'язання. Загалом, подальші результативні дослідження в царині анотування семантичних ролей є надзвичайно перспективними.

Переклад В. Коломієць

Toutanova, K. A Global Joint Model for Semantic Role Labeling [Глобальна об'єднана модель для розмітки семантичних ролей] / Kristina Toutanova, Aria Haghighi, Christopher D. Manning // Computational linguistics. – 2008. – Vol. 34. – No. 2. – Pages 161–191. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2008.34.2.161#.WH4bp33sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2008.34.2.161

У статті описана модель розмітки семантичних ролей, у якій відображена мовна здогадка про те, що фрейм семантичного аргументу є об'єднаною структурою зі стійкими залежностями між аргументами. Продемонстровано, як використати ці стійкі залежності у статистичній об'єднаній моделі з великим набором ознак словосполучень із множинними аргументами. Запропонована модель значно перевершує подібну найпродуктивнішу локальну модель, яка не включає залежності між різними аргументами.

Оцінено переваги застосування цієї комбінованої інформації в корпусі Propbank при використанні в якості вхідних даних безпомилкових і автоматично породжених дерев залежностей. До переваг належить зменшення до 24,1% кількості помилок для всіх аргументів і до 36,8% для

ядерних аргументів синтаксичних дерев золотого стандарту. При використанні автоматично породжених синтаксичних дерев кількість помилок зменшилась, відповідно, на 8,3% для всіх аргументів і на 10,3% для ядерних аргументів. Також описано результати на об'єднаному наборі даних конференції CoNLL 2005. На додаток, досліджено можливість застосування різних видів синтаксичного аналізу для подолання шуму і невизначеності синтаксичного аналізатора.

Переклад І. Снєгурова, М. Погребної

Moschitti, A. Tree Kernels for Semantic Role Labeling [Кернфункції для анотування семантичних ролей] / Alessandro Moschitti, Daniele Pighin, Roberto Basili // Computational linguistics. – 2008. – Vol. 34. – No. 2. – Pages 193–224. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2008.34.2.193#.WH4b633sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2008.34.2.193

Доступність масштабних наборів даних з анотованими вручну предикатно-аргументними структурами останнім часом сприяла використанню методів машинного навчання у розробці систем автоматичного анотування семантичних ролей. Головна увага дослідників у цій області прикута до вибору способів представлення ознак і способу ефективного декомпозування завдання в різних моделях навчання. Щодо першого способу, використовуються переважно структурні параметри повних синтаксичних розборів, оскільки вони представляють способи програмування різних принципів, підказаних теорією зв'язку між синтаксисом і семантикою. Другий спосіб пов'язаний з кількома навчальними схемами на основі загальних уявлень про синтаксичні аналізатори. Наприклад, етапи зміни ранжування на основі альтернативних предикатно-аргументних послідовностей того самого речення виявились дуже ефективними.

У статті запропоновано кілька кернфункцій для моделювання характеристик синтаксичних дерев у автоматах на основі кернфункцій, наприклад перцептронах або машинах опорних векторів. Зокрема, різні види ядер послідовностей на деревах описуються як загальні підходи до проектування ознак у анотуванні семантичних ролей. Більше того, проведено велику кількість експериментів з такими ядрами для дослідження їх ролі на окремих етапах структури анотування семантичних ролей, як окремо, так і разом з іншими ознаками, які традиційно анотуються вручну. Результати розпізнавання меж, класифікації і зміни ранжування свідчать про значний вплив кернфункцій на загальну точність, особливо якщо кількість тренувальних даних незначна. На закінчення, кернфункції уможливлюють загальний і портативний метод проектування ознак, який можна застосувати до великої кількості завдань обробки природної мови.

Переклад В. Коломієць

Xue, N. Labeling Chinese Predicates with Semantic Roles [Анотування семантичних ролей китайських присудків] / Nianwen Xue // Computational linguistics. – 2008. – Vol. 34. – No. 2. – Pages 225–255. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2008.34.2.225#.WH4cOn3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2008.34.2.225

У статті описано анотування семантичних ролей у китайській мові, виконане у двох щойно створених корпусах: китайському PropBank, семантично анотованому корпусі китайських дієслів, і китайському Nombank, супутньому корпусі, який містить розмітку предикатно-аргументних структур субстантивованих присудків. Оскільки поміти семантичних ролей присвоюються складникам синтаксичного дерева, у статті спочатку описано експерименти, в яких поміти семантичних ролей автоматично присвоювались побудованим вручну синтаксичним деревам із корпусу Chinese Treebank. Це дало уявлення про ефективність автоматичного визначення поміт семантичних ролей на основі синтаксичної анотації в банку синтаксичних дерев. Потім описано експерименти з використанням автоматичного синтаксичного розбору і зменшенням обсягу ручного анотування даних, які вводяться до синтаксичного аналізатора: автоматичний синтаксичний аналіз на основі золотого стандарту сегментації і частиномовної розмітки, автоматичний синтаксичний аналіз тільки на основі золотого стандарту сегментації і повністю автоматичний синтаксичний аналіз. Ці експерименти визначали, наскільки ефективною може бути анотування семантичних ролей у китайській мові в реальних ситуаціях. Отримані результати свідчать, що за умови застосування синтаксичних дерев, побудованих вручну, точність анотування семантичних ролей у китайській мові співставна з точністю сучасних систем анотування семантичних ролей в англійській мові, налаштованих і протестованих на англійському корпусі PropBank, хоча китайський PropBank значно менше за розміром. Проте, коли використовується автоматичний синтаксичний аналізатор, точність створеної системи значно нижче, ніж точність сучасних систем аналізу англійської мови. Це означає, що для підвищення ефективності анотування семантичних ролей у китайській мові необхідно удосконалити автоматичний синтаксичний аналіз китайської мови.

Переклад В. Коломієць

Punyakanok, V. The Importance of Syntactic Parsing and Inference in Semantic Role Labeling [Роль автоматичного синтаксичного аналізу і логічного виведення в анотуванні семантичних ролей] / Vasin Punyakanok, Dan Roth, Wen-tau Yih // Computational linguistics. – 2008. – Vol. 34. – No. 2. – Pages 257–287. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2008.34.2.257#.WH4cfn3sSGA – Режим доступу до повнотекстової статті:

http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2008.34.2.257

У статті описано загальний підхід до анотування семантичних ролей. Цей підхід поєднує машинне навчання з процедурою логічного виведення на основі цілочислового лінійного програмування, яке включає у загальний процес прийняття рішень лінгвістичні й структурні обмеження. У таких рамках розглядається роль даних автоматичного синтаксичного аналізу в анотуванні семантичних ролей. Продемонстровано, що повні дані автоматичного синтаксичного аналізу безумовно є найнеобхіднішими для визначення аргументу, особливо на найпершій стадії – стадії обрізки. Як не дивно, якість стадії обрізки не може визначатись виключно на основі її точності та повноти. Натомість вона залежить від характеристик можливих вихідних змінних, від яких залежить складність наступних проблем. Виходячи з цього спостереження, запропоновано ефективний і простий метод комбінування різних систем анотування семантичних ролей шляхом об'єднаного логічного виведення, який значно поліпшує його результативність.

Створена система була оцінена на об'єднаному наборі для анотування семантичних ролей конференції CoNLL-2005 і отримала найвищий показник F1 з 19 учасників.

Переклад В. Коломієць

Pradhan, S. Towards Robust Semantic Role Labeling [Створення робастної системи анотування семантичних ролей] / Sameer S. Pradhan, Wayne Ward, James H. Martin // Computational linguistics. – 2008. – Vol. 34. – No. 2. – Pages 289–310. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2008.34.2.289#.WH4fr33sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2008.34.2.289

Більшість досліджень анотування семантичних ролей присвячені тренуванню і оцінюванню на одному й тому ж корпусі. Такий підхід, хоч і є прийнятним для проведення нового дослідження, може призвести до перенавчання на одному корпусі. У статті описано принципи роботи новітньої системи анотування семантичних ролей ASSERT і проаналізовано надійність цієї системи, коли її тренують на одній категорії даних і використовують для анотування іншої категорії. Стаття починається з опису результатів тренування і тестування системи на корпусі PropBank, який містить анотовані тексти з газети Wall Street Journal (*скор.* WSJ). Потім описано експерименти для оцінки можливості перенесення системи на інше джерело даних. Ці експерименти полягають у порівнянні результатів при використанні матеріалів з WSJ і матеріалів з корпусу Brown Corpus, які містяться в корпусі PropBank. Результати свідчать, що хоча синтаксичний аналіз та ідентифікація аргументів переносяться на новий корпус порівняно добре, цього не можна сказати про класифікацію аргументів. Наведено аналіз

причин цієї ситуації, які загалом вказують на природу здебільшого лексичних/семантичних ознак, які переважають у завданні класифікації, в той час як у завданні ідентифікації аргументів переважають структурні ознаки загального характеру.

Переклад В. Коломієць

Jørgensen, F. A Minimal Recursion Semantic Analysis of Locatives [Семантичний аналіз локативів з мінімальною рекурсією] / Fredrik Jørgensen, Jan Tore Lønning // Computational linguistics. – 2009. – Vol. 35. – No. 2. – Pages 229–270. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.06-69-prep5#.WH4g133sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.06-69-prep5

У статті описана пілотна реалізація граматики, яка містить різні типи локативних прийменникових груп. А саме, досліджено різницю між статичними і директивними локативами, а також між різними видами директивних локативів. У залежності від синтаксичного оточення локативи можуть бути як обставинами, так і референціальними виразами. Ми застосовуємо до них єдиний підхід. Граматика реалізована на матеріалі норвезьких локативів, але в статті і аналізуються, і порівнюються з норвезькими англійські локативи. Семантичний аналіз здійснено на основі пропозиції Маркуса Крахта (Markus Kracht, 2002). Продемонстровано, як можна вбудувати цей аналіз у семантику з мінімальною рекурсією (англ. Minimal Recursion Semantics, скор. MRS) (Copestake et al., 2005). Показано, як можна застосувати отриману систему в трансферній системі машинного перекладу і як можна поверхневе нерекурсивне представлення семантики перетворити на глибше семантичне представлення.

Переклад В. Коломієць

Tsang, V. A Graph-Theoretic Framework for Semantic Distance [Теоретико-графічна модель семантичної відстані] / Vivian Tsang, Suzanne Stevenson // Computational linguistics. – 2010. – Vol. 36. – No. 1. – Pages 31–69. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2010.36.1.36101#.WH4hmH3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2010.36.1.36101

Багато програм для обробки природної мови потребують класифікації текстів на основі семантичної відстані між ними (наскільки схожими або різними є ці тексти). Наприклад, порівнюючи текст нового документу з текстами документів на відомі теми, можна визначити тему нового тексту. Як правило, для визначення імпліцитної семантичної відстані між двома частинами тексту використовується дистрибутивна відстань. Однак такі методи не враховують семантичні відносини між словами. У цій статті

описано альтернативний метод вимірювання семантичної відстані між текстами, який об'єднує інформацію про дистрибуцію та онтологічні знання у формалізмі мережевого трафіка. Спочатку кожен текст було представлено у вигляді колекції зважених за частотою концептів з онтології. Потім було використано модель мережевого трафіка, яка є ефективним способом експліцитного вимірювання зваженої за частотою онтологічної відстані між концептами у двох текстах. Шляхом тестування розробленого методу в різних завданнях обробки природної мови було з'ясовано, що він дає хороші результати в двох із трьох завдань. Для того щоб мати змогу пояснити різницю в результатах використання методу на трьох різних наборах даних, було розроблено нову міру семантичної когерентності, яка пролила світло на характеристики набору даних, який якнайкраще підходить для запропонованого методу.

Переклад А. Синящик

Baroni, M. Distributional Memory: A General Framework for Corpus-Based Semantics [Дистрибутивна пам'ять: загальна методика корпусно-базованих досліджень семантики] / Marco Baroni, Alessandro Lenci // Computational linguistics. – 2010. – Vol. 36. – No. 4. – Pages 673–721. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli_a_00016#.WH4iSH3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli_a_00016

Корпусно-базовані дослідження семантики зосереджені на розробці спеціальних моделей, які обробляють окремі завдання або набори тісно пов'язаних завдань як розрізнені задачі, для вирішення яких потрібно видобути з корпусу різноманітну інформацію про сполучуваність. Методика дистрибутивної пам'яті, яка є альтернативою цьому підходу «одне завдання, одна модель», раз і назавжди видобуває з корпусу інформацію про сполучуваність у формі набору зважених строк слово-зв'язка-слово, організованих у тензор третього рангу. Після цього за допомогою тензора генеруються різні матриці, у чиїх рядках і стовпчиках зручно розв'язувати різні семантичні завдання. Таким чином, одна й та сама інформація про сполучуваність може використовуватись у різних завданнях, таких як моделювання суджень про подібність слів, виявлення синонімів, категорізація концептів, прогнозування сполучуваності дієслів, розв'язання проблем аналогії, класифікація відношень між парами слів, визначення смислових структур за допомогою моделей або пар прикладів, прогнозування типових характеристик концептів і класифікація дієслів. Широкомасштабне емпіричне тестування в усіх цих предметних областях свідчить, що методика дистрибутивної пам'яті конкурує зі спеціалізованими алгоритмами для таких самих завдань, нещодавно описаними в літературі, і з кількома новітніми методами. Таким чином, показано, що метод дистрибутивної пам'яті є

прийнятним, незважаючи на обмеження, накладені його багатоцільову природу.

Переклад В. Коломієць

Clarke, D. A Context-Theoretic Framework for Compositionality in Distributional Semantics [Контекстно-теоретична концепція композиційності в дистрибутивній семантиці] / Daoud Clarke // Computational linguistics. – 2012. – Vol. 38. – No. 1. – Pages 41–71. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00084#.WH4ivn3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00084

Результатом математичної формалізації "значення у вигляді контексту" є нова, алгебраїчна теорія значення із двохлінійною і сполучувальною композицією. Ці характеристики притаманні іншим методам, описаним у літературі, зокрема тензорному твору, векторному складанню, точковому множенню і матричному множенню.

Логічне слідування може бути представлене векторно-решітковим упорядкуванням на основі посиленої форми дистрибутивної гіпотези, а рівень логічного слідування визначається у формі умовної вірогідності. Наша концепція дозволяє описати підходи до завдання розпізнавання логічного слідування у тексті, зокрема застосування сполучення підланцюгів, вірогідності лексичного логічного слідування і латентного розміщення Діріхле.

Переклад В. Коломієць

Berant, J. Learning Entailment Relations by Global Graph Structure Optimization [Виявлення відношень логічного слідування шляхом оптимізації загальної структури графів] / Jonathan Berant, Ido Dagan, Jacob Goldberger // Computational linguistics. – 2012. – Vol. 38. – No. 1. – Pages 73–111. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00085#.WH4i9n3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00085

Важливою складовою прикладного семантичного виводу є виявлення відношень логічного слідування між предикатами. У статті запропоновано універсальний алгоритм логічного виводу, який виявляє правила такого логічного слідування. Спочатку у статті визначено структуру графа над предикатами, у якому відношення логічного слідування представлені у вигляді орієнтованих ребер. Потім до графа застосовано універсальне обмеження транзитивності з метою визначення оптимального набору ребер, і завдання оптимізації сформульоване як цілочислове лінійне програмування. Алгоритм застосований в умовах, у яких за наявності цільового концепта

алгоритм оперативно вивчає всі правила логічного слідування між предикатами, які зустрічаються разом із цим концептом. Результати свідчать, що у порівнянні з базовими алгоритмами запропонований універсальний алгоритм поліпшує результативність більше, ніж на 10%.

Переклад В. Коломієць

Fürstenau, H. Semi-Supervised Semantic Role Labeling via Structural Alignment [Напівконтрольоване анотування семантичних ролей шляхом структурного вирівнювання] / Hagen Fürstenau, Mirella Lapata // Computational linguistics. – 2012. – Vol. 38. – No. 1. – Pages 135-171. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00087#.WH4jPn3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00087

Необхідною передумовою розробки високоефективних систем анотування семантичних ролей є масштабні розмічені корпуси текстів. На жаль, створення таких корпусів дорого коштує, вони недостатньо великі і не можуть бути репрезентативними. Мета нашого дослідження полягає у полегшенні анотування, потрібного для створення ресурсів для розмітки семантичних ролей, шляхом навчання з частковим залученням учителя. Головна ідея нашого підходу полягає в тому, щоб знайти нові зразки для тренування класифікатора на основі їх схожості на розмічені вручну вихідні зразки. В основі лежить припущення, що фреймовий семантичний аналіз речень, однакових за лексичним матеріалом і синтаксичною структурою, співпадатиме. Знаходження однакових речень і присвоєння міток ролей формалізовані у вигляді проблеми вирівнювання графа, яка успішно вирішена за допомогою цілочислового лінійного програмування. Експериментальна перевірка анотування семантичних ролей свідчить, що автоматичне анотування за нашим методом є ефективнішим, ніж використання виключно розмічених вручну зразків.

Переклад В. Коломієць

Velldal, E. Speculation and Negation: Rules, Rankers, and the Role of Syntax [Припущення і заперечення: правила, ранжувальники і роль синтаксису] / Erik Velldal, Lilja Øvrelid, Jonathon Read, Stephan Oepen // Computational linguistics. – 2012. – Vol. 38. – No. 2. – Pages 369-410. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00126#.WH4jj33sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00126

У статті розглядається сукупність глибоких і поверхневих підходів до проблеми визначення діапазону припущень і заперечень у реченні, зокрема в літературі, присвяченій медико-біологічним дослідженням. Перша частина

статті присвячена припущенням. Продемонструвавши спочатку як можна точно визначити маркери припущень за допомогою дуже простого класифікатора, що використовує тільки локальний лексичний контекст, ми аналізуємо два різних синтаксичних підходи до визначення діапазонів цих сигналів у реченні. У той час як один підхід використовує створені вручну правила, що оперують структурами залежностей, другий автоматично опановує диференційну функцію ранжування за допомогою вузлів у піддеревах. Ми здійснюємо глибокий аналіз помилок, обговорюємо різні лінгвістичні особливості проблеми, і показуємо, що хоча обидва підходи добре працюють самі по собі, застосовуючи їх разом, можна отримати навіть кращі результати, які є найкращими з опублікованих результатів конкурсного завдання конференції з машинного навчання і обробки природних мов (Computational Natural Language Learning, скор. CoNLL) CoNLL-2010. У останній частині статті описано, як можна використати нашу систему визначення діапазону припущень для визначення діапазону заперечень. За допомогою зовсім незначної модифікації вихідної структури система дозволяє отримати прекрасні результати також і у вирішенні цього завдання.

Переклад М. Драчової

Gerber, M. Semantic Role Labeling of Implicit Arguments for Nominal Predicates [Маркування семантичних ролей імпліцитних аргументів номінативних присудків] / Matthew Gerber, Joyce Y. Chai // Computational linguistics. – 2012. – Vol. 38. – No. 4. – Pages 755–798. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00110#.WH6GsH3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00110

Іменні присудки часто містять імпліцитні аргументи. Останні праці, присвячені маркуванню семантичних ролей, зосереджувалися на знаходженні аргументів у локальному контексті присудка, проте досліджень, присвячених власне прихованим аргументам, не проводилось. Для того щоб закрити цю прогалину, було здійснено ручну розмітку корпусу імпліцитних аргументів десяти присудків із NomBank. Проаналізувавши цей корпус, ми з'ясували, що імпліцитні аргументи складають 71% усіх наявних у NomBank аргументів. За допомогою корпусу здійснено навчання дискримінаційної моделі, здатної визначати імпліцитні аргументи зі значенням F1-міри 50%, що значно перевершує результати навченої базової моделі. У статті описано проведене дослідження, проаналізовано широкий спектр характеристик, важливих для виконання завдання і розглянуто майбутні напрямки роботи над визначенням імпліцитних аргументів.

Переклад В. Коломієць

Shutova, E. Statistical Metaphor Processing [Статистична обробка метафор] / Ekaterina Shutova, Simone Teufel, Anna Korhonen //

Computational linguistics. – 2013. – Vol. 39. – No. 2. – Pages 301–353. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00124#.WH4kcn3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00124

Оскільки метафори часто зустрічаються в мові, їх комп'ютерна обробка є невід'ємною частиною реальних систем семантичної обробки природної мови. Попередні підходи до моделювання метафори використовували спеціальні, закодовані вручну знання і застосовувались у обмежених предметних областях або до підгрупи явищ. У статті вперше описано інтегровану статистичну модель обробки метафор у довільних текстах без обмежень у предметній області. Запропонований метод спочатку виявляє метафоричні вирази у основному тексті, а потім перефразує їх, використовуючи їх буквальні парафрази. Така модель інтерпретації метафори шляхом перефразування тексту сумісна з іншими системами обробки природної мови, які можуть виграти від розв'язання метафори. Запропонований метод передбачає мінімальне залучення учителя, спирається на найсучасніші методи синтаксичного аналізу і видобування лексики (розподілену кластеризацію і виведення вибіркових преференцій) і демонструє високу точність.

Переклад В. Коломієць

Liang, P. Learning Dependency-Based Compositional Semantics [Навчання композиційної семантики на основі залежностей] / Percy Liang, Michael I. Jordan, Dan Klein // Computational linguistics. – 2013. – Vol. 39. – No. 2. – Pages 389–446. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00127#.WH4k0n3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00127#.WH4k_H3sSGA

Уявімо, що хочемо створити систему, яка відповідає на питання природною мовою шляхом репрезентації її семантики як логічної форми і обчислення відповіді з урахуванням структурованої бази даних фактів. Головною частиною такої системи є семантичний аналізатор, який пов'язує питання і логічні форми. Семантичні парсери звичайно тренуються на прикладах питань з помітами їх цільових логічних форм, але цей різновид маркування є дорогим.

Наша мета натомість полягає в тому, щоб навчити семантичний парсер за допомогою пар питання-відповідь, у яких логічна форма представлена як прихована змінна. Розроблено новий семантичний формалізм, композиційна семантика на основі залежностей (*англ.* dependency-based compositional semantics, *скор.* DCS), і визначено логлінійну дистрибуцію логічних форм DCS.

Параметри моделі оцінюються за допомогою простої процедури, яка являє собою чергування променевого пошуку і числової оптимізації. На прикладі двох стандартних еталонних тестів показано, що наша система не поступається за точністю навіть найновішим системам, які потребують маркування логічних форм.

Переклад В. Коломієць

Bhagat, R. What Is a Paraphrase? [Що таке перифраза?] / Rahul Bhagat, Eduard Hovy // Computational linguistics. – 2013. – Vol. 39. – No. 3. – Pages 463–472. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00166#.WIE6jX3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00166

Перифрази – це речення або словосполучення, які виражають одне й те саме значення, використовуючи при цьому різні слова. Хоча згідно з визначенням, яке використовується у логіці, перифраза передбачає повну семантичну еквівалентність, у лінгвістиці допускається наближена еквівалентність, що значно збільшує кількість випадків "квазі-перифраз". Проте наближену еквівалентність важко визначити. Через це складно дати характеристику явищу перифрази у лінгвістиці. У статті описано 25 операцій, які дозволяють виявити квазі-перифрази. Масштаб охоплення і точність цього списку перевірено емпіричним шляхом за допомогою ручного аналізу випадкових вибірок з двох наявних у вільному доступі корпусів перифраз. Наведено розподіл квазі-перифраз, які зустрічаються в англійському тексті.

Переклад М. Погребної

Zapirain, B. Selectional Preferences for Semantic Role Classification [Обмеження сполучуваності для класифікації семантичних ролей] **/ Beñat Zapirain, Eneko Agirre, Lluís Màrquez, Mihai Surdeanu // Computational linguistics. – 2013. – Vol. 39. – No. 3 – Pages 631–663. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00145#.WH4lpH3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00145**

Стаття присвячена добре відомому невирішенному питанню у дослідженнях класифікації семантичних ролей: обмеженій ролі і рідкості лексичних характеристик. Проблему мінімізовано завдяки використанню моделей, які об'єднують автоматично виявлені обмеження сполучуваності. Досліджено декілька моделей на основі WordNet і обмежень сполучуваності за схожістю дистрибуції. Крім того показано, що завдання класифікації семантичних ролей краще моделювати за допомогою моделей обмеження сполучуваності на основі як дієслів, так і прийменників, а не самих дієслів. Експерименти з ізольованими моделями на основі обмежень сполучуваності

продемонстрували, що вони перевершили базову лексичну модель на 20 пунктів F1 у предметній області і майже на 40 пунктів F2 поза предметною областю. Також показано, що сучасна система класифікації семантичних ролей з додаванням функцій на основі обмежень сполучуваності працює значно краще як у межах предметної області (зменьшення кількості помилок на 17%), так і поза межами предметної області (зменьшення кількості помилок на 13%). Нарешті, показано, що у комплексній системі маркування семантичних ролей було отримано невеликі, але статистично значимі покращення, незважаючи на те, що наша модифікована модель класифікації семантичних ролей задіює лише приблизно 4% кандидатів у аргументи. Апостеріорний аналіз помилок свідчить, що функції на основі обмежень сполучуваності допомагають переважно в ситуаціях, де синтаксична інформація є або невірною, або недостатньою для визначення точної ролі.

Переклад В. Коломієць

Das, D. Frame-Semantic Parsing [Фреймово-семантичний синтаксичний аналіз] **/ Dipanjan Das, Desai Chen, André F. T. Martins, Nathan Schneider, Noah A. Smith // Computational linguistics. – 2014. – Vol. 40. – No. 1. – Pp. 9–56. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00163#.WH6LzH3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00163**

Фреймова семантика є лінгвістичною теорією, яка отримала практичне втілення для англійської мови у лексиконі FrameNet. Ми вирішили проблему фреймово-семантичного аналізу, використавши двоступеневу статистичну модель, яка знаходить лексичні мішені (тобто значущі слова і словосполучення) у контекстах речень і прогнозує фреймово-семантичні структури. Якщо мішень у контексті знайдена, на першому етапі вона перетворюється у семантичний фрейм. Щоб удосконалити перетворення у фрейми мішеней, які не зустрілись під час навчання, у цій моделі використовуються приховані змінні і напівконтрольоване навчання. На другому етапі відшукуються локально виражені семантичні аргументи мішені. Під час виведення швидкий точний подвійний алгоритм розбиття вираховує відразу всі аргументи фрейму з метою дотримання декларативно заявлених лінгвістичних обмежень, генеруючи структури вищої якості, ніж ненавчені локальні предиктори. Обидва компоненти спеціалізовані та спеціально навчені на невеликому наборі анотованих фреймово-семантичних розборів. На тестовому наборі даних семінару SemEval 2007 даний підхід, разом із евристичним ідентифікатором мішеней, які можна представити у вигляді фреймів, значно перевершив найсучасніший попередній аналізатор. Крім того, ми повідомляємо результати експериментів на набагато більшому наборі даних FrameNet 1.5. Наш фреймово-семантичний аналізатор є програмним забезпеченням із відкритим вихідним кодом.

Переклад О. Мартинюк, М. Погребної

Ó Séaghdha, D. Probabilistic Distributional Semantics with Latent Variable Models [Імовірнісна дистрибутивна семантика та моделі латентних змінних] / **Diarmuid Ó Séaghdha, Anna Korhonen // Computational linguistics. – 2014. – Vol. 40. – No. 3. – Pages 587–631. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00194#.WH4l733sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00194**

У статті описано ймовірнісний підхід до виявлення переважної сполучуваності лінгвістичних предикатів та використання отриманих представлень у моделюванні впливу контексту на значення слів. Наш підхід базується на використанні моделей латентних змінних Баєса, створених під впливом і на основі добре відомої моделі тематичної структури документів під назвою Латентне розміщення Діріхле (англ. Latent Dirichlet Allocation, скор. LDA); при роботі з даними предикат-аргумент, тематичні моделі автоматично виводять семантичні класи аргументів і приписують кожному предикату дистрибуцію в цих класах. У статті розглянуто LDA і цілий ряд розширень цієї моделі та здійснено їх оцінку за допомогою різних завдань семантичного прогнозування. Показано, що наш підхід забезпечує сучасний рівень продуктивності. Загалом стверджується, що ймовірнісні методи забезпечують ефективні й гнучкі дослідження дистрибутивної семантики.

Переклад Т. Павлущенко, М. Погребної

Lang, J. Similarity-Driven Semantic Role Induction via Graph Partitioning [Виведення семантичних ролей на основі схожості шляхом розбиття графа] / **Joel Lang, Mirella Lapata // Computational linguistics. – 2014. – Vol. 40. – No. 3. – Pages 633–669. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00195#.WH4mPH3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00195**

Як у багатьох завданнях з обробки природної мови, основним методом розмітки семантичних ролей стали керовані даними моделі на основі навчання з учителем. Ці моделі забезпечують високу продуктивність при достатній кількості розмічених тренувальних даних. Створення цих даних дорого коштує і забирає багато часу, тому виникає питання: чи є навчання без учителя гідною альтернативою? Робоча гіпотеза цього дослідження полягає в тому, що семантичні ролі можна вивести індуктивним шляхом без учителя з корпусу синтаксично размічених речень, керуючись трьома лінгвістичними принципами: (1) аргументи в одній синтаксичній позиції (в межах конкретного зв'язку) мають однакові семантичні ролі, (2) аргументи в межах підрядного речення мають особливі семантичні ролі, і (3) кластери, які представляють одну семантичну роль, повинні мати більш або менш рівнозначні лексичні значення і дистрибуцію. У статті описано метод, в якому втілено ці принципи і формалізовано визначення семантичних ролей у

вигляді проблеми розділення графа, в рамках якої окремі аргументи дієслова представлені як вершини графа, ребра якого виражають схожості між цими аргументами. Цей граф складається з багатьох рівнів ребер, кожен з яких виражає новий аспект схожості окремих аргументів, і для розбиття такого багаторівневого графа розроблено розширення стандартних алгоритмів кластеризації. Експерименти з англійською і німецькою мовами свідчать, що наш підхід дозволяє вивести індуктивним шляхом кластери семантичних ролей, які перевершують всі базові показники і можуть конкурувати з сучасними методами.

Переклад М. Погребної, І. Снєгурова

Grefenstette, E. Concrete Models and Empirical Evaluations for the Categorical Compositional Distributional Model of Meaning [Конкретні моделі та емпіричні оцінки категоріальної композиційної дистрибутивної моделі значення] / Edward Grefenstette, Mehrnoosh Sadrzadeh // Computational linguistics. – 2015. – Vol. 41. – No. 1. – Pages 71–118. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00209 – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00209

Для комп'ютерних лінгвістів моделювання композиційного значення речень із застосуванням емпіричних дистрибутивних методів завжди було завданням підвищеної складності. Категоріальна модель Кларка, Кука і Садрзаде (Clark, D. et al., 2008) та Кука, Садрзаде і Кларка (Coecke, B. et al., 2010) пропонує виконувати його шляхом об'єднання категоріальної граматики та дистрибутивної моделі значення. Вона враховує синтаксичні відношення під час виконання операцій компонування семантичних векторів. Але налаштування моделі є абстрактним. Відсутня оцінка моделі на основі емпіричних даних, вона не застосовувалась до жодних завдань обробки мови. Авторами створено конкретні моделі для вказаного налаштування шляхом створення алгоритмів для побудови тензорів та лінійних карт та підкріплення абстрактних параметрів емпіричними даними. Потім здійснено порівняння цих конкретних моделей з кількома експериментами, як відомими, так і новими, шляхом визначення, наскільки добре моделі узгоджуються з людськими судженнями при знаходженні парафрази. Результати дослідження показують, що в цих експериментах конкретне втілення застосування цієї загальної абстрактної моделі не поступається за результативністю іншим провідним моделям або перевершує їх.

Переклад М. Дубка

Zanzotto F. When the Whole Is Not Greater Than the Combination of Its Parts: A "Decompositional" Look at Compositional Distributional Semantics [Коли ціле не більше, ніж комбінація його частин: "Декомпозиційний" погляд на композиційну дистрибутивну семантику] / Fabio Massimo

Zanzotto, Lorenzo Ferrone, Marco Baroni // Computational linguistics. – 2015. – Vol. 41. – No. 1. – P. 165–173 – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00215 – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00215

Дистрибутивна семантика поширилася на словосполучення та речення за допомогою операцій складання. У статті розглянуто, як ці операції впливають на вимірювання подібності, і виявлено, що рівняння подібності важливого класу методів компоновки можна розкласти на операції, які виконуються на складових частинах вхідних словосполучень. Таким чином встановлюється міцний зв'язок між цими моделями та ядрами згортки.

Переклад А. Шульги

Shutova, E. Design and Evaluation of Metaphor Processing Systems [Проектування та оцінювання систем опрацювання метафор] / Ekaterina Shutova // Computational linguistics. – 2015. – Vol. 41. – No. 4. – Pages 579–623. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00233 – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00233

В опрацюванні природної мови (ОПМ) приділяють велику увагу методам розробки та оцінки систем, при цьому вони зазвичай оцінюються на основі стандартного завдання та загальних наборів даних. Це дозволяє здійснювати безпосереднє порівняння систем і сприяє розвитку галузі. Проте обчислення метафор значно більше фрагментоване, ніж аналогічні дослідження в інших галузях ОПМ і семантики. Протягом останніх років зріс інтерес до комп'ютерного моделювання метафор і з'явилося багат нових статистичних методів, що уможливлюють підвищення точності та надійності систем. Однак, відсутність визначення стандартного завдання, спільного набору даних та стратегії оцінювання ускладнює порівняння методів і тому перешкоджає спільному прогресу в цій галузі досліджень. Метою статті є огляд характеристик системи та стратегій оцінювання, які були запропоновані для завдання з опрацювання метафор, а також аналіз їх переваг та недоліків для визначення необхідних характеристик систем опрацювання метафор та набору вимог до їхньої оцінки.

Переклад А. Шульги

Hill, F. SimLex-999: Evaluating Semantic Models With (Genuine) Similarity Estimation [SimLex-999: оцінювання семантичних методів шляхом оцінки (справжньої) схожості]/ Felix Hill, Roi Reichart, Anna Korhonen // Computational linguistics. – 2015. – Vol. 41. – No. 4. – Pages 665–695. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00237 – Режим

доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00237

У статті представлено SimLex-999, еталонний ресурс для оцінки дистрибутивних семантичних методів, який перевершує існуючі ресурси в кількох важливих аспектах. По-перше, на відміну від золотих стандартів, таких як WordSim-353 та MEN, він експліцитно вираховує схожість, а не асоціативність чи пов'язаність, отже пари об'єктів, які асоціюються, але фактично не є схожими (Фрейд, психологія), мають низький рейтинг. Показано, що завдяки зосередженню на схожості, SimLex-999 стимулює розробку методів з різним, і, можливо, ширшим, спектром застосувань, ніж у методів, що відображають концептуальну асоціативність. По-друге, SimLex-999 містить низку конкретних і абстрактних пар прикметників, іменників та дієслів, а також незалежний рейтинг конкретності та (вільної) сили асоціативності для кожної пари. Ця різноманітність уможливлює детальний аналіз ефективності методів з концептами різних типів і, як наслідок, дає краще уявлення про те, яким чином можна вдосконалити методи. Крім того, на відміну від існуючих еталонних оцінювань, чию межу узгодженості між розмітниками автоматичні методи вже досягли або перевищили, сучасні методи демонструють значно гірші результати з SimLex-999. Отже, SimLex-999 має великі резерви для кількісного вираження майбутніх вдосконалень дистрибутивних семантичних методів, що спрямовуватиме розвиток наступного покоління методів на основі репрезентаційного машинного навчання.

Переклад М. Дубка

Boleda, G. Formal Distributional Semantics: Introduction to the Special Issue [Формальна дистрибутивна семантика: передмова до спеціального випуску] / Gemma Boleda, Aurélie Herbelot // Computational linguistics. – 2016. – Vol. 42. – No. 4. – Pages 619–635. – Режим доступу до анотації: https://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00261 – Режим доступу до повнотекстової статті: https://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00261

Формальна семантика і дистрибутивна семантика – це два дуже важливі семантичні методи в комп'ютерній лінгвістиці. Формальна семантика базується на символьній традиції та зосереджена навколо визначених шляхом умовиводів властивостей мови. Дистрибутивна семантика ґрунтується на статистичних та фактичних даних і зосереджується на аспектах значення, пов’язаних з описовим змістом. Ці два методи доповнюють сильні сторони одне одного, що і викликало зацікавленість у їх об’єднанні в один комплексний семантичний метод − «формальну дистрибутивну семантику». Проте, через принципову відмінність двох парадигм, створення інтеграційного методу пов’язане із значними теоретичними і технічними труднощами. Цей випуск журналу Computational Linguistics висвітлює

сучасний стан справ у формальній дистрибутивній семантиці; ця вступна стаття пояснює, з якою метою її було створено і підсумовує значимість попередніх публікацій з теми, забезпечуючи необхідну основу для опублікованих у випуску статей.

Переклад А. Шульги

Kruszewski G. There Is No Logical Negation Here, But There Are Alternatives: Modeling Conversational Negation with Distributional Semantics [Тут немає логічного заперечення, але є альтернативи: моделювання усного заперечення за допомогою дистрибутивної семантики] / Germán Kruszewski, Denis Paperno, Raffaella Bernardi, Marco Baroni // Computational linguistics. – 2016. – Vol. 42. – No. 4. – Pages 637–660. – Режим доступу до анотації: https://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00262 – Режим доступу до повнотекстової статті: https://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00262

Логічне заперечення є складним завданням для дистрибутивної семантики, оскільки предикати та їхні заперечення, як правило, зустрічаються в дуже подібних контекстах, отже, їх дистрибутивні вектори дуже схожі. Дійсно, навіть не зрозуміло, які саме властивості повинен мати дистрибутивний вектор, «який заперечується». Проте, коли лінгвістичне заперечення розглядається в його фактичному вживанні в дискурсі, воно часто виконує роль, яка дуже відрізняється від простого логічного заперечення. Якщо посеред розмови хтось заявляє, що «це не собака», заперечення явно передбачає обмежений набір альтернативних предикатів, які можуть бути вірними стосовно обговорюваного об’єкта. Зокрема, прийнятними альтернативами є інші представники родини псових і ссавці середнього розміру; птахи менш імовірні; хмарочоси та інші великі будівлі є практично неможливими. Усне заперечення діє як ступінчаста функція подібності, того роду, який можна легко виявити за допомогою дистрибутивної семантики. У цій статті представлено великий набір альтернативних рейтингів правдоподібності для усних заперечень іменних предикатів, а також показано, що проста подібність у дистрибутивному семантичному просторі забезпечує ідеальну відповідність суб’єктам даних. З одного боку, це заповнює прогалину в публікаціях, присвячених усному запереченню, і пропонує дистрибутивну семантику в якості правильного інструменту для прямих передбачень потенційних альтернатив заперечуваним предикатам. З другого боку, при розгляді в ширшому прагматичному аспекті результати показують, що заперечення є зовсім не проблемою, а ідеальною областю для застосування методів дистрибутивної семантики.

Переклад А. Шульги

Rimell, L. RELPRON: A Relative Clause Evaluation Data Set for Compositional Distributional Semantics [Набір підрядних означальних

речень для оцінки композиційної дистрибутивної семантики RELPRON] / Laura Rimell, Jean Maillard, Tamara Polajnar, Stephen Clark // Computational linguistics. – 2016. – Vol. 42. – No. 4. – Pages 661–701. – Режим доступу до анотації: https://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00263 – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00263

У статті представлено RELPRON, великий набір підрядних означальних речень із сполучними словами в ролі підмета і додатка, призначений для оцінювання методів композиційної дистрибутивної семантики. RELPRON орієнтований на серединний рівень граматичної складності між парами повнозначних слів і повними реченнями. Завдання передбачає співставлення термінів, таких як "мудрість", з репрезентативними властивостями, таким як "якість, набута завдяки досвіду". Унікальною особливістю RELPRON є те, що набір складається з перевірених властивостей, які не обов'язково вживаються у формі підрядного означального речення у вихідному корпусі. У статті також представлено деякі початкові експерименти на матеріалі RELPRON, в яких використано різноманітні композиційні методи, зокрема прості, такі як метод простих мінімальних основ для порівняння, метод арифметичних операторів на векторах, і більш складні методи, в яких слова в ролі аргументів представлено у вигляді тензорів. Останні методи базуються на детально описаному категоріальному підході. Отримані результати свідчать, що додавання векторів складно перевершити, що відповідає опублікованим даним, але використання категоріального підходу, який базується на моделі практичної лексичної функції, може зрівнятися по ефективності з додаванням векторів. Стаття завершується детальним аналізом RELPRON, який показує, як відрізняються результати для підрядних означальних речень із сполучними словами в ролі підметів та додатків, для різних іменників у ролі головних слів, і як вказані методи виконують проміжні завдання, необхідні для розуміння семантики підрядних означальних речень, а також забезпечення якісного аналізу, що висвітлює деякі з найбільш поширених помилок. Очікується, що представлені в статті конкурентоспроможні результати, в яких найкращі системи в середньому правильно ранжують кожну другу властивість певного терміна, сприятимуть появі нових підходів до завдання ранжування RELPRON та інших завдань на основі цікавих з лінгвістичної точки зору конструкцій.

Переклад М. Дубка

Asher, N. Integrating Type Theory and Distributional Semantics: A Case Study on Adjective–Noun Compositions [Інтегрування теорії типів і дистрибутивної семантики: дослідження прикладів сполучень прикметник-іменник] / Nicholas Asher, Tim Van de Cruys, Antoine Bride, Márta Abrusán // Computational linguistics. – 2016. – Vol. 42. – No. 4. – Pages 703–725. – Режим доступу до анотації:

http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00264 – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00264

У статті розглянуто інтеграцію формального семантичного підходу до лексичного значення і підходу на основі дистрибутивних методів. Спочатку коротко викладено формальну семантичну теорію, яка поєднує переваги як формального, так і дистрибутивного підходів. Після цього розроблено алгебраїчну інтерпретацію цієї формальної семантичної теорії і показано, як принаймні два види дистрибутивних моделей конкретизують цю інтерпретацію. Зосередивши увагу на сполученні прикметник-іменник, здійснено порівняння декількох дистрибутивних моделей з точки зору семантичної інформації, яка могла б знадобитися для формальної семантичної теорії, і показано, як знову використати інформацію, надану дистрибутивними моделями, у формальному семантичному підході.
У статті розглянуто інтеграцію формального семантичного підходу до лексичного значення і підходу на основі дистрибутивних методів. Спочатку коротко викладено формальну семантичну теорію, яка поєднує переваги як формального, так і дистрибутивного підходів. Після цього розроблено алгебраїчну інтерпретацію цієї формальної семантичної теорії і показано, як принаймні два види дистрибутивних моделей конкретизують цю інтерпретацію. Зосередивши увагу на сполученні прикметник-іменник, здійснено порівняння декількох дистрибутивних моделей з точки зору семантичної інформації, яка могла б знадобитися для формальної семантичної теорії, і показано, як знову використати інформацію, надану дистрибутивними моделями, у формальному семантичному підході.

Переклад М. Дубка

Weir, D. Aligning Packed Dependency Trees: A Theory of Composition for Distributional Semantics [Вирівнювання упакованих дерев залежностей: теорія композиції для дистрибутивної семантики] / David Weir, Julie Weeds, Jeremy Reffin, and Thomas Kober // Computational linguistics. – 2016. – Vol. 42. – No. 4. – Pages 727–761. – Режим доступу до анотації: https://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00265 – Режим доступу до повнотекстової статті: https://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00265

У статті представлено новий підхід до композиційної дистрибутивної семантики, в якому дистрибутивні оточення лексем виражаються у формі вивірених упакованих дерев залежностей. Показано, що ці структури можуть розпізнати повне оточення лексеми в реченні і є стандартною базою для об’єднання інформації про дистрибуцію таким чином, щоб забезпечити як одночасне зняття лексичної неоднозначності, так і узагальнення.

Переклад А. Шульги

Beltagy, I. Representing Meaning with a Combination of Logical and Distributional Models [Представлення значення за допомогою поєднання логічних і дистрибутивних моделей] / I. Beltagy, Stephen Roller, Pengxiang Cheng, Katrin Erk, Raymond J. Mooney // Computational linguistics. – 2016. – Vol. 42. – No. 4. – Pages 763–808. – Режим доступу до анотації: https://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00266 – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00266

Завдання опрацювання природної мови відрізняються необхідною для них семантичною інформацією, і на цей час жодне семантичне представлення не відповідає всім вимогам. Логічні представлення характеризують структуру речення, але не відображають градуйований аспект значення. Дистрибутивні моделі дають градуйовані оцінки подібності слів і фраз, але не відображають структуру речень так само детально, як логічні підходи. Тому стверджується, що ці два підходи є взаємодоповняльними.

У цій розвідці застосовано гібридний підхід, який поєднує логічну та дистрибутивну семантику, використовуючи ймовірнісну логіку, а саме логічні мережі Маркова (ЛММ). У статті розглянуто три компоненти прикладної системи: 1) метою логічного представлення є представлення вхідних задач за допомогою ймовірнісної логіки; 2) укладання бази знань створює зважені правила логічного виводу шляхом інтеграції дистрибутивної інформації та інших джерел; 3) ймовірнісний логічний вивід передбачає ефективне вирішення отриманих задач логічного виводу ЛММ. Для оцінювання запропонованого підходу використано завдання видобування з тексту імпліцитної інформації, яке уможливлює використання переваг як логічних, так і дистрибутивних представлень. Зокрема, описано базу даних SICK, завдяки якій вдалося отримати відмінні результати. Також представлено цінний ресурс для оцінювання систем видобування імпліцитної інформації на лексичному рівні – набір даних для видобування імпліцитної інформації на лексичному рівні, який складається з 10 213 правил, видобутих з бази даних SICK.

Переклад М. Дубка

Shutova, E. Multilingual Metaphor Processing: Experiments with Semi-Supervised and Unsupervised Learning [Багатомовне опрацювання метафор: експерименти з навчанням з мінімальним залученням учителя і без учителя] / Ekaterina Shutova, Lin Sun, Elkin Darío Gutiérrez, Patricia Lichtenstein, Srini Narayanan // Computational linguistics. – 2017. – Vol. 43. – No. 1. – Pages 71–123. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00275 – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00275

Метафора, яка часто зустрічається у мові та спілкуванні, є серйозною

проблемою для програм опрацювання природної мови. Традиційно обчислення метафор базувалося на застосуванні виконаної вручну розмітки, що ускладнювало розширення систем. В останні роки спостерігалося застосування статистичних підходів до опрацювання метафор. Проте ці підходи часто потребують масштабного ручного маркування і оцінюються здебільшого в обмеженій царині. У цьому дослідженні, навпаки, застосовано методи з незначним залученням учителя і без учителя – з обмеженим маркуванням або без нього – для визначення загальних методів обробки метафори на основі дистрибутивних властивостей понять. Досліджено різні рівні та види методів з учителем (навчання на основі лінгвістичних прикладів, навчання на основі заданого набору метафоричних представлень, а також навчання без маркування) з плоскими та ієрархічними, необмеженими та обмеженими налаштуваннями кластеризації. За мету поставлено визначення оптимального типу контролю для алгоритму навчання, який виявляє в тексті шаблони метафоричної асоціації. Для того, щоб дослідити розширюваність та адаптивність запропонованих методів, їх було застосовано до даних на трьох мовах з різних мовних груп – англійської, іспанської та російської. Було отримано високі результати з навчанням практично без учителя. Нарешті, показано, що статистичні методи можуть полегшити та розширити порівняльні дослідження метафори.

Переклад М. Дубка

Rothe S. AutoExtend: Combining Word Embeddings with Semantic Resources [AutoExtend: поєднання векторів представлення слів з семантичними ресурсами] / Sascha Rothe, Hinrich Schütze // Computational linguistics. – 2017. – Vol. 43. – No. 3. – Pages 593–617. – Режим доступу до анотації: https://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00294 – Режим доступу до повнотекстової статті: https://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00294

У статті представлено систему *AutoExtend*, яка об’єднує вектори представлення слів і семантичні ресурси шляхом автоматичної побудови векторів представлення несловесних об’єктів, таких як синсети та логічні категорії, і автоматичної побудови векторів представлення слів, які включають семантичну інформацію з ресурсу. Цей метод ґрунтується на кодуванні та декодуванні векторів представлення слів і характеризується гнучкістю, оскільки може опрацьовувати як вхідну інформацію будь-які вектори представлення слів і не потребує додаткового тренувального корпусу. Вихідні вектори представлення знаходяться в одному і тому ж векторному просторі, що і вхідні. Розріджена формалізація тензора гарантує ефективність та придатність до запаралелювання. В якості семантичних ресурсів у дослідженні використано WordNet, GermaNet і Freebase.

AutoExtend відповідає сучасним вимогам у виконанні завдань із визначення схожості слів у контексті і зняття лексичної багатозначності.

Переклад А. Шульги

Vulić, I. HyperLex: A Large-Scale Evaluation of Graded Lexical Entailment [HyperLex: великомасштабне оцінювання градуйованого лексичного логічного слідування] / Ivan Vulić, Daniela Gerz, Douwe Kiela, Felix Hill, Anna Korhonen // Computational linguistics. – 2017. – Vol. 43. – No. 4. – Pages 781–835. – Режим доступу до анотації: https://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00301 – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00301

У статті представлено HyperLex – набір даних і аналітичний ресурс, який кількісно визначає ступінь належності до семантичної категорії, тобто тип відношень, також відомий як гіперо-гіпонімія або відношення лексичного логічного слідування (ЛЛС), між 2616 парами понять. Дослідження в галузі когнітивної психології визначили, що типовість і належність до категорії/класу обчислюються в семантичній пам'яті людини як градуальне, а не бінарне відношення. Проте в більшості досліджень в галузі опрацювання природної мови та в існуючих великомасштабних інвентарях належності до понятійних категорій (WordNet, DBPedia тощо) категоріальна приналежність та ЛЛС вважаються бінарними. Для вирішення цієї проблеми на платформі краудсорсингу сотням носіїв англійської мови було запропоновано визначити типовість та міцність категоріальної приналежності серед різноманітних пар понять. Отримані результати підтверджують, що категоріальна приналежність та ЛЛС дійсно є більш градуальними, ніж бінарними. Також здійснено порівняння експертних оцінок з прогнозами автоматичних систем, яке виявило значні розбіжності між результатами експертної оцінки і сучасними моделями дистрибуції і навчання представленням на основі ЛЛС, а також суттєві відмінності між самими моделями. Обговорено шляхи вдосконалення семантичних моделей для подолання цієї невідповідності та вказано майбутні області застосування вдосконалених градуйованих систем ЛЛС.

Переклад М. Дубка

Автоматичний синтаксичний аналіз

Roark, B. Probabilistic Top-Down Parsing and Language Modeling [Імовірнісний низхідний синтаксичний аналіз і мовне моделювання] / Brian Roark // Computational linguistics. – 2001. – Vol. 27. – No. 2. – Pp. 249–276. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120101750300526#.WH57w33sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120101750300526

У статті описується робота універсального імовірнісного низхідного синтаксичного аналізатора і його застосування у розробці мовних моделей для систем розпізнавання мовлення. Спочатку в статті розглядаються ключові поняття мовного моделювання та імовірнісного синтаксичного аналізу, а також коротко описуються деякі попередні підходи до використання синтаксичної структури у мовному моделюванні. Далі описується лексикалізований імовірнісний низхідний синтаксичний аналізатор, який показує дуже гарні результати у порівнянні з найкращими універсальними статистичними синтаксичними аналізаторами як з точки зору правильності отриманих граматичних розборів, так і з точки зору ефективності їх знаходження. Потім описується нова мовна модель на основі імовірнісного низхідного синтаксичного аналізу. Емпіричні результати свідчать, що на відміну від попередніх моделей вона має кращу перплексивність на основі тренувального корпусу. Інтерполяція з триграмною моделлю дозволяє досягти набагато помітнішого покращення результатів, аніж застосування будь-якої іншої моделі, демонструючи ступінь розбіжності між інформацією, отриманою за допомогою нашої моделі синтаксичного аналізу, та інформацією, отриманою за допомогою триграмної моделі. Практичну цінність моделі підтверджує також невеликий експеримент із розпізнавання.

Переклад М. Драчової

Tomuro, N. Nonminimal Derivations in Unification-based Parsing [Немінімальні деривати у синтаксичному аналізі на основі уніфікаційних граматик] / Noriko Tomuro, Steven L. Lytinen // Computational linguistics. – 2001. – Vol. 27. – No. 2. – Pages 277–285. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120101750300535#.WH57Nn3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120101750300535

Запропонований для уніфікаційних граматик Б. Шибером [Shieber, 1992] алгоритм абстрактного синтаксичного аналізу є розширенням алгоритму

Ерли [Earley, 1970] для контекстно-вільних граматик з метою виділення структур. У статті показано, що за певних умов алгоритм Шибера утворює так званий немінімальний дериват: синтаксичне дерево з додатковими елементами, відсутніми у лицензійних наборах правил. Хоча надане Б. Шибером визначення синтаксичного дерева не виключає подібні немінімальні деривати, стверджується, що їх потрібно уважати помилкою. Описано джерела проблеми немінімальної деривації і запропоновано чітке визначення мінімального синтаксичного дерева і модифікацію алгоритму Шибера, яка забезпечує мінімалізм, щоправда за рахунок продуктивності.

Переклад В. Коломієць

Johnson, M. The DOP Estimation Method Is Biased and Inconsistent [Метод оцінювання синтаксичного аналізу, керованого даними, є необ'єктивним і непослідовним] / Mark Johnson // Computational linguistics. – 2002. – Vol. 28. – No. 1. – Pages 71 – 76. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120102317341783#.WH59DH3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120102317341783

Синтаксичний аналіз, керований даними, або модель керованого даними синтаксичного аналізу, поєднує фрагменти лінгвістичних репрезентацій з чисельними вагами, визначеними шляхом нормалізації емпіричної частоти кожного фрагмента в тренувальному корпусі (див. працю Bod, 1998 і цитовані у ній роботи). У статті повідомляється, що даний метод оцінювання є необ'єктивним і суперечливим. Інакше кажучи, зі збільшенням обсягу тренувального корпусу передбачуваний розподіл загалом не співпадає з реальним розподілом.

Переклад І. Снєгурова

Nederhof, M. Weighted Deductive Parsing and Knuth's Algorithm [Зважений дедуктивний синтаксичний аналіз і алгоритм Нута] / Mark-Jan Nederhof // Computational linguistics. – 2003. – Vol. 29. – No. 1. – Pages 135–143. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120103321337467#.WH59VX3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120103321337467

У статті аналізується зважений дедуктивний синтаксичний аналіз і розглядається проблема знаходження дериватів із найнижчою вагою. Показано, що здійснене Нутом узагальнення алгоритму Дійкстра для знаходження найкоротшого шляху є загальним методом вирішення вказаної проблеми. Наш підхід є модульним у тому сенсі, що алгоритм Нута формулюється незалежно від зваженої дедуктивної системи.

Переклад В. Коломієць

Oflazer, K. Dependency Parsing with an Extended Finite-State Approach [Розбір залежностей на основі концепції розширеного скінченного автомату] / Kemal Oflazer // Computational linguistics. – 2003. – Vol. 29. – No. 4. – Pages 515–544. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120103322753338#.WH59on3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120103322753338

У статті представлена схема розбору залежностей на основі концепції розширеного скінченного автомату. Аналізатор додає до вхідного представлення «канали», щоб можна було розмістити ребра, які відображають зв'язки синтаксичної залежності між словами, і аналізує вхідні дані багато разів для того, щоб отримати певний результат. Проміжні конфігурації, що порушують різні вимоги до проективних дерев залежностей, такі як відсутність ребер, що перетинаються, і відсутність незалежних елементів за винятком кореня дерева, фільтруються за допомогою фільтрів із скінченним числом станів. Парсер використовувався для синтаксичного аналізу турецької мови на основі граматики залежностей.

Переклад А. Синящик

Collins, M. Head-Driven Statistical Models for Natural Language Parsing [Вершинні статистичні моделі синтаксичного аналізу природної мови] / Michael Collins // Computational linguistics. – 2003. – Vol. 29. – No. 4. – Pages 589–637. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120103322753356#.WH597n3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120103322753356

У статті описано три статистичні моделі синтаксичного аналізу природної мови. Ці моделі розширюють методи від імовірнісних безконтекстних граматик до лексикалізованих граматик, утворюючи підходи, у яких дерево розбору представлене як послідовність рішень, яка відповідає розбудові дерева згори донизу, починаючи з вершини. Потім припущення про незалежність дозволяють отримати параметри, які програмують Х-штрих-схему, підкласифікацію, послідовність додатків, розміщення обставинних слів, лексичні залежності біграмів, переміщення питальних слів і вибір близького суміщення. Всі ці преференції виражені ступенями імовірності, зумовленими лексичними вершинами. Моделі оцінено за допомогою корпусу Penn Wall Street Journal Treebank і з'ясовано, що за точністю вони не поступаються іншим моделям, описаним у літературі. Для кращого розуміння цих моделей наведено результати для різних типів складників, а також розбивка показників точності й повноти у виявленні різних типів залежностей. Проаналізовано різні характеристики моделей шляхом експериментів з точністю синтаксичного аналізу, шляхом збирання частот різних структур у банку дерев, а також шляхом аналізу цікавих з

лінгвістичної точки зору прикладів. Нарешті, досліджувані моделі порівняно з іншими, які застосовувались у синтаксичному розборі банку дерев, для того щоб якось пояснити різницю між продуктивністю різних моделей.

Переклад А. Синящик

Hwa, R. Sample Selection for Statistical Parsing [Формування вибірки для статистичного синтаксичного аналізу] / Rebecca Hwa // Computational linguistics. – 2004. – Vol. 30. – No. 3. – Pp. 253–267. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120104185089 4#.WH5-UX3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/0891201041850894

Корпусно-базований статистичний синтаксичний аналіз спирається на використання великих обсягів анотованого тексту в якості прикладів для тренування. Створення такого ресурсу є дорогим і трудомістким процесом. У статті пропонується використовувати формування вибірки для знаходження корисних прикладів для тренування та скорочення людських зусиль, що витрачаються на анотування менш інформативних прикладів. Ми використовуємо декілька критеріїв, щоб спрогнозувати, чи можуть немарковані дані бути корисними навчальними прикладами. Для порівняння ефективності різних критеріїв прогнозування проведено експерименти із використанням двох синтаксичних тренувальних завдань і двох моделей навчання у рамках одного завдання синтаксичного аналізу. Ми виявили, що формування вибірки може значно зменшити обсяг анотованих тренувальних корпусів, і що невизначеність є надійним критерієм прогнозування, який можна легко застосувати для різних моделей навчання.

Переклад М. Драчової, В. Туз

Bikel, D. Intricacies of Collins' Parsing Model [Тонкощі моделі синтаксичного аналізу Коллінза] / Daniel M. Bikel // Computational linguistics. – 2004. – Vol. 30. – No. 4. – Pp. 479–511. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120104254492 9#.WH5-qH3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/0891201042544929

У статті наведено таку велику кількість досі неопублікованих подробиць про синтаксичний аналізатор Коллінза, що разом із дисертацією Коллінза (1999) вона містить усю інформацію, необхідну для відтворення результатів тестів Коллінза. Дійсно, ці досі неопубліковані дані пояснюють відносне збільшення на 11% похибки від імплементації, включаючи всі подробиці, до остаточної імплементації моделі Коллінза. Ми також демонструємо чистіший і однаково добре функціонуючий метод обробки пунктуації і сполучників і розкриваємо деякі інші ймовірнісні причуди синтаксичного аналізатора

Коллінза. Ми не тільки проаналізували значимість неопублікованих подробиць, але також здійснили повторний аналіз значимості деяких добре відомих деталей, з'ясувавши, що у моделі практично не використовуються подвійні лексичні залежності і що вибір стрижневого слова впливає на загальну продуктивність синтаксичного розбору менше, аніж уважалося раніше. Нарешті, ми провели експерименти, які свідчать, що справжня дискримінаційна потужність лексикалізації можливо полягає в тому, що генерація нелексикалізованих синтаксичних структур відбувається відповідно частиномовної приналежності стрижневого слова.

Переклад М. Драчової

Collins, M. Discriminative Reranking for Natural Language Parsing [Диференціальне переранжування у синтаксичному аналізі природних мов] / Michael Collins, Terry Koo // Computational linguistics. – 2005. – Vol. 31. – No. 1. – Pp. 25–70. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/0891201053630273#.WH5_An3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/0891201053630273

У статті розглядаються підходи до переранжування виходу існуючого імовірнісного синтаксичного аналізатора. Основний синтаксичний аналізатор виводить набір можливих граматичних розборів для кожного вхідного речення разом із відповідними ймовірностями, які визначають початкове ранжування цих граматичних розборів. Друга модель потім намагається поліпшити початкове ранжування, використовуючи для цього додаткові характеристики дерева. Сильною стороною нашого підходу є те, що він дозволяє розглядати дерево як довільний набір характеристик, без урахування того, як ці характеристики взаємодіють або перетинаються, і без необхідності визначати деривaційну або породжувальну модель, яка враховує ці характеристики. Ми пропонуємо новий метод переранжування на основі форсуючого підходу до проблем ранжування, описаного у роботі Й. Фройнда та ін. (Freund et al., 1998). Ми застосували форсуючий метод для синтаксичного аналізу банку дерев речень із газети Wall Street Journal. Даний метод являє собою комбінацію логарифмічної функції правдоподібності, яка використовується у базовій моделі (запропонованій М. Коллінзом (Collins, 1999)), і інформації про додаткові 500 тисяч характеристик синтаксичних дерев, які не брались до уваги у вихідній моделі. Нова модель досягла значення F-міри 89,75 %, відносного зниження на 13 % похибки F-міри порівняно з показником базової моделі — 88,2 %. У статті також представлено новий алгоритм форсуючого методу, у якому використано переваги обмеженої кількості характеристик синтаксичних дерев. Експериментально підтверджено, що новий алгоритм забезпечує значне зростання продуктивності у процесі активного використання форсуючого підходу. Ми уважаємо, що запропонований метод є привабливою альтернативою – як з точки зору простоти використання, так і з точки зору

продуктивності – розробці методів відбору характеристик у межах логлінійних моделей (моделей максимальної ентропії). Хоча експерименти, описані у цій статті, пов'язані із автоматичним синтаксичним аналізом природних мов, даний підхід може бути застосований для вирішення багатьох інших проблем автоматичної обробки природних мов, які звичайно формулюються як завдання ранжування, наприклад, розпізнавання мови, машинного перекладу або синтезу мови.

Переклад М. Драчової

Garmallo, P. Clustering Syntactic Positions with Similar Semantic Requirements [Кластеризація синтаксичних позицій із подібними семантичними вимогами] / Pablo Gamallo , Alexandre Agustini , Gabriel P. Lopes // Computational linguistics. – 2005. – Vol. 31. – No. 1. – Pp. 107–146. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/0891201053630318#.WH5_b33sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/0891201053630318

У статті описується неконтрольований метод визначення синтактико-семантичних вимог іменників, дієслів та прикметників на основі корпусу з частковою синтаксичною розміткою. Лінгвістичне поняття вимоги, що лежить в основі цього методу, базується на двох конкретних припущеннях. По-перше, уважається, що два слова, які знаходяться у відношеннях залежності, вимагають одне одного. Тут це явище називається взаємовимогою. По-друге, стверджується, що набір слів, які вживаються у схожих позиціях, повністю визначає вимоги, асоційовані з цими позиціями. Основною метою представленого у статті методу навчання є визначення кластерів схожих позицій шляхом визначення слів, які повністю встановлюють свої вимоги. Вказана стратегія дозволяє вивчати синтаксичні та семантичні вимоги слів у різних позиціях. Ця інформація використовується для розв'язання синтаксичної омонімії. В кінці статті проаналізовано результати виконання цього конкретного завдання. Числені експерименти проводились на базі корпусів португальської мови.

Переклад О. Мартинюк, М. Погребної

Kallmeyer, L. Tree-Local Multicomponent Tree-Adjoining Grammars with Shared Nodes [Багатокомпонентні граматики об'єднання дерев із спільними вузлами на початковому дереві] / Laura Kallmeyer // Computational linguistics. – 2005. – Vol. 31. – No. 2. – Pages 187–225. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/0891201054223968#.WIemeX3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/0891201054223968

У статті обговорюється питання про те, що виражальна сила граматик

об’єднання дерев (*англ.* tree-adjoining grammars, *скор.* TAGs) занадто обмежена, щоб упоратися з певними синтаксичними явищами, зокрема з перестановками у мовах із вільним порядком слів. Варіанти TAG, які досі пропонувалися для пояснення перестановок, не є досконалими. Тому у статті пропонується альтернативне розширення TAG на основі поняття спільних вузлів, так звана (обмежена) багатокомпонентна TAG із спільними вузлами на початковому дереві (*англ.* (restricted) tree-local multicomponent TAG with shared nodes, *скор.* RSN-MCTAG). Щоб довести, що це розширення TAG може впоратися з перестановками, коротко описується аналіз деяких перестановок у німецькій мові. Потім демонструється, що для певного типу RSN-MCTAG-граматик можна створити еквівалентні прості граматики склеювання інтервалів (*англ.* range concatenation grammars). Як наслідок, такі RSN-MCTAG-граматики є слабо контекстно-залежними і при цьому аналізуються за поліноміальний час. Ці специфічні RSN-MCTAG-граматики, можливо, можуть упоратися не з усіма перестановками, але з достатньо великою підмножиною.

Переклад В. Коломієць

Merlo, P. The Notion of Argument in Prepositional Phrase Attachment [Поняття аргумента у приєднанні прийменникової групи] / Paola Merlo, Eva Esteve Ferrer // Computational linguistics. – 2006. – Vol. 32. – No. 3. – Pages 341–378. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2006.32.3.341#.WIEiRn3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2006.32.3.341

У статті уточнено формулювання проблеми приєднання прийменникової групи як завдання чотирьохкрокового розв’язання неоднозначності. Стверджується, що для аналізу прийменникових груп потрібні знання про місце приєднання (традиційне розрізнення між приєднанням іменників і дієслів) і про сутність приєднання (розрізнення аргументів і ад’юнктів). Описано метод розпізнавання аргументів і ад’юнктів на основі визначення аргументів як вектора характеристик. У серії контрольованих класифікаційних експериментів спочатку досліджуються характеристики, які дозволяють встановити різницю між аргументами і ад’юнктами. З’ясовано, що у пригоді можуть стати як лінгвістична діагностика аргументів, так і лексичні семантичні класи. По-друге, досліджено найкращий метод здійснення чотирьохкрокової класифікації потенційно неоднозначних прийменникових фраз. З’ясовано, що хоча загалом краще вирішувати проблему як єдину задачу чотирьохкрокової класифікації, аргументи дієслів іноді розпізнаються точніше, якщо класифікація виконується як двокроковий процес: спочатку вибирається місце приєднання, а потім воно маркується як аргумент або ад’юнкт.

Переклад В. Коломієць

Atterer, M. Prepositional Phrase Attachment without Oracles [Підпорядкування прийменникових груп без оракулів] / Michaela Atterer, Hinrich Schütze // Computational linguistics. – 2007. – Vol. 33. – No. 4. – Pages 469–476. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2007.33.4.469#.WH6A4n3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2007.33.4.469

Дослідники проблеми підпорядкування прийменникової групи загалом уважають, що існує оракул, який генерує дві гіпотетичні структури, між якими потрібно зробити вибір. Інформація про існування двох можливих способів підпорядкування і інформація про лексичні вершини цих груп звичайно видобувається з еталонних дерев синтаксичного розбору. Показано, що з оракулом продуктивність методу перепідпорядкування є вищою, ніж без нього. Оскільки в програмному забезпеченні для обробки природної мови оракули відсутні, показники продуктивності, отримані за допомогою сучасної методики оцінювання підпорядкування прийменникових груп не є об'єктивними. Стверджується, що підпорядкування прийменникових груп потрібно оцінювати не ізольовано, а як невід'ємну частину системи синтаксичного розбору, не користуючись інформацією від еталонного оракула.

Переклад В. Коломієць

Clark, S. Wide-Coverage Efficient Statistical Parsing with CCG and Log-Linear Models [Широкомасштабний високоефективний статистичний синтаксичний аналіз на основі комбінаторної категорійної граматики і логлінійних моделей] / Stephen Clark, James R. Curran // Computational linguistics. – 2007. – Vol. 33. – No. 4. – Pp. 493–552. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2007.33.4.493#.WH6Bd33sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2007.33.4.493

У статті описано велику кількість логлінійних моделей синтаксичного аналізу для автоматично зґенерованої лексикалізованої грамматики. Ці моделі синтаксичного аналізу є «повними» у тому смислі, що імовірності визначаються для завершених розборів, а не для незалежних подій, отриманих шляхом розщеплення дерева розбору. Для оцінки моделей використовувалось дискримінаційне навчання, яке вимагало наявності у тренувальних даних не тільки правильного, але й неправильного дерева розбору для кожного речення. У якості лексикалізованого граматичного формалізму використовувалась комбінаторна категорійна граматика (Combinatory Categorial Grammar, скор. CCG), автоматично отримана з банку CCG, версії банку дерев Penn Treebank на основі CCG. Комбінація дискримінаційного навчання і автоматично отриманої граматики вимагає

значного обсягу пам'яті (до 25 ГБ), який забезпечувався шляхом паралельної реалізації алгоритму оптимізації BFGS, що виконувався на кластері Beowulf. Динамічне програмування при упакованій схемі, у комбінації з паралельною реалізацією, дозволило вирішити одне з наймасштабніших завдань оцінювання у літературі зі статистичного синтаксичного аналізу менш ніж за три години.

Ключовим компонентом системи синтаксичного аналізу, як для тренування, так і для тестування, є першокласний розмітник на основі методу максимальної ентропії, який приписує словам у реченні лексичні категорії CCG. Цей розмітник робить можливим дискримінаційне навчання, а також високоефективний синтаксичний аналіз. Незважаючи на «уявну неоднозначність» CCG, швидкість синтаксичного аналізу є значно вищою, ніж швидкість подібних синтаксичних аналізаторів у літературі. Також було удосконалено існуючі методи синтаксичного аналізу на основі CCG шляхом розробки нової моделі і ефективного алгоритму синтаксичного аналізу, який використовує усі відхилення, зокрема нестандартні відхилення CCG. Разом із обмеженнями нормальної форми ці модель і алгоритм синтаксичного аналізу забезпечують високу точність знаходження залежностей предикат-аргумент у банку CCG. Синтаксичний аналізатор також був протестований на банку дерев залежностей DepBank і порівняний із синтаксичним аналізатором RASP. Він показав кращі загальні результати і кращі результати для більшості типів залежностей. Тестування на банку дерев залежностей DepBank виявило багато проблем, пов'язаних із оцінкою синтаксичного аналізатора.

Стаття містить детальні рекомендації щодо розробки широкомасштабного синтаксичного аналізатора на основі CCG. Показано, що CCG може забезпечити точний і вісокоефективний синтаксичний аналіз.

Переклад В. Коломієць

Cahill, A. Wide-Coverage Deep Statistical Parsing Using Automatic Dependency Structure Annotation [Широкомасштабний глибокий статистичний синтаксичний аналіз із використанням автоматичної розмітки структури залежностей] / Aoife Cahill, Michael Burke, Ruth O'Donovan, Stefan Riezler, Josef van Genabith, Andy Way // Computational linguistics. – 2008. – Vol. 34. – No. 1. – Pp. 81–124. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2008.34.1.81#.WH6B5n3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2008.34.1.81

Останнім часом багато дослідників здійснили експериментальну перевірку синтаксичних аналізаторів – «глибокого», налаштованого вручну, широкомасштабного і «поверхового», на основі машинного навчання – на рівні дерев залежностей, використовуючи прості і автоматичні методи трансформації дерев, ґенерованих поверховими синтаксичними

аналізаторами, у дерева залежностей. У статті повторно розглядаються такі експерименти, цього разу із використанням складних автоматичних методів розмітки f-структур ЛФГ із цікавими результатами. Здійснено порівняння різних синтаксичних аналізаторів на основі імовірнісних контекстно-вільних граматик і на основі історії створення моделі з метою визначення базової системи синтаксичного аналізу, яка найкраще вписується у нашу метод автоматичної розмітки структури залежностей. Ця комбінована система синтаксичного аналізатора і розмітки структури залежностей порівнювалась із двома налаштованими вручну глибокими аналізаторами на основі обмежень, RASP і XLE. Оцінювання здійснювалось із використанням золотих стандартів на основі граматики залежностей, а статистична значущість результатів визначалась за допомогою наближеного критерію рандомізації. Проведені експерименти свідчать, що поверхові граматики на основі машинного навчання, удосконалені додаванням складних методів автоматичної розмітки структури залежностей, є ефективнішими, аніж створені вручну, глибокі, широкомасштабні граматики на основі обмежень. Зараз наша найкраща система має f-міру 82,73% на банку синтаксичних дерев PARC 700, що є статистично значимим поліпшенням на 2,18% останніх результатів 80,55% створеної вручну граматики ЛФГ і системи синтаксичного аналізу XLE і f-міру 80,23% на банку синтаксичних дерев CBS 500, що є статистично значимим поліпшенням на 3,66% результатів 76,57%, досягнутих створеною вручну граматикою і системою синтаксичного аналізу RASP.

Переклад В. Коломієць

Eryiğit, G. Dependency Parsing of Turkish [Синтаксичний аналіз турецької мови на основі граматики залежностей] / Gülşen Eryiğit, Joakim Nivre, Kemal Oflazer // Computational linguistics. – 2008. – Vol. 34. – No. 3. – Pp. 357–389. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2008.07-017-R1-06-83#.WH6CNn3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2008.07-017-R1-06-83

Важливою темою у синтаксичному аналізі є відповідність різних методів синтаксичного аналізу різним мовам. Особливий інтерес у цьому відношенні являють менш досліджені мови, типологічно відмінні від мов, для яких були розроблені методи. У статті описано дослідження керованого даними синтаксичного аналізу на основі граматики залежностей турецької мови, аглютинативної мови із вільним порядком складників, яку можна уважати типовим представником ширшого класу мов подібного типу. Проведені дослідження свідчать, що важливу роль у знаходженні синтаксичних відносин у такій мові відіграє морфологічна структура. Зокрема, показано, що використання у якості основних одиниць синтаксичного аналізу не словоформ, а сублексичних одиниць, відомих як *флективні групи,* підвищує точність аналізу. Це твердження тестувалося з допомогою двох різних

методів синтаксичного аналізу: одного на основі імовірнісної моделі з променевим пошуком, а другого на основі диференційних класифікаторів і детермінованого синтаксичного аналізу. Продемонстровано, що корисність сублексичних одиниць не залежить від методу обробки. Ретельно проаналізовано значення морфологічної і лексичної інформації, і продемонстровано, що за умови грамотного використання така інформація може значно підвищити точність синтаксичного аналізу. Завдяки використанню описаних у статті методів, було перевершено досягнуту попередніми дослідниками точність синтаксичного аналізу банку дерев турецької мови.

Переклад В. Коломієць

Nivre, J. Algorithms for Deterministic Incremental Dependency Parsing [Алгоритми детермінованого поетапного синтаксичного аналізу на основі дерев залежностей] / Joakim Nivre // Computational linguistics. – 2008. – Vol. 34. – No. 4. – Pp. 513–553. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.07-056-R1-07-027#.WH6C-33sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.07-056-R1-07-027

Алгоритми синтаксичного аналізу, які обробляють вхідні дані зліва направо і створюють єдине виведення часто уважалися неадекватними для обробки природної мови через численні неоднозначності, звичайно притаманні граматикам природної мови. Проте було доведено, що такі алгоритми, у поєднанні із класифікаторами на основі банків дерев, можуть бути використані для створення високоточних синтаксичних аналізаторів для зняття омонімії, зокрема для синтаксичних розборів на основі дерев залежностей. У статті спершу описано загальні принципи опису й аналізу алгоритмів детермінованого поетапного синтаксичного аналізу на основі дерев залежностей, оформленого як системи переходів. Потім описано і проаналізовано дві сім'ї таких алгоритмів: стекові алгоритми і алгоритми на основі списків. У першій сім'ї, яка обмежується проективними структурами залежностей, описано дугоспрямований і дугостандартний варіанти, а в другій сім'ї – проективний і непроективний варіанти. Для кожного з чотирьох алгоритмів наведено докази точності й складності. Крім того здійснено експериментальну перевірку всіх алгоритмів у комбінації з класифікаторами на основі методу опорних векторів для прогнозування наступної операції синтаксичного аналізу, використовуючи дані тринадцяти мов. Показано, що усі чотири алгоритми мають конкурентноспроможну точність, хоча непроективний алгоритм на основі списку звичайно перевершує проективні алгоритми для мов із значною долею непроективних конструкцій. Проте проективні алгоритми часто дають аналогічні результати у комбінації з методом, відомим як псевдо-проективний синтаксичний аналіз.

Лінійна часова складність стекових алгоритмів робить їх ефективнішими у навчанні і синтаксичному аналізі, але на практиці проективні алгоритми на

основі списків виявляються не менш ефективними. Більше того, коли проективні алгоритми використовуються для того, щоб здійснити псевдо-проективний синтаксичний аналіз, вони іноді стають менш ефективними у синтаксичному аналізі (але не у навчанні), ніж непроективні алгоритми на основі списків. Хоча більшість алгоритмів були частково описані у літературі раніше, це перший всебічний аналіз і оцінка алгоритмів у рамках єдиної концепції.

Переклад В. Коломієць

Schuler, W. Broad-Coverage Parsing Using Human-Like Memory Constraints [Використання властивих людині обмежень обсягу пам'яті у синтаксичному аналізаторі з широким покриттям] / William Schuler, Samir AbdelRahman, Tim Miller, Lane Schwartz // Computational linguistics. – 2010. – Vol. 36. – No. 1. – Pages 1–30. – Режим доступу до анотації : http://www.mitpressjournals.org/doi/abs/10.1162/coli.2010.36.1.36100#.WH6DYn3sSGA – Режим доступу до повнотекстової статті : http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2010.36.1.36100

За багатьма ознаками синтаксичний аналіз, який виконується людиною, здійснюється у короткотривалій пам'яті загального призначення. Проте відомо, що цей вид пам'яті має дуже малий обсяг, можливо обмежений всього трьома або чотирма окремими елементами. У статті описано модель синтаксичного аналізу, який успішно здійснюється в рамках таких жорстких обмежень шляхом розпізнавання складників у трансформованому представленні у правому куті (різновиді синтаксичного аналізу за лівим кутом) і з'єднання цього представлення з випадковими величинами у ієрархічній прихованій марківській моделі, зваженій послідовній моделі, яка прогнозує зміст обмеженого сховища пам'яті на тривалий час. Оцінка ефективності даної моделі за допомогою великого синтаксично анотованого корпусу англійських речень, а також точність створеної на основі цієї моделі методики синтаксичного аналізу з використанням обмеженої пам'яті дозволяють уважати модель цілком реальною.

Переклад В. Коломієць

Zhang, Y. Syntactic Processing Using the Generalized Perceptron and Beam Search [Синтаксичний аналіз за допомогою універсального перцептрона і променевого пошуку] / Yue Zhang, Stephen Clark // Computational linguistics. – 2011. – Vol. 37. – No. 1. – Pages 105–151. – Режим доступу до анотації : http://www.mitpressjournals.org/doi/abs/10.1162/coli_a_00037#.WH6EuX3sSGA – Режим доступу до повнотекстової статті : http://www.mitpressjournals.org/doi/pdf/10.1162/coli_a_00037

За допомогою загальної статистичної методики, яка складається з

універсальної лінійної моделі, навченої універсальним перцептроном і універсальним декодером з променевим пошуком, досліджено низку завдань синтаксичного аналізу. Цю методику застосовано до сегментування слів, одночасного сегментування і морфологічного розмічування, синтаксичного аналізу на основі граматики залежностей і синтаксичного аналізу структури словосполучення. Обидва компоненти методики є дуже простими у концептуальному і обчислювальному планах. Декодер з променевим пошуком вимагає тільки, щоб завдання синтаксичного аналізу було розділене на послідовність рішень для того, щоб на кожній стадії процесу декодер мав можливість розглянути перші N кандидатів і згенерувати усі можливі варіанти для наступної стадії. Відразу після налаштування декодер застосовується до тренувальних даних, використовуючи несуттєві оновлення відповідно універсального перцептрону для виведення моделі. Ця проста методика є дуже ефективною і за точністю результатів співставна з результатами діючих методик для всіх завдань, які ми розглянули.

Обчислювальна простота декодера і тренувального алгоритму забезпечила значно вищу швидкість тестування і менший час тренування, ніж їх основні альтернативи, зокрема логлінійний алгоритм, навчальний алгоритм із великим ступенем свободи і динамічне програмування для декодування. Крім того, запропонований метод дозволяє визначати довільні характеристики, які можуть неприпустимо уповільнювати альтернативні алгоритми навчання і декодування. Проаналізовано застосування загального методу до кожної з досліджуваних у статті проблем у порівнянні з альтернативними алгоритмами навчання і декодування. Також показано, що важливим фактором процесу є співставність кандидатів, які аналізуються променем. Стверджується, що концептуальна і обчислювальна простота та універсальність методу роблять його вигідним варіантом для виконання низки завдань синтаксичного аналізу і методом, який повинен обиратися для порівняння розробниками альтернативних підходів.

Переклад В. Коломієць

McDonald, R. Analyzing and Integrating Dependency Parsers [Аналіз і інтеграція синтаксичних аналізаторів залежностей] / Ryan McDonald, Joakim Nivre // Computational linguistics. – 2011. – Vol. 37. – No. 1. – Pp. 197–230. – Режим доступу до анотації : http://www.mitpressjournals.org/doi/abs/10.1162/coli_a_00039#.WH6E-H3sSGA – Режим доступу до повнотекстової статті : http://www.mitpressjournals.org/doi/pdf/10.1162/coli_a_00039

За останні п'ять років значно зросла кількість досліджень синтаксичних аналізаторів на основі граматики залежностей, які навчаються на прикладах із банків синтаксичних дерев. Це зростання було спричинене доступністю банків дерев для великої кількості мов – здебільшого завдяки конкурсним завданням конференції з машинного навчання і обробки природних мов (Computational Natural Language Learning, скор. CoNLL) – і зрозумілими

методами кодування складних явищ у мовах із вільним порядком слів, які використовуються у синтаксичних теоріях залежностей. Метою нашої статті є об'єктивна оцінка результатів цих досліджень шляхом аналізу двох основних парадигм синтаксичного аналізу на основі граматики залежностей, що керується даними, які часто називають синтаксичним аналізом на основі графів і синтаксичним аналізом на основі машин станів. Ми аналізуємо як теоретичні, так і емпіричні аспекти досліджень, і проливаємо світло на типи помилок, які роблять обидва типи синтаксичних аналізаторів, і їх обумовленість теоретичними припущеннями. Використовуючи ці спостереження, ми описуємо комбіновану систему на основі машинного навчання з використанням стекінгу і доводимо, що така система може навчитися позбавлятися недоліків кожної окремої системи.

Переклад М. Драчової

Gómez-Rodríguez, C. Dependency Parsing Schemata and Mildly Non-Projective Dependency Parsing [Схеми синтаксичного аналізу на основі граматики залежностей і слабо непроективний синтаксичний аналіз на основі граматики залежностей] / Carlos Gómez-Rodríguez, John Carroll, David Weir // Computational linguistics. – 2011. – Vol. 37. – No. 3. – Pages 541–586. – Режим доступу до анотації : http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00060#.WH6FoH3sSGA – Режим доступу до повнотекстової статті : http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00060

У статті описано схему синтаксичного аналізу на основі граматики залежностей, формальний метод на основі схеми синтаксичного аналізу К. Сіккела для синтаксичних аналізаторів на основі граматики складників, яку можна використати для того, щоб описати, проаналізувати і порівняти алгоритми синтаксичного аналізу на основі граматики залежностей. Даний метод було використано для опису кількох добре відомих проективних і непроективних синтаксичних аналізаторів на основі граматики залежностей, розробки доказів правильності і встановлення формального зв'язку між ними. Потім метод було використано для визначення нових поліноміальних алгоритмів синтаксичного аналізу для різних слабо непроективних граматик залежностей, зокрема глибоко вкладених структур, величина відстані між якими обмежена константою k за час O(n5+2k), і нового класу, який включає всі k структури величини відстані, наявні у кількох банках синтаксичних дерев природніх мов (які ми називаємо недостатньо глибоко вкладеними структурами величини відстані k), за час O(n4+3k). Нарешті, проілюстровано, як можна застосувати метод на основі схеми синтаксичного аналізу до граматики зв'язків, формалізму на основі залежностей.

Переклад В. Коломієць

Vadas, D. Parsing Noun Phrases in the Penn Treebank [Синтаксичний аналіз іменних груп у банку дерев Penn Treebank**] / David Vadas, James R.**

Curran // Computational linguistics. – 2011. – Vol. 37. – No. 4. – Pp. 753–809. – Режим доступу до анотації : http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00076#.WH6F6n3sSGA – Режим доступу до повнотекстової статті : http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00076

Іменні групи (ІГ) є важливою частиною природної мови і можуть мати дуже складну структуру. Проте ця структура ІГ здебільшого ігнорується у галузі статистичного синтаксичного аналізу, оскільки вони не марковані у корпусі, який використовується найчастіше. Ця відсутність золотого стандарту обмежувала попередні спроби здійснити синтаксичний аналіз ІГ, унеможливлюючи проведення контрольованих експериментів, які досягли високої ефективності у такій великій кількості завдань з обробки природної мови.

Ми повністю вирішили цю проблему шляхом ручного маркування структури ІГ у всій частині банку дерев Penn Treebank, яка складається із статей газети Wall Street Journal. Отримані нами показники міри узгодженості між маркувальниками спростовують переконання, що це занадто важке завдання, і демонструють, що послідовна розмітка ІГ можлива. Зараз наш золотий стандарт може бути використаний у всіх синтаксичних аналізаторах.

Ми експериментували з цими новими даними, використовуючи модель синтаксичного аналізу М. Коллінза [Collins, 2003], і з'ясували, що ефективність розпізнавання структури ІГ є значно нижчою, аніж загальна продуктивність моделі. F-міра цього аналізатора майже на 5,69 % нижча, аніж базового, який використовує детерміновані правила. Шляхом багатьох експериментів встановлено, що такий результат спричинений, у першу чергу, відсутністю лексичної інформації.

Щоб вирішити цю проблему, була створена широкозахватна, повномасштабна програма, яка бере ІГ у дужки. За допомогою нашої бази даних з банку дерев Penn Treebank, на кілька порядків більшої, ніж ті, що використовувалися раніше, ми створили контрольовану модель, яка демонструє чудових результатів. Наша модель має показник F-міри 93,8% при виконанні простих завдань, які виконувались у більшості попередніх досліджень, і на додаток бере у дужки довші, складніші ІГ, які рідко згадуються у літературі. Для цього складнішого завдання досягнутий показник F-міри 89,14%. Нарешті, упроваджено модель наступної обробки, яка бере у дужки ІГ, визначені аналізатором Д. Бікеля [Bikel, 2004]. Розроблена нами модель бракетування ІГ включає широкий спектр характеристик, які забезпечують лексичну інформацію, яка була відсутня у тестуваннях парсерів, і, в результаті, ми перевищуємо показник F-міри парсера 9,04%.

Вказані експерименти демонструють корисність корпусу і показують, що структура ІГ може зараз використовуватись у великій кількості програм обробки природної мови.

Переклад В. Коломієць

Nederhof, M. Splittability of Bilexical Context-Free Grammars is Undecidable [Розщеплюваність білексичних контекстно-незалежних граматик є нерозв'язною] / Mark-Jan Nederhof, Giorgio Satta // Computational linguistics. – 2011. – Vol. 37. – No. 4. – Pages 867–879. – Режим доступу до анотації : http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00079#.WH6GNX3sSGA – Режим доступу до повнотекстової статті : http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00079

Білексичні контекстно-незалежні граматики виявилися точними моделями для статистичного синтаксичного аналізу природної мови. Час обробки за допомогою існуючих алгоритмів динамічного програмування, які використовуються для синтаксичного розбору речень на основі цих моделей, складає O(□w□4), де w є строкою вводу.

Білексична контекстно-незалежна граматика є розщеплюваною, якщо ліві аргументи вершини завжди незалежні від правих аргументів і навпаки. Коли білексична контекстно-незалежна граматика є розщеплюваною, швидкість синтаксичного розбору можна асимптотично поліпшити до O(□w□3). Отже, дослідження цієї характеристики має надзвичайно важливе значення для ефективності синтаксичного аналізу. Але у статті показано негативний результат: розщеплюваність білексичних контекстно-незалежних граматик є нерозв'язною.

Переклад В. Коломієць

Roark, B. Finite-State Chart Constraints for Reduced Complexity Context-Free Parsing Pipelines [Скінченні табличні обмеження для спрощення конвейерного контекстно-незалежного синтаксичного аналізу] / Brian Roark, Kristy Hollingshead, Nathan Bodenstab // Computational linguistics. – 2012. – Vol. 38. – No. 4. – Pp. 719–753. – Режим доступу до анотації : http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00109#.WH6HQX3sSGA – Режим доступу до повнотекстової статті : http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00109

Ми описуємо методи зменшення найбільших і типових труднощів конвейерного контекстно-незалежного синтаксичного аналізу завдяки жорстким обмеженням, визначеним під час кінцевої попередньої обробки. Ми робили $O(n)$ прогнозів, щоб визначити, починає чи закінчує кожне слово вхідного речення у комірках таблиць багатослівний складник з двох або більше слів або чи допускає воно у комірках таблиць однослівні складники, представлені самим словом. Такі обмеження попередньої обробки прискорюють пошук будь-якого алгоритму синтаксичного аналізу на основі таблиць і істотно скорочують час декодування. У багатьох випадках, які ми назвали «закриттям» комірки таблиці, наповнення комірок зменшилося до нуля. Ми описуємо методи закриття достатньої кількості комірок таблиць для забезпечення переконливої квадратичної або навіть лінійної найбільшої

проблеми контекстно-незалежного висновку. Крім того, ми використовуємо високоточні обмеження для досягнення значних стандартних прискорень і об'єднуємо обмеження високої точності та обмеження для найскладніших проблем, щоб досягти найкращих результатів при обробці як коротких, так і довгих послідовностей. Такі обмеження обробки досягаються без зменшення точності обробки, а в деяких випадках точність підвищується. Ми показуємо, що наш метод підходить для численних граматик і є додатковим для інших методів скорочення, описуючи емпіричні результати як для точного, так і для приблизного висновку завдяки вичерпному алгоритму Кока — Янгера — Касамі, синтаксичному аналізатору Чарняка і берклійському синтаксичному аналізатору. Ми також повідомляємо результати аналізу китайської мови, де ми досягли найкращих зафіксованих результатів для окремої моделі на часто згадуваному наборі даних.

Переклад М. Драчової

Ballesteros, M. Going to the Roots of Dependency Parsing [Звернення до коренів синтаксичного аналізу на основі граматики залежностей] / Miguel Ballesteros, Joakim Nivre // Computational linguistics. – 2013. – Vol. 39. – No. 1. – Pp. 5–13. – Режим доступу до анотації : http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00132#.WH6H4X3sSGA – Режим доступу до повнотекстової статті : http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00132

Дерева залежностей, які використовуються у синтаксичному аналізі, часто включають корінь, представлений у вигляді пустого слова, приєднаного до початку або кінця речення, засобу, який зазвичай вважається звичайним технічим прийомом і не впливає на емпіричні результати. Ми показали, що це припущення є хибним і що точність керованих даними синтаксичних аналізаторів на основі дерев залежностей насправді може залежати від наявності й розташування пустого кореня. Зокрема, ми продемонстрували, що жадібний дугоспрямований синтаксичний аналізатор на основі машин станів, який здійснює обробку зліва направо, завжди працює гірше, коли пустий корінь розташований на початку речення (як прийнято зараз у керованому даними синтаксичному аналізі на основі дерев залежноестей), ніж коли він розташований у кінці або відсутній. Контрольні експерименти із дуговим синтаксичним аналізатором на основі машин станів і аналізатором на основі графу не виявили постійних преференцій, але, тим не менш, показали, що розташування кореня суттєво впливає на результати. Ми зробили висновок, що розташування пустих кореневих вузлів у керованому даними синтаксичному аналізі на основі дерев залежностей є недооціненим джерелом розбіжностей у експериментах і також може бути параметром, який потрібно налаштовувати для деяких синтаксичних аналізаторів.

Переклад І. Снєгурова

Tsarfaty, R. Parsing Morphologically Rich Languages: Introduction to the Special Issue [Автоматичний синтаксичний аналіз мов із розвиненою морфологією: передмова до спеціального випуску] / Reut Tsarfaty, Djamé Seddah, Sandra Kübler, Joakim Nivre // Computational linguistics. – 2013. – Vol. 39. – No. 1. – Pages 15–22. – Режим доступу до анотації : http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00133#.WH6IF33sSGA – Режим доступу до повнотекстової статті : http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00133

Автоматичний синтаксичний аналіз є головним завданням обробки природної мови. Він включає визначення для кожного речення природною мовою абстрактного представлення граматичних об'єктів у реченні і взаємовідносин між цими об'єктами. Це представлення забезпечує зв'язок з композиційною семантикою і з поняттями «хто кому що зробив». Протягом останніх двох десятиліть було досягнуто значних успіхів у автоматичному синтаксичному аналізі англійської мови, які призвели до значного покращення якості програм, основною частиною яких є синтаксичні аналізатори, таких як системи видобування інформації, аналізу тональності, реферування і машинного перекладу. Спроби відтворити успіх автоматичного синтаксичного аналізу англійської мови для других мов часто давали незадовільні результати. Зокрема, з'ясувалось, що автоматичний синтаксичний аналіз мов із складною будовою слова і вільним порядком слів потребує значної адаптації. У цьому спеціальному випуску повідомляється про методи успішного вирішення проблем, пов'язаних із синтаксичним аналізом різних мов із розвиненою морфологією. У передмові дається характеристика мов із розвиненою морфологією, описуються проблеми автоматичного синтаксичного аналізу мов із розвиненою морфологією і окреслюється основний зміст статей у спеціальному випуску. У статтях описано останні дослідження, присвячені автоматичному синтаксичному аналізу у різних міжмовних середовищах. Вони свідчать, що автоматичний синтаксичний аналіз мов із розвиненою морфологією стикається з проблемами, які виходять за рамки вибору конкретної репрезентації і алгоритму.

Переклад В. Коломієць

Seeker, W. Morphological and Syntactic Case in Statistical Dependency Parsing [Морфологічні і синтаксичні відмінки у статистичному синтаксичному аналізі на основі дерев залежностей] / Wolfgang Seeker, Jonas Kuhn // Computational linguistics. – 2013. – Vol. 39. – No. 1. – Pp. 23–55. – Режим доступу до анотації : http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00134#.WH6JL33sSGA – Режим доступу до повнотекстової статті : http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00134

Більшість морфологічно багатих мов із вільним порядком слів

використовують системи відмінків для позначення граматичної функції іменних елементів, особливо основних аргументів дієслова. Стандартний поетапний підхід до синтаксичного аналізу на основі дерев залежностей передбачає повне зняття морфологічної (відмінкової) омонімії до здійснення автоматичного синтаксичного аналізу. Експериментальний синтаксичний аналіз чеської, німецької та угорської мов показав, що цей підхід може привести до помилок у морфологічній розмітці під час синтаксичного аналізу мов, для яких характерний синкретизм у морфологічних відмінкових парадигмах. Ми розробили іншу модель, у якій відмінок використовується як можливо недостатньо визначений фільтруючий механізм, що обмежує варіанти синтаксичного аналізу. Ретельно розроблені морфо-синтаксичні обмеження можуть обмежити пошуковий простір статистичного синтаксичного аналізатора на основі дерев залежностей і виключати рішення, які порушили б обмеження, явно зазначені у частиномовній приналежності слів у даному реченні. Ми експериментально доводимо, що обмежена система перевершує найсучаснішу поетапну модель на основі даних, а, також, що вивід синтаксичного аналізатора є гарантовано вільним від локальних і глобальних морфо-синтаксичних помилок, що може бути корисним для наступних прикладних програм.

Переклад В. Туз

Fraser, A. Knowledge Sources for Constituent Parsing of German, a Morphologically Rich and Less-Configurational Language [Набори правил і процедур для заснованого на складниках синтаксичного аналізу німецької мови, морфологічно багатої мови з менш усталеним порядком слів] / Alexander Fraser, Helmut Schmid, Richárd Farkas, Renjing Wang, Hinrich Schütze // Computational linguistics. – 2013. – Vol. 39. – No. 1. – Pp. 57–85. – Режим доступу до анотації : http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00135#.WH6Ji33sSGA – Режим доступу до повнотекстової статті : http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00135

Досліджувався синтаксичний аналіз на основі складників німецької мови, яка є морфологічно багатою мовою з менш усталеним порядком слів. Використовувалась імовірнісна контекстно-вільна граматика на основі банку дерев, адаптована до числених морфологічних особливостей німецької мови шляхом марковізації і додавання спеціальних характеристик до її продукцій. Здійснена оцінка результативності додавання лексичної інформації. Також проаналізовано як монолінгвальний, так і білінгвальний підходи до переранжування розбору. Запропонована система переранжування є новою найсучаснішою системою у заснованому на складниках синтаксичному аналізі банку дерев Tiger. Здійснено аналіз, який завершується висновками, що стосуються синтаксичного аналізу інших морфологічно багатих мов з менш усталеним порядком слів**.**

Переклад В. Коломієць

Kallmeyer, L. Data-Driven Parsing using Probabilistic Linear Context-Free Rewriting Systems [Керований даними синтаксичний аналіз за допомогою імовірнісних лінійних контекстно-незалежних систем переписування] / Laura Kallmeyer, Wolfgang Maier // Computational linguistics. – 2013. – Vol. 39. – No. 1. – Pp. 87–119. – Режим доступу до анотації : http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00136#.WH6Jwn3sSGA – Режим доступу до повнотекстової статті : http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00136

У статті представлена перша ефективна реалізація виваженого дедуктивного синтаксичного аналізатора на основі алгоритму Кока — Янгера — Касамі для імовірнісних лінійних контекстно-незалежних систем переписування (Probabilistic Linear Context-Free Rewriting System — PLCFRS). Лінійна контекстно-незалежна система переписування (LCFRS), розширення контекстно-незалежної граматики (Context-Free Grammar — CFG), може ефективно описувати порушення однорідності і тому ідеально підходить для використання у синтаксичному аналізі, який керується даними. Для прискорення процесу синтаксичного аналізу ми використовували різні розрахунки об'єктів аналізу на основі контексту, деякі з яких допускали синтаксичний аналіз A*. Аналізатор тестувався за допомогою граматик, отриманих із німецького банку дерев NeGra. Наші експерименти свідчать, що керований даними синтаксичний аналіз для лінійної контекстно-незалежної системи переписування є здійснимим і дає результати конкурентоспроможної якості.

Переклад М. Драчової

Goldberg, Y. Word Segmentation, Unknown-word Resolution, and Morphological Agreement in a Hebrew Parsing System [Сегментування слів, розпізнавання незнайомих слів і морфологічне узгодження у синтаксичному аналізаторі івриту] / Yoav Goldberg, Michael Elhadad // Computational linguistics. – 2013. – Vol. 39. – No. 1. – Pp. 121–160. – Режим доступу до анотації : http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00137#.WH6KBX3sSGA – Режим доступу до повнотекстової статті : http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00137

У статті описано систему синтаксичного аналізу на основі граматики складників для сучасного івриту. Система заснована на запропонованому С. Петровим та ін. (Petrov et al., 2006) методі синтаксичного аналізу на основі імовірнісної контекстно-незалежної граматики з прихованими анотаціями (Probabilistic Context-Free Grammar With Latent Annotations, скор. PCFG-LA), який зазнав різноманітних уточнень з метою врахування особливостей івриту як морфологічно багатої мови з невеликим банком дерев. Ми показуємо, що результати синтаксичного аналізу можна поліпшити завдяки використанню лінгвістичного ресурсу, відмінного від банку дерев, а саме морфологічного

аналізатора на основі лексикону. Ми описуємо комбіновану обчислювальну модель зовнішнього лексикону і синтаксичного аналізатора на основі банку дерев, також у типовому випадку, коли у лексиконі і банку дерев використовуються різні схеми анотування. Ми показуємо, що можна одночасно здійснювати сегментування слів івриту і синтаксичний аналіз на основі граматики складників, використовуючи ґратчастий синтаксичний аналіз на основі алгоритму Кока — Янгера — Касамі. Одночасне виконання завдань ефективне і істотно перевершує показники конвейерної моделі. Ми пропонуємо моделювати граматичне узгодження у синтаксичному аналізаторі на основі граматики складників як ортогональний граматиці механізм фільтра і представляємо конкретну реалізацію цього методу. Хоча синтаксичний аналізатор на основі граматики складників не робить великої кількості помилок в узгодженні, механізм фільтра ефективно виправляє ті помилки узгодження, які аналізатор таки допускає.

Отримані результати виходять за рамки обробки івриту і можуть бути широко застосовані у обробці природної мови. Іврит є конкретним прикладом морфологічно багатої мови і ідеї, висунуті у цій роботі, також корисні для обробки інших мов, зокрема англійської. Методика ґратчастого синтаксичного аналізу корисна у будь-яких випадках, коли інформація на вході неоднозначна. Розширення лексичного покриття синтаксичного аналізатора на основі банку дерев завдяки використанню зовнішнього лексикону потрібне для будь-якої мови із невеликим банком дерев.

Переклад М. Драчової

Marton, Y. Dependency Parsing of Modern Standard Arabic with Lexical and Inflectional Features [Синтаксичний аналіз сучасної літературної арабської мови на основі граматики залежностей за допомогою лексичних і флексійних характеристик] / Yuval Marton, Nizar Habash, Owen Rambow // Computational linguistics. – 2013. – Vol. 39. – No. 1. – Pp. 161–194. – Режим доступу до анотації : http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00138#.WH6Hj33sSGA – Режим доступу до повнотекстової статті : http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00138

Ми досліджували вплив особливостей лексичної і флективної морфології на синтаксичний аналіз на основі граматики залежностей арабської мови, морфологічно багатої мови зі складними моделями узгодження. Використовуючи контрольовані експерименти, ми співставили використання різних наборів частиномовних тегів і морфологічних характеристик у двох вхідних станах: машинно-передбаченому стані (у якому частиномовні теги і значення морфологічних характеристик присвоюються автоматично) і золотому стані (в якому їх справжні значення відомі). Ми з'ясували, що більш інформативні (точні) набори тегів корисні у золотому стані, але можуть бути згубними у машинно-передбаченому стані, у якому більш ефективними є прості, але точніше передбачені теги. Ми визначили набір

характеристик (означеність, особа, число, рід і неогласована лема), який покращує якість синтаксичного аналізу у машинно-передбаченому стані, в той час як інші характеристики корисніші у золотому стані. Ми вперше продемонстрували, що у синтаксичному аналізі корисніші функціональні характеристики роду і числа (наприклад, «ламана множина») і, можливо, близька характеристика розумності («людськості»), аніж форми роду і числа. Нарешті, ми довели, що якість синтаксичного аналізу в передбаченому стані можна значно покращити навчанням у комбінованому золотому+передбаченому стані. Ми експериментували з двома синтаксичними аналізаторами, які працюють на основі машин станів, MaltParser і Easy-First Parser. Наші висновки стабільні і не залежать від аналізаторів, моделей і вхідних станів. Це наводить на думку, що вплив лінгвістичної теорії у формі наборів тегів і виділених нами характеристик не обмежений рамками конкретних експериментальних досліджень і може бути корисним для інших синтаксичних аналізаторів і морфологічно багатих мов.

Переклад М. Драчової

Green, S. Parsing Models for Identifying Multiword Expressions [Моделі синтаксичного аналізу для розпізнавання багатослівних виразів] / Spence Green, Marie-Catherine de Marneffe, Christopher D. Manning // Computational linguistics. – 2013. – Vol. 39. – No. 1. – Pp. 195–227. – Режим доступу до анотації : http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00139#.WH6KW33sSGA – Режим доступу до повнотекстової статті : http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00139

Багатослівні вирази знаходяться на межі синтаксису і семантики і є причиною появи альтернативних теорій синтаксису, наприклад конструкційної граматики. Проте у обробці природної мови синтаксичний аналіз і розпізнавання багатослівних виразів досі моделювались окремо. Ми розробили дві структуровані прогностичні моделі для одночасного синтаксичного аналізу і розпізнавання багатослівних виразів. Перша заснована на контекстно-вільних граматиках, а друга використовує граматики заміщення дерев, формалізм, що дозволяє зберігати синтаксичні фрагменти більшого обсягу. Наші експерименти показують, що обидві моделі можуть розпізнавати багатослівні вирази набагато точніше, ніж найсучасніша система, заснована на статистичних даних про спільну появу слів.

Ми експериментували з арабською та французькою мовами, для кожної з яких характерні багатослівні вирази. На відміну від англійської мови, вони також мають багатшу морфологію, яка є причиною лексичної розрідженості у обмежених корпусах. Щоб подолати цю розрідженість, ми розробили просте факторне лексичне представлення контекстно-вільної моделі синтаксичного аналізу. Результати морфологічного аналізу автоматично перетворюються на теги з великою кількістю характеристик, прикріплені до

лексичних одиниць. Цей метод, який ми називаємо факторною лексикою, покращує як точність стандартного синтаксичного аналізу, так і точність розпізнавання багатослівних виразів.

Переклад М. Драчової

Gómez-Rodríguez, C. Divisible Transition Systems and Multiplanar Dependency Parsing [Ділимі системи переходів і мультипланарний синтаксичний аналіз на основі дерев залежностей] / Carlos Gómez-Rodríguez, Joakim Nivre // Computational linguistics. – 2013. – Vol. 39. – No. 4. – Pp. 799–845. – Режим доступу до анотації : http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00150#.WH6LCH3sSGA – Режим доступу до повнотекстової статті : http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00150

Синтаксичний аналіз на основі машин станів є популярним підходом у синтаксичному аналізі на основі дерев залежностей, який забезпечує високу ефективність вузькоспеціалізованих аналізаторів. Існує багато різних аналізаторів на основі машин станів, часто формалізованих в рамках дещо різних теорій. У статті показано, що велику кількість відомих систем проективного синтаксичного аналізу на основі дерев залежностей можна розглядати як варіанти однієї стекової системи з невеликим набором елементарних переходів, які можуть бути об'єднані у складні переходи і обмежені різними способами. Ми називаємо такі системи стековими системами переходів і підтверджуємо велику кількість теоретичних висновків стосовно їх точності та складності. Зокрема, ми описуємо важливий підклас, відомий як ефективні ділимі системи переходів, які аналізують планарні графи залежностей у лінійному часі. Далі ми показуємо, по-перше, як можна обмежити цю систему, щоб вона аналізувала саме набір планарних дерев залежностей, і по-друге, як можна узагальнити цю систему до *k*-планарних дерев, використовуючи численні стеки. Використовуючи перший відомий ефективний тест *k*-планарності, ми досліджуємо, як система знаходить *k*-планарні дерева у доступних банках дерев і виявили, що вона дуже добре працює з 2-планарними деревами. В кінці ми здійснюємо експериментальну перевірку і показуємо, що наш 2-планарний синтаксичний аналізатор дозволяє досягти істотного поліпшення якості синтаксичного аналізу у порівнянні з відповідними 1-планарним і проективним синтаксичними аналізаторами для масивів даних із непроективними деревами залежностей і працює нарівні з широко використовуваним дугоспрямованим псевдопроективним синтаксичним аналізатором.

Переклад М. Драчової

Henderson, J. Multilingual Joint Parsing of Syntactic and Semantic Dependencies with a Latent Variable Model [Багатомовний об’єднаний синтаксичний аналіз синтаксичних і семантичних залежностей за допомогою моделі з прихованою змінною] / James Henderson, Paola

Merlo, Ivan Titov, Gabriele Musillo // Computational linguistics. – 2013. – Vol. 39. – No. 4. – Pp. 949–998. – Режим доступу до анотації : http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00158#.WH6LSH3sSGA – Режим доступу до повнотекстової статті : http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00158

Сучасні дослідження моделей синтаксичного аналізу, керованих даними, перейшли від виключно синтаксичного аналізу до інтенсивніших семантичних представлень, показуючи, що успішне розуміння смислу тексту вимагає структурованого аналізу як його граматики, так і його семантики. У статті повідомляється про об'єднану породжувальну модель на основі передісторії для прогнозування найвірогіднішого дерева виведення синтаксичного аналізатора на основі дерев залежностей як для синтаксичних, так і для семантичних залежностей у різних мовах. Оскільки ці дві структури залежностей не ізоморфні, ми пропонуємо слабку синхронизацію на рівні значущих підпослідовностей двох дерев виведення. Ці синхронізовані підпослідовності містять інформацію про ліве оточення кожного окремого слова. Ми також пропонуємо інноваційні виведення семантичних структур залежностей, які відповідають відносно вільній природі цих графів. Для навчання об'єднаної моделі цих синхронизованих виведень ми використовуємо модель синтаксичного аналізу із прихованою змінною – модель Incremental Sigmoid Belief Network (ISBN). Ця модель продукує представлення прихованих властивостей у деревах виведень, які використовуються для виявлення взаємозв'язків як усередині двох дерев виведення, так і між ними, вперше використовуючи ISBN для розв'язання проблеми багатозадачного навчання. Ця об'єднана модель демонструє конкурентноздатний рівень як синтаксичного, так і семантичного аналізу різних мов. Завдяки загальному характеру нашого методу, вказане застосування моделі ISBN для аналізу слабо синхронізованих синтактико-семантичних дерев виведення також свідчить про можливість її застосування для вирішення інших проблем, коли йдеться про опанування двома незалежними, але спорідненими представленнями.

Переклад І. Снегурова, М. Погребної

Demberg, V. Incremental, Predictive Parsing with Psycholinguistically Motivated Tree-Adjoining Grammar [Поетапний, прогностичний синтаксичний аналіз на основі психолінгвістично обумовленої граматики з'єднання дерев] / Vera Demberg, Frank Keller, Alexander Koller // Computational linguistics. – 2013. – Vol. 39. – No. 4. – Pp. 1025–1066. – Режим доступу до анотації : http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00160#.WH6Li33sSGA – Режим доступу до повнотекстової статті : http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00160

Психолінгвістичні дослідження свідчать, що головними характеристиками

обробки речення людиною є поетапність, зв'язність (спрощені дерева не мають неприєднаних вузлів) і прогнозування (наступна синтаксична структура є очікуваною). Проте універсальної моделі синтаксичного аналізу із вказаними характеристиками поки немає. У статті описано перший універсальний імовірнісний синтаксичний аналізатор на основі психолінгвістично мотивованої граматики з'єднання дерев (PsychoLinguistically motivated Tree-Adjoining Grammar, *скор.* PLTAG), модифікованої граматики з'єднання дерев (Tree-Adjoining Grammar, *скор.* TAG), яка задовольняє всім трьом умовам. Тренування аналізатора здійснювалось на модифікованій за правилами граматики з'єднання дерев версії синтаксично анотованого корпусу Penn Treebank. Продемонстровано, що він працює так само, як існуючі аналізатори на основі TAG, які є поетапними, але не мають прогностичної сили. Запропонована модель PLTAG також використовувалась для прогнозування швидкості обробки тексту людиною і показала кращі результати на відеоокулографічному корпусі Данді, ніж стандартна модель несподіваності.

Переклад В. О. Коломієць

Nivre, J. Constrained Arc-Eager Dependency Parsing [Обмежений дугоспрямований синтаксичний аналіз на основі граматики залежностей] / Joakim Nivre, Yoav Goldberg, Ryan McDonald // Computational linguistics. – 2014. – Vol. 40. – No. 2. – Pp. 249–257. – Режим доступу до анотації : http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00184#.WH6MUn3sSGA – Режим доступу до повнотекстової статті : http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00184

Дугоспрямовані синтаксичні аналізатори на основі граматики залежностей обробляють речення за один перегляд вхідних даних зліва направо і характеризуються лінійною часовою складністю із жадібним декодуванням або променевим пошуком. Ми показуємо, як можна обмежити такі аналізатори, щоб ураховувати два різні типи умов до обмежень кістяка вихідного графа залежностей, що вимагають, щоб певні кістякові дерева відповідали піддеревам графа, і до обмежень дуг, що вимагають наявності у графі певних дуг. Обмеження вбудовано у дугоспрямований аналізатор на основі машин станів як набір вихідних умов для кожного переходу, вони зберігають лінійну часову складність синтаксичного аналізатора.

Переклад М. Драчової

Nivre, J. Arc-Eager Parsing with the Tree Constraint [Дугоспрямований синтаксичний аналіз із обмеженням дерев] / Joakim Nivre, Daniel Fernández-González // Computational linguistics. – 2014. – Vol. 40. – No. 2. – Pages 259–267. – Режим доступу до анотації : http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00185#.WH6Mhn3sSGA – Режим доступу до повнотекстової статті :

http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00185

Дугоспрямована система синтаксичного аналізу на основі машин станів широко використовується в обробці природних мов незважаючи на те, що вона не гарантує отримання правильно побудованого дерева залежностей на виході. Ми пропонуємо нескладну модифікацію оригінальної системи, яка забезпечує обмеження дерев без внесення жодних змін до процедури навчання синтаксичного аналізатора. Експерименти з обробки різних мов свідчать, що цей метод зменшує кількість помилок у середньому на 72 % і незмінно перевершує результати стандартного евристичного алгоритму, який використовується нині.

Переклад М. Драчової

Gardent, C. Multiple Adjunction in Feature-Based Tree-Adjoining Grammar [Множинна ад'юнкція у категоріальній граматиці з'єднання дерев] / Claire Gardent, Shashi Narayan // Computational Linguistics. – 2015. – Vol. 41. – No. 1. – Pages 41–70. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00217 – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00217

Шейбс та Шібер (1994) продемонстрували, що в автоматичному синтаксичному аналізі за допомогою граматики з'єднання дерев (ГЗД) коректна підтримка синтаксичного аналізу, семантичної інтерпретації та статистичного моделювання мови неможлива без незалежних дериватів. Втім, запропонований ними алгоритм синтаксичного аналізу не можна прямо застосувати до категоріальних ГЗД (КГЗД). У статті запропоновано алгоритм розпізнавання для КГЗД, який працює як із залежними, так і з незалежними дериватами. Отриманий алгоритм поєднує переваги незалежних дериватів з перевагами категоріальних граматик. Зокрема, показано, що він пояснює, з одного боку, цілий ряд взаємодій між залежними і незалежними дериватами, а з другого боку, синтаксичні обмеження, лінійне упорядкування і локальні та глобальні семантичні залежності.

Переклад М. Дубка

Mirroshandel S. A. Integrating Selectional Constraints and Subcategorization Frames in a Dependency Parser [Використання обмежень у сполучуваності та субкатегорійних фреймів у синтаксичних аналізаторах на основі граматики залежностей] / Seyed Abolghasem Mirroshandel, Alexis Nasr // Computational linguistics. – 2016. – Vol. 42. – No. 1. – Pages 55–90. – Режим доступу до анотації: https://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00242 – Режим доступу до повнотекстової статті: https://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00242

Статистичні синтаксичні аналізатори навчаються на банках синтаксичних дерев, які складаються з декількох тисяч речень. Для запобігання розрідженості даних та складності обчислень такі аналізатори висувають вагомі гіпотези про незалежність рішень, прийнятих для побудови синтаксичного дерева. Ці гіпотези про незалежність призводять до членування синтаксичних структур на невеликі фрагменти, що в свою чергу не дозволяє синтаксичному аналізатору адекватно змоделювати багато лексико-синтаксичних явищ, наприклад обмеження в сполучуваності та субкатегорійні фрейми. Крім того, банки синтаксичних дерев надто малі для дослідження багатьох лексико-синтаксичних закономірностей, таких як обмеження в сполучуваності та субкатегорійні фрейми. У статті запропоновано рішення для обох проблем: як вирахувати шаблони, що перевищують розмір фрагментів, які моделюються в синтаксичному аналізаторі; і як отримати субкатегорійні фрейми та обмеження в сполучуваності з нерозмічених корпусів і вбудувати їх у процес автоматичного синтаксичного аналізу. Запропонований метод було апробовано на французькій та англійській мовах. Експерименти з французькою мовою показали зменшення порушень обмежень у сполучуваності на 41,6% і зменшення помилок у виділенні субкатегорійних фреймів на 22%. Ці показники нижчі для англійської мови: 16,21% у першому випадку та 8,83% у другому.

Переклад А. Шульги

Gildea, D. Synchronous Context-Free Grammars and Optimal Parsing Strategies [Синхронні контекстно-вільні граматики та стратегії оптимального автоматичного синтаксичного аналізу] / Daniel Gildea, Giorgio Satta // Computational linguistics. – 2016. – Vol. 42. – No. 2. – Pages 207–243. – Режим доступу до анотації: https://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00246 – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00246

Складність автоматичного синтаксичного аналізу з синхронними контекстно-вільними граматиками є багаточленом по довжині речення для закріпленої граматики, але ступінь багаточлена залежить від граматики. Зокрема, ступінь залежить від довжини правил, представлених правилами перестановок, і стратегії автоматичного синтаксичного аналізу, прийнятої для розкладання розпізнавання правила на дрібніші кроки. Проблему пошуку найкращої стратегії автоматичного синтаксичного аналізу для правила розглянуто з точки зору складності простору та часу. Продемонстровано, що знаходження двійкової стратегії з найнижчою просторовою складністю є NP-складною задачею. Продемонстровано також, що будь-який алгоритм пошуку стратегії з найнижчою часовою складністю передбачає вдосконалення алгоритмів апроксимації для визначення деревної ширини загальних графів.

Переклад М. Дубка

Zhang X. Transition-Based Parsing for Deep Dependency Structures [Автоматичний синтаксичний аналіз глибинних структур залежностей на основі переходів] / Xun Zhang , Yantao Du , Weiwei Sun and Xiaojun Wan // Computational linguistics. – 2016. – Vol. 42. – No. 3. – Pages 353–389. – Режим доступу до анотації: https://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00252 – Режим доступу до повнотекстової статті: https://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00252

Деривації в різних граматичних формалізмах дозволяють видобувати різноманітні структури залежностей. Зокрема, завдяки лінгвістичному аналізу на основі комбінаторної категоріальної граматики (ККГ), лексико-функціональної граматики (ЛФГ) та верховинної граматики складників (ВГС) можна на додаток до представлення поверхневої структури дерева видобути білексичні глибинні структури залежностей. Традиційно ці структури залежностей отримують як вторинний продукт граматично-орієнтованих синтаксичних аналізаторів. А в цій статті досліджується альтернативний керований даними підхід до побудови загальних графів залежностей на основі переходів, який успішно використовується в автоматичному синтаксичному аналізі. Представлено дві нові системи на основі переходів, які об'єднують існуючі методи опрацювання синтаксичних дерев і які можуть поетапно генерувати довільні орієнтовані графи. На основі цих систем переходів можна побудувати статистичні синтаксичні аналізатори, які є конкурентоспроможними як за точністю, так і за ефективністю. Крім того, різнорідна будова систем переходів забезпечує різноманітність сумісних моделей синтаксичного аналізу, що значно підвищує ефективність синтаксичного аналізатора. Для зняття лексичної багатозначності запропоновано два нові методи поліпшення якості аналізу, а саме: комбінацію переходів і спрощення дерев. Завдяки комбінації переходів кожна дія, яка виконується синтаксичним аналізатором, суттєво змінює конфігурації. Отже, для зняття статистичної неоднозначності можна виділити чіткіші категорії. Для визначення цих інформативних категорій метод спрощення дерев виводить основи дерев із графів залежностей і повторно використовує методи синтаксичного аналізу дерев для отримання категорій на основі дерев. Здійснено функторно-аргументний аналіз на основі ККГ, аналіз граматичних зв'язків на основі ЛФГ та аналіз семантичної залежності на основі ВГС англійської та китайської мов. Проведені експерименти свідчать, що керовані даними моделі з відповідними системами переходів можуть забезпечити високоякісний аналіз глибинних структур залежностей, нарівні з більш складними граматичними моделями. Експерименти також свідчать про ефективність гетерогенної будови систем синтаксичного аналізу на основі переходів, комбінації переходів і спрощення дерев для зняття статистичної неоднозначності.

Переклад А. Шульги

Gómez-Rodríguez, C. Restricted Non-Projectivity: Coverage vs. Efficiency [Обмежена непроективність: охоплення чи ефективність] / Carlos Gómez-Rodríguez // Computational linguistics. – 2016. – Vol. 42. – No. 4. – Pages 809–817. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00267 – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00267

Щоб забезпечити оптимальне співвідношення ефективності автоматичного синтаксичного аналізу та охоплення притаманних природним мовам синтаксичних структур, протягом останнього десятиліття було запропоновано різні обмежені класи непроективних дерев залежностей. Метою цього масштабного дослідження було оцінювання охоплення широкого кола таких класів у корпусах 30 мов за допомогою двох різних мірил синтаксичної розмітки. Результати свідчать, що серед відомих нині послаблень проективності найкраще співвідношення охоплення та обчислювальної складності точного автоматичного синтаксичного аналізу досягається або за допомогою дерев, які перетинаються в одній кінцевій точці, або за допомогою багатовузлових дерев, залежно від бажаного рівня охоплення. Також описано деякі особливості зв'язку багатовузлових дерев з іншими відповідними класами дерев.

Переклад М. Дубка

Ballesteros, M. Greedy Transition-Based Dependency Parsing with Stack LSTMs [Жадібний автоматичний синтаксичний аналіз залежностей на основі переходів за допомогою ТКПС] / Miguel Ballesteros, Chris Dyer, Yoav Goldberg, Noah A. Smith // Computational linguistics. – 2017. – Vol. 43. – No. 2. – Pages 311–347. – Режим доступу до анотації: https://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00285 – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00285

У статті представлено жадібний автоматичний синтаксичний аналізатор на основі переходів, який автоматично навчається репрезентувати стани автоматичного синтаксичного аналізатора за допомогою рекурентних нейронних мереж. Основним нововведенням, яке дозволяє робити це ефективно, є нова структура управління для послідовних нейронних мереж – стековий модуль тривалої короткочасної пам'яті (ТКЧП). Як і в звичайних стекових структурах даних, які використовуються в автоматичних синтаксичних аналізаторах на основі переходів, елементи можуть додаватися до або видалятися з вершини стека за сталий проміжок часу, але, крім цього, ТКЧП підтримує безперервне просторове розміщення вмісту стека. Запропонована модель фіксує три аспекти стану автоматичного

синтаксичного аналізатора: (i) необмежений перегляд буфера вхідних слів; (ii) повну історію виконаних автоматичним синтаксичним аналізатором переходів; (iii) повний вміст стека фрагментів частково побудованого дерева, зокрема їхні внутрішні структури. Крім того, здійснено порівняння двох різних представлень слова: (i) стандартних векторів слів на основі довідкових таблиць і (ii) символьних моделей слів. Хоча стандартні моделі додавання слів добре працюють на всіх мовах, символьні моделі покращують опрацювання слів, відсутніх у словнику, особливо в морфологічно багатих мовах. Нарешті, обговорено використання динамічних оракулів у навчанні автоматичного синтаксичного аналізатора. Під час навчання динамічні оракули по черзі отримують зразки станів автоматичного синтаксичного аналізатора з навчальних даних та з автоматично створюваної моделі, що робить цю модель більш стійкою до тих видів помилок, які матимуть місце під час тестування. Результатом автоматичного навчання запропонованої моделі за допомогою динамічних оракулів є дійсно конкурентоспроможний лінійний жадібний аналізатор.

Переклад М. Дубка

Gebhardt K. Hybrid Grammars for Parsing of Discontinuous Phrase Structures and Non-Projective Dependency Structures [Гібридні граматики для синтаксичного аналізу перерваних фразових структур і непроективних структур залежностей] / Kilian Gebhardt, Mark-Jan Nederhof, Heiko Vogler // Computational linguistics. – 2017. – Vol. 43. – No. 3. – Pages 465–520. – Режим доступу до анотації: https://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00291 – Режим доступу до повнотекстової статті: https://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00291

У статті досліджується поняття гібридних граматик, які формалізують і узагальнюють низку існуючих методів для опрацювання перерваних синтаксичних структур. Розглянуто як перервані фразові структури, так і непроективні структури залежностей. Формально гібридні граматики пов’язані з синхронними граматиками, в яких один компонент генерує лінійні структури, а інший – ієрархічні. Результатом об’єднання лексичних елементів обох компонентів є перервані структури. Описано декілька типів гібридних граматик. Також, розглянуто виведення граматик з банків синтаксичних дерев. Основною перевагою гібридних граматик над існуючими методами є можливість розмежувати переривність необхідних структур і часову складність автоматичного синтаксичного аналізу. Це дозволяє проаналізувати застосування для аналізу перерваних структур різноманітних алгоритмів автоматичного синтаксичного аналізу з різними властивостями. Це підтверджується представленими експериментальними результатами, які демонструють широкий діапазон тривалості роботи, точності та частоти збоїв автоматичного синтаксичного аналізу.

Переклад А. Шульги

Аналіз дискурсу

Marcu, D. The Rhetorical Parsing of Unrestricted Texts: A Surface-based Approach [Аналіз риторичної структури необмежених текстів: поверхневий підхід] / Daniel Marcu // Computational linguistics. – 2000. – Vol. 26. – No. 3. – Pages 395–448. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120100561755#.WIEF5H3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120100561755

Зв'язні тексти – це не просто послідовності речень та їх частин, а достатньо складні утворення, що мають дуже непросту риторичну структуру. У статті досліджуються можливості автоматичного отримання коректно утворених риторичних структур за допомогою алгоритмів поверхневого аналізу. Ці алгоритми визначають ключові фрази дискурсу та розбивають речення на клаузи, будують гіпотези про риторичні відносини між текстовими одиницями і створюють надійні дерева риторичної структури для необмежених текстів природною мовою. Емпіричним підґрунтям алгоритмів є корпусне дослідження ключових фраз, у побудові дерев риторичної структури застосовується формалізація першого порядку.

Здійснено як внутрішнє, так і зовнішнє оцінювання алгоритмів. За допомогою внутрішнього оцінювання визначена схожість між деревами риторичної структури, побудованими автоматично та вручну. Зовнішнє оцінювання показало, що автоматично отримані риторичні структури можна успішно використовувати у процесі автоматичного реферування текстів.

Переклад О. Мартинюк

Pulman, G.S. Bidirectional Contextual Resolution [Двостороннє розв'язання контексту] / Stephen G. Pulman // Computational linguistics. – 2000. – Vol. 26. – No. 4. – Pages 497–537. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120100750105939#.WIKMJn3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120100750105939

У статті описано застосування формалізму для інтерпретації і генерації речень, які містять контекстно-залежні конструкти, такі як детермінанти, займенники, фокус і еліпсис. У якості представлення нечітко вираженого значення, пов'язаного з визначеними логічними формами за допомогою умовних еквівалентностей, використовується варіант квазілогічної форми. Умовні еквівалентності визначають інтерпретацію контекстуально залежних конструктів з урахуванням даного контексту. При співвіднесенні виразів із контекстами використовуються об'єднання і роз'єднання вищого порядку.

Умовні еквівалентності можуть бути без змін використані як для інтерпретації, так і для генерації.

Переклад К. Погорєлова

Vieira, R. An Empirically Based System for Processing Definite Descriptions [Система обробки визначених дескрипцій на основі дослідних даних] / Renata Vieira, Massimo Poesio // Computational linguistics. – 2000. – Vol. 26. – No. 4. – Pages 539–593. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120100750105948#.WIE1tH3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120100750105948

У статті описана діюча система обробки визначених дескрипцій у довільних областях. Розробка системи здійснювалась на основі опублікованих раніше результатів корпусного аналізу, який виявив широке використання у корпусі газетних текстів нових для дискурсу дескрипцій для всебічної оцінки запропонованих методів вирівнювання визначених дескрипцій з їх антецедентами, сегментування дискурсу, розпізнавання нових для дискурсу дескрипцій і генерування анкорів для зв'язаних дескрипцій.

Переклад В. Коломієць

Tür, G. Integrating Prosodic and Lexical Cues for Automatic Topic Segmentation [Інтеграція просодичної і лексичної інформації для автоматичної тематичного сегментування] / Gökhan Tür, Dilek Hakkani-Tür, Andreas Stolcke, Elizabeth Shriberg // Computational linguistics. – 2001. – Vol. 27. – No. 1. – Pages 31–57. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120101300346796#.WIEGR33sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120101300346796

Описано вірогіднісну модель, яка використовує і просодичну, і лексичну інформацію для автоматичного сегментування мовлення на тематично споріднені одиниці. Запропоновано два методи об'єднання лексичної і просодичної інформації за допомогою прихованих марківських моделей і дерев рішень. Лексична інформація отримувалась із розпізнавача мовлення, а просодичні риси автоматично видобувались із коливань частоти основного тону. Для оцінювання методу використовувався корпус випусків новин, застосовувався показник DARPA-TDT. Результати свідчать, що просодична модель сама по собі може скласти конкуренцію методам сегментування на основі слів. Більше того, було досягнуто значного зменшення помилок завдяки об'єднанню просодичних джерел знань на основі слів і на основі просодії.

Переклад В. Коломієць

Kibble, R. A Reformulation of Rule 2 of Centering Theory [Переформулювання правила 2 теорії центрування] / Rodger Kibble //

Computational linguistics. – 2001. – Vol. 27. – No. 4. – Pages 579–587. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120101753342680#.WIEHCH3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120101753342680

Стверджується, що стандартне ранжування преференцій добре відомих переходів центрування Continue, Retain, Shift є необгрунтованим: часткове, контекстнозалежне ранжування є результатом взаємодії принципів дубльованої **зв'язності** (збереження попереднього центру уваги) і **значимості** (реалізація центру уваги як найбільш значимої іменної групи). Пропонується нове формулювання правила 2 теорії центрування, яке враховує ці принципи і спрощену версію поняття **дешевизни** [M. Strube and U. Hahn, 1999]. Стверджується, що це формулювання дозволяє природним шляхом упоратися зі "змінами тем", які можуть порушити традиційне ранжування преференцій.

Переклад В. Коломієць

Webber, B. Anaphora and Discourse Structure [Анафора і структура дискурсу] / Bonnie Webber, Matthew Stone, Aravind Joshi, Alistair Knott // Computational linguistics. – 2003. – Vol. 29. – No. 4. – Pages 545–587. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120103322753347#.WIPQY33sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120103322753347

У статті стверджується, що багато частотних прислівникових груп, які зазвичай сприймаються як сигнал дискурсного зв'язку між синтаксично пов'язаними елементами у структурі дискурсу, натомість функціонують анафорично, передаючи граматичне значення, і тільки опосередковано залежать від структури дискурсу. Таким чином, підтримка композиційної семантики забезпечується простішою структурою дискурсу і розкриваються численні шляхи взаємодії між граматичним значенням, що передається прислівниковими групами, і значенням, що асоціюється зі структурою дискурсу. У заключній частині статті викладається авторське бачення лексикалізованої граматики дискурсу, яка полегшує інтерпретацію дискурсу завдяки композиційним правилам, розв'язанню анафори і виведенню.

Переклад В. Коломієць

Wolf, F. Representing Discourse Coherence: A Corpus-Based Study [Представлення зв'язності дискурсу: корпусне дослідження] / Florian Wolf, Edward Gibson // Computational linguistics. – 2005. – Vol. 31. – No. 2. – Pages 249–287. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120104223977#.WIEIXn3sSGA – Режим доступу до повнотекстової статті:

http://www.mitpressjournals.org/doi/pdf/10.1162/0891201054223977

Метою статті є опис набору структурних відносин дискурсу, які легко кодувати, і розробка критеріїв для належної структури даних для представлення цих відносин. Під структурою дискурсу у статті розуміються інформаційні відносини між реченнями у дискурсі. Описаний набір відносин дискурсу запозичено з праці Т. Гоббса (T. Hobbs, 1985).

У статті описано метод анотування структур зв'язності дискурсу, який було використано для ручного анотування бази даних з 135 текстів з газети Wall Street Journal і стрічки новин агенства Associated Press. Усі тексти були незалежно анотовані двома анотаторами. Показник коефіцієнта каппа більше 0,8 свідчить про дуже високий ступінь узгодженості між анотаторами.

У статті також доведено, що в описовому плані дерева не є належною структурою даних для представлення структури дискурсу. В структурах зв'язності автентичних текстів було виявлено багато різних видів перехресних залежностей, а також багато вузлів з численими «господарями». Висновки підтверджені статистичними даними з анотованої вручну бази даних обсягом 135 текстів.

Переклад В. Коломієць

Barzilay, R. Modeling Local Coherence: An Entity-Based Approach [Моделювання локальної когерентності на основі референтів] / Regina Barzilay, Mirella Lapata // Computational linguistics. – 2008. – Vol. 34. – No. 1. – Pages 1–34. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2008.34.1.1#.WIEJZH3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2008.34.1.1

У статті запропоновано новітній підхід до представлення і вимірювання локальної когерентності. Головним у цьому підході є представлення дискурсу у вигляді таблиці референтів, яка відображає особливості розподілу референтів у тексті. Запропонований у статті алгоритм автоматично представляє текст у вигляді набору референціальних ланцюжків і реєструє дистрибутивну, синтаксичну і референціальну інформацію про референти дискурсу. Оцінка когерентності представлена як завдання машинного навчання і показано, що репрезентація на основі референтів добре підходить для генерування і класифікації текстів на основі ранжування. За допомогою запропонованої репрезентації були отримані хороші показники у класифікації текстів, оцінюванні когерентності анотацій і легкості сприйняття.

Переклад К. Погорєлова, М. Драчової

Elsner, M. Disentangling Chat [Розпутування чату] / Micha Elsner, Eugene Charniak // Computational linguistics. – 2010. – Vol. 36. – No. 3. – Pages 389–409. – Режим доступу до анотації:

http://www.mitpressjournals.org/doi/abs/10.1162/coli_a_00003#.WITMQ33sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli_a_00003

Коли одночасно ведеться кілька розмов, слухач повинен вирішити, частиною якої розмови є кожне висловлення, щоб зрозуміти і належним чином відреагувати на нього. Це завдання називається розпутуванням. У статті описано корпус діалогів з мережі Internet Relay Chat, у якому різні розмови були розпутані вручну, і оцінено якість анотування. Запропоновано кластерну модель розпутування на основі графа, яка враховує лексичні, часові і дискурсивні характеристики. Виконані за допомогою моделі розпутування тісно корелюють із ручним анотуванням. На завершення обговорено два розширення моделі, індивідуальні налаштування і визначення початку розмови, які є обіцяючими, але поки ще не дали практичних результатів.

Переклад В. Коломієць

Yang, F. An Investigation of Interruptions and Resumptions in Multi-Tasking Dialogues [Дослідження перебивання і відновлення у багатоцільових діалогах] / Fan Yang, Peter A. Heeman, Andrew L. Kun // Computational linguistics. – 2011. – Vol. 37. – No. 1. – Pages 75–104. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli_a_00036#.WIELOX3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli_a_00036

Стаття присвячена багатоцільовим діалогам між людьми, у яких пари співрозмовників, користуючись мовою, працюють над поточним завданням, іноді завершуючи оперативні завдання. Поточним завданням є сеанс гри в покер, у якому співрозмовникам потрібно зібрати покерну руку, а оперативним завданням є гра з картинками, у якій співрозмовники мають з'ясувати, чи є на їхніх дисплеях певна картинка. Для того щоб зрозуміти складні механізми, які використовуються людьми для контролю цих складних інтеракцій, було використано емпіричні корпусні дослідження і експерименти з машинного навчання. По-перше, було проаналізовано перебивання виконання завдань: переключення з поточного завдання на оперативне. З'ясовано, що загалом співрозмовники стараються за можливості переривати виконання поточного завдання у менш несприятливий момент. Також з'ясовано, що дискурсивні маркери oh (*о*) і wait (*стривай*) вживаються для переривання завдання удвічі частіше, ніж у розмові про поточне завдання. Крім того, виявлено, що висота тону статистично корелює з перериванням завдання; фактично, чим більш дезорганізуючим є переривання, тим вище висота тону. По-друге, проаналізовано відновлення виконання завдання: повернення до поточного завдання після завершення оперативного завдання, яке перебило його виконання. З'ясовано, що

співрозмовники можуть просто продовжити розмову з того місця, в якому вона була перервана, але іноді вони повторюють останнє висловлювання або підсумовують важливу інформацію, якою вони обмінялися до перебивання. По-третє, для визначення наскільки точно можуть перебивання виконання завдання бути розпізнані автоматично і для визначення ефективності ключових слів, виявлених у корпусному дослідженні, застосовано машинне навчання. З'ясовано, що контекст дискурсу, висота тону і дискурсивні маркери oh і wait є важливими характеристиками, які забезпечують надійне розпізнавання перебивань виконання завдання, і за допомогою нелексичних характеристик можна підняти ефективність розпізнавання перебивань, зменшивши відносну кількість помилок більше, ніж на 50% у порівнянні з базовим рівнем. Нарешті, проаналізовано значення отриманих результатів для створення мовного інтерфейсу для підтримки багатоцільових діалогів.

Переклад В. Коломієць

Morante, R. Modality and Negation: An Introduction to the Special Issue [Модальність і заперечення: вступ до спеціального видання] / Roser Morante, Caroline Sporleder // Computational linguistics. – 2012. – Vol. 38. – No. 2. – Pages 223–260. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00095#.WIETM33sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00095

Раніше переважна більшість досліджень у галузі обробки природної мови зосереджувались на пропозиційних аспектах значення. Проте для справжнього розуміння мови не менш важливі екстрапропозиційні аспекти. Зазвичай важливими компонентами цих екстрапропозиційних аспектів значення є модальність і заперечення. Хоча більшість комп'ютерних лінгвістів часто ігнорували модальність і заперечення, протягом останніх років інтерес до них виріс, про що свідчить їх розмітка у декількох корпусах. Дослідники почали працювати над моделюванням фактичності, переконання і визначеності, знаходженням гіпотетичних висловлень і обмежень, виявленням суперечностей і визначенням сукупності виразів модальності і заперечення. У статті вміщено огляд способів моделювання модальності і заперечення у комп'ютерній лінгвістиці.

Переклад І . Снєгурова

Saurí, R. Are You Sure That This Happened? Assessing the Factuality Degree of Events in Text [Ви впевнені, що це правда? Оцінка ступеня достовірності подій у тексті] / Roser Saurí, James Pustejovsky // Computational linguistics. – 2012. – Vol. 38. – No. 2. – Pages 261–299. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00096#.WITCmn3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00096

Визначіння реальності, або дійсності, поданої у тексті інформації про події є важливою передумовою міркувань про події у дискурсі. Висновки, зроблені на основі подій, які здаються нереальними або лише можливими, відрізняються від висновків, зроблених на основі подій, які уважаються реальними. Достовірність подій включає два окремі шари інформації. З одного боку, вона пов'язана з полярністю, яка розрізняє позитивні і негативні втілення подій. З другого боку, вона має справу зі ступенем упевненості (наприклад, вірогідний, можливий), інформаційним рівнем, який відноситься до категорії епістемічної модальності. Мета статті – допомогти краще зрозуміти, як реальність подій виражається у природній мові. Для цього пропонується лінгвістично орієнтована обчислювальна модель, в основі якої лежить алгоритм, який пов'язує ефект відношень реальності з рівнями синтаксичної інтеграції. Для перевірки концепції запропонована модель була реалізована в De Facto, профайлері реальності згаданих у тексті подій, і протестована на матеріалі спеціально створеного для цієї мети корпусу з результатами F1-міри 0,70 (макроусереднення) і 0,80 (мікроусереднення). Ці два показники взаємно компенсують характерне для кожного з них надлишкове акцентування (чи то на менше, чи то на більше заповнених категоріях) і тому можуть уважатися нижньою і верхньою межами результативності системи De Facto.

Переклад В. Коломієць

de Marneffe, M.-C. Did It Happen? The Pragmatic Complexity of Veridicality Assessment [Чи це правда? Прагматична складність оцінювання адекватності сприйняття] / Marie-Catherine de Marneffe, Christopher D. Manning, Christopher Potts // Computational linguistics. – 2012. – Vol. 38. – No. 2. – Pages 301–333. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00097#.WITDLn3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00097

Розуміння природної мови значною мірою залежить від оцінки істинності – чи розглядаються згадані в тексті події як реальні, чи ні; проте в сучасних системах видобування відносин і подій цій характеристиці приділяється мало уваги. Крім того, у проведених дослідженнях загалом припускалося, що істинність виражається семантичним значенням слів, у статті ж показано, що значну роль у формуванні істинності грає контекст і загальні знання про світ. Ми розширили корпус FactBank, який містить розмітку істинності на основі семантики, додавши розмітку істинності на основі прагматики. Наші мітки складніші, ніж розмітка на основі лексичних значень, але достатньо систематичні, щоб використовуватися у комп'ютерних дослідженнях автоматичного розуміння тексту. Вони також свідчать, що судження про істинність не завжди є категоричними, а тому повинні моделюватися у вигляді дистрибуцій. Нами розроблено класифікатор для автоматичного

приписування дистрибуції реальності подій на основі наших нових міток. Класифікатор спирається не тільки на лексичні характеристики, такі як сумнів або заперечення, але й на синтаксичні особливості і наближення до загальних знань про світ, створюючи, таким чином, складну картину різноманітних факторів, які впливають на реальність.

Переклад В. Коломієць

Szarvas, G. Cross-Genre and Cross-Domain Detection of Semantic Uncertainty [Незалежне від жанру і тематичної області визначення семантичної невпевненості] / György Szarvas, Veronika Vincze, Richárd Farkas, György Móra, Iryna Gurevych // Computational linguistics. – 2012. – Vol. 38. – No. 2. – Pages 335–367. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00098#.WITD7H3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00098

Невпевненість є важливим лінгвістичним явищем, актуальним у різних програмах обробки природної мови у різноманітних жанрах, від медичних до соціальних, від стрічок новин до наукового дискурсу, і тематичних областях, від наукових до гуманітарних. Семантичну невпевненість пропозиції у багатьох випадках можна ідентифікувати, користуючись вихідним словником (тобто, лексичними сигналами), і основні етапи ідентифікації невпевненості у програмі включають етапи знаходження лексичних сигналів, характерних для жанру і тематичної області, зняття лексичної омонімії і зв'язування їх із одиницями, які становлять інтерес для конкретної програми (наприклад, розпізнаними подіями у видобуванні інформації). Основна увага у даному дослідженні приділена особливостям розпізнавання контекстно-залежних семантичних сигналів невпевненості у різних жанрах і тематичних областях.

Оскільки у програмах для різних тематичних областей можуть використовуватися різні категорії невпевненості, у дослідженні застосована єдина підкатегоризація семантичної невпевненості. На основі цієї підкатегоризації було нормалізовано анотацію трьох корпусів і отримано результати для чотирьох дуже точних категорій семантичної невпевненості за допомогою сучасної моделі розпізнавання сигналів невпевненості.

Отримані результати свідчать про залежність проблеми від жанру і тематичної області, проте також показано, що навіть набір даних із віддаленої тематичної області може сприяти розпізнаванню і вирішенню неоднозначності сигналів невпевненості, ефективно зменшуючи затрати на анотування, необхідні для роботи з новою тематичною областю. Отже, об'єднана субкатегоризація і адаптація предметної області для тренування моделей є ефективним рішенням незалежного від тематичної області й жанру розпізнавання семантичної невпевненості.

Переклад В. Коломієць

Joty S. CODRA: A Novel Discriminative Framework for Rhetorical Analysis [ЗАДРА: новий диференційований підхід до риторичного аналізу] / Shafiq Joty, Giuseppe Carenini and Raymond T. Ng // Computational linguistics. – 2015. – Vol. 41. – No. 3. – Pages 385–435. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00226 – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00226

Частини складного речення і речення рідко функціонують в реальному дискурсі самостійно. Скоріше, взаємозв'язок між ними несе важливу інформацію, яка дозволяє дискурсу виражати значення як ціле, а не суму окремих частин. Мета риторичного аналізу полягає у з'ясуванні структури цієї узгодженості. У статті представлено ЗАвершений ймовірнісний Диференційований підхід до здійснення Риторичного Аналізу згідно теорії риторичної структури (ЗАДРА), яка постулює представлення дискурсу у вигляді дерева.

ЗАДРА складається з сегментатора та аналізатора дискурсу. Спочатку сегментатор дискурсу на основі двійкового класифікатора визначає елементарні одиниці дискурсу в заданому тексті. Потім аналізатор будує дерево дискурсу, застосовуючи оптимальний алгоритм автоматичного синтаксичного аналізу до ймовірностей, виведених із двох умовних випадкових полів: один – для синтаксичного аналізу окремих речень, а другий – для синтаксичного розбору сукупностей речень. У статті описано два підходи, метою яких є ефективне об'єднання обох етапів аналізу. Шляхом проведення низки експериментів на основі двох різних наборів даних, продемонстровано, що ЗАДРА перевершує сучасні досягнення, часто з великим відривом. Також показано, що точність може бути покращена далі за допомогою перерозподілу k-найкращих гіпотез, згенерованих ЗАДРА.

Переклад А. Шульги

Roth, M. Inducing Implicit Arguments from Comparable Texts: A Framework and Its Applications [Видобування імпліцитних аргументів з порівняльних текстів: метод і його застосування]/ Michael Roth, Anette Frank // Computational linguistics. – 2015. – Vol. 41. – No. 4. – Pages 625–664. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00236 – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00236

У статті досліджено аспекти значення речень, які не виражені в локативних предикатно-аргументних структурах. Зокрема, проаналізовано приклади семантичних аргументів, які можна вивести виключно з контексту дискурсу. Метою цього дослідження є автоматичне видобування та опрацювання таких випадків, названих «імпліцитними аргументами», для вдосконалення комп'ютерних моделей мови. Щоб досягти цієї мети, запропоновано

ефективний підхід для складного завдання видобування імпліцитних аргументів і їхніх антецедентів з дискурсу та емпірично продемонстровано важливість моделювання цього явища в завданнях на рівні дискурсу.

В основу запропонованого підходу покладено інноваційний проективний підхід, який дозволяє точно виявляти імпліцитні аргументи шляхом вирівнювання та порівняння предикатно-аргументних структур у парах порівняльних текстів. В рамках цього підходу створено метод вирівнювання за предикатами на основі графів, який значно перевершує попередні підходи. За допомогою такого вирівнювання показано, що можна автоматично видобувати й застосовувати окремі імпліцитні аргументи для покращення чинної моделі зв'язування імпліцитних аргументів в дискурсі. Також підтверджено, що хоча рішення щодо реалізації аргументів в більшості випадків є невловимим явищем, вони можуть суттєво вплинути на сприйняття зв’язності тексту. Проведені експерименти показали, що попередні моделі зв’язності не можуть прогнозувати цей вплив. Отже, розроблено нову модель зв’язності, яка вчиться точно прогнозувати предикатно-аргументні структури на основі автоматично вирівняних пар імпліцитних і експліцитних аргументів.

Переклад М. Дубка

Nguyen, D. Computational Sociolinguistics: A Survey [Комп’ютерна соціолінгвістика: огляд] / Dong Nguyen, A. Seza Doğruöz, Carolyn P. Rosé, Franciska de Jong // Computational linguistics. – 2016. – Vol. 42. – No. 3. – Pages 537–593. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00258 – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00258

Мова – це соціальне явище, соціальній природі якого притаманна варіативність. Останнім часом у галузі комп’ютерної лінгвістики (КЛ) спостерігається зростання інтересу до соціального виміру мови. У статті розглядається нова галузь, яка відображає цей підвищений інтерес, – "комп’ютерна соціолінгвістика". Мета огляду – дати вичерпне уявлення про виконані комп’ютерними лінгвістами соціолінгвістичні дослідження таких проблем як співвідношення мови та соціальної ідентичності, використання мови в побутовому спілкуванні та багатомовне спілкування. Крім того, показано, як масштабні, керовані даними методи, які широко використовуються в комп’ютерній лінгвістиці, можуть доповнити існуючі соціолінгвістичні студії, і як соціолінгвістика може вдосконалювати та спростовувати методи та припущення, які використовуються в дослідженнях з комп’ютерної лінгвістики, тобто продемонстровано потенційні можливості співпраці зацікавлених наукових спільнот. Завдання огляду − висвітлити потенційні переваги тіснішої співпраці двох галузей. У заключній частині статті розглядаються недосліджені проблеми.

Переклад М. Дубка

Habernal I. Argumentation Mining in User-Generated Web Discourse [Глибинний аналіз аргументування у створеному користувачем веб-дискурсі] / Ivan Habernal, Iryna Gurevych // Computational linguistics. – 2017. – Vol. 43. – No. 1. – Pages 125–179. – Режим доступу до анотації: https://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00276 – Режим доступу до повнотекстової статті: https://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00276

Метою глибинного аналізу аргументування, постійно еволюціонуючої дослідницької галузі комп'ютерної лінгвістики, є розробка методів, здатних аналізувати аргументування людини. Ця стаття з кількох поглядів виходить за рамки сучасних досліджень. (i) Матеріалом цього дослідження є фактичні дані із Всесвітньої мережі, що вимагає вирішення проблем, спричинених різноманіттям стилів, розмаїттям тематики, необмеженим зашумленим мережевим дискурсом, створеним користувачами. (ii) Шляхом адаптування моделі аргументації, протестованої у широкомасштабному дослідженні маркування було заповнено прогалину між нормативними теоріями аргументування та особливостями аргументування, які зустрічаються у фактичних даних. (iii) Створено новий корпус "золотого стандарту" (340 документів обсягом 90 тисяч словоформ) і проведено експерименти з кількома методами машинного навчання з метою визначення компонентів аргументів. Забезпечено вільний доступ загалу до даних, вихідних кодів і принципів маркування. Результати дослідження свідчать, що глибинний аналіз аргументування в створеному користувачами веб-дискурсі є можливим, але складним завданням.

Переклад А. Шульги

Stab C. Parsing Argumentation Structures in Persuasive Essays [Автоматичний синтаксичний аналіз структур аргументації в есе-переконаннях] / Christian Stab, Iryna Gurevych// Computational linguistics. – 2017. – Vol. 43. – No. 3. – Pages 619–659. – Режим доступу до анотації: https://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00295 – Режим доступу до повнотекстової статті: https://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00295

У статті представлено новий підхід до автоматичного синтаксичного розбору структур аргументації. Компоненти аргументів визначаються шляхом маркування послідовностей на рівні лексем, а нова об'єднана модель застосовується для виявлення структур аргументації. Запропонований метод на глобальному рівні оптимізує типи компонентів аргументів та аргументативних відношень за допомогою цілочисельного лінійного програмування. Доведено, що цей метод значно перевершує високі евристичні вихідні показники у двох різних типах дискурсу. Крім того, у статті описано новий корпус есе-переконань з маркуванням структур аргументації. Продемонстровано, що схема маркування і рекомендації щодо

маркування забезпечують високий ступінь узгодженості між маркувальниками.

Переклад А. Шульги

Аналіз і синтез мовлення

Marchand, Y. A Multistrategy Approach to Improving Pronunciation by Analogy [Багатостратегійний підхід до покращення вимови за аналогією] / Yannick Marchand, Robert I. Damper // Computational linguistics. – 2000. – Vol. 26. – No. 2. – Pages 195–219. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120100561674#.WITHzn3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120100561674

Вимова за аналогією (ВзА) це керований даними метод співвіднесення літер та звуків, який може бути застосований у наступному поколінні систем озвучування письмового тексту. Ця стаття розвиває попередні дослідження з ВзА у декількох напрямах. По-перше, ми включили «повний» шаблон, який вирівнює рядок із вхідними літерами та словникові статті та враховує лексичний наголос при перетворенні літери у фонему. По-друге, ми застосували даний метод для перетворення фонеми у літеру. По-третє, і найголовніше, ми проекспериментували з багатьма різними стратегіями кількісної оцінки варіантів вимови. Окремі показники для кожної стратегії отримуються на основі рангу і множаться або додаються для отримання остаточного, загального результату. Було досліджено п’ять стратегій та отримано результати з усіх 31 можливих комбінацій. Обидва методи комбінування працюють аналогічно, при зовсім незначній перевазі правила множення вірогідностей над правилом складання. Непараметричний статистичний аналіз свідчить, що продуктивність підвищується, коли комбінація включає більше стратегій: ця тенденція дуже виразна ($p < 0:0005$). Так само і при перетворенні літер у фонеми, найкращі результати одержуються при комбінуванні усіх п’яти методів: точність транскрипції підвищується до 65.5% порівняно з 61.7% для нашого найкращого попереднього результату та 63.0% для найефективнішої окремої стратегії. Ці покращення дуже істотні (відповідно $p \sim 0$ і $p < 0:00011$). Подібні результати були отримані при перетворенні фонем у літери і літер у наголос, хоча для ВзА перше перетворення було легшим завданням, ніж перетворення літер у фонеми, а друге перетворення було важчим завданням. Основні причини помилок у багатостратегійному підході мало відрізняються від основних причин помилок у найкращій окремій стратегії, включаючи здебільшого голосні літери і фонеми.

Переклад Д. Попової

Ke, J. Optimization Models of Sound Systems Using Genetic Algorithms Summarization [Моделі оптимізації систем звуків на основі реферування генетичних алгоритмів] / Jinyun Ke, Mieko Ogura, William S.-Y. Wang // Computational linguistics. – 2003. – Vol. 29. – No. 1. – Pages 1–18. – Режим

доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120103321337412#.WIEHb33sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120103321337412

У цьому дослідженні пропонуються моделі оптимізації, які використовують генетичні алгоритми (ГА) для аналізу конфігурації систем голосних і інтонаційних систем. Як і в попередніх пояснювальних моделях, які використовувалися для аналізу систем голосних, для прогнозування оптимальних систем голосних і інтонаційних систем використовуються певні критерії, які уважаються принципами, що визначають структуру систем звуків. У більшості попередніх досліджень ураховувався лище один критерій. Коли враховуються два критерії, вони часто об'єднуються в одну скалярну функцію. Запропонована для аналізу інтонаційних систем модель ГА використовує ранжувальний метод Pareto, який дуже підходить для розв'язання проблем оптимізації з багатьма критеріями. З метою оптимізації інтонаційних систем, перцептивний контраст і складність вираженості розглядаються одночасно. Хоча узгодженість між спрогнозованими і наявними системами не настільки значна, як для систем голосних, подальші дослідження у цьому напрямку є перспективними.

Переклад В. Коломієць

Deemter, K. V. Real versus Template-Based Natural Language Generation: A False Opposition? [Справжнє і шаблонне генерування природної мови: хибне протиставлення?] / Kees van Deemter, Mariët Theune, Emiel Krahmer // Computational linguistics. – 2005. – Vol. 31. – No. 1. – Pages 15–24. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/0891201053630291#.WITJOn3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/0891201053630291

У статті ставиться під сумнів існуюча думка про те, що генерування природної мови на основі шаблонів завжди програє іншим підходам з точки зору зручності експлуатації, лінгвістичної обгрунтованості і якості виведення. Висловлені претензії проілюстровано за допомогою деяких сучасних систем генерування природної мови, які називаються "шаблонними".

Переклад В. Коломієць

Schuler, W. A Framework for Fast Incremental Interpretation during Speech Decoding [Модель швидкої покрокової інтерпретації під час розпізнавання мовлення] / William Schuler, Stephen Wu, Lane Schwartz // Computational linguistics. – 2009. – Vol. 35. – No. 3. – Pages 313–343. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.08-011-R2-07-

021#.WITLTn3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.08-011-R2-07-021

У статті описано метод вбудовування референціальної семантичної інформації з моделі світу або онтології прямо у вірогіднісну модель мови, яка звичайно використовується у розпізнаванні мовлення, де її можна вірогіднісно оцінити разом із фонологічними і синтаксичними факторами як інтегральну частину процесу розпізнавання. Застосування у процесі розпізнавання референтів з моделі світу значно розширює простір пошуку, але застосовуючи єдину інтегровану фонологічну, синтаксичну і референціальну семантичну модель мови, перетворювач коду може покроково спростити цей пошук на основі вірогідностей, які асоціюються з цими об'єднаними контекстами. Результатом є єдина уніфікована референціальна семантична вірогіднісна модель, яка передбачає використання у розпізнаванні мовлення кількох різновидів контекстів і забезпечує точне розпізнавання у реальному часі у великих доменах за відсутності еталонних тренувальних речень з домену.

Переклад В. Коломієць

White, M. Generating Tailored, Comparative Descriptions with Contextually Appropriate Intonation [Генерування індивідуалізованих порівняльних описів із відповідною ситуації інтонацією] / Michael White, Robert A. J. Clark, Johanna D. Moore // Computational linguistics. – 2010. – Vol. 36. – No. 2. – Pages 159-201. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.09-023-R1-08-002#.WITLtX3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.09-023-R1-08-002

Генерування відповідей, які відповідають уподобанням користувачів, вимагає адаптації на всіх рівнях процесу генерування. У статті описано багаторівневий підхід до представлення адаптованої під користувача інформації в усних діалогах, який уперше об'єднує багатофакторні моделі рішень, стратегічне планування змісту, поверхневу реалізацію, яка включає прогнозування інтонації, і синтез вибору об'єкта, який враховує отриману інтонаційну структуру. Система вибирає найважливіші варіанти для повідомлення і фактори, які є найнеобхіднішими для здійснення вибору між ними, з урахуванням моделі користувача. Кілька варіантів обираються в тому випадку, коли кожен з них передбачає значний компроміс. Щоб повідомити про ці компроміси, система використовує новітній спосіб представлення, який прямо дозволяє визначати структуру інформації і зміст референціальних виразів. Просодична структура виводиться під час поверхневої реалізації із структури інформації, використовуючи комбінаторну категоріальну граматику для гнучкого, керованого даними визначення меж фраз. Показано, що такий підхід до вибору тонічного наголосу і крайових тонів дозволяє отримати просодичні структури, які за оцінками експертів є значно більш

прийнятними, аніж базові моделі прогнозування просодії. Потім продемонстровано, що у порівнянні з двома базовими синтетичними голосами ці просодичні структури уможливлюють синтез, який звучить значно природніше завдяки синтезованому методом вибору звукових елементів голосу для відтворення потрібних тонів. Експертна оцінка і аналіз f0 підтвердили вищість керованої генератором інтонації і її вплив на оцінки слухачів.

Переклад В. Коломієць

Аналіз тональності

Wiebe, J. Learning Subjective Language [Виявлення мовних показників суб'єктивності] / Janyce Wiebe , Theresa Wilson , Rebecca Bruce , Matthew Bell , Melanie Martin // Computational linguistics. – 2004. – Vol. 30. – No. 3. – Pages 277–308. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/0891201041850885#.WH4XO33sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/0891201041850885

У природній мові суб'єктивність відноситься до аспектів мови, за допомогою яких виражають думки, оцінки і здогадки. Існує велика кількість додатків для обробки природної мови, які потребують аналізу суб'єктивності, зокрема видобування знань і категоризація текстів. Метою цього дослідження було автоматичне виявлення мовних засобів вираження модальності у корпусах текстів. Були розроблені й протестовані показники суб'єктивності, зокрема низькочастотні слова, коллокації, а також прикметники і дієслова визначені за допомогою дистрибутивної схожості. Функції також аналізувалися у процесі спільної роботи. Показники, виявлені за допомогою різних методів на основі різних наборів даних, демонструють узгодженість проявів, тобто всі вони дають хороші і погані результати на однакових наборах даних. Крім того, у статті показано, что щільність показників суб'єктивності у оточуючому контексті має значний вплив на вірогідність суб'єктивності слова, і вміщено результати дослідження анотування, метою якого була оцінка суб'єктивності речень з високою щільністю показників. Нарешті, щоб продемонструвати корисність отриманих у дослідженні знань, показники були використані для розпізнавання вираження думки (різновид категорізації текстів і розпізнавання жанру).

Переклад В. Коломієць

Wilson, T. Recognizing Contextual Polarity: An Exploration of Features for Phrase-Level Sentiment Analysis [Розпізнавання контекстуальної полярності: дослідження ознак для аналізу модальності на рівні словосполучення] / Theresa Wilson, Janyce Wiebe, Paul Hoffmann // Computational linguistics. – 2009. – Vol. 35. – No. 3. – Pages 399–433. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.08-012-R1-06-90#.WIERDH3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.08-012-R1-06-90

В основі багатьох методів автоматичного аналізу модальності лежить великий лексикон, де у слів помічена їх апріорна полярність (яка також

називається семантичною орієнтацією). Проте контекстуальна полярність словосполучення, у якому вживається окреме слово, може дуже відрізнятись від апріорної полярності цього слова. Позитивні слова вживаються у фразах, які виражають негативні емоції, або навпаки. Також, досить часто слова, які є позитивними або негативними поза контекстом, є нейтральними у контексті, тобто їх вживають зовсім не для того, щоб виразити емоцію. Мета цього дослідження полягає у автоматичному розрізненні апріорної і контекстуальної полярності з акцентом на з'ясуванні важливих для вирішення цього завдання ознак. Оскільки важливим аспектом проблеми є з'ясування, коли емоційно забарвлені слова вживаються у нейтральних контекстах, проаналізовано ознаки найтрального і емоційно забарвленого значення, а також ознаки позитивної і негативної контекстуальної полярності. Аналіз включав оцінку продуктивності ознак у різних алгоритмах машинного навчання. В усіх алгоритмах машинного навчання, за винятком одного, найкращі результати досягаються шляхом комбінування усіх ознак. Іншим аспектом аналізу було з'ясування впливу нейтральних уживань на продуктивність ознак позитивної і негативної полярності. Ці експерименти свідчать, що присутність нейтральних уживань значно погіршує продуктивність цих ознак і що можливо найкращим способом підвищення результатів розпізнавання усіх видів полярності є удосконалення здатності системи розпізнавати нейтральні слововживання.

Переклад В. Коломієць

Qiu, G. Opinion Word Expansion and Target Extraction through Double Propagation [Розширення словника оціночної лексики та виявлення об'єкта оцінювання шляхом подвійного розповсюдження] / Guang Qiu, Bing Liu, Jiajun Bu, Chun Chen // Computational linguistics. – 2011. – Vol. 37. – No. 1. – Pages 9–27. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli_a_00034#.WIERoX3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli_a_00034

Останнім часом завдяки частому практичному застосуванню і складним дослідницьким завданням аналіз оцінкових суджень, відомий як видобування оцінкових суджень або аналіз тональності, привертає дуже багато уваги. У статті розглядаються дві важливі проблеми, а саме: розширення словника оціночної лексики та виявлення об'єкта оцінювання. Об'єкти оцінювання (скорочено *об'єкти*) – це сутності та їхні характерні ознаки, щодо яких виражаються оціночні судження. Щоб виконати ці завдання, ми з'ясували, що є декілька синтаксичних відношень, які поєднують оціночні слова і об'єкти. Ці відношення можуть бути визначені за допомогою синтаксичного аналізатора на основі граматики залежностей, а потім використані для розширення вихідного словника оціночної лексики та для видобування об'єктів. В основі запропонованого методу лежить бутстрепінг. Ми називаємо його подвійним розповсюдженням, оскільки він розповсюджує

інформацію між оціночними словами та об'єктами. Основною перевагою запропонованого методу є те, що для запуску процесу бутстрепінга потрібен лише вихідний словник оціночної лексики. Отже, завдяки використанню вихідної оціночної лексики метод є напівконтрольованим. На етапі оцінювання запропонований метод був порівняний із кількома найсучаснішими методами за допомогою стандартного набору тестів для оцінки продуктів. Результати свідчать, що наш метод значно результативніший, аніж уже існуючі методи.

Переклад Д. Попової

Taboada, M. Lexicon-Based Methods for Sentiment Analysis [Словникові методи аналізу тональності] / Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, Manfred Stede // Computational linguistics. – 2011. – Vol. 37. – No. 2. – Pages 267–307. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00049#.WIESBH3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00049

У статті описано словниковий метод визначення емоційно забарвленої лексики в текстах. Програма Semantic Orientation CALculator (SO-CAL) використовує словники слів із вказівкою їх семантичної орієнтації (полярності і інтенсивності) і враховує підсилення і заперечення. Програма SO-CAL застосовувалась у процесі класифікації полярності, тобто приписування тексту оцінки «позитивний або негативний», яка відображає ставлення автора до основної теми тексту. Показано, що SO-CAL однаково ефективна для різних тематик і для абсолютно нових даних. Крім того, описано процес укладання словників і використання сервісу Mechanical Turk для перевірки їх одноманітності та надійності.

Переклад Д. Попової

Wan, X. Bilingual Co-Training for Sentiment Classification of Chinese Product Reviews [Спільне двомовне навчання для класифікації тональності китайських відгуків на товари] / Xiaojun Wan // Computational linguistics. – 2011. – Vol. 37. – No. 3. – Pages 587–616. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00061#.WIES1H3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00061

Відсутність вивірених тональних словників і корпусів китайської мови уповільнює проведення досліджень, присвячених класифікації тональності китайських текстів. Проте у відкритому доступі в інтернеті є багато англомовних тональних ресурсів. У статті розглядається проблема міжмовної класифікації тональності, яка використовує лише доступні англомовні ресурси для класифікації тональності китайських текстів. Спочатку, просто

використовуючи служби машинного перекладу для подолання мовного бар'єру, здійснюється аналіз декількох базових (у тому числі словникових і корпусних) методів міжмовної класифікації тональності, а потім пропонується метод спільного двомовного навчання, який використовує як емоційні оцінки англійських авторів, так і емоційні оцінки китайських авторів, вилучені з додаткових нерозмічених китайських текстів. Результати експерименту із застосуванням двох наборів тестів, свідчать про ефективність запропонованого методу, який може перевершити базові та трансдуктивні методи.

Переклад М. Драчової

Johansson, R. Relational Features in Fine-Grained Opinion Analysis [Реляційні характеристики у точному аналізі думок] / Richard Johansson, Alessandro Moschitti // Computational linguistics. – 2013. – Vol. 39. – No. 3 – Pages 473–509. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00141#.WIETuX3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00141

Точні методи аналізу думок часто використовують лінгвістичні характеристики, але не беруть до уваги взаємодію між думками. У статті описано серію експериментів, які свідчать, що реляційні характеристики, які здебільшого є похідними від структур синтаксичної залежності і семантичних ролей, можуть суттєво підвищити продуктивність автоматичних систем у різноманітних завданнях точного аналізу думок: розмітці виразів емоційного ставлення, знаходженні власників думок і визначенні полярностей виразів емоційного ставлення. Ці характеристики дозволяють моделювати способи взаємодії у реченні на довільних відстанях думок, виражених у дискурсі природною мовою. Використання відношень вимагає одночасного розгляду кількох думок, що ускладнює пошук оптимального аналізу. Проте в якості достатньо точного і надійного наближення може бути використаний переранжувальник.

Здійснено оцінювання великої кількості наборів характеристик і підходів до машинного навчання. У завданні видобування виразів емоційного ставлення найкраща модель показала загальне поліпшення на 10 балів у повноті на корпусі MPQA у порівнянні зі стандартним розмітником послідовностей на основі локальних контекстуальних характеристик, а точність знизилася дуже мало. Значне покращення також спостерігалося у розширенних завданнях, у яких бралися до уваги власники і полярності: відповідно 10 і 7 балів у повноті. Крім того, системи поліпшили опубліковані раніше результати для видобування немаркованих (6 балів по F-метриці) і маркованих за полярністю (10-15 балів) виразів емоційного ставлення. Нарешті, в якості зовнішнього оцінювання видобуті з корпусу MPQA вирази емоційного ставлення були використані у реальних завданнях видобування думок. В усіх розглянутих сценаріях компоненти машинного навчання на

основі виразів емоційного ставлення забезпечують статистично значиме поліпшення результатів.

Переклад В. Коломієць

Hassan, A. A Random Walk–Based Model for Identifying Semantic Orientation [Модель встановлення семантичної орієнтації на основі випадкового блукання] / Ahmed Hassan, Amjad Abu-Jbara, Wanchen Lu, Dragomir Radev // Computational linguistics. – 2014. – Vol. 40. – No. 3. – Pages 539–562. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00192#.WIEVLn3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00192

Автоматичне встановлення тональності слів є дуже важливою задачею, яка є основним компонентом багатьох систем обробки природної мови, таких як системи класифікації текстів, фільтрування текстів, аналізу оглядів товарів, аналізу результатів опитувань та глибинного аналізу дискусій в режимі онлайн. У статті представлено метод встановлення тональності слів, який визначає полярність будь-якого заданого слова шляхом застосування моделі випадкового блукання Маркова і великого графа співвіднесеності слів. Модель здатна точно і швидко визначити полярність кожного слова та її інтенсивність. Вона може застосовуватися як у напівконтрольованих умовах з використанням навчальної вибірки розмічених слів, так і в слабоконтрольованих умовах з використанням лише невеликої кількості відібраних слів для встановлення двох класів полярності. Метод протестуваний експериментально із використанням золотеталонного набору позитивно та негативно забарвлених слів із лексикону системи General Inquirer. Також продемонстровано як можна використовувати запропонований метод для класифікації за трьома ознаками, яка окрім позитивно та негативно забарвлених слів визначає нейтральні слова. Проведені експерименти свідчать, що запропонований метод перевершує сучасні методи у напівконтрольованих умовах та досягає тих самих показників, що й найкращі методи, у слабоконтрольованих умовах. На додаток до цього, запропонований метод швидший і не потребує великого корпусу. Також описано модифікації запропонованих методів для визначення полярності іноземних слів та слів, які не входять до вокабулярію.

Переклад М. Погребної

Dong, L. A Statistical Parsing Framework for Sentiment Classification [Статистична модель синтаксичного аналізу для класифікації тональності] / Li Dong, Furu Wei, Shujie Liu, Ming Zhou, Ke Xu // Computational linguistics. – 2015. – Vol. 41. – No. 2. – Pages 293–336. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00221 – Режим доступу до повнотекстової статті:

http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00221

У статті представлено статистичну модель синтаксичного аналізу для класифікації тональності на рівні речення. На відміну від попередніх досліджень, у яких для аналізу тональності застосовуються результати синтаксичного аналізу, автори статті створили статистичний аналізатор для безпосереднього аналізу структури тональності речення. Показано, що в аналізі тональності складні явища (наприклад, заперечення, підсилення та контрастність) можна обробляти в такий самий комплексний та ймовірнісним спосіб, що й прості та прямі вирази тональності. Розроблено граматику тональності на основі контекстно-незалежних граматик (КНГ) та подано формальний опис моделі аналізу тональності. Створено модель синтаксичного аналізу, щоб отримати можливі синтаксичні дерева тональності речення, на основі яких пропонується модель полярності для визначення сили тональності та її полярності. Вибір найкращого дерева тональності виконується моделлю ранжування. Тренування синтаксичного аналізатора здійснюється безпосередньо на прикладах речень, розмічених лише мітками модальної полярності, без жодних міток синтаксичної структури або полярності складників речення. Завдяки цьому можна легко отримати навчальні дані. Зокрема, тренування синтаксичного аналізатора тональності здійснюється на великій кількості оціночних речень з рейтингами користувачів в ролі маркерів полярності. Обширні експерименти з існуючими наборами даних для порівняльного аналізу демонструють суттєві покращення в порівнянні з базовими підходами до класифікації тональності.

Переклад М. Дубка

Dras, M. Evaluating Human Pairwise Preference Judgments [Оцінювання парних оціночних суджень експертів] / Mark Dras // Computational linguistics. – 2015. – Vol. 41. – No. 2. – Pages 337–345. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00222 – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00222

Важливу роль в опрацюванні природної мови відіграє людська оцінка, часто представлена у формі оціночних суджень. Незважаючи на деякі застосування класичних непараметричних і вузькоспеціалізованих підходів до оцінювання цих видів суджень, існує ціла низка їх досліджень у контексті оцінювання сенсорного розрізнення та людських суджень, які є ключовими в ньому, підкріплена строгою статистичною теорією і програмним забезпеченням у вільному доступі, яку можна використати в опрацюванні природної мови. Досліджено один з підходів, логарифмічні лінійні моделі Бредлі-Террі, який застосовано до вибіркових даних для опрацювання природної мови.

Переклад М. Дубка

Benamara, F. Evaluative Language Beyond Bags of Words: Linguistic Insights and Computational Applications [Оцінювальна лексика поза мішками слів: лінгвістична інформація і комп'ютерні програми] / Farah Benamara, Maite Taboada, Yannick Mathieu // Computational linguistics. – 2017. – Vol. 43. – No. 1. – Pages 201–264. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00278 – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00278

Дослідження оцінювання, впливу та суб'єктивності є міждисциплінарним завданням, яке поєднує соціологію, психологію, економіку, лінгвістику та інформатику. Існує низка високоякісних оглядів галузі, виконаних комп'ютерними лінгвістами і мовознавцями. Проте дуже мало оглядів поєднують дві згадані дисципліни, щоб показати користь від лінгвістичних методів для автоматичних систем аналізу тональності. У цьому огляді продемонстровано, що поєднання лінгвістичної, дискурсивної та іншої контекстуальної інформації, разом із статистичним опрацюванням даних, може мати перевагу над підходами, які використовують лише один із цих аспектів. Спочатку подано вичерпне уявлення про оцінювальну лексику як з лінгвістичної, так і з обчислювальної точки зору. Після цього висловлено переконання, що загальноприйнятне обчислювальне визначення поняття оцінювальної лексики не враховує динамічний характер оцінювання, в якому тлумачення певної оцінки залежить від лінгвістичних та позалінгвістичних контекстуальних факторів. Отже, запропоновано динамічне визначення, що включає функції оновлення. Функції оновлення дозволяють включати в обчислення тональності оцінювальних слів або виразів різні контекстуальні аспекти і застосувати їх на всіх рівнях дискурсу. Досліджено кожний рівень і визначено, які мовні аспекти сприяють точному визначенню тональності. Огляд завершено коротким описом можливих майбутніх напрямів аналізу тональності, а також ролі, яку має відігравати дискурсивна і контекстуальна інформація.

Переклад М. Дубка

Встановлення референції

Pineda, L. A Model for Multimodal Reference Resolution [Модель мультимодального встановлення референції] / Luis Pineda, Gabriela Garza // Computational linguistics. – 2000. – Vol. 26. – No. 2. – Pages 139–193. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120100561665#.WIKLkH3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120100561665

Важливим аспектом інтерпретації мультимодального повідомлення є здатність визначити, коли один і той самий реальний об'єкт є референтом символів у різних модальностях. Наприклад, щоб зрозуміти підпис до картинки, потрібно розпізнати графічні символи, до яких відсилають іменники і займенники у тексті природною мовою. Цю проблему можна розглядати як проблему анафори; однак, на відміну від встановлення лінгвістичної анафори, коли антецеденти займенників вибираються із лінгвістичного контексту, під час інтерпретації текстової частини мультимодальних повідомлень антецеденти вибираються із графічного контексту. З цієї точки зору, встановлення мультимодальної референції є схожим на розв'язання анафори у різних модальностях. Інший погляд на цю проблему – уважати займенники у підписах до картинок дейктичними. При такому підході контекст інтерпретації терміну природної мови визначається як набір виразів графічної мови із добре визначеними синтаксисом і семантикою. Природна мова і графічні терміни розглядаються як переклади один одного подібні перекладам з однієї природної мови на іншу. Цей другий підхід покладено в основу представленої у статті теорії. У рамках цієї теорії розглядається відношення між мультимодальним представленням і просторовим дейксисом з одного боку і між мультимодальним міркуванням і розв'язанням дейксису з другого боку. Також розглядається інтегрована модель розв'язання анафори і дейксису у контексті інтерпретації мультимодального дискурсу.

Переклад Д. Попової, М. Погребної

van Deemter, K. On Coreferring: Coreference in MUC and Related Annotation Schemes [Про кореферентність: кореферентність у системі розмітки конференції з розуміння повідомлень і споріднених системах] / Kees van Deemter, Rodger Kibble // Computational linguistics. – 2000. – Vol. 26. – No. 4. – Pages 639–637. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120100750105966#.WIKMf33sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120100750105966

У статті стверджується, що розмітка «кореферентності», яка виконується, наприклад, учасниками конференції з розуміння повідомлень, виходить далеко за межі розмітки власне кореферентності. В результаті не завжди зрозуміло, які семантичні відносини маркує ця розмітка. У статті проаналізовано численні проблеми з цією розміткою і зроблено висновок про необхідність переосмислення завдання розмітки кореференції до її широкого застосування. Зокрема, запропоновано розділити завдання, виділяючи розмітку відносин власне кореференції з-поміж інших завдань, таких як розмітка зв'язаної анафори і взаємовідносин між підметом і предикативною іменною групою.

Переклад В. Коломієць

Mitkov, R. Introduction to the Special Issue on Computational Anaphora Resolution [Передмова до спеціального випуску, присвяченого розв'язанню анафори] / Ruslan Mitkov, Branimir Boguraev, Shalom Lappin // Computational linguistics. – 2001. – Vol. 27. – No. 4. – Pages 473–477. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120101753342626#.WIPBMn3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120101753342626

Анафора забезпечує зв'язність тексту і є феноменом, який активно вивчається і у формальній, і у комп'ютерній лінгвістиці. Правильна інтерпретація анафори є вкрай важливою для обробки природної мови. Наприклад, розв'язання анафори є ключовим завданням у інтерфейсах на базі природної мови, машинному перекладі, реферуванні текстів, видобуванні інформації, питально-відповідних системах і великій кількості інших прикладних програм з обробки природної мови.

Переклад В. Коломієць

Stuckardt, R. Design and Enhanced Evaluation of a Robust Anaphor Resolution Algorithm [Розробка і вдосконалене оцінювання робастного алгоритму розв'язання анафори] / Roland Stuckardt // Computational linguistics. – 2001. – Vol. 27. – No. 4. – Pages 479–506. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120101753342635#.WIPBQn3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120101753342635

Уже відомо, що обмеження синтаксичної коіндексації має вирішальне значення для практичних підходів до розв'язання анафори. Оскільки, зокрема через синтакстичну неоднозначність, припущення про існування однозначного тлумачення синтаксису виявилося нереалістичним, у робастному розв'язанні анафори використовуються методи подолання цього недоліку.

У статті описано алгоритм ROSANA, який узагальнює перевірку обмежень коіндексації, щоб застосувати її до недосконалих синтаксичних описів, створених робастним новітнім парсером. За допомогою формальної оцінки на двох корпусах текстів різних жанрів і тематики показано, що ROSANA забезпечує ефективне робастне встановлення кореферентності. Крім того, за допомогою поглибленого аналізу доведено, що робастне впровадження заборони на кореферентність є майже оптимальним. Проведене дослідження свідчить, що у порівнянні з підходами на основі поверхневої попередньої обробки переважно нееврістична алгоритмізація заборони на кореферентність відкриває можливості для деякого поліпшення результатів. Більше того, показано що більш значного поліпшення результатів слід очікувати на інших ланках, особливо завдяки урахуванню жанру текстів при виборі стратегій ранжування.

Дослідження ефективності системи ROSANA значною мірою спирається на удосконалену методику оцінювання систем встановлення кореферентності, розробка якої є другим важливим доробком автора. На додаток до теоретико-модельної системи оцінювання, розробленої для оцінювання на конференції з розуміння повідомлень, визначено додаткові показники оцінювання, які, з одного боку, підтримують розробника систем встановлення анафори, а з другого боку, проливають світло на прикладні аспекти інтерпретації займенників.

Переклад В. Коломієць

Tetreault, R. J. A Corpus-Based Evaluation of Centering and Pronoun Resolution [Корпусно-базоване оцінювання центрування і встановлення референції займенників] / Joel R. Tetreault // Computational linguistics. – 2001. – Vol. 27. – No. 4. – Pages 507–520. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120101753342644#.WIPBSX3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120101753342644

У статті порівнюються алгоритми встановлення референції займенників і описується алгоритм центрування (центрування зліва направо), в основу якого покладено обмеження і правила теорії центрування і який є альтернативою алгоритму Бренанна, Фрідмана і Поларда (1987). Цей алгоритм центрування зліва направо було використано для того, щоб перевірити чи дійсно дві психолінгвістичні теорії ранжування Cf-списку поліпшують точність встановлення референції займенників. Результатом дослідження стала розробка нового ранжування Cf-списку на основі синтаксису і корпусно-базовані дані, що суперечать згаданим психолінгвістичним
теоріям.

Переклад В. Туз, М. Погребної

Meng Soon, W. A Machine Learning Approach to Coreference Resolution of Noun Phrases [Встановлення кореферентності іменних груп на основі машинного навчання] / Wee Meng Soon, Hwee Tou Ng, Daniel Chung Yong Lim // Computational linguistics. – 2001. – Vol. 27. – No. 4. – Pages 521–544. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120101753342653#.WIPBS33sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120101753342653

У статті описано підхід на основі машинного навчання до встановлення кореферентності іменних груп у необмеженому тексті. Модель навчається на невеликому за обсягом анотованому корпусі і встановлює кореферентність не лише певного типу іменних груп (наприклад, займенників), а радше загальних іменних груп. Вона також не накладає обмежень на типи іменних груп за значенням; тобто кореферентність встановлюється незалежно від типу іменної групи («організація», «особа» тощо). Модель тестувалась на загальних наборах даних (а саме, корпусах з розміткою кореферентності MUC-6 і MUC-7), отримано обнадійливі результати, які свідчать, що підхід на основі машинного навчання є перспективним для встановленнння кореферентності загальних іменних груп і за точністю є рівноцінним підходам без машинного навчання. Наша система є першою системою на основі машинного навчання, аналогічною найкращим сучасним системам, у яких не використовується машинне навчання, за ефективністю на цих наборах даних.

Переклад В. Туз, М. Погребної

Palomar, M. An Algorithm for Anaphora Resolution in Spanish Texts [Алгоритм розв'язання анафори у текстах іспанською мовою] / Manuel Palomar, Antonio Ferrández, Lidia Moreno, Patricio Martínez-Barco, Jesús Peral, Maximiliano Saiz-Noeda, Rafael Muñoz // Computational linguistics. – 2001. – Vol. 27. – No. 4. – Pages 545–567. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120101753342662#.WIPDrX3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120101753342662

У статті представлено алгоритм розпізнавання іменних груп, які виступають у ролі антецедентів особових займенників третьої особи, вказівних, зворотних та пропущених (нульових) займенників у необмежених текстах іспанською мовою. Визначено список обмежень та преференцій для різних типів займенникових виразів, а також детально задокументувано важливість кожного виду знань (лексичних, морфологічних, синтаксичних та статистичних) для розв'язання анафори у іспанській мові. У статті наведено визначення синтаксичних умов відсутності кореферентності типу іменна група-займенник у іспанській мові з використанням часткового синтаксичного аналізу. Алгоритм оцінювався на корпусі, який містить 1,677

займенників, було досягнуто коефіцієнт успішності 76,8%. Також було використано чотири конкуруючі алгоритми, ефективність яких була перевірена за допомогою оцінки наосліп на тому ж тестовому корпусі. Цей новий підхід можна легко застосувати для інших мов, таких як англійська, португальська, італійська чи японська.

Переклад В. Туз, М. Погребної

Byron, D. K. The Uncommon Denominator: A Proposal for Consistent Reporting of Pronoun Resolution Results [Відсутність єдиного підходу: пропозиція узгодити звітність про результати встановлення референції займенників] / Donna K. Byron // Computational linguistics. – 2001. – Vol. 27. – No. 4 – Pages 569–577. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120101753342671#.WIPDwH3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120101753342671

При встановленні референції займенників дослідники непослідовно обчислюють коефіцієнт успішності й не надають повного опису результатів. Пропонується нова норма складання звітності, яка підвищує якість представлення окремих результатів і збільшує шанси читачів порівняти методи, використані у різних дослідженнях. Також запропоновано поряд із точністю і повнотою використовувати нову інформативну метрику ефективності – відсоток розв'язання.

Переклад В. Туз, М. Погребної

Branco, A. Binding Machines [Автоматичне зв'язування] / António Branco // Computational linguistics. – 2002. – Vol. 28. – No. 1. – Pages 1–18. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120102317341747#.WIPFE33sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120102317341747

Зв'язуючі обмеження утворюють один з найбільш робастних модулів граматичних знань. Незважаючи на свою крослінгвістичну універсальність і практичну значимість для розв'язання анафори, вони опиралися повній інтеграції у автоматичний граматичний аналіз. Основною причиною цього є вихідний принцип всебічної коіндексації для їх специфікації і верифікації. В якості альтернативи пропонується підхід, який дозволяє специфікацію зв'язуючих обмежень на основі уніфікації, але передбачає методологію верифікації, яка допомагає позбутися існуючих недоліків. Цей альтернативний підхід базується на уявленні про те, що анафоричні імена можна уважати біндерами.

Переклад В. Коломієць

Miltsakaki, E. Toward an Aposynthesis of Topic Continuity and Intrasentential Anaphora [Про апосинтез* тематичної цілісності і

міжреченнєву анафору] / Eleni Miltsakaki // Computational linguistics. – 2002. – Vol. 28. – No. 3. – Pages 319–355. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120102760276009#.WIPO8H3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120102760276009

Проблема підбору референтів для анафоричних виразів аналізувалась у великій кількості літературних джерел і за допомогою різноманітних підходів були досягнуті вагомі результати. Проте жодна окрема модель не може упоратися з усіма випадками. Ми уважаємо, що це спричинене неспроможністю моделей розрізнити два окремі процеси. На основі теоретичних висновків та емпіричних даних з різних мов пропонується апосинтетична* модель дискурсу, де тематична цілісність, обчислена для всіх одиниць, і притаманні цим одиницям пріоритети фокусування знаходяться під дією різних механізмів. Виявлені для всіх одиниць (тобто у різних реченнях) пріоритети фокусування найкраще моделювати за допомогою структурного методу згідно з теорією центрування. Механізм фокусування всередині одиниці, як стверджується у працях із семантичного/прагматичного фокусування, керується пріоритетами, спроектованими семантикою дієслів і сполучних слів, які містить одиниця. Показано, що таке розмежування не тільки вирішило важливі проблеми розв'язання анафори, але й зблизило суперечливі, на перший погляд, результати, представлені в літературі. Детально описано модель вирішення анафори, що чергує ці два механізми. Основні гіпотези запропонованої моделі перевірено у експериментальному дослідженні з англійської мови та корпусно-базованому дослідженні з грецької мови.

*Апосинтез – це грецьке слово, яке означає "декомпозиція", тобто виокремлення компонентів того, що виглядає єдиним цілим.

Переклад В. Туз, М. Погребної

Bos, J. Implementing the Binding and Accommodation Theory for Anaphora Resolution and Presupposition Projection [Використання теорії прив'язування і пристосування для розв'язання анафори і проекції пресупозиції] / Johan Bos // Computational linguistics. – 2003. – Vol. 29. – No. 2. – Pages 179–210. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120103322145306#.WIPPrn3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120103322145306

У статті розглядаються обчислювальні аспекти запропонованої Ван дер Сандтом теорії прив'язування і пристосування (binding and accommodation theory, *скор.* BAT) для проекції пресупозицій і розв'язання анафори. BAT переформульована відповідно до вимог комп'ютерного впровадження, яке передбачає операції зі структурами репрезентації дискурсу (переіменування і злиття), репрезентацію пресупозицій (можливість вибіркового зв'язування і

визначення вільних і зв'язаних змінних) і формулювання обмежень прийнятності, накладених ВАТ. Описано ефективний алгоритм розв'язання пресупозицій, представлені й інтегровані у вказаний алгоритм кілька подальших удосконалень, таких як першість у прив'язуванні і пристосуванні. Нарешті, проаналізовано інноваційне застосування високоточних програм для доведення теорем для контролю узгодженості репрезентацій дискурсу.

Переклад В. Коломієць

Markert, K. Comparing Knowledge Sources for Nominal Anaphora Resolution [Порівняння джерел знань для розв'язання іменної анафори] / Katja Markert, Malvina Nissim // Computational linguistics. – 2005. – Vol. 31. – No. 3. – Pages 367–402. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120105774321064#.WIPQ9n3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120105774321064

Порівнюються два шляхи отримання лексичної інформації для вибору антецедента у порівняльній анафорі та кореферентності означених іменникових груп. Зокрема, порівнюються алгоритм, який спирається на посилання, закодовані у створеній вручну лексичній ієрархії WordNet, і алгоритм, що видобуває знання з корпусів за допомогою спрощених лексико-семантичних моделей. В якості корпусів використано Британський Національний Корпус (БНК), а також Інтернет, який раніше для цього завдання не використовувався. Отримані результати свідчать, що (а) знань, закодованих у WordNet, часто недостатньо, особливо для анафоричних відношень, які використовують суб'єктивні або асоціативні знання; (б) у розв'язанні порівняльної анафори алгоритм на основі Інтернету перевершує алгоритм на основі WordNet; (в) у розв'язанні кореферентності означених іменникових груп алгоритм на основі Інтернету дає такі самі результати, як і алгоритм на основі WordNet, при використанні всього корпусу, але перевершує алгоритм на основі WordNet при використанні підкорпусів; (г) в обох дослідженнях алгоритм на основі БНК виявився менш ефективним через розрідженість даних. Отже у проведених дослідженнях алгоритм на основі Інтернету частково компенсував відсутність лексичних знань, яка часто дається взнаки у розв'язанні анафори, і впорався з прикладами з контекстно-залежними анафоричними відношеннями. Завдяки своїй дешевизні і відсутності потреби у ручному моделюванні лексичних знань, він є перспективним джерелом знань для інтеграції в системи розв'язання анафори.

Переклад А. Синящик

Yang, X. A Twin-Candidate Model for Learning-Based Anaphora Resolution [Метод розв'язання анафори з використанням машинного навчання з одним кандидатом] / Xiaofeng Yang, Jian Su, Chew Lim Tan // Computational linguistics. – 2008. – Vol. 34. – No. 3. – Pages 327–356. –

Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2008.07-004-R2-06-57#.WIPT133sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2008.07-004-R2-06-57

Традиційний метод розв'язання анафори з використанням машинного навчання з одним кандидатом розглядає потенційні антецеденти анафора ізольовано і тому не може точно ранжувати потенційні антецеденти для навчання і розв'язання анафори. Для вирішення даної проблеми пропонується метод розв'язання анафори з двома кандидатами. Головна ідея, покладена в основу методу, полягає у перетворенні проблеми розв'язання анафори у проблему ранжування. Конкретніше, модель вибудовує класифікатор, який ранжує потенційні антецеденти і під час розв'язання вибирає антецедент певного анафора, виходячи із ранжування кандидатів. У статті представлено детальний опис методу розв'язання анафори з двома кандидатами. Крім того, розглянуто способи використання методу у складнішому завданні встановлення кореференції. Здійснено оцінювання методу з двома кандидатами за допомогою наборів даних для автоматичного видобування змісту. Результати експерименту свідчать, що запропонований метод із двома кандидатами є ефективнішим, ніж метод із одним кандидатом, у розв'язанні займенникової анафори. Він також однаково ефективно або ефективніше справляється із встановленням кореференції.

Переклад В. Коломієць

Recasens, M. On Paraphrase and Coreference [Про перифраз і кореференцію] / Marta Recasens, Marta Vila // Computational linguistics. – 2010. – Vol. 36. – No. 4. – Pages 639–647. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli_a_00014#.WIPUUX3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli_a_00014

Завдяки забезпеченню кращого розуміння перифрази і кореференції з точки зору схожості і розбіжностей у їх лінгвістичній природі у статті уточнюються цілі видобування перифраз і встановлення кореференції і наскільки вони можуть допомогти одне одному. Стверджується, що це обговорення має безпосереднє відношення до обробки природної мови.

Переклад В.Коломієць

Siddharthan, A. Information Status Distinctions and Referring Expressions: An Empirical Study of References to People in News Summaries [Ознаки статусу інформації і референційні вирази: емпіричне дослідження посилань на людей у зведеннях новин] / Advaith Siddharthan, Ani Nenkova, Kathleen McKeown // Computational linguistics. – 2011. – Vol. 37. – No. 4. – Pages 811–842. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00077#.WIPUwH3

sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00077

Хоча існує багато теоретичних праць, присвячених використанню різних ознак статусу інформації для пояснення форм посилань у письмових текстах, досліджень, у яких здіснена спроба автоматично виявляти ці ознаки для генерування посилань у контексті автоматично відтвореного тексту, проведено недостатньо. У статті описано модель генерування посилань на людей у випусках новин, у якій використано ідеї як з теоретичних праць, так і з корпусного аналізу інформаційних випусків, написаних людьми. Зокрема, запропонована модель виявляє, як дві характеристики згаданої у випуску новин особи – чи відома вона читачеві і яка її загальна роль у випуску новин – впливають на зміст і форму першого посилання на ту особу в інформаційному випуску. Показано, що ці дві ознаки можна виявити у типовому введенні для багатодокументного реферування і що їх можна використати у процесі генерування для підвищення якості екстрактивних рефератів.

Переклад В. Коломієць

Sapena, E. A Constraint-Based Hypergraph Partitioning Approach to Coreference Resolution [Метод встановлення кореференції шляхом розбиття гіперграфів на основі обмежень] / Emili Sapena, Lluís Padró, Jordi Turmo // Computational linguistics. – 2013. – Vol. 39. – No. 4. – Pages 847–884. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00151#.WIPVrn3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00151

Стаття присвячена дослідженню машинного навчання для встановлення кореференції. Встановлення кореференції — це завдання обробки природньої мови, яке полягає у визначенні виразів у дискурсі, які відносяться до одного об’єкта.

Головними положеннями цієї статті є 1) новий підхід до встановлення кореференції на основі дотримання обмежень, у якому проблема представлена у вигляді гіперграфа і вирішується шляхом розмитої розмітки і 2) дослідження підвищення результативності встановлення кореференції шляхом використання знань про світ, отриманих з Вікіпедії.

Розроблений метод може з більшою виразністю, ніж методи на основі пар, використовувати модель класифікації згадувань об’єктів і долати слабкі місця попередніх підходів у сучасних системах, такі як зв’язування суперечностей, класифікації без контексту і оцінювання пар за нестачі інформації. Крім того, запропонований метод дозволяє вбудовувати нову інформацію шляхом додавання обмежень. Також здійснено дослідження для того, щоб використовувати знання про світ з метою підвищенння результативності.

Програма RelaxCor, яка є втіленням запропонованого підходу, за результативністю не поступається сучасним системам і брала участь у міжнародних змаганнях SemEval-2010 і CoNLL-2011. RelaxCor зайняла друге місце у змаганні CoNLL-2011.

Переклад М. Драчової і К. Погорєлова

Lee, H. Deterministic Coreference Resolution Based on Entity-Centric, Precision-Ranked Rules [Детерміністичний підхід до встановлення кореференції на основі об'єктно-орієнтованих, ранжованих за точністю правил] / Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky // Computational linguistics. – 2013. – Vol. 39. – No. 4. – Pages 885–916. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00152#.WIRq333sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00152

У статті описано новий детерміністичний підхід до встановлення кореференції, який поєднує загальну інформацію і конкретні характеристики сучасних моделей на основі машинного навчання з прозорістю і модульним принципом організації детерміністичних систем на основі правил. Наша архітектура фільтрації по черзі застосовує комплекс детерміністичних моделей кореференції від найвищої до найнижчої точності, у якому кожна модель базується на результатах кластера попередньої моделі. Два рівні розробленої архітектури фільтрації: рівень визначення згадування, орієнтований переважно на повноту, за яким слідують фільтри референції, орієнтовані на точність – це надійний спосіб досягнення як високої точності, так і високої повноти. Крім того, у запропонованому підході використовується глобальна інформація за допомогою об'єктно-орієнтованої моделі, яка сприяє уніфікації характеристик усіх згадувань того самого реального об'єкта. Незважаючи на свою простоту, запропонований метод дозволив досягти конкурентноспроможних результатів на кількох корпусах і жанрах і був вбудований у сучасні гібридні системи встановлення кореференції для китайської і арабської мов. Отже, у розробленій системі втілена нова парадигма об'єднання знань у системах на основі правил, яка матиме наслідки для комп'ютерної лінгвістики загалом.

Переклад В. Коломієць

Bejan, C. A. Unsupervised Event Coreference Resolution [Неконтрольоване виявлення кореферентних номінацій події] / Cosmin Adrian Bejan, Sanda Harabagiu // Computational linguistics. – 2014. – Vol. 40. – No. 2. – Pages 311–347. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00174#.WIRpyH3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00174

Завдання виявлення кореферентних номінацій події відіграє ключову роль у багатьох системах обробки природної мови, таких як видобування інформації, питання-відповідь, визначення і контроль тематики. У статті описано новий класс неконтрольованих, непараметричних байесовських моделей для вірогіднісного визначення кореферентних сукупностей номінацій події у наборі нерозмічених документів. Для визначення цих сукупностей із набору документів автоматично видобуваються лексичні, синтаксичні і семантичні характеристики кожної номінації події. Створення великого набору характеристик кожної номінації події дозволяє розглядати завдання виявлення кореферентних номінацій події як завдання групування номінацій із однаковими характеристиками (вони мають одних і тих же учасників, відбуваються в одному й тому ж місці, в один і той же час тощо).

Деякі з найскладніших проблем неконтрольованого виявлення кореферентних номінацій події пов'язані з (а) вибором представлення номінацій події у вигляді великого набору характеристик і (б) можливістю моделювати події, описані в одному й тому ж і в багатьох документах. Наша перша некотрольована модель, яка вирішує ці проблеми, являє собою узагальнення ієрархічного процесу Діріхле. Це нове доповнення демонструє здатність ієрархічного процесу Діріхле виявляти невизначеність кількості компонентів сукупності і крім того враховує будь-яке кінечне число характеристик, які асоціюються з кожною номінацією події. Більше того, щоб подолати деякі обмеження цього доповнення, створена нова гібридна модель, яка об'єднує необмежену латентно-класову модель і модель для дискретних часових рядів. Головною перевагою цієї гібридної моделі є її здатність автоматично робити на основі даних висновок про кількість характеристик, які асоціюються із кожною номінацією події, і водночас автоматично відбирати найінформативніші характеристики для виявлення кореферентних номінацій події. Оцінка виявлення кореферентних номінацій події у одному і різних документах свідчить про значне удосконалення цих моделей у порівнянні з вихідними показниками для цього завдання.

Переклад В. Коломієць

Fernandes, E. R. Latent Trees for Coreference Resolution [Приховані дерева для встановлення кореференції] / Eraldo Rezende Fernandes, Cícero Nogueira dos Santos, Ruy Luiz Milidiú // Computational linguistics. – 2014. – Vol. 40. – No. 4. – Pages 801–835. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00200#.WIRpR33sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00200

У статті описано систему структурного аналізу на основі машинного навчання для необмеженого встановлення кореференції, яка використовує дві ключові методики моделювання: приховані дерева кореференційних зв'язків і автоматичне виведення ознак на основі ентропії. Моделювання на основі прихованих дерев уможливлює комп'ютерне навчання, бо включає

інформативну приховану структуру. Крім того, користуючись методом автоматичного виведення ознак, можна ефективно створювати удосконалені нелінійні моделі за допомогою автоматичних алгоритмів лінійних моделей. У статті наведено емпіричні результати, які висвітлюють роль кожного методу моделювання, використаного у розробленій системі. Емпірична оцінка зійснювалась за допомогою багатомовних необмежених баз даних для змагань у рамках конференції CoNLL-2012, які охоплюють три мови: арабську, китайську і англійську. До усіх мов застосовувалась одна й та ж система за винятком незначних адаптацій до специфічних характеристик мови, таких як вкладені посилання і спеціальні статичні списки займенників. Попередня версія цієї системи була представлена на закритих змаганнях конференції CoNLL-2012 і стала найкращою серед конкурсантів із офіційним результатом 58,69. Єдиним удосконаленням останньої версії системи є додавання можливих дуг, які зв'язують вкладені посилання для китайської мови. Завдяки додаванню таких дуг показники для цієї мови зросли майже на 4,5 пункти. Результат існуючої системи становить 60,15 і відповідає зменшенню кількості помилок на 3,5%. Це найрезультативніша система для кожної з трьох мов.

Переклад В. Коломієць

Генерування тексту

Rubinoff, R. Integrating Text Planning and Linguistic Choice Without Abandoning Modularity: The IGEN Generator [Інтеграція планування тексту та лінгвістичного вибору без відмови від модульності: генератор IGEN] / Robert Rubinoff // Computational linguistics. – 2000. – Vol. 26. – No. 2. – Pages 107–138. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120100561656#.WITGzX3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120100561656

Як правило, генерування природної мови ділиться на компонент планування тексту і лінгвістичний компонент. Проте, цей поділ базується на припущенні, що два зазначені компоненти можуть діяти незалежно один від одного, що не завжди так. Генератор IGEN усуває необхідність подібного припущення; він управляє взаємодією компонентів, зберігаючи переваги модульності. IGEN робить це за допомогою коментарів, які його лінгвістичний компонент розміщує на структурах, які він будує; ці коментарі містять формальний опис наслідків конкретних лінгвістичних рішень, що дозволяє планувальнику оцінити ці рішення, не маючи жодних лінгвістичних знань. Цей підхід дозволяє IGEN вносити зміни в роботу, виконану окремо кожним компонентом, навіть у випадках, коли кінцевий результат залежить від взаємодії між ними. Крім того, оскільки IGEN моделює усі можливі наслідки лінгвістичних рішень, він може ефективно працювати в умовах обмеженого часу або мовних ресурсів.

Переклад Д. Попової

Reiter, E. Pipelines and Size Constraints [Програмні конвеєри і обмеження обсягу] / Ehud Reiter // Computational linguistics. – 2000. – Vol. 26. – No. 2. – Pages 251–259. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120100561692#.WIS2aH3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120100561692

Деякі типи документів повинні відповідати вимогам до обсягу, наприклад не перевищувати обмеження кількості сторінок. У конвеєрній системі генерування природної мови (ГПМ) виконати цю вимогу може бути складно, тому що обсяг залежить в основному від змісту, який визначається на початковому етапі програмного конвеєру, але обсяг не може бути точно визначений, доки система ГПМ не завершить оброблення документа. В статті представлено результати експериментальної перевірки здатності однофазового конвеєру, багатофазового конвеєру та модернізованих варіантів системи СТОП (яка генерує індивідуалізовані заклики до відмови

від куріння) задовільнити обмеження обсягу. Ці дані свідчать, що багатофазовий програмний конвеєр працює набагато краще, ніж однофазовий, а найкращий результат показує модернізована система.

Переклад О. Мартинюк

Bateman, J. Towards Constructive Text, Diagram, and Layout Generation for Information Presentation [Представлення інформації шляхом створення оригінального тексту, діаграм та зовнішнього вигляду сторінки] / John Bateman, Thomas Kamps, Jörg Kleinz, Klaus Reichenberger // Computational linguistics. – 2001. – Vol. 27. – No. 3. – Pages 409–449. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120101317066131#.WIS2xX3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120101317066131

Належне поєднання елементів, що не виходить за рамки гармонійного оформлення сторінки, є широко відомою та невід'ємною частиною презентації складної інформації. Проте у комп'ютерних презентаціях на питання про точну функцію і природу оформлення сторінки зверталось недостатньо уваги, дослідники часто обмежуються відносно локальними проблемами шрифтів і форматування тексту, залишаючи без уваги важливіше питання оформлення сторінок. Стаття присвячена вибору і функції оформлення сторінок, яке правильно поєднує текстові та графічні способи подання інформації для створення гармонійного дизайну презентації. Продемонстровано, що поряд із більш традиційними інструментами, такими як форматування тексту та внутрішньотекстова розмітка зв'язків дискурсу, багаті можливості досягнення гармонійності презентації криються у оформленні сторінки. У генерації зовнішнього вигляду сторінки, тексту і діаграм використано інтегративний підхід. Наш метод було розроблено на основі попереднього емпіричного дослідження професійно створених стилів сторінок і реалізовано в експериментальній інформаційній системі в галузі історії мистецтва.

Переклад І. Снєгурова

van Deemter, K. Generating Referring Expressions: Boolean Extensions of the Incremental Algorithm [Генерування референційних виразів: булеві розширення інкрементального алгоритму] / Kees van Deemter // Computational linguistics. – 2002. – Vol. 28. – No. 1. – Pages 37–52. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120102317341765#.WJxU1_LsSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120102317341765

У цій статті розглядається підхід до генерування референційних виразів і приділяється увага незавершеності існуючих алгоритмів у даній сфері. Після

ознайомлення з посиланнями на індивідуальні об’єкти, ми обговорюємо посилання на множини, в тому числі булевські описи, в яких використовуються властивості заперечення і роз’єднання. Для того, щоб забезпечити генерування відрізняючого опису кожен раз коли такі описи зустрічаються, у статті запропоновано узагальнення і розширення інкрементального алгоритму Дейла й Рейтера (1995).

Переклад І. Снєгурова

Reiter, E. Human Variation and Lexical Choice [Відмінності між людьми і вибір лексики] / Ehud Reiter, Somayajulu Sripada // Computational linguistics. – 2002. – Vol. 28. – No. 4. – Pages 545–553. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120102762671981#.WITF3n3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120102762671981

У багатьох дослідженнях обробки природної мови імпліцитно припускається, що у мовній спільноті значення слів є чітко визначеними, проте насправді є багато доказів того, що різні люди асоціюють слова з дещо різними значеннями. У статті узагальнено докази цього твердження з літератури і зі здійснюваних дослідницьких проектів і проаналізовано його значення для генерування природної мови, особливо для вибору лексики, тобто підбору слів для генерованого тексту.

Переклад В. Коломієць

Krahmer, E. Graph-Based Generation of Referring Expressions [Генерування референційних виразів на основі графів] / Emiel Krahmer, Sebastiaan van Erk, André Verleg // Computational linguistics. – 2003. – Vol. 29. – No. 1. – Pages 53–72. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120103321337430#.WIS3Zn3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120103321337430

У статті описано новий підхід до генерування референційних виразів. Пропонується формалізувати середовище (що складається з набору об'єктів з різними характеристиками і відносинами) як розмічений орієнтований граф і описувати вибір змісту (які характеристики включати у референційний вираз) як задачу створення підграфа. Для управління процесом пошуку і вибору певних рішень з-поміж інших використовуються функції затрат. Запропонований підхід має чотири основні переваги: (1) графові структури вивчались досить широко, тому використання графів відкриває прямий доступ до багатьох теорій і алгоритмів для роботи з графами; (2) багато алгоритмів нинішнього покоління можуть бути переформульовані мовою графів, завдяки чому полегшується порівняння та інтеграція різних підходів; (3) використання графів дозволяє розв’язати низку проблем, від яких

страждали попередні алгоритми генерації референційних виразів; і (4) спільне використання графів і функцій затрат прокладає шлях до інтеграції методів на основі правил з новішими стохастичними підходами.

Переклад К. Погорєлова

Power, R. Document Structure [Структура документа] / Richard Power, Donia Scott, Nadjet Bouayad-Agha // Computational linguistics. – 2003. – Vol. 29. – No. 2. – Pages 211–260. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120103322145315#.WIE26n3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120103322145315

Ми уважаємо абстрактну структуру документа окремим описовим рівнем аналізу та генерування письмових текстів. Мета такої схеми – бути сполучною ланкою між змістом тексту (тобто структурою його дискурсу) і його формою (тобто поділом на графічні складові, як-от розділи, абзаци, речення, марковані списки, цифри і примітки). Абстрактну структуру документа можна розглядати як компонент «граматики тексту» Дж. Нанберга; вона також тісно пов'язана з «логічною» розміткою у таких мовах як HTML і LaTEX. Ми демонструємо, що використовуючи це проміжне представлення, можна чіткіше визначити декілька підзадач у генеруванні та розумінні мови.

Переклад А. Синящик

Kibble, R. Optimizing Referential Coherence in Text Generation [Підвищення референційної когерентності у генеруванні текстів] / Rodger Kibble , Richard Power // Computational linguistics. – 2004. – Vol. 30. – No. 4. – Pages 401–416. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/0891201042544893#.WIS3_H3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/0891201042544893

У статті описана діюча система, яка використовує теорію центрування для планування зв'язних текстів і вибору референційних виразів. Стверджується, що однією з цілей планування текстів і речень має бути підтримання референційної цілісності і, як наслідок, спрощення встановлення займенникової референції. Можливість неоднозначного вживання займенників можна зменшити, забезпечивши відповідну послідовність клауз і аргументів усередині клауз. Основою для такого інтегрованого підходу є теорія центрування. Згідно теорії центрування генерування зв'язних текстів розглядається як завдання врахування обмежень. Добре відоме правило 2 теорії центрування переформульоване як набір обмежень – зв'язність, виразність, дешевизна і нерозривність. Показано зразки виведень, отримані завдяки певному врахуванню цих обмежень. Цей метод полегшує детальне дослідження метрик оцінювання і тому стане ефективним дослідницьким

інструментом на додаток до миттєвої практичної вигоди у вигляді пришвидшення і полегшення сприйняття згенерованих текстів. Метод застосовується у системах генерування природної мови, які здійснюють ієрархічне структурування текстів на основі теорії когерентних відносин з певними додатковими припущеннями.

Переклад В. Коломієць

Van Deemter, K. Generating Referring Expressions that Involve Gradable Properties [Генерування референційних виразів з градуальними характеристиками] / Kees van Deemter // Computational linguistics. – 2006. – Vol. 32. – No. 2. – Pages 195–222. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2006.32.2.195#.WIS5EX3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2006.32.2.195

У статті досліджена роль градуальних параметрів референційних виразів з точки зору генерування природної мови. Спочатку описано простий семантичний аналіз нечітких описів (тобто референційних виразів, до складу яких входять градуальні прикметники), який відображає у них контекстно-залежне значення прикметників. Потім показано, як цей різновид аналізу може використовуватися у алгоритмах генерування нечітких описів на основі числових даних. Нарешті, розглянуто питання про те, коли потрібно використовувати такі описи. У заключній частині статті розглядаються виділеність і націленість, які аналізуються так, ніби вони є градуальними прикметниками.

Переклад В. Коломієць

Lapata, M. Automatic Evaluation of Information Ordering: Kendall's Tau [Автоматична оцінка упорядкування інформації: тау Кендала] / Mirella Lapata // Computational linguistics. – 2006. – Vol. 32. – No. 4. – Pages 471–484. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2006.32.4.471#.WIEJEH3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2006.32.4.471

У статті розглядається питання упорядкування інформації, яка є основою багатьох програм опрацювання текстів природною мовою, таких як генерування тексту з концептуальних представлень і багатодокументне реферування. Запропоновано метод оцінювання на основі коефіцієнта рангової кореляції τ Кендала. Цей метод є недорогим, надійним і незалежним від представлення. Продемонстровано, що коефіцієнт рангової кореляції τ Кендала надійно корелює з експертними оцінками і часом зчитування.

Переклад В. Коломієць

Paraboni, I. Generating Referring Expressions: Making Referents Easy to Identify [Генерування референційних виразів: спрощення ідентифікації

референтів] / Ivandré Paraboni, Kees van Deemter, Judith Masthoff // Computational linguistics. – 2007. – Vol. 33. – No. 2. – Pages 229–254. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2007.33.2.229 – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2007.33.2.229

Референційні вирази часто потрібно вибирати таким чином, щоб було легко розпізнати їх референти. Стаття присвячена рефернційним виразам у ієрархічно структурованих тематичних областях і досліджує гіпотезу про те, що референційні вирази можна удосконалити, включивши до них логічно надлишкову інформацію, якщо таким чином можна значно пришвидшити знаходження та ідентифікацію референта. Описано загальні алгоритми, які втілюють цю ідею, шляхом включення у загальний вираз логічно надлишкової інформації у деяких чітко окреслених ситуаціях. Для перевірки висунутої гіпотези і для оцінки продуктивності запропонованих алгоритмів було проведено два керовані експерименти з участю людей. Перший експеримент підтвердив, що експерти віддають перевагу логічно надлишковим виразам у випадках, у яких це було передбачено нашим алгоритмом. Другий ескперимент свідчить, що створена нашим алгоритмом логічна надлишковість іде на користь читачам з точки зору зусиль, потрібних на ідентицікацію референта виразу.

Переклад В. Коломієць

Karamanis, N. Evaluating Centering for Information Ordering Using Corpora [Оцінка застосування центрування в упорядкуванні інформації за допомогою корпусів] / Nikiforos Karamanis, Chris Mellish, Massimo Poesio, Jon Oberlander // Computational linguistics. – 2009. – Vol. 35. – No. 1. – Pages 29–46. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.07-036-R2-06-22#.WIS9cH3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.07-036-R2-06-22

У статті розглядаються кілька мір когерентності, визначених за допомогою теорії центрування, і досліджується придатність таких мір для упорядкування інформації у автоматичному генеруванні текстів. Емпірично виявлено найперспективнішу міру і перевірено її ефективність шляхом застосування загальної методики до кількох корпусів. Головний висновок полягає в тому, що найпростіша міра (яка спирається виключно на переходи NOCB) встановлює надійний вихідний рівень, який не можуть перевершити інші міри, які користуються додатковими рисами центрування. Цей вихідний рівень можна застосувати у розробці систем генерування тексту як на основі тексту, так і на основі концептуальних представлень.

Переклад В. Коломієць

Reiter, E. An Investigation into the Validity of Some Metrics for Automatically Evaluating Natural Language Generation Systems [Дослідження валідності деяких метрик автоматичного оцінювання систем генерування природної мови] / Ehud Reiter, Anja Belz // Computational linguistics. – 2009. – Vol. 35. – No. 4. – Pages 529–558. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2009.35.4.35405#.WIS-p33sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2009.35.4.35405

Зростає зацікавленість у використанні автоматично обчислених метрик оцінювання для оцінювання систем генерування природної мови (ГПМ), адже вони часом значно дешевші, ніж оцінки експертів, які традиційно використовуються у ГПМ. У статті вміщено аналіз попередніх досліджень оцінювання ГПМ і валідації автоматичних метрик у опрацюванні природної мови, а потім представлено результати двох досліджень того, наскільки деякі метрики, популярні в інших областях опрацювання природної мови (особливо BLEU і ROUGE), корелюють із судженнями експертів у предметній області згенерованих комп'ютером прогнозів погоди. Отримані результати свідчать, що принаймні у цій предметній області метрики можуть бути корисною міркою якості мови, хоча докази цього не такі беззаперечні, як нам хотілося б у ідеалі; втім вони не є корисною метрикою якості змісту. Також проаналізовано велику кількість застережень, які потрібно пам'ятати під час інтерпретації результатів цього та інших валідаційних досліджень.

Переклад В. Коломієць

Madnani, N. Generating Phrasal and Sentential Paraphrases: A Survey of Data-Driven Methods [Генерування перефразувань словосполучень і речень: огляд методів, керованих даними] / Nitin Madnani, Bonnie J. Dorr // Computational linguistics. – 2010. – Vol. 36. – No. 3. – Pages 341–387. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli_a_00002#.WITBRX3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli_a_00002

Завдання перефразування дуже добре знайоме носіям усіх мов. Більше того, завдання автоматичної генерування або видобування семантичних еквівалентів різних одиниць мови – слів, словосполучень і речень – є важливим компонентом опрацювання природної мови і все частіше використовується для підвищення ефективності різного програмного забезпечення для опрацювання природної мови. У статті зроблена спроба здійснити всебічний і незалежний від комп'ютерних програм аналіз керованих даними методів генерування перефразування словосполучень і речень, одночасно демонструючи розуміння важливості і потенційного використання перефразування в дослідженнях опрацювання природної мови.

Також проаналізовано досягнення в ручному і автоматичному створенні корпусів перефразувань. Нарешті, обговорено стратегії оцінювання методів генерування перефразувань і коротко розглянуто деякі новітні тенденції у генеруванні перефразувань.

Переклад В. Коломієць

Mairesse, F. Controlling User Perceptions of Linguistic Style: Trainable Generation of Personality Traits Segmentation [Контроль сприйняття користувачем мовного стилю: генерування сегментування характеристик особистості на основі машинного навчання] / François Mairesse, Marilyn A. Walker // Computational linguistics. – 2011. – Vol. 37. – No. 3. – Pages 455–488. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00063#.WITNUH3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00063

Останнім часом дослідники генерування природної мови почали ураховувати мовне варіювання, створюючи алгоритми, здатні модифікувати мовний стиль системи в залежності від мовного стилю користувача або інших факторів, таких як індивідуальні особливості або ввічливість. Хоча контроль стилю завжди спирався на правила, розроблені вручну, для створення системи генерування, здатної відтворити широкий діапазон варіювання, характерний для людського діалогу, знадобляться статистичні методи. Досягнення в розробці статистичного підходу до генерування природної мови свідчать, що граматичну правильність і природність загальних висловлювань можна поліпшити за допомогою даних, проте ці керовані даними методи не спроможні забезпечити стилістичну варіативність, яка справлятиме на людей потрібне системі враження. У статті описано Personage, генератор мови з високим ступенем параметризації, параметри якого визначено на основі психологічних даних про індивідуальні мовні рефлекси. Представлено інноваційний метод генерування природної мови на основі статистичного підходу, який прогнозує рішення генератора, потрібні для передачі будь-якої комбінації скалярних значень в рамках п’яти основних вимірів особистості. Експертна оцінка свідчить, що запропоновані моделі визначення параметрів забезпечують безперервне явно виражене стилістичне варіювання за багатьма параметрами без обчислювальної вартості методів «повторного генерування».

Переклад В. Коломієць

Power, R. Generating Numerical Approximations [Генерування числових апроксимацій] / Richard Power, Sandra Williams // Computational linguistics. – 2012. – Vol. 38. – No. 1. – Pages 113–134. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00086#.WITCLn3sSGA – Режим доступу до повнотекстової статті:

http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00086

У статті описано обчислювальну модель для планування словосполучень типу “понад чверть” і “25,9 відсотків”, які описують частини при різних рівнях точності. Модель пропонує основні варіанти у плануванні числового опису, використовуючи формальні визначення математичної форми (наприклад, різницю між долями і відсотками) і закругленість, адаптовану з попередніх досліджень. Завдання змодельоване у вигляді задачі задоволення обмежень з рішеннями, які послідовно розсортовані за преференціями (наприклад, закругленості). Деталізовані обмеження визначені за допомогою корпусу числових виразів, укладеного у проекті NumGen*, і оцінені за допомогою емпіричних досліджень, у яких інформантів просили утворити (або завершити) числові вирази у заданих умовах.

*NumGen: Генерування грамотних описів числових величин для людей з різними рівнями математичної грамотності (http://mcs.open.ac.uk/sw6629/numgen). NumGen був профінансований Радою з економічних і соціальних досліджень шляхом виділення гранту з вих. номером RES-000-22-2760.

Переклад В. Коломієць

Krahmer, E. Computational Generation of Referring Expressions: A Survey [Автоматичне генерування референційних виразів: огляд] / Emiel Krahmer, Kees van Deemter // Computational linguistics. – 2012. – Vol. 38. – No. 1. – Pages 173–218. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00088#.WIPVO33sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00088

У статті представлено огляд комп’ютерних досліджень генерування референційних виразів (*англ.* referring expression generation, *скор.* REG). Стаття знайомить із проблемою REG і описує перші дослідження у цій області, аналізуючи основні припущення, які лежать у їх основі, і демонструючи, як розширились за останні роки їх напрями. У статті проаналізовано обчислювальні платформи, які лежать в основі REG, і показано нову тенденцію, яка намагається поєднати алгоритми REG з добре усталеними методами представлення знань. Значну увагу приділено останнім спробам оцінювання алгоритмів REG і висновкам, які можна зробити на їх основі. Стаття завершується аналізом майбутніх напрямів досліджень в області REG, зосереджених на посиланнях у ширших і більш реалістичних контекстах.

Переклад В. Коломієць

Chali, Y. Towards Topic-to-Question Generation [На шляху до автоматичного генерування питань на задану тему] / Yllias Chali, Sadid

A. Hasan // Computational linguistics. – 2015. – Vol. 41. – No. 1. – Pages 1–20. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00206 – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00206

Ця стаття присвячена автоматичному генеруванню всіх можливих запитань на задану тему. Зокрема, вважаємо, що кожна тема асоціюється з корпусом текстів, які містять корисну інформацію про тему. Крім того, генерування запитань відбувається шляхом використання відомостей про носіїв власних назв та аргументно-предикативних структур речень з корпусу текстів. Значущість згенерованих запитань оцінюється за допомогою латентного розміщення Діріхле шляхом визначення підтем (тісно пов'язаних з основною темою) у конкретному корпусі текстів і застосування розширеного ядра строкових підпослідовностей для обчислення їхньої схожості з запитаннями. Також, у статті пропонується використання ядер синтаксичних дерев для автоматичної оцінки синтаксичної правильності запитань. Запитання ранжуються з урахуванням їхнього значення (в контексті конкретного корпусу текстів) та синтаксичної правильності. Подібний спосіб виконання вказаного завдання не використовувався в жодному з попередніх досліджень. Як свідчить низка експериментів, запропонований підхід до генерування запитань на задану тему може значно перевершити результати останніх досліджень.

Переклад А. Шульги

Zhang, Y. Discriminative Syntax-Based Word Ordering for Text Generation [Диференційоване впорядкування слів на основі синтаксису для генерування текстів]/ Yue Zhang, Stephen Clark // Computational linguistics. – 2015. – Vol. 41. – No. 3. – Pages 503–538. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00229 – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00229

Основною проблемою генерування текстів є упорядкування слів. У статті описано упорядкування слів на основі синтаксичного підходу та диференціальної моделі. Розглянуто два граматичні формалізми: комбінаторну категорійну граматику (ККГ) і граматику залежностей. При пошуці ймовірної послідовності слів та синтаксичному аналізі простір пошуку є дуже великим, що ускладнює автоматичне диференціювання. Автори статті розробили орієнтовану на навчання пошукову систему, що базується на першому найкращому результаті пошуку і проаналізували кілька альтернативних алгоритмів навчання.

Представлена система є гнучкою, оскільки дозволяє встановлювати обмеження на вихідні послідовності слів. Для демонстрації цієї гнучкості розглядаються різні умови вводу інформації. По-перше, досліджено так зване

«чисте» завдання з упорядкування слів, у якому вхідними даними є мультимножина слів, а завдання полягає в їх упорядкуванні у граматично правильне речення. Таке завдання вже розв'язувалось, і в статті повідомляється, що отримані результати кращі, ніж результати існуючих систем на базі стандартної тестової вибірки Wall Street Journal. По-друге, розглянуто ту саму проблему перевпорядкування, але з різними умовами вводу інформації: від «голого» набору даних без міток залежностей чи частиномовної розмітки до виключного випадку, коли вхідними даними є повна частиномовна розмітка та неупорядковані, немарковані залежності (а також різноманітні проміжні умови). При розв'язанні завдання з розділеними ресурсами конференції з генерації природної мови 2011 року за допомогою розробленої системи було отримано результати, які конкурують з результатами найкращих систем, що також підтверджує практичну цінність розробленої системи.

Переклад А. Шульги

Gardent, C. A Statistical, Grammar-Based Approach to Microplanning [Статистичний підхід до мікропланування на основі граматики] / Claire Gardent, Laura Perez-Beltrachini // Computational linguistics. – 2017. – Vol. 43. – No. 1. – Pages 1–30. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00273 – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00273

Незважаючи на те, що протягом останніх років було здійснено велику кількість досліджень, присвячених керованому даними генеруванню природної мови, залишились поза увагою дрібномодульні залежності, які з'являються під час мікропланування між агрегуванням, поверхневою реалізацією та сегментуванням речень. У статті запропоновано гібридний символьно-статистичний підхід для одночасного моделювання правил, які регулюють ці залежності. Запропонований підхід поєднує невелику створену вручну граматику, статистичний гіперрозмітник і алгоритм поверхневої реалізації. Підхід застосовано для вербалізації запитів до баз знань і протестовано на 13 базах знань, щоб показати його незалежність від галузі. Запропонований підхід оцінено кількома способами. Кількісний аналіз свідчить, що гібридний підхід перевершує суто символьний підхід як за швидкістю, так і за покриттям. Результати експертного оцінювання свідчать про те, що користувачі вважають вивід цієї гібридної статистично-символьної системи більш природним, ніж вивід суто шаблонного підходу і суто символьного підходу на основі граматики. Нарешті, на прикладах показано, що запропонований підхід може враховувати різні фактори, які впливають на агрегування, сегментування речень і поверхневу реалізацію.

Переклад М. Дубка

Paraboni, I. Effects of Cognitive Effort on the Resolution of Overspecified Descriptions [Вплив когнітивного зусилля на розуміння надмірно конкретизованих описів] / Ivandré Paraboni, Alex Gwo Jen Lan, Matheus Mendes de Sant'Ana, Flávio Luiz Coutinho. – 2017. – Vol. 43. – No. 2. – Pages 451–459. – Режим доступу до анотації: https://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00288 – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00288

Дослідження генерування референційних виразів (ГРВ) виявили різний вплив надмірної конкретизації референтів на розуміння певних описів. Для глибшого вивчення подібного впливу у статті описано два експерименти з відстеженням руху очей, у яких вимірювався час, необхідний для розпізнавання цільових об'єктів на основі різних типів інформації. Результати свідчать, що надмірна конкретизація референтів може або сприяти, або перешкоджати ідентифікації, залежно від того, яка саме інформація занадто конкретизована. Це спостереження може стати у пригоді в розробці складніших алгоритмів ГРВ, орієнтованих на слухача.

Переклад М. Дубка

Зняття лексичної багатозначності

Stevenson, M. The Interaction of Knowledge Sources in Word Sense Disambiguation [Взаємодія баз даних у знятті лексичної неоднозначності] / Mark Stevenson, Yorick Wilks // Computational linguistics. – 2001. – Vol. 27. – No. 3. – Pages 321–349. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120101317066104#.WIEg1H3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120101317066104

Ефективність вирішення проблеми лексичної неоднозначності (Word sense disambiguation, WSD) можна підвищити, перейнявши практику об'єднання різних баз даних з досліджень штучного інтелекту. Для перевірки цієї гіпотези потрібно визначити, які лексичні бази даних є найкориснішими, і з'ясувати, чи дозволяє їх об'єднання отримати кращі результати. У статті представлено систему семантичної розмітки, яка використовує декілька баз даних. Оцінка системи за допомогою нашого корпусу виявила точність понад 94%.

Наша система не обмежується обробкою обмеженого списку слів, а намагається зняти омонімію усіх повнозначних слів у тексті. Ми уважаємо, що такий підхід є більш підходящим для створення практичних систем.

Переклад М. Погребної

Edmonds, P. Near-Synonymy and Lexical Choice [Неточна синонімія і лексичний вибір] / Philip Edmonds, Graeme Hirst // Computational linguistics. – 2002. – Vol. 28. – No. 2. – Pages 105–144. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120102760173625#.WIT4Jn3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120102760173625

Створено нову обчислювальну модель представлення точних значень неточних синонімів та відмінностей між ними. Також змодельовано процес лексичного вибору, який може вирішити, котрий із декількох неточних синонімів найкраще вжити у певній ситуації. Це дослідження отримало практичне застосування у машинному перекладі і генерації тексту.

Спочатку було визначено проблеми представлення неточних синонімів у комп'ютерному лексиконі та продемонстровано, що жодна з попередніх моделей не враховує належним чином неточну синонімію. Потім було висунуто гіпотезу, яка пояснює неточну синонімію, спираючись головним чином на поняття деталізації репрезентації, згідно якого значення слова є результатом залежного від контексту поєднання контекстно-незалежного основного значення та сукупності його очевидних відмінностей від його

неточних синонімів. Таким чином, неточні синоніми утворюють кластери.

Потім на основі стандартної онтологічної моделі було розроблено кластеризовану модель лексичних знань. Модель відсікає онтологію на рівні великих структурних одиниць, уникаючи таким чином небажаного збільшення в онтології кількості залежних від мови концептів, але зберігаючи при цьому переваги ефективного обчислення і аргументування. Модель ділить неточні синоніми на субконцептуальні кластери, які з'єднуються з онтологією. Кластер розмежовує неточні синоніми у плані деталізації значення, імплікації, вираженого відношення і стилю. Модель є достатньо загальною, щоб пояснити інші типи варіацій, наприклад, у особливостях сполучуваності.

Результатом роботи кластерізованої моделі лексичних даних є ефективний, надійний і гнучкий процес точного лексичного вибору. Для того щоб модель працювала, критерії лексичного вибору було формалізовано як налаштування вираження певних концептів із різним рівнем прямоти, вираження відношення і створення певних стилів. Власне процес лексичного вибору складається з двох рівнів: між кластерами і між неточними синонімами кластерів. Описано застосування прототипа системи, який називається I-Saurus.

Переклад І. Снєгурова

Lapata, M. The Disambiguation of Nominalizations [Зняття неоднозначності номіналізацій] / Maria Lapata // Computational linguistics. – 2002. – Vol. 28. – No. 3. – Pages 357–388. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120102760276018#.WIEhXn3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120102760276018

У цій статті висвітлюється інтерпретація номіналізацій, особливого класу іменних груп, у яких ядро виражене віддієслівним іменником, а модифікатор є аргументом вихідного дієслова. При спробі автоматично інтерпретувати номіналізації потрібно завжди брати до уваги: (а) обмеження сполучуваності, накладені субстантивованим ядром іменної групи, (б) той факт, що відношення між модифікатором і ядром можуть бути неоднозначними, і (в) той факт, що ці неоднозначності можна легко вирішити завдяки контексту або прагматичним чинникам. Інтерпретація номіналізацій створює додаткову проблему для імовірнісних підходів, оскільки аргументні відношення між ядром і модифікатором у корпусі виявити непросто. Навіть наближення, що встановлює вихіднє дієслово, від якого утворене ядро іменної групи, не забезпечує достатніх даних. Ми пропонуємо розглядати інтерпретацію як вирішення проблеми неоднозначності і показуємо, як можна "відтворити" відсутні дані про дистрибуцію, використовуючи частковий синтаксичний аналіз, методи згладжування даних, і контекст. Ми об'єднали ці окремі джерела інформації, використовуючи програму Ripper, яка видобуває набори

правил з даних, і досягали точності 86,1% (при стандарті 61,5%) на матеріалі Британського національного корпусу.

Переклад Т. Павлущенко, М. Погребної

McCarthy, D. Disambiguating Nouns, Verbs, and Adjectives Using Automatically Acquired Selectional Preferences [Зняття омонімії іменників, дієслів і прикметників за допомогою автоматично визначених селекційних преференцій] / Diana McCarthy, John Carroll // Computational linguistics. – 2003. – Vol. 29. – No. 4. – Pages 639–654. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120103322753365#.WIe_w33sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120103322753365

Системи зняття лексичної багатозначності (ЗЛБ) використовують селекційні преференції як джерело інформації, необхідної для вирішення проблеми лексичної багатозначності. Ми оцінюємо ЗЛБ, використовуючи селекційні преференції, отримані для граматичних відносин англійських прикметника-іменника, підмета і прямого додатка із стандартного тестового корпусу. Селекційні преференції характеризують класи дієслів або прикметників, а не окремі словоформи, отже їх можна використати для того, щоб вирішити омонімію супутніх прикметників і дієслів, а не лише іменних вершин аргументів. Також досліджено використання евристики «одне значення на дискурс» з метою присвоєння смислової мітки для певного слова іншим уживанням цього слова у документі з метою збільшення охоплення. Хоча у порівнянні з іншими системами ЗЛБ без учителя преференції дають хороші результати на одному й тому корпусі, результати дослідження свідчать, що багатьом програмам необхідні додаткові джерела інформації для досягнення прийнятного рівня точності й охоплення. Крім кількісної оцінки результатів, їх проаналізовано з метою визначення ситуацій, у яких селекційні преференції дають найточніший результат і в яких евристика «одне значення на дискурс» підвищує продуктивність.

Переклад А. Синящик

Lapata, M. Verb Class Disambiguation Using Informative Priors [Визначення класу дієслова за допомогою інформативних пріоритетів] / Mirella Lapata, Chris Brew // Computational linguistics. – 2004. – Vol. 30. – No. 1. – Pages 45–73. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120104773633385#.WIEh233sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120104773633385

У лексичній семантиці широко використовується дослідження класів дієслова Левін (1993). Згідно її типології, деякі дієслова, такі як give, належать до одного класу. Але інші дієслова, такі як write, можуть входити

до кількох альтернативних класів. Ми розширили список Левін до простої статистичної моделі омонімії дієслова. Використовуючи цю модель, можна генерувати найкращі рішення для багатозначних дієслів без використання корпусу зі знятою омонімією. У статті також показано, що ці найкращі рішення можна використовувати в якості пріоритетів для програми зняття омонімії дієслова.

Переклад Т. Павлущенко, М. Погребної

McCarthy, D. Unsupervised Acquisition of Predominant Word Senses [Алгоритм неконтрольованого встановлення переважаючих значень слова] / Diana McCarthy, Rob Koeling, Julie Weeds, John Carroll // Computational linguistics. – 2007. – Vol. 33. – No. 4. – Pages 553–590. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2007.33.4.553#.WIEryH3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2007.33.4.553

Останнім часом здійснено багато досліджень, присвячених усуненню лексичної омонімії, особливо після появи тестових наборів конференції Senseval. Оскільки слово часто має більше одного значення, зняття лексичної неоднозначності може підвищити продуктивність програм, які вимагають семантичної інтерпретації уведених мовних даних. Головна проблема полягає в тому, що точність вирішення проблеми лексичної омонімії значною мірою залежить від обсягу доступних даних з виконаною вручну семантичною розміткою, і що навіть найкращі системи, здійснюючи анотування кожного слововживання у документі, рідко перевершують результати простого евристичного алгоритму, який використовує перше, або переважаюче, значення слова в усіх контекстах. Ефективність запропонованого евристичного алгоритму пояснюється асиметричною природою дистрибуцій лексичних значень. Дані для евристичного алгоритму можна брати як із словників, так і з набору даних із семантичною розміткою. Проте, кількість останніх обмежена, а дистрибуція значень і переважаюче значення слова може залежати від предметної області і джерела тексту. (Наприклад, у популярних і наукових журналах перше значення слова «зірка» буде різним). У статті докладно проаналізовано запропонований раніше метод автоматичного визначення переважаючого значення слова у сирому тексті. Розглянувши велику кількість джерел даних і параметризацій цього методу і проаналізувавши результати оцінювання і аналіз помилок, визначено, у яких випадках цей метод є ефективним, а в яких ні. Зокрема з'ясовано, що цей метод дає кращі результати для іменників і прикметників, ніж для дієслів і прислівників, але на відміну від дуже популярного корпусу SemCor дає точнішу інформацію про переважаюче значення іменників з низькою частотністю у вказаному корпусі. Також показано, що цей метод можна успішно адаптувати для предметних областей, використовуючи в

якості вводу спеціальні корпуси текстів з конкретної предметної області з ручним анотуванням предметної області або класифікованих автоматично.

Переклад В. Коломієць

O'Hara, T. Exploiting Semantic Role Resources for Preposition Disambiguation [Використання розмітки семантичних ролей у знятті прийменникової омонімії] / Tom O'Hara, Janyce Wiebe // Computational linguistics. – 2009. – Vol. 35. – No. 2. – Pages 151–184. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.06-79-prep15#.WIElk33sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.06-79-prep15

У статті описано як можна використати семантичні ресурси для зняття прийменникової омонімії. Основними ресурсами є корпуси Penn Treebank і FrameNet із розміткою семантичних ролей. Ресурси також включають твердження з бази знань Factotum, а також інформацію з онтології Cyc і концептуальних графів. На основі цих ресурсів створено спільний інвентар для аналізу визначень, який є метою цього дослідження.

Зняття омонімії зосереджене на відносинах, позначених прийменниковими групами, і розглядається як усунення омонімії конкретного прийменника. Запропоновано новий підхід до зняття лексичної омонімії, шляхом використання гиперонімів WordNet як коллокацій, а не просто слів. Описано різні експерименти з даними з корпусів Penn Treebank і FrameNet, які ілюструють наслідки фільтрації, зокрема класифікацію прийменників разом і окремо. Подібні експерименти проведені і з даними з Factotum, зокрема метод прогнозування вірогідного використання прийменників у корпусах, оскільки бази знань як правило не містять інформації про способи вираження відносин у англійській мові (на відміну від детальних поміт із цією інформацією у корпусах Penn Treebank і FrameNet). Також описано експерименти з даними з FrameNet, включеними у розроблений для аналізу визначень спільний інвентар відносин, які демонструють, як можна застосувати зняття прийменникової омонімії у засвоєнні лексики.

Переклад В. Коломієць

Giuliano, C. Kernel Methods for Minimally Supervised WSD [Ядерні методи для мінімально контрольованого зняття багатозначності] / Claudio Giuliano, Alfio Massimiliano Gliozzo, Carlo Strapparava // Computational linguistics. – 2009. – Vol. 35. – No. 4. – Pages 513–528. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2009.35.4.35407#.WIEnAn3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2009.35.4.35407

У статті описано напівконтрольований метод розв'язання анафори, який використовує зовнішні, отримані без будь-якого контролю знання. Зокрема,

використовуються базові функції ядра для незалежної оцінки синтагматичної і тематичної схожості, створюючи набір класифікаторів слів, які використовують модель спільного домену, отриману з великого корпусу нерозмічених даних. Результати свідчать, що запропонований підхід дозволив досягти сучасного рівня продуктивності у завданнях конференції Senseval-3 для різних обмежених наборів слів і всіх слів корпусу, хоча він використовує значно меншу кількість тренувальних прикладів, ніж інші методи.

Переклад В. Коломієць

Yuret, D. The Noisy Channel Model for Unsupervised Word Sense Disambiguation [Модель каналу з перешкодами для неконтрольованого зняття лексичної багатозначності] / Deniz Yuret, Mehmet Ali Yatbaz // Computational linguistics. – 2010. – Vol. 36. – No. 1. – Pages 111–127. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2010.36.1.36103#.WIEnoX3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2010.36.1.36103

У статті описано генеративну імовірнісну модель, модель каналу з перешкодами, для неконтрольованого зняття лексичної багатозначності. У запропонованій моделі кожний контекст C змодельований як окремий канал, через який мовець планує передати певне значення S, використовуючи потенційно неоднозначне слово W. Щоб вибрати потрібне значення, слухач використовує дистрибуцію можливих значень у даному контексті P(S|C) і можливі слова, які можуть виразити кожне значення P(W|C). Ми виходили з того, що P(W|C) є незалежним від контексту і вираховували його, використовуючи частоти значень у тезаурусі WordNet. Головною проблемою неконтрольованого зняття лексичної багатозначності є визначення обумовленого контекстом значення без доступу до жодного тексту із семантичною розміткою. Наведено один із способів вирішення цієї проблеми за допомогою статистичної мовної моделі, яка спирається на великий обсяг нерозміченого тексту. У середині моделі використовуються великі семантичні класи S. Досліджено вплив різних рівнів деталізації на ефективність зняття лексичної багатозначності. Запропонована система продукує дуже точні значення для оцінювання, а за ефективністю зняття неоднозначності іменників вона перевершила більшість описаних у літературі систем і наблизилась до найкращих контрольованих систем.

Переклад Т. Павлущенко, М. Погребної

Erk, K. Measuring Word Meaning in Context [Визначення значення слова у контексті] / Katrin Erk, Diana McCarthy, Nicholas Gaylord // Computational linguistics. – 2013. – Vol. 39. – No. 3. – Pages 511–554. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00142#.WIEoSH3s

SGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00142#.VRGBVfmsU5E

Зняття лексичної багатозначності (Word Sense Disambiguation, WSD) є давнім і важливим завданням комп'ютерної лінгвістики, розв'язання якого все ще є складним як для комп'ютерів, так і для людей-анотаторів. Останнім часом було запропоновано декілька способів представлення значення слова у контексті, які відрізняються від традиційного використання одного найбільш підходящого значення для кожного випадку. Вони пояснюють значення слова у контексті за допомогою кількох перифраз як крапок у векторному просторі чи розподілу прихованих значень. Потрібні нові методи оцінки і порівняння цих різних представлень.

У цій статті запропоновано дві нові схеми анотування, які ранжують значення слова у контексті. При анотуванні за схемою *W*ssim оцінюється прийнятність кожного словникового значення за допомогою порядкової шкали. При застосуванні схеми *U*sim безпосередньо оцінюється подібність пар вживання однієї леми, знову за допомогою шкали. Показано, що нові схеми анотування дозволяють отримати високі показники узгодженості між анотаторами, а також демонструють стійку кореляцію з традиційним анотуванням одного значення та з анотуванням кількох лексичних перифраз. Анотатори використовують увесь масштаб порядкової шкали і роблять дуже точні висновки, які «змішують та співставляють» значення для кожного окремого вживання. Також продемонстровано, шо ранжування за схемою *U*sim підкоряється аксіомі трикутника, що свідчить на користь моделей, які розглядають подібність вживання як мірку.

Останнім часом здійснено велику роботу по грубій класифікації значень. У статті показано, що можна використовувати ранжування за схемами *W*ssim і *U*sim з метою аналізу існуючої грубої класифікації, щоб визначити групи значень, які можуть суперечити інтуїції непідготовлених носіїв мови. Також у ході порівняння продемонстровано, що показники *W*ssim не входять до будь-якої статичної класифікації значень.

Переклад Д. Попової

Agirre, E. Random Walks for Knowledge-Based Word Sense Disambiguation [Використання методу випадкових блукань у вирішенні проблеми лексичної багатозначності на основі знань] / Eneko Agirre, Oier López de Lacalle, Aitor Soroa // Computational linguistics. – 2014. – Vol. 40. – No. 1. – Pages 57–84. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00164#.WIEpdX3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00164#.VRGAy_msU5E

Системи зняття лексичної багатозначності (Word Sense Disambiguation,

WSD) автоматично обирають потрібне значення слова у контексті. У цій статті ми представляємо алгоритм WSD на основі випадкових блукань у великих лексичних базах даних (ЛБД). Ми демонструємо, що наш алгоритм працює краще, ніж інші графові моделі, коли за основу береться граф, побудований за допомогою WordNet та eXtended WordNet. Поєднання нашого алгоритму та ЛБД вигідно відрізняється від інших відомих методів на основі знань, які застосовують подібні знання до різноманітних баз даних англійської мови та бази даних іспанської мови. Ми додаємо детальний аналіз факторів, що впливають на алгоритм. Алгоритм та використовувані лексичні бази даних знаходяться у відкритому доступі і результати можна легко перевірити.

Переклад Д. Попової

Pilehvar, M. T. A Large-Scale Pseudoword-Based Evaluation Framework for State-of-the-Art Word Sense Disambiguation [Широкомасштабна оцінка сучасних методів зняття лексичної багатозначності на основі псевдослів] / Mohammad Taher Pilehvar, Roberto Navigli // Computational linguistics. – 2014. – Vol. 40. – No. 4. – Pages 837–881. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00202#.WIEp5n3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00202#.VRF_GfmsU5E

Оцінка деяких завдань у лексичній семантиці нерідко є обмеженою через брак великої кількості ручних анотацій не лише для навчальних цілей, але також для тестування. Одним із таких завдань є зняття лексичної багатозначності (Word Sense Disambiguation, WSD), оскільки ручна розмітка баз даних є дуже складною і займає багато часу. Як наслідок, оцінювання, як правило, виконується у невеликих масштабах, що не дозволяє здійснити ретельний аналіз факторів, від яких залежить продуктивність системи.

У цій роботі ми досліджуємо це питання шляхом реалістичного моделювання великомасштабної оцінки завдання WSD за допомогою двох головних нововведень. По-перше, ми пропонуємо два нові підходи до широкомасштабної генерації багатозначних псевдослів (тобто штучних слів, здатних моделювати реальні багатозначні слова); по-друге, ми використовуємо найбільш підходящий тип псевдослова, щоб створити великі корпуси з розміткою псевдозначень, які можна використати в ролі великомасштабної експериментальної бази для порівняння новітніх методів на основі навчання з учителем і на основі знань. Використовуючи цю експериментальну базу, ми досліджуємо вплив навчання з учителем і знань на два основні методи зняття лексичної багатозначності та здійснюємо ретельний аналіз факторів, які впливають на їх продуктивність.

Переклад Д. Попової

Word Sense Clustering and Clusterability [Кластеризація значень слів і здатність до кластеризації] / Diana McCarthy, Marianna Apidianaki, Katrin Erk // Computational linguistics. – 2016. – Vol. 42. – No. 2. – Pages 245–275. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00247 – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00247

Зняття лексичної багатозначності та споріднена галузь автоматичного видобування значень слів традиційно виходять з припущення, що випадки вживання леми можна розділити на значення. Але для певних лем це завдання є значно легшим, аніж для інших. Ця праця ґрунтується на останніх дослідженнях, які пропонують описувати значення слів шляхом градуювання, а не строгого розподілу на значення; у статті стверджується, що не всі леми потребують складнішого градуйованого аналізу, залежно від їхньої здатності до розподілу. Хоча завдяки попереднім дослідженням і лінгвістичній літературі існує багато доказів існування спектру розчленовуваності значень слів, це перша спроба виміряти вказане явище та об'єднати публікації з машинного навчання, присвячені здатності до кластеризації, з даними про вживання слів, які використовуються в комп'ютерній лінгвістиці.

Автори вирішили реалізувати розчленовуваність у вигляді здатності до кластеризації, міри того, наскільки легко кластеризуються вживання лем. Перевірено два способи вимірювання здатності до кластеризації: (1) описані в публікаціях з машинного навчання методи, метою яких є оцінка якості оптимальних кластерних рішень, отриманих методом k-середніх, і (2) припущення, що якщо лема більш схильна до кластеризації, два кластерні рішення, що базуються на двох різних «поглядах» на ті самі дані, будуть більш схожими. Два погляди, використані в дослідженні, – це два різні набори створених вручну лексичних замінників цільової леми: з одного боку – одномовні перефразування, а з другого – переклади. Автоматичну кластеризацію застосовано до виконаних вручну маркувань. Використання ручного маркування зумовлене бажанням отримати максимально інформативні й «чисті» репрезентації випадків, які кластеризуються. Показано, що за умови контролю над полісемією, запропоновані міри здатності до кластеризації, як правило, корелюють з розчленовуваністю, зокрема деякі міри здатності до кластеризації типу (1), і що ці міри перевершують базовий поріг, що визначається обсягом перекриття при м'якій кластеризації.

Переклад М. Дубка

Stilo, G. Hashtag Sense Clustering Based on Temporal Similarity [Кластеризація значень хештегів за часовою подібністю] / Giovanni Stilo, Paola Velardi // Computational linguistics. – 2017. – Vol. 43. – No. 1. – Pages 181–200. – Режим доступу до анотації:

https://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00277 – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00277

Хештеги – це вигадливі мітки, які використовуються у мікроблогах, щоб охарактеризувати тему повідомлення/обговорення. Незважаючи на своє первинне призначення, хештеги не можуть використовуватися як засіб кластеризації повідомлень із подібним вмістом. По-перше, оскільки користувачі активно і спонтанно створюють хештеги багатьма мовами, одна й та ж тема може асоціюватись із різними хештегами, і навпаки, один і той же хештег може стосуватися різних тем у різні періоди часу. По-друге, на відміну від загальновживаних слів зняття лексичної багатозначності хештегів ускладнюється відсутністю доступних каталогів значень (наприклад, Вікіпедії або WordNet); і, крім того, мітки хештегів складно аналізувати, оскільки вони часто складаються з абревіатур, складених слів тощо. Загальноприйнятий спосіб визначення значення хештегів – це аналіз їхнього контексту, але, як зазначено вище, хештеги можуть мати багато різних значень. У статті запропоновано алгоритм кластеризації за часовими значеннями, який базується на ідеї про те, що семантично пов'язані хештеги використовуються аналогічно й одночасно.

Переклад М. Дубка

Tripodi, R. A Game-Theoretic Approach to Word Sense Disambiguation [Метод зняття лексичної багатозначності на основі теорії ігор] / Rocco Tripodi, Marcello Pelillo // Computational linguistics. – 2017. – Vol. 43. – No. 1. – Pages 31–70. Режим доступу до анотації: https://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00274 – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00274

У статті представлено нову модель для зняття лексичної багатозначності, сформульовану в термінах еволюційної теорії ігор, де кожне багатозначне слово представлено у вигляді вузла графа, ребра якого представляють відношення між словами, а значення представлені як класи. Вірогідності належності слів до класів міняються одночасно, відповідно до потенційних значень сусідніх слів. Для вимірювання впливу кожного слова на вибір інших слів використано інформацію про дистрибуцію, а для вимірювання міцності сумісності між виборами – інформацію про семантичну подібність. Ця інформація може допомогти сформулювати проблему багатозначності слів як проблему дотримання обмежень і вирішити її за допомогою інструментів, запозичених з теорії ігор, зберігши цілісність тексту. В основу метода покладено дві ідеї: подібні слова відносять до подібних класів і значення слова залежить не від усіх слів у тексті, а лише від деяких з них. У статті викладено детальне обґрунтування ідеї моделювання проблеми зняття лексичної багатозначності в термінах теорії ігор, проілюстроване прикладом.

У висновку наведено всебічний аналіз сукупності показників подібності для використання у методі та порівняння з найновішими системами. Результати свідчать, що запропонована модель перевершує найсучасніші алгоритми і може бути застосована до різних завдань і в різних ситуаціях.

Переклад М. Дубка

Комп'ютерна лексикографія

Daciuk, J. Incremental Construction of Minimal Acyclic Finite-State Automata [Покрокова побудова мінімальних ациклічних скінченних автоматів] / Jan Daciuk, Stoyan Mihov, Bruce W. Watson, Richard E. Watson // Computational linguistics. – 2000. – Vol. 26. – No. 1. – Pages 3–16. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120100561601#.WIUkGH3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120100561601

У статті описано новий метод побудови мінімальних детермінованих ациклічних скінченних автоматів із набору рядків. Традиційні методи складаються з двох етапів: на першому будується префіксне дерево, на другому воно мінімізується. Запропонований метод дозволяє побудувати мінімальний автомат за один етап шляхом додавання один за одним нових рядків і одночасної мінімізації отриманого автомата. Описано загальний алгоритм та спеціалізацію, яка спирається на лексикографічне упорядкування вхідних рядків. Запропонований метод швидкий і у порівнянні з іншими методами значно зменшує вимоги до пам’яті.

Переклад Д. Попової

Bozsahin, C. The Combinatory Morphemic Lexicon [Словник сполучуваності морфем] / Cem Bozsahin // Computational linguistics. – 2002. – Vol. 28. – No. 2. – Pages 145–186. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120102760173634#.WIT58n3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120102760173634

Граматики, які розпізнають слова із лексикону, можуть бути несумісними з прозорою проекцією предметних семантико-синтаксичних відношень між меншими мовними одиницями. Для укладання морфемного граматичного словника розроблена морфосинтаксична модель на основі комбінаторної категоріальної граматики, яка забезпечує універсальні складники, вільне узгодження категорій, і лексичну проекцію морфосинтаксичних характеристик та прив’язаність до граматики. Ці механізми мають достатню експресивну силу для того, щоб сформулювати у словнику семантично прозорі характеристики без обов’язкового обмеження створення структур словами і словосполученнями. Наприклад, зв'язані морфеми у якості лексичних одиниць можуть вживатися в межах словосполучення або слова, незалежно від їх приєднувальних властивостей, але відповідно до їх семантики. Налаштування словника можна змінити відповідно до характеристик певної мови. Розроблений словник є прозорою комбінацією

флективної морфології, синтаксису та семантики. У статті описано комп'ютерну систему і продемонстровано практичне застосування моделі на прикладі англійської та турецької мов.

Переклад І. Снєгурова

Carrasco, C. R. Incremental Construction and Maintenance of Minimal Finite-State Automata [Покрокова побудова і супроводження мінімальних скінченних автоматів] / Rafael C. Carrasco, Mikel L. Forcada // Computational linguistics. – 2002. – Vol. 28. – No. 2. – Pages 207–216. – Режим доступу до аннотация: http://www.mitpressjournals.org/doi/abs/10.1162/089120102760173652#.WIT6YX3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120102760173652

Дацюк та ін. (Daciuk et al.) [Computational Linguistics 26(1):3–16 (2000)] описують метод покрокової побудови мінімальних, детермінованих, нециклічних скінченних автоматів (словників) з наборів рядків. Проте, нециклічні скінченні автомати мають обмеження. Наприклад, якщо хтось хоче, щоб лінгвістична програма приймала всі можливі цілі числа або Інтернет адреси, відповідний скінченний автомат повинен бути циклічним. У статті описано простий і не менш ефективний метод модифікації будь-якого мінімального скінченного автомата (незалежно від того циклічний він чи ні), щоб можна було додавати до або вилучати з мови, яку допускає автомат, рядок. Обидві операції є дуже важливими при обслуговуванні словника, вони вирішують проблему створення словника, яку розглядали як особливий випадок Дацюк та ін. Запропоновані у статті алгоритми можна вивести безпосередньо із поданих у будь-якому підручнику пояснень стосовно перетину і доповнення скінченних автоматів. Ці алгоритми використовують особливі властивості автоматів, які є результатом операції перетину, коли один із скінченних автоматів приймає один рядок.

Переклад І. Снєгурова

Ploux, S. A Model for Matching Semantic Maps between Languages (French/English, English/French) [Модель співставлення семантичних карт різних мов (французька/англійська, англійська/французька)] / Sabine Ploux, Hyungsuk Ji // Computational linguistics. – 2003. – Vol. 29. – No. 2. – Pages 155–178. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120103322145298#.WIIJ1X3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120103322145298

У статті описано просторову модель співставлення семантичних значень у двох мовах, французькій і англійській. Використовуючи зв'язки семантичної схожості, модель створює карту, яка представляє слово у вихідній мові. Потім модель проектує значення з карти на простір у цільовій мові. Новий

простір зберігає зв'язки семантичної схожості, характерні для другої мови. Після цього обидві карти проектуються на одну площину, щоб виявити співпадаючі значення. З навчальною метою опис усіх кроків у статті проілюстровано кількома прикладами. Повний комплект розроблених додатків знаходиться за адресою http://dico.isc.cnrs.fr.

Переклад Т. Павлущенко, М. Погребної

Santamar, C. Automatic Association of Web Directories with Word Senses [Автоматичне зв'язування веб-каталогів зі значеннями слів] / Celina Santamar, Julio Gonzalo, Felisa Verdejo // Computational linguistics. – 2003. – Vol. 29. – No. 3. – Pages 485–502. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120103322711613#.WIHl433sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120103322711613

Ми описуємо алгоритм, який зв'язує лексичну інформацію (з семантичної мережі WordNet 1.7) із веб-каталогами (із проекту Відкритий Каталог), для того щоб пов'язати значення слів із такими каталогами. Такі зв'язки можуть бути використані як детальні описи для автоматичного отримання корпусу з семантичною розміткою, кластеризації тематично пов'язаних значень та виявлення особливостей значень. Алгоритм протестовано на 29 іменниках (147 значень), використаних у змаганні Senseval 2, що дозволило отримати 148 зв'язків (значення слів, веб-каталог), котрі охоплюють 88% значень з однієї галузі в тестових даних із точністю 86%. Глибина деталізації описів значень у веб-каталогах проаналізована в процесі котрольованого розв'язання лексичної омонімії з використанням тестового набору під назвою Senseval 2. Результати свідчать, що якщо зв'язок каталог/значення слів правильний, то зразки, отримані автоматично з веб-каталогів, майже так само придатні для тренування, як і оригінальні навчальні приклади з Senseval 2. Отримані результати підтвердили нашу гіпотезу, що веб-каталоги є цінним ресурсом лексичної інформації з меншою кількістю помилок, надійнішим та краще структурованим, аніж Всесвітня мережа в цілому як корпус.

Переклад В. Туз

Navigli, R. Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites [Видобування онтологій предметних областей із сховищ документів і спеціалізованих веб-сайтів] / Roberto Navigli, Paola Velardi // Computational linguistics. – 2004. – Vol. 30. – No. 2. – Pages 151–179. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120104323093276#.WIT7DH3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120104323093276

У статті описано метод і інструмент, OntoLearn, призначений для видобування онтологій предметних областей із веб-сайтів і, загалом, з

документів, які розповсюджуються серед членів віртуальних організацій. OntoLearn спочатку видобуває термінологію предметної області із наявних документів. Після цього здійснюється семантична інтерпретація складних термінів предметної області і вони розташовуються за ієрархічним принципом. Нарешті, виявленими концептами предметної області коригується і збагачується онтологія загального призначення, WordNet. Новизна запропонованого підходу полягає, насамперед, у семантичній інтерпретації, тобто співставленні складних концептів із складними термінами. Для цього у WordNet знаходяться відповідний концепт для кожного слова у низці термінів і відповідні концептуальні відносини, які об'єднують компоненти концепту. Семантична інтерпретація здійснюється на основі нового алгоритму розв'язання лексичної багатозначності під назвою структурно-семантичні взаємозв'язки.

Переклад В. Коломієць

Daciuk, J. Comments on "Incremental Construction and Maintenance of Minimal Finite-State Automata," by Rafael C. Carrasco and Mikel L. Forcada [Коментар до статті Р. Карраско і М. Форкади "Покрокова побудова і супроводження мінімальних скінченних автоматів"] / Jan Daciuk // Computational linguistics. – 2004. – Vol. 30. – No. 2. – Pages 227–235. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120104323093302#.WIXMUH3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120104323093302

У нещодавній статті Р. Карраско і М. Форкади [Carrasco and Forcada, June 2002] описано два алгоритми: один для поетапного додавання рядків до мови мінімального детермінованого циклічного автомату, а другий для поетапного видалення рядків з автомату. Перший алгоритм є узагальненням "алгоритму для несортованих даних", другого з двох покрокових алгоритмів для створення мінімальних детермінованих циклічних автоматів, описаних у роботі Д. Дасюка та ін. [Daciuk et al., 2000]. Показано, що другий алгоритм із старішої публікації − "алгоритм для сортованих даних" – може бути узагальнений подібним чином. Новий алгоритм є швидшим, ніж алгоритм для додавання рядків, описаний у статті Р. Карраско і М. Форкади, оскільки він обробляє кожний стан лише один раз.

Переклад В. Коломієць

Mihov, S. Fast Approximate Search in Large Dictionaries [Швидкий приблизний пошук у великих словниках] / Stoyan Mihov, Klaus U. Schulz // Computational linguistics. – 2004. – Vol. 30. – No. 4. – Pages 451–477. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/0891201042544938#.WIeIzn3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/0891201042544938

Необхідність виправляти спотворені рядки існує у багатьох галузях обробки природної мови. За наявності словника, який містить усі можливі вхідні слова, підхожим набором кандидатів для виправлення спотвореного уведення *P* є набір усіх слів у словнику, для яких відстань Левенштейна до *P* не перевищує заданого (маленького) порога *k*. У статті описані методи ефективного відбору таких наборів кандидатів. Спочатку представлено базовий метод виправлення на основі концепції «універсального автомата Левенштейна», потім продемонстровано, як можна суттєво удосконалити базову процедуру, використовуючи два методи фільтрації, запозичені з області приблизного текстового пошуку. Перший метод, який використовує стандартні словники і словники з оберненими словами, забезпечує дуже швидке виправлення більшості видів вхідних рядків. Результати проведеного тестування свідчать, що час виправлення для порогів фіксованої величини залежить від очікуваної кількості кандидатів на виправлення, яка зменшується для довших вхідних слів. Вибір оптимального методу фільтрації також залежить від довжини вхідних слів.

Переклад А. Синящик

Cooper, M. A Mathematical Model of Historical Semantics and the Grouping of Word Meanings into Concepts [Математична модель історичної семантики і групування значень слова у концепти] / Martin C. Cooper // Computational linguistics. – 2005. – Vol. 31. – No. 2. – Pages 227–248. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/0891201054223995#.WH4XyH3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/0891201054223995

Статистичний аналіз багатозначності на матеріалі шістнадцяти англійських і французьких словників виявив, що в кожному словнику кількість значень слова має майже експоненціальний розподіл. Представлена ймовірнісна модель історичної семантики, яка пояснює цей розподіл. Ця математична модель також слугує засобом визначення середньої кількості різних концептів для слова, яка виявилася значно меншою, ніж середня кількість вказаних у словнику значень слова. Групування значень слова у концепти ґрунтується на їх здатності породжувати однакові нові значення (шляхом метафори, метонімії тощо), тобто на їх потенційному майбутньому, а не на їхній історії.

Переклад О. Мартинюк, М. Погребної

O'Donovan, R. Large-Scale Induction and Evaluation of Lexical Resources from the Penn-II and Penn-III Treebanks [Широкомасштабне отримання і оцінювання лексичних ресурсів із банків дерев Penn-II і Penn-III] / Ruth O'Donovan, Michael Burke, Aoife Cahill, Josef van Genabith, Andy Way // Computational linguistics. – 2005. – Vol. 31. – No. 3. – Pp. 329–366. – Режим

доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120105774321073#.WH5_vn3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120105774321073

Ми описуємо методику отримання фреймів субкатегорізації на основі алгоритму автоматичного анотування f-структур у термінології лексико-функціональної граматики (lexical-functional grammar, скор. LFG) для банків дерев Penn-II і Penn-III. Ми отримуємо фрейми субкатегорізації на основі синтаксичних функцій (семантичні форми LFG) і традиційні фрейми субкатегорізації на основі категорій контекстно-вільної граматики, а також змішані фрейми на основі функцій/категорій, разом із інформацією про прийменники для непрямих відмінків і інформацією про прийменники або частки для фразових дієслів або без такої інформації. Наш метод пов'язує імовірності з фреймами відповідно до леми, розмежовує активні і пасивні фрейми і ретельно враховує результати розірваних залежностей у структурах вихідних даних. На відміну від багатьох інших методів, наш метод не визначає заздалегідь, які типи фреймів субкатегоризації будуть отримані, а дізнається про них із вихідних даних. Разом із частками і прийменниками ми отримали 21 005 типів фреймів лем для 4 362 лем дієслів, загальна кількість типів фреймів – 577, в середньому 4,8 типів фреймів для дієслова. Ми представляємо широкомасштабне оцінювання повного набору отриманих форм шляхом порівняння із укладеним вручну словником COMLEX. Наскільки нам відомо, це найбільше і найповніше оцінювання автоматично отриманих фреймів субкатегоризації для англійської мови.

Переклад І. Снєгурова

Miller, G. A. WordNet Nouns: Classes and Instances [Іменники у WordNet: класи і представники класів] / George A. Miller, Florentina Hristea // Computational linguistics. – 2006. – Vol. 32. – No. 1. – Pages 1–3. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2006.32.1.1#.WIUSWn3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2006.32.1.1

У лексичній базі даних для англійської мови WordNet, яка достатньо активно використовується комп'ютерними лінгвістами, раніше не виокремлювались гіпоніми як класи і гіпоніми як представники класу. У статті описано спробу здійснити таке розмежування і запропоновано простий спосіб додати результати до майбутніх версій WordNet.

Переклад В. Коломієць

Budanitsky, A. Evaluating WordNet-based Measures of Lexical Semantic Relatedness [Оцінювання метрик лексико-семантичної спорідненості на основі WordNet] / Alexander Budanitsky, Graeme Hirst // Computational

linguistics. – 2006. – Vol. 32. – No. 1. – Pages 13–47. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2006.32.1.13#.WH4YeH3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2006.32.1.13

Квантифікация лексико-семантичної спорідненості широко застосовується у обробці природної мови і було запропоновано багато різних метрик. Нами оцінено п'ять із цих метрик, кожна з яких використовує у якості основного ресурсу WordNet, шляхом оцінки їх ефективності у виявленні і виправленні реальних орфографічних помилок. Було з'ясовано, що метрика на основі інформаційного змісту (Jiang-Conrath) дає кращі результати, ніж метрики Hirst-St-Onge, Leacock-Chodorow, Lin і Resnik. Крім того, пояснюється, чому дистрибутивна схожість не є адекватною заміною лексично-семантичній спорідненості.

Переклад В. Коломієць

Inkpen, D. Building and Using a Lexical Knowledge Base of Near-Synonym Differences [Створення і використання лексичної бази даних про розбіжності між неточними синонімами] / Diana Inkpen, Graeme Hirst // Computational linguistics. – 2006. – Vol. 32. – No. 2. – Pages 223–262. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2006.32.2.223#.WIUSp33sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2006.32.2.223

Неправильний вибір слова системою машинного перекладу або генерації природної мови може спричинити небажані конотації, імплікації або відношення. Вибір із сукупності неточних синонімів, таких як **неточність**, **помилка**, **погрішність**, і **огріх** (слів, які мають спільну частину значень, але відрізняються за деякими ознаками), можна зробити тільки за наявності інформації про відмінності між ними.

У статті описано метод автоматичного створення нового різновиду лексичних ресурсів: базу даних про відмінності між неточними синонімами. Розроблено алгоритм неконтрольованого навчання списків рішень, який генерує правила видобування знань із спеціального словника відмінностей між синонімами. Ці правила потім були використані для видобування інформації з тексту словника.

Після цього вихідна база даних була збагачена інформацією з інших машиночитаних словників. Інформація про сполучуваність неточних синонімів видобувалась із довільних текстів. Створена база даних була використана у системі генерації природної мови Xenon, було продемонстровано як можна використати новий лексичний ресурс для вибору неточного синоніма, який найкраще відповідає певній ситуації.

Переклад В. Коломієць

Li, P. A Sketch Algorithm for Estimating Two-Way and Multi-Way Associations [Алгоритм-ескіз для оцінки двосторонніх і багатосторонніх асоціацій] / Ping Li, Kenneth W. Church // Computational linguistics. – 2007. – Vol. 33. – No. 3. – Pages 305–354. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2007.33.3.305#.WH4a8H3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2007.33.3.305

Немає потреби аналізувати увесь корпус (наприклад, Інтернет), щоб з'ясувати, чи існує тісна асоціація між двома (або більше) словами. Можна отримати оцінку асоційованості за допомогою невеликої вибірки. Розроблено алгоритм-ескіз, який створює таблиці спряженості для вибірки. Оцінка усіх даних таблиці спряженості може здійснюватися за допомогою простого масштабування. Проте можна поліпшити результати, скориставшись частотами документів. Запропонований метод удвічі зменшує кількість помилок у порівнянні з ескізами Бродера.

Переклад В. Коломієць

Fazly, A. Unsupervised Type and Token Identification of Idiomatic Expressions [Неконтрольоване розпізнавання ідіоматичних виразів на основі типів і вживань] / Afsaneh Fazly, Paul Cook, Suzanne Stevenson // Computational linguistics. – 2009. – Vol. 35. – No. 1. – Pages 61–103. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.08-010-R1-07-048#.WH4f933sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.08-010-R1-07-048

У розмовній мові велика кількість ідіоматичних виразів, проте вони залишаються загадкою, оскільки достовірно не відомо, як люди їх вивчають і розуміють. Вони особливо цікавлять лінгвістів, психолінгвістів і лексикографів, головним чином завдяки своїм синтаксичним і семантичним характеристикам, а також нечіткому лексичному статусу. Незважаючи на велику кількість досліджень характерних особливостей ідіом у лінгвістичній літературі, немає єдиної думки про те, які саме характеристики притаманні цим виразам. Через свої особливості ідіоматичні вирази здебільшого ігнорувалися комп'ютерними лінгвістами. У статті розглядається придатність деяких виявлених характеристик ідіом для автоматичного розпізнавання. Конкретніше, розроблено статистичні міри, кожна із яких моделює конкретну характеристику ідіоматичних виразів на основі особливостей їх реального вживання у тексті. Ці статистичні міри було використано у класифікації за типами, яка передбачала автоматичне розмежування ідіоматичних виразів (виразів із можливою ідіоматичною інтерпретацією) від схожих на них за формою буквальних виразів (ідіоматична інтерпретація яких є неможливою). Крім того, деякі міри використано у розпізнаванні слів,

у процесі якого розмежовуються ідіоматичне і буквальне вживання потенційно ідіоматичних виразів у контексті.

Переклад В. Коломієць

Girju, R. The Syntax and Semantics of Prepositions in the Task of Automatic Interpretation of Nominal Phrases and Compounds: A Cross-Linguistic Study [Синтаксис і семантика прийменників у автоматичній інтерпретації іменних груп і складних слів: порівняльне дослідження] / Roxana Girju // Computational linguistics. – 2009. – Vol. 35. – No. 2. – Pages 185–228. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.06-77-prep13#.WH4ggH3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.06-77-prep13

У статті досліджено синтаксичні і семантичні параметри прийменників у контексті семантичної інтерпретації іменних груп і складних слів. Дослідження проведене на основі багатомовних даних з набору з шести мов: англійської, іспанської, італійської, французької, португальської і румунської. Акцент на англійській мові і романських мовах добре вмотивований. Англійські іменні групи і складні слова здебільшого перекладаються конструкціями типу N P N (іменник прийменник іменник), у яких P (прийменник) може варіювати залежно від семантики. Таким чином, у статті описано емпіричне дослідження дистрибуції іменних груп і складних слів і дистрибуції їх значень у двох різних корпусах на основі двох наборів новітніх класифікаційних міток: набору із восьми прийменників Лауера і нашого списку 22 семантичних відносин. Також показано зв'язок між двома наборами міток. Крім того, за наявності тренувального набору англійських іменних груп і складних іменників і їх перекладів на п'ять романських мов, запропонований алгоритм автоматично визначає правила класифікації і застосовує їх до нових тестових даних для семантичної інтерпретації. Експериментальні результати порівнюються з результатами двох новітніх методів, описаних у літературі.

Переклад В. Коломієць

Zhitomirsky-Geffet, M. Bootstrapping Distributional Feature Vector Quality [Самоналаштування якості дистрибутивного вектора ознак] / Maayan Zhitomirsky-Geffet, Ido Dagan // Computational linguistics. – 2009. – Vol. 35. – No. 3. – Pages 435–461. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.08-032-R1-06-96#.WH4hTn3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.08-032-R1-06-96

У статті описано новітній метод самоналаштування для підвищення якості оцінювання вектора ознак в умовах дистрибутивної схожості слів. Цей метод було створено завдяки спробам використати дистрибутивну схожість для

визначення конкретних семантичних відносин лексичного слідування. Здійснений аналіз виявив, що основною причиною досить низького ступеня семантичної схожості, виявленої за допомогою методів дистрибутивної схожості, є недостатня якість векторів ознак слів, спричинена недосконалою оцінкою ознак. Завдяки цим даним було визначено алгоритм самоналаштування, який забезпечує вдосконалену оцінку ознак, а отже вищу якість векторів ознак. В основі запропонованого підходу лежить ідея, що ознаки, спільні для подібних слів, також є найхарактернішими для їх значень, а отже повинні бути активізовані. Ця ідея реалізована через етап самоналаштування, який було застосовано до вихідної стандартної апроксимації простору схожості. Висока ефективність методу самоналаштування оцінювалась у двох різних експериментах: на основі створеної вручну еталонної анотації і на основі автоматично створеного набору даних для зняття омонімії. Ці результати були потім підтверджені шляхом застосування новітнього квантитативного вимірювання якості вагових функцій ознак. Вдосконалена вагова функція також уможливлює масштабний відбір ознак, що означає, що найхарактерніші ознаки слова справді сконцентровані у верхніх рангах його вектора. Нарешті, експерименти з трьома значимими мірами схожості і двома ваговими функціями ознак показали, що схема самоналаштування є обгрунтованою і не залежить від вихідних функцій, до яких її застосовують.

Переклад В. Коломієць

Cook, P. Automatically Identifying the Source Words of Lexical Blends in English [Автоматичне визначення вихідних компонентів лексичних стягнень у англійській мові] / Paul Cook, Suzanne Stevenson // Computational linguistics. – 2010. – Vol. 36. – No. 1. – Pages 129-149. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2010.36.1.36104#.WH4h4H3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2010.36.1.36104

Неологізми створюють проблеми для систем обробки природної мови, тому що їх немає у лексиконі системи, і як наслідок, відсутня лексична інформація про такі слова. Поширеним способом створення нових слів є лексичне стягнення, прикладом якого є cosmeceutical, стягнення слів cosmetic і pharmaceutical. У статті запропоновано статистичну модель для виведення вихідних компонентів лексичного стягнення на основі виявлених лінгвістичних характеристик стягнень. Ці характеристики переважно залежать від впізнаваності вихідних слів у стягненні. Було анотовано набір із 1186 неологізмів, який включав 515 стягнень, і здійснено тестування розроблених методів за допомогою частини набору із 324 одиниць. У цьому першому дослідженні нових стягнень точність визначення вихідних компонентів стягнення становила 40%, що відповідає зниженню частоти помилок на 39% понад відомим базовим рівнем. У статті також наведено

попередні результати, які свідчать, що використані для ідентифікації вихідних компонентів характеристики можуть бути використані для розрізнення стягнень та інших типів неологізмів.

Переклад М. Андрєєва

Fengxiang, F. An Asymptotic Model for the English Hapax/Vocabulary Ratio [Асимптотична модель співвідношення гапакс/вокабулярій у англійській мові] / Fan Fengxiang // Computational linguistics. – 2010. – Vol. 36. – No. 4. – Pages 631–637. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli_a_00013#.WIHrgX3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli_a_00013

У існуючій літературі зазначається, що приблизно 50% вокабулярію англійського тексту чи колекції текстів становлять гапакси. Подібна постійність дещо спантеличує. Для дослідження цього явища було використано Британський національний корпус, який містить 100 мільйонів слів. Результати свідчать, що співвідношення гапаксів і вокабулярію виглядає як U-подібна крива. Спочатку зі збільшенням обсягу тексту співвідношення гапакси/вокабулярій зменшується; однак після того, як обсяг тексту сягає близько трьох мільйонів слів, співвідношення гапакси/вокабулярій починає неухильно зростати. Комп'ютерне моделювання показує, що зі збільшенням обсягу тексту вищезгадане співвідношення може досягти 1.

Переклад А. Синящик

Boleda, G. Modeling Regular Polysemy: A Study on the Semantic Classification of Catalan Adjectives [Моделювання регулярної багатозначності: дослідження семантичної класифікації каталонських прикметників] / Gemma Boleda, Sabine Schulte im Walde, Toni Badia // Computational linguistics. – 2012. – Vol. 38. – No. 3. – Pages 575–616. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00093#.WH4j7X3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00093

У статті описано дослідження автоматичного визначення семантичних класів каталонских прикметників на основі дистрибуції і морфологічної інформації, з особливим акцентом на багатозначних прикметниках. Мета дослідження полягає у виділенні і описі широких класів, наприклад, якісних (gran "великий") і відносних (pulmonar "легеневий") прикметників, а також у виявленні багатозначних прикметників, таких як econòmic ("економічний | дешевий"). Безпесередньою метою є моделювання регулярної багатозначності, тобто типів чергування значень, які є спільними для всіх лем. Поки що і семантичні класи прикметників, і регулярна багатозначність

рідко привертали увагу в емпіричній комп'ютерній лінгвістиці.

У статті розглядаються два основних конкретних питання. По-перше, якою є адекватна широка семантична класифікація прикметників? Наведено емпіричне обгрунтування якісного і відносного класів, визначених у теоретичних працях, і відкрито тип прикметників, якому не приділялось достатньо уваги, а саме, клас на позначення подій. По-друге, як з точки зору обчислень найкраще моделювати регулярну багатозначність? У статті описано дві моделі і стверджується, що і в теоретичному, і в емпіричному плані друга з них, яка моделює регулярну багатозначність у значенні одночасного членства у різних базових класах, є більш адекватною, ніж перша, яка намагаєтьс визначити незалежні багатозначні класи. Наш найкращий класифікатор досягає точності 69,1% у порівнянні із стандартом 51%.

Переклад М. Погребної

Peris, A. Empirical Methods for the Study of Denotation in Nominalizations in Spanish [Емпіричні методи дослідження денотації у номіналізаціях] / Aina Peris, Mariona Taulé, Horacio Rodríguez // Computational linguistics. – 2012. – Vol. 38. – No. 4. – Pages 827–865. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00112#.WH4kNn3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00112

Стаття присвячена віддієслівним іменникам у іспанській мові, а саме денотативним відмінностям між номіналізаціями подій і результатів. Вона має дві цілі: по-перше, виявити найголовніші характеристики, потрібні для такого денотативного розрізнення і, по-друге, створити систему автоматичної класифікації віддієслівних іменників за їх денотацією. Дослідження базується на теоретичних гіпотезах, які стосуються цих семантичних відмінностей. Здійснено їх емпіричний аналіз за допомогою методів машинного навчання, які є основою класифікатора ADN-Classifier. Це перший інструмент, призначений для автоматичної класифікації віддієслівних іменників у іспанській мові на події, результати або недостатньо визначені типи. ADN-Classifier допоміг здійснити кількісну оцінку істинності наших тверджень стосовно віддієслівних іменників. Проведено серію експериментів для тестування ADN-Classifier за допомогою різних моделей і в різних реалістичних ситуаціях, які відрізнялися наявними ресурсами знань і програмами обробки природної мови. ADN-Classifier продемострував хороші результати (точність 87,20%).

Переклад В. Коломієць

Mohammad, S. M. Computing Lexical Contrast [Обчислення лексичного контрасту] / Saif M. Mohammad, Bonnie J. Dorr, Graeme Hirst, Peter D. Turney // Computational linguistics. – 2013. – Vol. 39. – No. 3. – Pages 555–590. – Режим доступу до анотації:

http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00143#.WH4lUH3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00143

Знання ступеня семантичного контрасту між словами широко застосовується у обробці природної мови, зокрема машинному перекладі, видобуванні інформації і діалогових системах. Створені вручну лексикони містять антоніми, такі як *гарячий* і *холодний*. Антоніми бувають різних видів, наприклад антиподні, комплементарні і ті, що градуюються. Проте існуючі лексикони рідко класифікують антоніми на різні типи. Вони також не містять пар слів, які не є антонімами, але в тій чи іншій мірі є протилежними за значенням, таких як *теплий* і *холодний* або *тропічний* і *морозний*. У статті запропоновано автоматичний метод ідентифікації протилежних в тій чи іншій мірі пар слів, який базується на припущенні, що якщо пара слів, А і Б, є в тій чи іншій мірі протилежними за значенням, то існує така пара антонімів, В і Г, у якій В є тісно пов'язаним із А, а Г є тісно пов'язаним із Б. (Наприклад, існує пара антонімів *гарячий* і *холодний*, і *тропічний* пов'язаний із *гарячим*, а *морозний* пов'язаний із *холодним*.) Це називається припущенням протилежності.

Спочатку за допомогою масштабного інтернет-артельного експерименту було визначено ступінь згоди між людьми щодо поняття антонімії і її різновидів. Потім описано автоматичну і емпіричну міру лексичного контрасту, яка базується на припущенні протилежності, корпусній статистиці і структурі тезаурусу типу Роже. Показано, як за допомогою чотирьох різних наборів даних здійснено оцінювання нашого методу на двох різних завданнях: вирішенні питань про найбільш контрастуюче слово і розрізнення синонімів і антонімів. Результати аналізувалися по чотирьом частинам мови і п'яти різним видам антонімів. Показано, що запропонована міра лексичного контрасту перевершує існуючі методи, дозволяючи досягти високої точності і широкої покриваючої здатності.

Переклад В. Коломієць

Velardi, P. OntoLearn Reloaded: A Graph-Based Algorithm for Taxonomy Induction [OntoLearn Reloaded: алгоритм на основі графа для генерації таксономії] / Paola Velardi, Stefano Faralli, Roberto Navigli // Computational linguistics. – 2013. – Vol. 39. – No. 3. – Pages 665–707. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00146#.WIUd5n3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00146

У 2004 році в цьому журналі була опублікована наша стаття з описом OntoLearn, однієї з перших систем для автоматичної генерації таксономії з документів і веб-сайтів. З того часу наша група продовжувала активні дослідження системи OntoLearn, яка стала довідником для дослідників. У цій

статті описано оновлений метод генерації таксономії, який називається OntoLearn Reloaded. На відміну від методів генерації таксономії, описаних у літературі, наш новий алгоритм генерує концепти і відношення повністю з нуля шляхом автоматичного видобування термінів, визначень і гіперонімів. Результатом є дуже густий, циклічний і потенційно незв'язаний граф гіперонімів. Потім алгоритм генерує з цього графа таксономію шляхом оптимального галуження і нової процедури зважування. Виконані експерименти свідчать про отримання високоякісних результатів як під час створення зовсім нових таксономій, так і під час реконструювання гілок ієрархій існуючих таксономій.

Переклад В. Коломієць

Li, L. Improved Estimation of Entropy for Evaluation of Word Sense Induction [Удосконалена оцінка ентропії для оцінювання виведення значень слів] / Linlin Li, Ivan Titov, Caroline Sporleder // Computational linguistics. – 2014. – Vol. 40. – No. 3. – Pages 671–685. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00196#.WH4miX3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00196

Теоретико-інформаційні підходи є найтиповішими способами оцінки методів кластерного аналізу, зокрема систем виведення значень слів (*англ.* word sense induction, *скор.* WSI). Такі підходи базуються на статистичних оцінках ентропії. Проте стандартна оцінка методом максимальної вірогідності є дуже упередженою і упередженість залежить, поміж іншого, від числа кластерів і розміру вибірки. Через це вказані підходи є ненадійними і необ'єктивними у випадках, коли різні системи створюють різну кількість кластерів, а обсяг вибірки не є занадто великим. А це якраз повністю відповідає умовам оцінки WSI, за яких експериментально визначеної кількості значень у кластері не існує, а типовий сценарій оцінювання передбачає використання невеликої кількості вживань кожного слова для обчислення кількісних показників. У статті описано точніші алгоритми оцінювання ентропії і проаналізовано їхню продуктивність як у моделюванні, так і в оцінці систем WSI .

Переклад В. Коломієць

Gao, D. Cross-lingual Sentiment Lexicon Learning With Bilingual Word Graph Label Propagation [Автоматичне створення двомовного словника емоційно-оціночної лексики шляхом використання двомовного маркування графів слів] / Dehong Gao, Furu Wei, Wenjie Li, Xiaohua Liu, Ming Zhou // Computational linguistics. – 2015. – Vol. 41. – No. 1. – Pages 20–40. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00207 – Режим доступу до повнотекстової статті:

http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00207

У статті розглядається завдання автоматичного створення двомовного словника емоційно-оціночної лексики, що має на меті автоматичне генерування словників емоційно-оціночної лексики для цільових мов за допомогою наявних англомовних словників емоційно-оціночної лексики. Завдання формалізовано як проблему машинного навчання на двомовному графі слів, на якому коректно репрезентовані внутрішньомовні зв'язки між словами однієї мови та міжмовні зв'язки між словами у різних мовах. Розглядаючи слова англомовного емоційно-оціночного лексикону як вихідні, запропоновано метод використання розмітки двомовного графу слів з метою визначення полярності нерозмічених оціночних слів у цільовій мові. Зокрема, показано, що для побудови внутрішньомовного відношення можуть бути використані як синонімічні, так і антонімічні зв'язки між словами, а також, що для побудови міжмовних зв'язків може бути успішно використана інформація про вирівнювання слів, одержана з двомовних паралельних речень. Оцінка автоматичного створення словника емоційно-оціночної лексики китайської мови показує, що запропонований підхід перевершує існуючі підходи і за точністю, і за повнотою. Експерименти на матеріалі набору даних проекта NTCIR також підтверджують ефективність автоматично згенерованого словника емоційно-оціночної лексики у класифікації емотивності на рівні речень.

Переклад М. Дубка

Irvine A. A Comprehensive Analysis of Bilingual Lexicon Induction [Комплексний аналіз виведення двомовного словника] / Ann Irvine, Chris Callison-Burch // Computational linguistics. – 2017. – Vol. 43. – No. 2. – Pages 273–310. – Режим доступу до анотації: https://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00284 – Режим доступу до повнотекстової статті: https://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00284

Завдання виведення двомовного словника передбачає виведення перекладів слів з одномовних корпусів двома мовами. У статті представлено найповніший на сьогодні аналіз створення двомовного словника. Проведено експерименти на широкому спектрі мов та різних обсягах даних. Проаналізовано англійські переклади з 25 іноземних мов: албанської, азербайджанської, бенгальської, боснійської, болгарської, себуанської, гуджараті, гінді, угорської, індонезійської, латвійської, непальської, румунської, сербської, словацької, сомалійської, іспанської, шведської, тамільської, телегу, турецької, української, узбецької, в'єтнамської та уельської. Особливості виведення двомовного словника проаналізовано не лише на високочастотних словах, як робили попередні дослідники, а й на низькочастотних словах. Низькочастотні слова більш важливі для систем статистичного машинного перекладу, в яких, як правило, відсутні переклади

рідко вживаних слів, яких бракує в їх навчальних даних. Здійснено систематичний аналіз широкого спектру особливостей та явищ, які впливають на якість перекладів, отриманих шляхом виведення двомовного словника. Наведено ілюстративні приклади найкращих перекладів для ортогональних показників еквівалентності перекладів, таких як контекстна та темпоральна схожість. Проаналізовано впливи частотності та нерівномірності даних, обсяги початкових двомовних словників та одномовних навчальних корпусів. Крім того, введено новий дискримінаційний підхід до виведення двомовного словника. Ця дискримінаційна модель здатна поєднувати в собі різноманітні характеристики, які поодинці є лише слабкими ознаками еквівалентності перекладів. Коли вагові коефіцієнти ознак встановлюються дискримінаційно, ці сигнали забезпечують переклади значно вищої якості, ніж попередні підходи, які поєднували сигнали без учителя (наприклад, використовуючи мінімальний інвертований ранг). Також, здійснено пряме порівняння продуктивності запропонованого методу з передовим генеративним підходом – алгоритмом аналізу відповідностей канонічних кореляцій (АВКК), який використовує Хаґіґі та ін. (Haghighi et al., 2008). Точність запропонованого алгоритму досягає 42% на відміну від 15% точності АВКК.

Переклад А. Шульги

Корпусна лінгвістика

Yamamoto, M. Using Suffix Arrays to Compute Term Frequency and Document Frequency for All Substrings in a Corpus [Використання списків суфіксів для обчислення частоти термінів і частоти документів у всіх підрядках корпусу] / Mikio Yamamoto, Kenneth W. Church // Computational linguistics. – 2001. – Vol. 27. – No. 1. – Pages 1–30. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120101300346787#.WIHkXn3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120101300346787

У статистичній обробці природної мови звичайно використовуються біграми і триграми; у статті описано методи обробки значно довших n-грамів. Списки суфіксів (Manber, U. and Myers, G., 1990) були спочатку створені для того, щоб обчислити частотність та розташування підрядка (n-грама) у послідовності (корпусі) довжиною N. Для обчислення частот усіх N(N+1)/2 підрядків у корпусі, підрядки були згруповані у прийнятну кількість класів еквівалентності. Таким чином забороняюче обчислення підрядків було скорочене до практичного обчислення класів. У статті описано як алгоритм, так і програму, які використовувались для обчислення частотності термінів (term frequency, *скор.* tf) і частотності документів (document frequency, *скор.* df) для усіх n-грамів у двох великих корпусах, 50-мільйонному англійському корпусі текстів з газети Wall Street Journal обсягом 50 мільйонів слів і японському корпусі текстів з газети Mainichi Shimbun обсягом 216 мільйонів ієрогліфів.

У другій частині статті ці частоти використано для знаходження «цікавих» підрядків. Лексикографів цікавили n-грами з високим рівнем спільної інформації (СП), у яких об'єднана частота термінів є вищою за випадкову, за умови що частини n-граму об'єднані незалежно. Остаточна зворотна частота документу (ОЗЧД) порівнює частоту документу з іншою випадковою моделлю, у якій терміни з певною частотою розподілені по всій колекції у випадковому порядку. СП, як правило, відбирає словосполучення з некомпозиційною семантикою (що часто порушує припущення про незалежність), у той час як ОЗЧД зазвичай виявляє технічну термінологію, імена і ключові слова, придатні для видобування інформації (яка, зазвичай, має невипадковий розподіл у документах). Комбінація СП і ОЗЧД дає кращі результати, ніж будь-яка окрема складова, у виокремленні японських слів.

Переклад В. Коломієць

Kilgarriff, A. Introduction to the Special Issue on the Web as Corpus [Вступне слово до спеціального випуску, присвяченого Всесвітній мережі як корпусу текстів] / Adam Kilgarriff, Gregory Grefenstette //

Computational linguistics. – 2003. – Vol. 29. – No. 3. – Pages 333–347. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120103322711569#.VStSk1ChGCA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120103322711569

Всесвітня мережа, завдяки величезній кількості найрізноманітніших лінгвістичних даних на різних мовах у вільному доступі, є омріяним поприщем для лінгвістів. У цьому спеціальному випуску журналу «Комп'ютерна лінгвістика» розглядаються шляхи дослідження цієї мрії.

Переклад В. Туз

Resnik, P. The Web as a Parallel Corpus [Інтернет як паралельний корпус] / Philip Resnik, Noah A. Smith // Computational linguistics. – 2003. – Vol. 29. – No. 3. – Pages 349–380. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120103322711578#.VStTRVChGCA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120103322711578

Паралельні корпуси стали невід’ємним ресурсом у роботі в галузі багатомовної обробки природної мови. У цій статті доповідається про досвід використання системи STRAND для знаходження паралельних текстів у Всесвітній мережі. Спочатку розглядаються вихідний алгоритм та результати, а потім представляється низка важливих удосконалень. Ці вдосконалення включають використання контрольованого тренування на основі структурних особливостей документів з метою покращення результатів класифікації, новий критерій перекладацької еквівалентності, який базується на змісті, та адаптацію системи для користування Архівом Інтернету для широкомасштабного пошуку паралельних текстів із Всесвітньої мережі. На завершення демонструється корисність цих методів у створенні великого паралельного корпусу для мовної пари з обмеженою кількістю електронних ресурсів.

Переклад Д. Попової

Keller, F. Using the Web to Obtain Frequencies for Unseen Bigrams [Використання Всесвітньої мережі для отримання частот прихованих біграм] / Frank Keller, Mirella Lapata // Computational linguistics. – 2003. – Vol. 29. – No. 3. – Pages 459–484. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120103322711604#.VStUoVChGCA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120103322711604

У цій статті продемонстровано, що Всесвітню мережу можна використовувати для отримання частот біграм, невидимих у певному корпусі. Ми описуємо метод отримання підрахунків для біграм прикметника-

іменника, іменника-іменника та дієслова-додатка із Всесвітньої мережі за допомогою запиту у пошуковій системі. Ми оцінюємо цей метод, демонструючи: (а) високу кореляцію частот у Всесвітній мережі та в корпусі; (б) достовірний кореляційний зв'язок між частотами у Всесвітній мережі та оцінками достовірності; (в) достовірний кореляційний зв'язок між частотами у Всесвітній мережі та частотами, відтвореними за допомогою згладжування на основі класів; (г) високу ефективність частот, характерних для Всесвітньої мережі, при пробному знятті багатозначності.

Переклад Д. Попової

Fais, L. Inferable Centers, Centering Transitions, and the Notion of Coherence [Вивідні центри, центровані переходи і поняття когерентності] / Laurel Fais // Computational linguistics. – 2003. – Vol. 30. – No. 2. – Pages 119–150. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120104323093267#.WIHn0n3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120104323093267

Виконане на основі теорії центрування дослідження корпусу японських електронних листів, яке аналізується у статті, значною мірою спирається на урахування вивідних центрів. Проте використання цього різновиду центрів призводить до високого ступеня неоднозначності у розмітці переходів і, як наслідок, у характеристиці когерентності корпусу. Складність полягає у вимозі ідентифікації референтів дискурсу у дефініціях перехідних станів. Замість висновків, підказаних використанням вивідних центрів, пропонується лексична когезія як цілком кокретне і усталене поняття. Два нові переходи, основані на лексичній спорідненості, а не на ідентичності, доповнюють стандартні визначення і більш адекватно характеризують когерентність цього корпусу. Проаналізовано наслідки і перспективи висунутої пропозиції.

Переклад В. Коломієць

Palmer, M. The Proposition Bank: An Annotated Corpus of Semantic Roles [Банк пропозицій: анотований корпус семантичних ролей] / Martha Palmer , Daniel Gildea , Paul Kingsbury // Computational linguistics. – 2005. – Vol. 31. – No. 1. – Pages 71–106. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/0891201053630264#.WIHoPH3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/0891201053630264

У проекті Банк пропозицій застосовується практичний підхід до семантичного представлення, який додає до синтаксичних структур із корпусу Penn Treebank шар інформації про аргументи предикатів або мітки семантичних ролей. Створений ресурс можна уважати поверхневим, оскільки у ньому не представлені кореференція, квантифікація і багато інших явищ

вищого порядку, але також всеосяжним, оскільки він ураховує кожне вживання кожного дієслова у корпусі і дозволяє отримувати репрезентативні статистичні дані.

Обговорено критерії, які використовуються для визначення наборів семантичних ролей, які використовуються у процесі розмітки і для аналізу частоти синтаксичних/синтаксичних чергувань у корпусі. Описано автоматичну систему розмітки семантичних ролей, навчену на корпусних даних, і проаналізовано вплив на її продуктивність різних типів інформації, зокрема порівняння повного синтаксичного розбору із лінійним зображенням і роль пустих категорій («слідів») банку синтаксичних дерев.

Переклад В. Коломієць

Ringlstetter, C. Orthographic Errors in Web Pages: Toward Cleaner Web Corpora [Орфографічні помилки на веб-сторінках: на шляху до зменшення кількості помилок у веб-корпусах] / Christoph Ringlstetter, Klaus U. Schulz, Stoyan Mihov // Computational linguistics. – 2006. – Vol. 32. – No. 3. – Pages 295–340. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2006.32.3.295#.VStQ_lChGCA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2006.32.3.295

Оскільки Всесвітня мережа безумовно є найбільшим публічним сховищем текстів природною мовою, сучасні експерименти, методи та інструменти в галузі корпусної лінгвістики часто використовують Інтернет як корпус. Для забезпечення роботи прикладних програм, для яких відсутність помилок має критичне значення, потрібно впоратися із проблемою великої кількості орфографічних і граматичних помилок у веб-документах. У цій статті ми досліджуємо розподіл різних типів орфографічних помилок на веб-сторінках. Як побічний продукт розробляються методи для ефективного виявлення сторінок із помилками та для маркування орфографічних помилок у прийнятних веб-документах, зменшуючи, таким чином, кількість помилок у корпусах та базах лінгвістичних знань, які автоматично вилучаються з Інтернету.

Переклад Д. Попової

Hockenmaier, J. CCGbank: A Corpus of CCG Derivations and Dependency Structures Extracted from the Penn Treebank [Корпус дериватів і структур залежностей, видобутих з корпусу Penn Treebank на основі ККГ] / Julia Hockenmaier, Mark Steedman // Computational linguistics. – 2007. – Vol. 33. – No. 3. – Pages 355–396. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2007.33.3.355#.WIHpyn3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2007.33.3.355

У статті описано алгоритм перетворення корпусу Penn Treebank у корпус

дериватів комбінаторної категоріальної граматики (ККГ), доповнений суміжними і віддаленими залежностями слів. Отриманий корпус, ККГбанк, включає 99,4% речень з корпусу Penn Treebank. Доступ до корпусу, який використовується для тренування широкозахватних статистичних парсерів з сучасним рівнем визначення залежностей надає Консорціум лінгвістичних даних.

Для отримання лингвістично достовірних результатів досліджень на основі ККГ і видалення невідповідностей у вихідному анотуванні знадобились детальний аналіз конструкцій і анотування у корпусі Penn Treebank і велика кількість виправлень у корпусі Treebank. У статті аналізується вплив результатів дослідження на видобування інших лінгвістично виразних граматик з корпусу Treebank і на структуру майбутніх банків синтаксичних дерев.

Переклад В. Коломієць

Cohn, T. Constructing Corpora for the Development and Evaluation of Paraphrase Systems [Укладання корпусів для розробки і оцінки систем перефразування] / Trevor Cohn, Chris Callison-Burch, Mirella Lapata // Computational linguistics. – 2008. – Vol. 34. – No. 4. – Pages 597–614. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.08-003-R1-07-044#.WIHqpX3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.08-003-R1-07-044

Важливим компонентом багатьох завдань обробки природної мови є автоматичне перефразування. У статті описано новий паралельний корпус із розміткою перефразувань. У дослідженні використовується визначення перефразування на основі вирівнювання слів. Показано, що воно дозволяє досягти високої міри узгодженості між анотаторами. Оскільки коефіцієнт каппа призначений для номінальних даних, у дослідженні використано альтернативний критерій узгодженості, прийнятний для структурованих завдань вирівнювання. Проаналізовано шляхи ефективного використання корпусу у автоматичному оцінюванні перефразування (наприклад, шляхом визначення точності, повноти і F1), а також у розробці лінгвістично багатих моделей перефразування на основі синтаксичної структури.

Переклад В. Коломієць

Marom, Y. An Empirical Study of Corpus-Based Response Automation Methods for an E-mail-Based Help-Desk Domain [Емпіричне вивчення корпусних методів автоматизації відповіді для домену електронної служби технічної підтримки] / Yuval Marom, Ingrid Zukerman // Computational linguistics. – 2009. – Vol. 35. – No. 4. – Pages 597–635. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2009.35.4.35404#.WIHrA33sSGA – Режим доступу до повнотекстової статті:

http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2009.35.4.35404

У даній статті описано дослідження корпусних методів автоматизації відповідей електронної служби технічної підтримки. Точніше кажучи, ми досліджуємо два практичні аспекти цієї проблеми: (1) пошук інформації та (2) рівень деталізації інформації. Ми розглядаємо застосування двох методів збору інформації (вивід та передбачення) до інформації, представленої на двох рівнях деталізації (на рівні тексту та на рівні речення). До методів текстового рівня належить повторне використання наявного електронного листа-відповіді для відповіді на нові запити. Методи на рівні речення включають використання методів екстрактивного багатотекстового реферування з метою поєднання інформаційних блоків більше ніж з одного електронного листа. Оцінка ефективності різних методів показує, що при поєднанні вони здатні успішно автоматизувати створення відповідей для значної частини запитів електронною поштою у нашому корпусі. Ми також досліджуємо процес метапідбору, який навчається обирати один метод для обробки нового запиту електронною поштою, забезпечуючи, таким чином, єдине вирішення питання автоматизації відповідей.

Переклад Д. Попової, М. Погребної

Zaidan, O. F. Arabic Dialect Identification [Розпізнавання діалектів арабської мови] / Omar F. Zaidan, Chris Callison-Burch // Computational linguistics. – 2014. – Vol. 40. – No. 1. – Pages 171–202. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00169#.WIHzlH3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00169

Письмова форма арабської мови, сучасна стандартна арабська мова (Modern Standard Arabic, *скор.* MSA), значно відрізняється від різних розмовних регіональних діалектів арабської мови – справжніх «рідних мов» носіїв арабської мови. Ці діалекти у свою чергу дуже відрізняються один від одного. Проте, оскільки письмові тексти переважно пишуться стандартною арабською мовою, майже всі корпуси арабської мови складаються переважно з текстів на стандарній арабській мові. У статті описано створення новітнього ресурсу арабської мови з анотуванням діалектів. Створено великий одномовний корпус під назвою Анотований онлайн-корпус арабської мови (O. F. Zaidan and C. Callison-Burch, 2011), який містить велику кількість текстів на арабських діалектах. Описано спробу анотування, метою якого було розпізнавання діалектизмів (і самого діалекту) у кожному з понад 100 000 речень у корпусі, виконану шляхом краудсорсингу, і проаналізовано цікаві варіанти поведінки анотаторів (наприклад, переважне розпізнавання їх власних діалектів). Цей новий анотований корпус було використано для розпізнавання арабських діалектів − визначення діалекту, на якому написано речення, на основі особливостей послідовності слів у ньому. Корпусні дані

було використано для навчання і оцінки автоматичних класифікаторів для визначення діалектів і було встановлено, що класифікатори, які використовують діалектні дані, значно перевершують контрольні результати, отримані за допомогою даних виключно стандартною арабською мовою, і демонструють точність визначення, близьку до експертної. Нарешті, створені класифікатори були використані для пошуку діалектних даних у результатах масштабного інтернет-пошуку, які складалися з 3,5 мільйонів сторінок, видобутих із електронних арабських газет.

Переклад В. Коломієць

Tsvetkov, Y. Identification of Multiword Expressions by Combining Multiple Linguistic Information Sources [Знаходження багатослівних словосполучень шляхом комбінування різних джерел лінгвістичної інформації] / Yulia Tsvetkov, Shuly Wintner // Computational linguistics. – 2014. – Vol. 40. – No. 2. – Pages 449–468. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00177#.WIHsen3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00177

У статті сформульовано загальні принципи використання різних джерел лінгвістичної інформації з метою знаходження багатослівних словосполучень в текстах природною мовою. Визначено різні лінгвістично обгрунтовані класифікаційні ознаки і запропоновано нові методи їх обчислення. Потім вручну визначено взаємозв'язки між цими ознаками і представлено їх у вигляді Байесовій мережі. В результаті отримано потужний класифікатор, який може знаходити багатослівні словосполучення різних типів і різні синтаксичні конструкції у корпусах текстів. Запропонований метод є неконтрольованим і незалежним від мови, він потребує відносно мало лінгвістичних ресурсів і завдяки цьому підходить для великої кількості мов. Наведено результати для англійської, французької та ідиш і продемонстровано значне підвищення точності знаходження словосполучень у порівнянні з вихідними даними меншої складності.

Переклад В. Коломієць

Marimon, M. Automatic Selection of HPSG-Parsed Sentences for Treebank Construction [Автоматичний відбір синтаксичних дерев, створених аналізатором на основі граматики HPSG, для побудови банку дерев] / Montserrat Marimon, Núria Bel, Lluís Padró // Computational linguistics. – 2014. – Vol. 40. – No. 3. – Pages 523–531. – Режим доступу до анотації http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00190#.WH6Mxn3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00190

У статті описується комплексний підхід до визначення і відбору синтаксичних дерев високої якості за допомогою створеної вручну HPSG

граматики для іспанської мови, втіленої у системі створення лінгвістичних знань. У даному підході використовується повне узгодження (тобто точне синтаксичне співпадіння) разом із алгоритмом вибору дерева розбору і синтаксичним аналізатором на основі дерев залежності, навченим на тих самих даних. Головна мета полягає у створенні гібридної методики розмітки корпусу, яка є комбінацією виключно автоматичної розмітки і ручного відбору дерев розбору, з метою підвищення ефективності анотування і одночасно підтримання високої якості та узгодженості, необхідних для будь-якого з передбачених застосувань банку синтаксичних дерев.

Переклад В. Коломієць

Prasad, R. Reflections on the Penn Discourse TreeBank, Comparable Corpora, and Complementary Annotation [Декілька міркувань про корпус Penn Discourse TreeBank, порівняльні корпуси та додаткове анотування] / Rashmi Prasad, Bonnie Webber, Aravind Joshi // Computational linguistics. – 2014. – Vol. 40. – No. 4. – Pages 921–950. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00204#.VStPVFChGCA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdfplus/10.1162/COLI_a_00204

Корпус Penn Discourse Treebank (PDTB) став загальнодоступним у 2008 році. Він і досі залишається найбільшим анотованим вручну корпусом риторичної структури дискурсу. Використана в корпусі розмітка риторичної структури, яка виражена певними лексичними засобами дискурсивної зв'язності або асоціюється із суміжністю речень, не тільки полегшила його використання у прикладній лінгвістиці та психолінгвістиці, а й сприяла анотуванню порівняльних корпусів інших мов і жанрів.

У зв'язку з цим наша стаття переслідує чотири цілі: (1) надати вичерпну інформацію про PDTB для тих, хто про нього не знає; (2) виправити деякі помилкові (чи, можливо, легковажні) припущення щодо PDTB та його анотування, які могли знизити вагомість отриманих результатів чи пригальмувати виконання процедур прийняття рішень, підказаних даними; (3) пояснити розбіжності в анотуванні порівняльних ресурсів у інших мовах і жанрах, які повинні допомогти розробникам майбутніх порівняльних корпусів зрозуміти, чи мають ці розбіжності значення; і (4) перелічити і пояснити відношення між анотуванням PDTB та додатковим анотуванням інших мовних явищ. В статті використано дослідження, як наші, так і інших дослідників, виконані після появи корпусу.

Переклад Д. Попової

Gimenes P. Spelling Error Patterns in Brazilian Portuguese [Розподіл орфографічних помилок у бразильському варіанті португальської мови] / Priscila A. Gimenes, Norton T. Roman, Ariadne M. B. R. Carvalho // Computational linguistics. – 2015. – Vol. 41. – No. 1. – Pages 175–183. – Режим доступу до анотації:

http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00216 – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00216

Статистика розподілу помилок у друкованих текстах, виведена Дамерау п’ятдесят років тому, і досі використовується у великій кількості різних мов. Оскільки ці статистичні дані були отримані з текстів англійською мовою, було порушено питання про те, чи можна їх застосувати до інших мов. У статті це питання розв’язується шляхом аналізу набору друкованих текстів бразильською португальською і виведення статистики саме для цієї мови. Результати показують, що важливу роль відіграють діакритичні знаки, про що свідчить частота помилок, пов'язаних із ними. Тому первинні висновки Дамерау є здебільшого непридатними для систем перевірки орфографії, хоча й можуть бути корисними, якщо не зважати на діакритичні знаки. Крім того, порівняння цих результатів із опублікованими результатами для іспанської мови не виявило статистично значущих відмінностей між двома мовами. Це означає, що розподіл орфографічних помилок залежить від прийнятого набору символів, а не від самої мови.

Переклад А. Шульги

Barbot N. Large Linguistic Corpus Reduction with SCP Algorithms [Скорочення великих лінгвістичних корпусів за допомогою алгоритмів ЗМП] / Nelly Barbot, Olivier Boëffard, Jonathan Chevelu and Arnaud Delhay // Computational linguistics. – 2015. – Vol. 41. – No. 3. – Pages 355–383. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00225 – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00225

Розбудова лінгвістичного корпусу є важливим завданням у створенні великих розмічених корпусів, необхідних для створення різних видів додатків. Наприклад, такі технології усного мовлення, як автоматичне розпізнавання або синтез мовлення потребують величезної кількості мовних даних навчання моделей на основі даних або для синтезу мовлення. Збір даних завжди пов’язаний із витратами (запис мовлення, перевірка поміток тощо), і, як правило, чим більше даних зібрано, тим дорожчим є додаток. В цьому контексті у статті описано способи скорочення обсягу текстових корпусів із збереженням достатнього рівня мовного різноманіття, необхідного для моделі або додатку. Ця проблема може бути формалізована як завдання покриття множини (ЗПМ). У статті оцінюються дві алгоритмічні дослідницькі установки, які застосовувались для розробки великих корпусів текстів англійської та французької мов для пошуку фонологічної інформації або частиномовного розмічування. Перший розглянутий алгоритм є стандартним, «жадібним» рішенням з агломеруючою стратегією, а автори пропонують другий алгоритм на основі релаксації Лагранжа. Другий підхід

передбачає нижчий рівень витрат для кожної розв'язаної проблеми. Цей рівень можна використати як метрику для оцінки якості скороченого корпусу незалежно від застосованого алгоритму. Експерименти показують, що умовно оптимальний алгоритм, подібний до «жадібного», дає хороші результати; вартість його рішень близька до нижнього рівня (близько 4.35% для трифонемних рішень). Зазвичай обмеження в ЗПМ бінарні, але у статті запропоновано узагальнення, в якому обмеження на кожну використану категорію можуть бути багатоелементними.

Переклад А. Шульги

Ling, W. Mining Parallel Corpora from Sina Weibo and Twitter [Видобування паралельних корпусів з мікроблогів Sina Weibo та Twitter]/ Wang Ling, Luís Marujo, Chris Dyer, Alan W. Black, Isabel Trancoso // Computational linguistics. – 2016. – Vol. 42. – No. 2. – Pages 307–343. – Режим доступу до анотації: https://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00249 – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00249

Мікроблоги, наприклад Twitter, Facebook і Sina Weibo (китайський еквівалент Twitter), є неабияким лінгвістичним ресурсом. На відміну від текстів жанрів, які редагуються, таких як стрічка новин, мікроблоги містять різностильові обговорення практично будь-якої теми великою кількістю людей різними мовами і діалектами. У статті показано, що деякі користувачі мікроблогів публікують "самостійно перекладені" повідомлення, призначені для читачів, які розмовляють іншими мовами, створюючи один і той же запис кількома мовами або повторно публікуючи переклади оригінальних записів іншою мовою. У статті представлено метод пошуку та видобування таких природних паралельних даних. Для вирішення проблеми вирівнювання, якого вимагає пошук паралельних текстів, запропоновано високоефективний алгоритм динамічного програмування. Застосувавши цей метод, було отримано приблизно три мільйони паралельних сегментів китайською і англійською мовами з мікроблогу Sina Weibo шляхом цільового моніторингу користувачів Weibo, які роблять записи кількома мовами. Крім цього, з довільної вибірки записів у Twitter було отримано великий обсяг паралельних даних для різних мовних пар. Оцінювання проведено шляхом оцінки точності запропонованого методу видобування даних по відношенню до ручного маркування, а також з точки зору корисності в якості тренувальних даних для китайсько-англійської системи машинного перекладу. На відміну від традиційних ресурсів паралельних даних автоматично видобуті паралельні дані забезпечують значне покращення якості перекладу записів у мікроблогах і незначні покращення перекладу відредагованих текстів новин.

Переклад М. Дубка

Лінгвістичне анотування

Eugenio, B. D. The Kappa Statistic: A Second Look [Коєфіцієнт Каппа: новий погляд] / Barbara Di Eugenio, Michael Glass // Computational linguistics. – 2004. – Vol. 30. – No. 1. – Pages 95–101. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120104773633402#.WIHnYH3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120104773633402

Протягом останніх років коефіцієнт узгодженості Каппа фактично став стандартом у оцінюванні узгодженості між анотаторами у завданнях анотування. У статті розглядаються фактори, які впливають на к і які здебільшого ігнорувались дослідниками. По-перше, аналізуються припущення, покладені в основу різних обчислень очікуваного компонента узгодженості к. По-друге, проаналізовано, як показник к залежить від розповсюдженості й упередженості.

Переклад В. Коломієць

Craggs, R. Evaluating Discourse and Dialogue Coding Schemes [Оцінювання схем анотування дискурсу і діалогу] / Richard Craggs, Mary McGee Wood // Computational linguistics. – 2005. – Vol. 31. – No. 3. – Pages 289–296. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120105774321109#.WIHovn3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120105774321109

В оцінюванні схем анотування дискурсу і діалогу важливу роль грає статистика узгодженості. Проте відповідні методи оцінювання узгодженості між анотаторами і способи інтерпретації їх результатів потребують глибшого розуміння. У статті описується роль методів оцінювання узгодженості між анотаторами і стверджується, що у дослідженнях надійності для оцінювання узгодженності підходять лише методи з поправкою на випадковість, які передбачають стандартний розподіл міток для всіх анотаторів. Потім наводяться рекомендації, як робити висновки про надійність на основі результатів статистики узгодженості.

Переклад В. Коломієць

Navigli, R. Consistent Validation of Manual and Automatic Sense Annotations with the Aid of Semantic Graphs [Послідовна перевірка валідності ручного і автоматичного анотування значень за допомогою семантичних графів] / Roberto Navigli // Computational linguistics. – 2006. – Vol. 32. – No. 2. – Pages 273–281. – Режим доступу до анотації:

http://www.mitpressjournals.org/doi/abs/10.1162/coli.2006.32.2.273#.WH4Z633sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2006.32.2.273

Загальновизнано, що завдання анотування текстів значеннями із електронного словника є складним і часто суб'єктивним. Хоча для подолання розбіжностей між анотуванням значень можуть використовуватись методики типу міри узгодженості між анотаторами і голосування, немає гарантії послідовності у виборі значень відносно словника посилань.

У статті описано візуальний інструмент для перевірки ручного і автоматичного анотування значень під назвою Valido, який вирівнює можливі розбіжності і забезпечує послідовність рішень за допомогою моделей семантичних взаємовідносин.

Переклад В. Коломієць

Bayerl, P. S. Identifying Sources of Disagreement: Generalizability Theory in Manual Annotation Studies [Пошук причин неузгодженості: теорія узагальнюваності у дослідженнях ручного анотування] / Petra Saskia Bayerl, Karsten Ingmar Paul // Computational linguistics. – 2007. – Vol. 33. – No. 1. – Pages 3–8. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2007.33.1.3#.WIHpZ33sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2007.33.1.3

Багато проектів, пов'язаних з анотуванням, показали, що якість здійсненого вручну анотування часто є нижчою, ніж потрібно для надійного аналізу даних. Тому актуальним завданням є визначення основних причин низької якості анотування. Цінним інструментом для його вирішення є теорія узагальнення, адже вона дозволяє диференціювати і детально аналізувати фактори, від яких залежить якість анотації. У статті розглядаються основні поняття теорії узагальнення і наводиться приклад її застосування на основі опублікованих матеріалів.

Переклад А. Синящик

Reidsma, D. Reliability Measurement without Limits [Оцінювання надійності без обмежень] / Dennis Reidsma, Jean Carletta // Computational linguistics. – 2008. – Vol. 34. – No. 3. – Pages 319–326. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2008.34.3.319#.WIEKjn3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2008.34.3.319

У комп'ютерній лінгвістиці уважається, що кількісна оцінка надійності якогось статистичного показника, наприклад κ, 0,8 гарантує придатність закодованих вручну даних для певної мети, від 0,67 до 0,8 є достатньою, а

нижчі значення − сумнівними. У статті показано, що основне застосування цих даних, машинне навчання, допускає дані з низьким рівнем надійності, якщо будь-яка неузгодженість між анотаторами виглядає як випадковий шум. Проте, коли неузгодженість починає виникати регулярно, комп'ютер може врахувати їх так само, як враховує справжні закономірності у даних, через що результати виглядатимуть краще, ніж вони є насправді. Через велику кількість показників надійності, які зараз визнаються в цій області, неузгодженість може викликати значне підвищення результатів і навіть показник 0,8 не зможе гарантувати, що результати, які здаються хорошими, дійсно є такими. Хоча це висновок на основі здорового глузду, він впливає на особливості роботи комп'ютерних лінгвістів. Вони, принаймні, повинні шукати закономірності у неузгодженості між анотаторами і оцінювати їх наслідки.

Переклад В. Коломієць

Artstein, R. Inter-Coder Agreement for Computational Linguistics [Оцінка узгодженості між анотаторами у комп'ютерній лінгвістиці] / Ron Artstein, Massimo Poesio // Computational linguistics. – 2008. – Vol. 34. – No. 4. – Pages 555–596. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.07-034-R2#.WIHqOX3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.07-034-R2

У статті досліджено методи оцінки узгодженості між анотаторами корпусів. Показано математику та базові припущення коефіцієнтів узгодженості, зокрема альфи Криппендорфа, а також пі Скотта і каппи Коена, проаналізовано використання коефіцієнтів у кількох завданнях анотування. Стверджується, що зважені, альфаподібні коефіцієнти, які традиційно застосовуються у комп'ютерній лінгвістиці рідше, ніж каппаподібні мірки, можуть бути більш прийнятними для багатьох анотувань корпусів, але їх використання ще більше ускладнює інтерпретацію значення коефіцієнта.

Переклад В. Коломієць

Klebanov, B. B. From Annotator Agreement to Noise Models [Від показника узгодженості між розмітниками до шумових моделей] / Beata Beigman Klebanov, Eyal Beigman // Computational linguistics. – 2009. – Vol. 35. – No. 4. – Pages 495–503. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2009.35.4.35402#.WIXUQH3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2009.35.4.35402

У статті обговорюється перехід від анотованих даних до золотого стандарту, тобто до підвибірки, яка з високою достовірністю є достатньо вільною від шумів. За відсутності відповідного повторного тлумачення показники узгодженості свідчать, що за якістю набір даних не є еталоном.

Високий показник узгодженості не є ані достатнім, ані необхідним для виокремлення з анотованого матеріалу деякої кількості достовірних даних. Розроблено математичну базу для оцінки рівня шуму узгодженої підвибірки анотованих даних, що допомагає зважено підходити до виділення еталону.

Переклад М. Погребної, І. Снєгурова

Bayerl, S. P. What Determines Inter-Coder Agreement in Manual Annotations? A Meta-Analytic Investigation [Що впливає на узгодженість між розмітниками? Металінгвістичне дослідження] / Petra Saskia Bayerl, Karsten Ingmar Paul // Computational linguistics. – 2011. – Vol. 37. – No. 4. – Pages 699–725. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00074#.WIHr0X3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00074

Останні дослідження узгодженості розмітників в основному стосувалися обчислень і інтерпретації, а також правильного вибору індексів. Хоча такі дослідження важливі, вони враховують лише «кінець» історії, а саме, що робити, коли зібрано дані. На нашу думку, не менш важливо знати, перш за все, як досягається узгодженість і які фактори впливають на узгодженість розмітників у процесі анотування, оскільки ця інформація може лягти в основу конкретних рекомендацій щодо планування і організації проектів анотування. Для того щоб з'ясувати, чи існують фактори, які постійно впливають на узгодженість розмітників, було виконано метааналітичний аналіз досліджень анотування, які містили відсотки узгодженості. Метааналіз, здійснений на основі 346 індексів узгодженості, узагальнив фактори, згадані у 96 дослідженнях анотування з трьох предметних областей (розв'язання семантичної неоднозначности, просодичних транскрипцій і фонетичних транскрипцій). Проведений аналіз виявив сім факторів, які впливають на опубліковані показники узгодженості: предметна область анотування, число категорій у схемі анотування, кількість розмітників у проекті, попереднє навчання розмітників, інтенсивність навчання розмітників, мета анотування, а також метод, використаний для підрахунку розбіжностей процентних долей. На основі отриманих результатів розроблено практичні рекомендації щодо оцінювання, інтерпретації, обчислення і опису узгодженості розмітників. Також коротко проаналізовано теоретичне значення поняття якості анотування.

Переклад В. Коломієць

Jiang, W. Automatic Adaptation of Annotations [Автоматичне адаптування розмітки] / Wenbin Jiang, Yajuan Lü, Liang Huang, Qun Liu // Computational linguistics. – 2015. – Vol. 41. – No. 1. – Pages 119–147. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00210 – Режим доступу до повнотекстової статті:

http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00210

Розмічені вручну корпуси текстів є незамінними ресурсами, проте для багатьох завдань розмічування, таких як створення банків дерев, існує чимало корпусів з несумісними принципами розмічування. Це призводить до неефективного використання людського досвіду, проте проблему можна вирішити шляхом інтеграції знань у корпуси з різними принципами розмічування. У статті описано проблему адаптування розміток і внутрішні принципи її розв'язання і представлено серію послідовно вдосконалених моделей, які можуть автоматично адаптувати розміток.

Створені алгоритми оцінено на завданнях із сегментації слів китайської мови та синтаксичного аналізу на основі граматики залежностей. Оскільки немає універсальних правил сегментації через відсутність морфології у китайській мові, для сегментації слів адаптовано розмітку із значно більшого корпусу People's Daily для меншого, але більш популярного корпусу Penn Chinese Treebank. Для синтаксичного аналізу на основі граматики залежностей адаптовано розмітку з корпусу Penn Chinese Treebank для семантично-орієнтованого корпусу Dependency Treebank, анотованого з використанням суттєво відмінних принципів розмічування. В обох експериментах автоматичне адаптування розміток дало позитивні результати, забезпечивши сучасний рівень ефективності, незважаючи на використання виключно локальних категорій у машинному навчанні.

Переклад М. Дубка

Mathet, Y. The Unified and Holistic Method Gamma (γ) for Inter-Annotator Agreement Measure and Alignment [Гамма (γ) уніфікованого й цілісного методу визначення та вирівнювання узгодженості між розмітниками] / Yann Mathet, Antoine Widlöcher, Jean-Philippe Métivier // Computational linguistics. – 2015. – Vol. 41. – No. 3. – Pages 437–479. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00227 – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00227

Вже більше 15 років для перевірки достовірності процесів розмічування в комп'ютерній лінгвістиці широко використовуються міри узгодженості. Хоча категоризуванню було присвячено багато уваги, уніфікування розглядається в меншій кількості досліджень, а при об'єднанні обох парадигм є доступними і згадуються ще менше методів. Стаття має три цілі. По-перше, стверджується, що для того, щоб впоратися з уніфікуванням, міри вирівнювання та узгодженості слід розглядати як єдиний процес, оскільки відповідна міра повинна спиратися на вирівнювання одиниць, запропоноване різними розмітниками, і це вирівнювання повинне вираховуватись згідно з принципами конкретної міри. По-друге, запропоновано нову універсальну міру γ, яка відповідає цій вимозі і враховує обидві парадигми, і описано її впровадження. По-третє, показано, що при одночасному застосуванні двох

парадигм цей новий метод працює так само добре, як і інші спеціалізовані методи категоризування або сегментування, а може навіть краще.

Переклад М. Дубка

Mathet, Y. The Agreement Measure γcat a Complement to γ Focused on Categorization of a Continuum [Показник узгодженості γcat, додаток до γ, призначений для категоризації континууму] / Yann Mathet // Computational linguistics. – 2017. – Vol. 43. – No. 3. – Pages 661–681. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00296 – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00296

Оцінювання узгодженості, коли кілька розмітників вільно розташовують одиниці різних розмірів і категорій на континуумі, є складним завданням через розбіжності як у розташуванні, так і в класифікації за категоріями. Новий показник узгодженостіγ пропонує комплексне рішення, яке враховує і розташування, і категорії. У статті запропоновано додатковий коефіцієнт γcat, який доповнює γ оцінювання узгодженості у категоризації континууму, ігноруючи при цьому розбіжності в розташуванні. При застосуванні виключно до класифікації за категоріями (з попередньо визначеними одиницями) γcat діє так само, як і відомий спеціальний показник α Кріппендорфа, навіть при відсутніх значеннях, що доводить його сталість. Також запропоновано варіацію γcat, яка забезпечує всебічне оцінювання класифікації на категорії для кожної окремої категорії. Всю множину коефіцієнтів γ реалізовано у вільному програмному забезпеченні.

Переклад М. Дубка

Проблеми машинного навчання

Siegel, V. E. Learning Methods to Combine Linguistic Indicators: Improving Aspectual Classification and Revealing Linguistic Insights [Методи навчання для об'єднання лінгвістичних показників: поліпшення видової класифікації та виявлення лінгвістичної інформації] / Eric V. Siegel, Kathleen R. McKeown // Computational linguistics. – 2000. – Vol. 26. – No. 4. – Pages 595–628. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120100750105957#.WH3xNn3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120100750105957

Для того щоб судити про час, видова класифікація ділить дієслова на невеликий набір базових категорій. Ця класифікація необхідна для розуміння часових модифікаторів та оцінки часових відносин і тому є обов'язковим компонентом багатьох додатків для обробки природної мови.

Видова категорія дієслова може бути спрогнозована за допомогою частот одночасного вживання дієслова і певних лінгвістичних модифікаторів. Ці показники частоти, які називаються лінгвістичними показниками, відбираються за допомогою лінгвістичної інформації. Проте самі по собі лінгвістичні показники мають невелику прогностичну цінність, і тому є недостатніми, якщо використовуються окремо.

У статті порівнюються три методи машинного навчання з учителем для об'єднання декількох лінгвістичних показників для видової класифікації: дерев рішень, генетичного програмування та логістичної регресії. Створено набір з 14 показників для класифікації за двома видовими відмінностями. Як показала оцінка довільних наборів дієслів, які зустрічаються у двох корпусах, такий підхід підвищує якість класифікації для обох відмінностей. Це свідчить про ефективність лінгвістичних показників і дозволяє створити такий необхідний всеохоплюючий метод для автоматичної видової класифікації. Крім того, автоматично створені моделі виявили декілька лінгвістичних показників, потрібних для видової класифікації. Також здійснено порівняння методів контрольованого і неконтрольованого навчання для цього завдання.

Переклад К. Погорєлова

Higgins, D. A Machine Learning Approach to Modeling Scope Preferences [Застосування машинного навчання для моделювання налаштувань області дії] / Derrick Higgins, Jerrold M. Sadock // Computational linguistics. – 2003. – Vol. 29. – No. 1. – Pages 73–96. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120103321337449#.WH4WlH3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120103321337449

У статті описане корпусно-базоване дослідження налаштувань області дії кванторів. Керуючись останніми працями з мультимодальних граматичних систем у теоретичній лінгвістиці і давньою традицією об'єднання різних інформаційних ресурсів у обробці природної мови, область дії визначається як окремий граматичний модуль синтаксису модуля граматики. Цей модуль включає численні джерела доказів щодо найбільш вірогідної інтерпретації області дії для речення і повністю керується даними. Здатність запропонованих моделей спрогнозувати найбільш вірогідну інтерпретацію області дії для конкретного речення було оцінено в ході описаних у статті експериментів, на основі даних корпусу Penn Treebank, як із синтаксичною анотацією, так і без неї. Ми прагнемо привернути увагу до питання визначення налаштувань області дії, яке загалом ще й досі ігнорується у теоретичній лінгвістиці, а також розглянути різні моделі взаємодії між синтаксисом і областю дії квантора.

Переклад К. Погорєлова

Abney, S. Understanding the Yarowsky Algorithm [Осмислення алгоритму Яровського] / Steven Abney // Computational linguistics. – 2004. – Vol. 30. – No. 3. – Pages 365–395. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/0891201041850876#.WIXPe33sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/0891201041850876

Велика кількість завдань у комп'ютерній лінгвістиці може бути вирішена методами бутстрепінгу (напівконтрольованого навчання). Добре відомим алгоритмом бутстрепінгу є алгоритм Яровського, але він недостатньо досліджений математично. У статті він аналізується як оптимізація об'єктивної функції. Конкретніше, показано, що велика кількість варіантів алгоритму Яровського (правда, не сам оригінальний алгоритм) оптимізують або вірогідність, або тісно споріднену об'єктивну функцію К.

Переклад В. Коломієць

Fernández, R. Classifying Non-Sentential Utterances in Dialogue: A Machine Learning Approach [Класифікація діалогічних реплік, які не є реченнями: метод машинного навчання] / Raquel Fernández, Jonathan Ginzburg, Shalom Lappin // Computational linguistics. – 2007. – Vol. 33. – No. 3. – Pages 397–427. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2007.33.3.397#.WH4bPn3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2007.33.3.397

У дослідженні використано добре відомі методи машинного навчання для розв'язання новітнього завдання – класифікації реплік діалогу, які не є реченнями. Запропонована детальна таксономія класів висловлень, які не є реченнями, а також викладено результати кількох експериментів з машинним

навчанням. Спочатку описано пілотне дослідження одного з класів висловлень, які не є реченнями, у таксономії – голих питальних слів або «шлюзів» - і розглянуто проблему роз'язання двозначності у прочитанні, яку можуть спричинити шлюзи. Потім цей підхід поширено на класифікацію усього набору класів висловлювань, які не є реченнями, із збалансованою f-мірою результатів біля 87%. Отже проведені експерименти засвідчили, що за допомогою таксономії можна успішно вирішити завдання правильного визначення классу висловлень, які не є реченнями, і таким чином створити обнадійливу основу для більш загального завдання повноцінної обробки висловлень, які не є реченнями.

Переклад В. Коломієць

Graehl, J. Training Tree Transducers [Навчання перетворювачів дерев] / Jonathan Graehl, Kevin Knight, Jonathan May // Computational linguistics. – 2008. – Vol. 34. – No. 3. – Pp. 391–427. – Режим доступу до анотації http://www.mitpressjournals.org/doi/abs/10.1162/coli.2008.07-051-R2-03-57#.WH6Con3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2008.07-051-R2-03-57

Багато імовірнісних моделей для природніх мов пишуться зараз у вигляді ієрархічних дерев. Моделювання на основі дерев поки що не має багатьох базових інструментів, які є стандартними у моделюванні на основі строк із використанням кінцевих станів. Можливою теоретичною базою може стати теорія автоматичного перетворення дерев, оскільки вона розроблялася у великій кількості літературних джерел. Ми обґрунтовуємо використання перетворювачів дерев для природних мов і розглядаємо проблему тренування імовірнісних перетворювачів дерево-дерево і дерево-строка.

Переклад В. Коломієць

Gildea, D. Grammar Factorization by Tree Decomposition [Факторизація грамматики шляхом роз'єднання дерев] / Daniel Gildea // Computational linguistics. – 2011. – Vol. 37. – No. 1. – Pp. 231–248. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli_a_00040#.WH6FSH3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli_a_00040

У статті описується використання такого поняття теорії графів як ширина дерева у знаходженні ефективних алгоритмів синтаксичного аналізу. Цей метод, як і алгоритм дерев з'єднань, який використовується у графічних моделях машинного навчання, дозволяє автоматично знаходити ефективні алгоритми, такі як алгоритм O(n4) для білексичних граматик, запропонованих Д. Айзнером і Д. Саттою. Проаналізовано труднощі використання цього методу у алгоритмах синтаксичного аналізу для загальних лінійних контекстно-незалежних систем переписування.

Продемонстровано, що будь-який поліноміальний алгоритм для цієї задачі передбачає удосконалений алгоритм апроксимації для добре дослідженої проблеми ширини дерева у загальних графах.

Переклад В. Коломієць

Huang, F. Learning Representations for Weakly Supervised Natural Language Processing Tasks [Навчання представлень для завдань обробки природної мови з незначним залученням учителя] / Fei Huang, Arun Ahuja, Doug Downey, Yi Yang, Yuhong Guo, Alexander Yates // Computational linguistics. – 2014. – Vol. 40. – No. 1. – Pages 85–120. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00167#.WIE9Q33sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00167

Важливе значення для створення точних систем обробки природної мови при недостатній кількості анотованих даних із відповідної предметної області має знаходження правильних представлень слів. У статті досліджено нові методи виявлення ознак за допомогою n-грамних моделей, прихованих марківських моделей та інших статистичних мовних моделей, зокрема нової моделі марківського випадкового поля на частковій решітці. Експерименти з різними завданнями, зокрема із частиномовною розміткою і видобуванням інформації, свідчать, що ознаки, отримані за допомогою статистичних моделей, у комбінації з більш традиційними ознаками дають кращі результати, ніж традиційні представлення самі по собі, і що представлення у вигляді графічних моделей дають кращі результати, ніж n-грамні моделі, особливо для малочастотних і багатозначних слів.

Переклад В. Коломієць

Sun, X. Feature-Frequency–Adaptive On-line Training for Fast and Accurate Natural Language Processing [Збільшення швидкості й точності обробки природної мови за допомогою онлайн навчання, яке адаптується до частоти ознак] / Xu Sun, Wenjie Li, Houfeng Wang, Qin Lu // Computational linguistics. – 2014. – Vol. 40. – No. 3. – Pages 563–586. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00193#.WIENNH3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00193

Швидкість і точність навчання – це дві головні цілі масштабних систем обробки природної мови. Як правило, приходиться вибирати між швидкістю і точністю. Легко збільшити швидкість навчання за рахунок точності або підвищити точність за рахунок швидкості. Проте нелегко одночасно підвищити швидкість навчання і точність, що і є метою даного дослідження. Для досягнення цієї мети розроблено новий метод навчання − онлайн

навчання, яке адаптується до частоти ознак − для швидкого і точного навчання систем обробки природної мови. Метод базується на ідеї про те, що частотніші характеристики повинні мати швидкість навчання, яка швидше уповільнюється. Теоретичний аналіз свідчить, що запропонований метод є конвергентним, з високою швидкістю збіжності. Експерименти проводились на основі добре відомих контрольних завдань, зокрема таких як розпізнавання власних назв, сегментація слів і словосполучень і аналіз тональності. Ці завдання складаються із трьох структурованих класифікаційних завдань і одного неструктурованого класифікаційного завдання, відповідно з бінарними ознаками і з речовими ознаками. Результати експерименту свідчать, що запропонований метод є швидшим і точнішим, ніж існуючі методи, даючи конкурентноспроможні результати у завданнях із різними характеристиками.

Переклад В. Коломієць

Berant J. Efficient Global Learning of Entailment Graphs [Ефективне комплексне автоматичне створення графів логічного слідування] / Jonathan Berant, Noga Alon, Ido Dagan, Jacob Goldberger // Computational linguistics. – 2015. – Vol. 41. – No. 2. – Pages. 249–291 – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00220 – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00220

Правила логічного слідування між предикатами є невід'ємною частиною багатьох семантичних додатків. Тому протягом останніх років автоматичне виведення таких правил було активною галуззю дослідження. Було доведено, що методи автоматичного виведення правил логічного слідування між предикатами, які враховують залежності між різними правилами (наприклад, логічне слідування є транзитивним відношенням), покращують якість правил, але не масштабуються, тобто кількість предикатів, які опрацьовуються, часто досить мала. У статті представлено методи для автоматичного створення транзитивних графів (що називаються графами логічного слідування), які містять десятки тисяч вузлів: вузли представляють предикати, а ребра відповідають правилам логічного слідування. Запропоновані методи можуть масштабуватись до великої кількості предикатів, використовуючи структурні властивості графів логічного слідування, зокрема той факт, що вони мають «деревоподібні» властивості. Ці методи були застосовані до двох наборів даних і було показано, що вони знаходять якісні рішення швидше, ніж методи, запропоновані раніше. Крім того, вони вперше масштабуються до великих графів, які налічують 20 000 вузлів і понад 100 000 ребер.

Переклад А. Шульги

Rozovskaya, A. Adapting to Learner Errors with Minimal Supervision [Адаптування до помилок не носіїв мови з мінімальним залученням

учителя] / Alla Rozovskaya, Dan Roth, Mark Sammons // Computational linguistics. – 2017. – Vol. 43. – No. 4. – Pages 723–760. – Режим доступу до анотації: https://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00299 – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00299

У статті розглянуто проблему виправлення помилок у письмових текстах, створених тими, для кого англійська є нерідною мовою, з точки зору машинного навчання і розглянуто важливе питання розробки для цього завдання відповідної парадигми навчання, яка визначає типові письмові помилкиу не носіїв мови за допомогою часткового залучення учителя. Існуючі підходи до навчання є компромісом між великою кількістю немаркованих даних, отриманих завдяки моделям, навчання яких забезпечували носії мови, та додатковими знаннями про типові помилки при вивченні мови, отриманими за допомогою дорожчого методу навчання на анотованих навчальних даних.

Запропоновано новий метод навчання, який поєднує переваги обох стандартних парадигм навчання – навчання на текстах носіїв мови або на розмічених текстах тих, хто вивчає мову – та перевершує обидва ці стандартні методи. Використовуючи ключовий висновок про відносну простоту параметрів, які стосуються закономірностей письмових помилок у англійській як іноземній мові, розроблено методи, які можуть включати знання про закономірності помилок на основі невеликої розміченої вибірки, але які, з іншого боку, розроблені на основі текстів носіїв англійської мови.

Основним внеском статті є представлення та аналіз двох методів адаптування автоматично визначених моделей до типових письмових помилок не носіїв мови; перший метод застосовується у генеративних класифікаторах, другий − у дискримінативних класифікаторах. Обидва методи продемонстрували сучасні результати у кількох змаганнях з редагування текстів. Зокрема, система Іллінойс, у якій реалізовано ці методи, посіла перше місце у двох останніх спільних завданнях з редагування текстів конференції CoNLL. Проводиться подальше оцінювання запропонованих методів, у якому досліджується вплив використання даних про помилки носіїв однієї мови, близькоспоріднених мов і неспоріднених мов.

Переклад М. Дубка

Ákos, K. Representation of Linguistic Form and Function in Recurrent Neural Networks [Представлення мовної форми і функції в рекурентних нейронних мережах] / Ákos Kádár, Grzegorz Chrupała, Afra Alishahi. – 2017. – Vol. 43. – No. 4. – Pages 761–780. – Режим доступу до анотації: https://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00300 – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00300

У статті представлено інноваційні методи аналізу шаблонів активації

рекурентних нейронних мереж з лінгвістичної точки зору та досліджено типи мовних структур, які вони автоматично вивчають. В якості прикладу використано стандартну автономну модель мови та архітектуру багатозадачної закритої рекурентної мережі, що складається з двох паралельних шин із спільними вставними словами. Візуальна шина автоматично навчається шляхом прогнозування представлень візуального оточення, що відповідає вхідному реченню, а Текстова шина автоматично навчається передбачати наступне слово в тому самому реченні. Запропоновано метод оцінки ролі окремих слововживань у вхідному реченні в остаточному прогнозі мереж. За допомогою цього методу показано, що Візуальна шина вибірково приділяє увагу лексичним категоріям і граматичним функціям, що містять семантичну інформацію, і автоматично навчається по-різному обробляти типи слів залежно від їхньої граматичної функції та їхньої позиції в послідовній структурі речення. На противагу цьому, моделі мови, навпаки, відносно чутливіші до слів із синтаксичною функцією. Подальший аналіз найбільш інформативних контекстів кожної моделі, представлених n-грамами, показує, що в порівнянні з Візуальною шиною, моделі мови сильніше реагують на абстрактні контексти, які представляють синтаксичні конструкції.

Переклад М. Дубка

Сегментування тексту

Teahan, W. J. A Compression-based Algorithm for Chinese Word Segmentation [Алгоритм для сегментування китайських текстів на слова на основі стиснення] / W. J. Teahan, Yingying Wen, Rodger McNab, Ian H. Witten // Computational linguistics. – 2000. – Vol. 26. – No. 3. – Pages 375–393. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120100561746#.WIEqvn3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120100561746

У китайському письмі немає пробілів або інших розмежувань слів. Хоча текст можна уважати відповідною послідовністю слів, є багато проблем у встановленні меж. Тлумачення тексту як послідовності слів корисне для певних завдань інформаційного пошуку та збереження даних, наприклад повнотекстового пошуку, стиснення текстів на основі слів, та виокремлення ключових фраз. У статті описується алгоритм виведення правильного розміщення меж слова за допомогою стандартної для стиснення текстів адаптивної мовної моделі. Алгоритм навчається на корпусі попередньо сегментованого тексту, і при застосуванні до нового тексту вставляє межі слова так, щоб максимально збільшити отримане стиснення. Цей простий і загальний метод добре працює у спеціальних алгоритмах сегментування китайської мови.

Переклад О. Мартинюк

Venkataraman, A. A Statistical Model for Word Discovery in Transcribed Speech [Статистична модель встановлення меж слів у транскрибованому мовленні] / Anand Venkataraman // Computational linguistics. – 2001. – Vol. 27. – No. 3. – Pages 351–372. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120101317066113#.WIEGq33sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120101317066113

Представлено статистичну модель сегментування і встановлення меж слів у потоці мовлення. Описано поетапний неконтрольований алгоритм навчання для встановлення меж слів на основі цієї моделі. Також наведено результати емпіричних перевірок, які свідчать, що цей алгоритм може конкурувати з іншими моделями, які використовувались для подібних завдань.

Переклад В. Коломієць

Pevzner, L. A Critique and Improvement of an Evaluation Metric for Text Segmentation [Критичний аналіз і вдосконалення метрики оцінювання сегментування тексту] / Lev Pevzner, Marti A. Hearst // Computational

linguistics. – 2002. – Vol. 28. – No. 1. – Pages 19–36. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120102317341756#.WH4VbX3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120102317341756

Метрика оцінки Pk, вперше запропонована Бееферманом, Бергером і Лафферті у 1997 році, стає стандартною мірою оцінювання алгоритмів сегментування тексту. Однак, теоретичний аналіз метрики виявив декілька проблем: метрика звертає більше уваги на пропущені, ніж хибно визначені межі, надає забагато значення незначним помилкам і залежить від варіювання розподілу розміру сегментів. Для вирішення вказаних проблем запропонована проста модифікація метрики Pk. Нова метрика, яка називається Window Diff, пересуває по тексту вікно фіксованого розміру і щоразу, коли кількість меж у вікні не співпадає з дійсною кількістю меж для того вікна з текстом, фіксує помилку алгоритму.

Переклад І. Снєгурова

Mikheev, A. Periods, Capitalized Words, etc. [Крапки, слова з великої літери тощо] / Andrei Mikheev // Computational linguistics. – 2002. – Vol. 28. – No. 3. – Pages 289–318. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120102760275992#.WIE2eH3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120102760275992

У статті описано підхід до вирішення трьох важливих проблем нормалізації тексту: виявлення меж речення, зняття багатозначності слів, написаних з великих літер, у позиціях, де очікується велика літера, та ідентифікація скорочень. На відміну від двох популярних методів обчислювальної статистики і написання спеціалізованих граматик, наш тексто-орієнтований підхід враховує промовисті локальні контексти і повторення окремих слів у межах документа. Цей метод не втрачає ефективності при зміні тематики та появі нової лексики і за продуктивістю не поступається найкращим опублікованим результатам. Після вбудовування у морфологічний аналізатор він допоміг значно знизити рівень помилок у обробці слів з великої літери і встановленні меж речень. Досліджено можливість застосування методу для інших мов і отримано обнадійливі результати.

Переклад К. Погорєлова

Feng, H. Accessor Variety Criteria for Chinese Word Extraction [Критерії варіативності засобів доступу для встановлення меж китайських слів] / Haodi Feng, Kang Chen, Xiaotie Deng, Weimin Zheng // Computational linguistics. – 2003. – Vol. 30. – No. 1. – Pages 75–93. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120104773633394#.WIE

3Mn3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120104773633394

Стаття присвячена проблемі встановлення меж слів у китайських корпусах. Під словом розуміється послідовність кількох китайських ієрогліфів, яка несе певне значення. Наприклад, з точки зору деяких людей «відсоток» і «більше і більше» не є традиційними китайськими словами. Проте у даному дослідженні вони є словами, тому що широко вживаються і мають конкретні значення. Ми виходимо з того, що слово є самостійною мовною одиницею, яку можна використовувати у багатьох різних мовних середовищах. Ми уважаємо ієрогліфи, які знаходяться безпосередньо перед послідовністю (попередники), і ієрогліфи, які знаходяться безпосередньо після послідовності (наступники), важливими факторами для визначення незалежного характеру послідовності. Ми назвали такі ієрогліфи засобами доступу до послідовності, проаналізували кількість окремих попередників і наступників послідовності у великому корпусі (документи TREC 5 і TREC 6) і використали їх для визначення контекстуальної незалежності послідовності від решти речень у документі. Проведені експерименти підтвердили нашу гіпотезу і показали, що це просте правило дає хороші результати у встановленні меж китайських слів і не поступається іншим ітеративним методам, а для довгих слів навіть перевершує їх.

Переклад В. Коломієць

Gao, J. Chinese Word Segmentation and Named Entity Recognition: A Pragmatic Approach [Сегментування китайських текстів на слова і розпізнавання власних назв: прагматичний підхід] / Jianfeng Gao, Mu Li, Chang-Ning Huang, Andi Wu // Computational linguistics. – 2005. – Vol. 31. – No. 4. – Pages 531–574. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120105775299177#.WIE4q33sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120105775299177

У статті описано прагматичний підхід до встановлення меж китайських слів. Він відрізняється від попередніх підходів у основному в трьох аспектах. По-перше, хоча у теоретичній лінгвістиці китайські слова визначені за допомогою різних лінгвістичних критеріїв, у даному дослідженні китайські слова визначаються прагматично як одиниці сегментування, визначення яких залежить від способів їх використання і обробки у реальних комп'ютерних додатках. По-друге, запропоновано прагматичний математичний підхід, у якому встановлення меж відомих слів і виявлення невідомих слів різних типів (наприклад, слів, утворених морфологічними способами, чисел і адрес, власних назв та інших відсутніх у списку слів) може здійснюватися одночасно у єдиному форматі. У інших системах ці завдання звичайно виконуються окремо. Нарешті, ми не допускаємо існування універсального стандарту встановлення меж слів, який не залежить від додатку. Навпаки,

через той прагматичний факт, що різні додатки для обробки природної мови можуть використовувати різні характеристики китайських слів, ми наголошуємо на необхідності багатьох стандартів сегментування.

Ці прагматичні підходи були втілені у детально описаному адаптивному сегментаторі китайських текстів під назвою MSRSeg. Він складається з двох компонентів: 1) універсального сегментатора на основі лінійних змішаних моделей, який забезпечує єдиний підхід до п'яти основних функцій обробки китайської мови на рівні слів: обробки слів лексикону, морфологічного аналізу, виявлення чисел і адрес, розпізнавання власних назв та ідентифікації нових слів, і 2) набору адаптерів виведення, для адаптації виведення універсального сегментатора до стандартів для різних додатків. Оцінка за допомогою п'яти тестових наборів з різними стандартами показала, що адаптивна система відповідає сучасним вимогам на усіх тестових наборах.

Переклад В. Коломієць

Bestgen, Y. Improving Text Segmentation Using Latent Semantic Analysis: A Reanalysis of Choi, Wiemer-Hastings, and Moore (2001) [Удосконалення сегментування тексту за допомогою латентного семантичного аналізу: повторний аналіз статті Ф. Чой, П. Вімер-Гастінгс і Д. Мур [Choi, Wiemer-Hastings, and Moore, 2001] / Yves Bestgen // Computational linguistics. – 2006. – Vol. 32. – No. 1. – Pages 5–12. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2006.32.1.5#.WH4YHn3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2006.32.1.5

Ф. Чой, П. Вімер-Гастінгс і Д. Мур [Choi, Wiemer-Hastings, and Moore, 2001] запропонували використовувати латентний семантичний аналіз (англ. Latent Semantic Analysis, скор. LSA) для видобування семантичної інформації з корпусів, для того щоб удосконалити точність алгоритму сегментування тексту. Порівнявши точність того самого алгоритму за умови урахування або неврахування додаткової семантичної інформації, вони змогли показати переваги, отримані завдяки такій інформації. Проте у їхніх експериментах семантичну інформацію було отримано з корпусу текстів, які повинні були бути сегментовані під час пілотної фази. Якщо більша частина отриманих переваг пояснюється цією унікальною особливістю корпусу LSA, можна поставити під сумнів можливість використання LSA для отримання загальної семантичної інформації, за допомогою якої можна сегментувати нові тексти. Обидва описані у статті експерименти свідчать, що присутність у корпусі LSA пілотних матеріалів має серйозні наслідки, але також що загальна семантична інформація, отримана з великих корпусів, явно поліпшує точність сегментування.

Переклад В. Коломієць

Kiss, T. Unsupervised Multilingual Sentence Boundary Detection [Спонтанне визначення меж речень різними мовами] / Tibor Kiss, Jan Strunk // Computational linguistics. – 2006. – Vol. 32. – No. 4. – Pp. 485–525. – Режим доступу до анотації http://www.mitpressjournals.org/doi/abs/10.1162/coli.2006.32.4.485#.WH6AG33sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2006.32.4.485

У статті описується незалежний від мови, спонтанний підхід до визначення меж речення. Він базується на припущенні, що значна частина проблем, пов'язаних із визначенням меж речень, вирішується відразу після розпізнання абревіатур. Замість використання орфографічних підказок запропонована система здатна з високою точністю виявляти абревіатури, використовуючи три критерії, які вимагають тільки інформації про тип кандидата у абревіатуру і не залежать від контексту: абревіатури можна визначити як високочастотні колокації, що складаються із скороченого слова і кінцевої крапки, абревіатури зазвичай короткі і абревіатури можуть містити внутрішні крапки. Також продемонстровано можливість використання двох інших важливих підзадач визначення меж речення, а саме знаходження заголовних букв та порядкових числівників, для знаходження коллокацій. Запропонована система була ретельно протестована на текстах різних жанрів одинадцятьма різними мовами. Вона досягає хороших результатів без будь-яких додаткових поправок або ресурсів, які відображають специфіку мови. Для оцінки роботи системи використано три критерії. Запропонована система порівняно з іншими системами для визначення меж речень, описаних у літературі.

Переклад К. Погорєлова

Li, Z. Punctuation as Implicit Annotations for Chinese Word Segmentation [Пунктуація як імпліцитна розмітка для сегментування китайських текстів] / Zhongguo Li, Maosong Sun // Computational linguistics. – 2009. – Vol. 35. – No. 4. – Pages 505–512. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2009.35.4.35403#.WIE5ZH3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2009.35.4.35403

У статті описана модель автоматичного встановлення меж китайських слів на основі знаків пунктуації, які є прекрасними роздільниками слів. Навчання здійснюється за допомогою сегментованого вручну корпусу. Запропонований метод є набагато ефективнішим, ніж попередні методи у розпізнаванні невідомих слів. Це крок до розв'язання однієї з найскладніших поблем у сегментування китайських текстів.

Переклад В. Коломієць

Wang, H. A New Unsupervised Approach to Word Segmentation [Новий неконтрольований метод визначення меж слів] / Hanshi Wang, Jian Zhu, Shiping Tang, Xiaozhong Fan // Computational linguistics. – 2011. – Vol. 37. – No. 3. – Pages 421–454. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00058#.WIE6HH3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00058

У статті описано ESA, новий неконтрольований метод визначення меж слів. ESA – це ітеративний процес, який складається з трьох етапів: оцінювання (Evaluation), вибору (Selection) і корегування (Adjustment). На етапі оцінювання як обов'язкова, так і можлива поява у корпусі послідовності символів уважається статистичним підтвердженням якості аналізу. Крім того, статистичні дані про послідовності символів різної довжини стають співставними один із одним завдяки простій обробці під назвою збалансовування (Balancing). На етапі вибору обирається стратегія відносного максимуму без порогових обмежень, яка може бути реалізована за допомогою динамічного програмування. На етапі корегування, частина статистичних даних оновлюється для підвищення якості нових результатів. У проведеному експерименті оцінювання ESA було здійснене за допомогою набору даних SIGHAN Bakeoff-2. Результати свідчать про ефективність ESA для корпусів китайської мови. Варто зазначити, що F-міри результатів переважно монотонно зростають і можуть швидко наблизитися до відносно високих показників. Крім того, емпіричні формули на основі отриманих результатів можуть використовуватися для прогнозування параметрів ESA, щоб обійтися без визначення параметрів, яке зазвичай забирає багато часу.

Переклад В. Коломієць

Формальні моделі мови і їх застосування у комп'ютерній лінгвістиці

Karttunen, L. Introduction to the Special Issue on Finite-State Methods in NLP [Вступ до спеціального випуску, присвяченого методам скінченних станів у обробці природної мови] / Lauri Karttunen, Kemal Oflazer // Computational linguistics. – 2000. – Vol. 26. – No. 1. – Pages 1–2. – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120100561593

На перших етапах становлення сучасної лінгвістики граматики скінченних станів ігнорувалися як абсолютно непридатні, але протягом останнього десятиліття спостерігається значне зростання їх використання у різних додатках для обробки природної мови. У п’яти статтях, які увійшли до спеціального випуску, розглядаються різні аспекти теорії скінченних станів і її практичне застосування.

В. Коломієць

Nederhof, M. J. Practical Experiments with Regular Approximation of Context-Free Languages [Практичні експерименти з формального ототожнення безконтекстних мов] / Mark-Jan Nederhof // Computational linguistics. – 2000. – Vol. 26. – No. 1. – Pages 17–44. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120100561610#.WIUkaH3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120100561610

Розглянуто декілька методів побудови кінцевих автоматів із використанням контекстно-вільної граматики, у тому числі обидва методи, які дозволяють отримати підмножини, і ті, які дають розширені множини вихідної контекстно-вільної мови. Деякі з цих методів регулярного наближення є новими, а деякі інші являють собою удосконалення того, що описано в літературі. Проведено практичні експерименти з різними методами регулярного наближення для усних вхідних даних: гіпотези розпізнавача усного мовлення фільтруються кінцевим автоматом.

Переклад Д. Попової

Alshawi, H. Learning Dependency Translation Models as Collections of Finite-State Head Transducers [Алгоритми навчання перекладу на основі залежностей у вигляді наборів скінченних перетворювачів] / Hiyan Alshawi, Srinivas Bangalore, Shona Douglas // Computational linguistics. – 2000. – Vol. 26. – No. 1. – Pages 45–60. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120100561629#.WIUk13

3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120100561629

У статті дається визначення перетворювачів з ваговою обробкою, скінченних автоматів, які виконують перетворення, починаючи з середини рядка. Ці пертворювачі явно більш ефективні, ніж конкретний стандартний скінченний перетворювач, який обробляє рядки зліва направо. Далі дається визначення алоритмів перетворення на основі залежностей як колекцій перетворювачів з ваговою обробкою, які застосовуються в ієрархічному порядку. Описано алгоритм пошуку на основі динамічного програмування для знаходження оптимального перетворення вхідного рядка відповідно алгоритму перетворення на основі залежностей. Описано метод автоматичного тренування алгоритму перетворення на основі залежностей з використанням набору прикладів вхідних та вихідних рядків. Спочатку, керучись статистикою кореляцій, алгоритм шукає ієрархічні пари тренувальних прикладів, а потім створює переходи перетворювачів відповідно цим парам. Описано результати експериментального застосування цього методу навчання при перекладі з англійської на іспанську та японську мови.

Переклад Д. Попової, М. Погребної

van Noord, G. Treatment of Epsilon Moves in Subset Construction [Використання епсилон-переходів у створенні підмножин] / Gertjan van Noord // Computational linguistics. – 2000. – Vol. 26. – No. 1. – Pages 61–76. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120100561638#.WIUlVn 3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120100561638

Стаття присвячена проблемі детермінування кінцевих автоматів з великою кількістю епсілон-переходів. Експерименти з кінцевими наближеннями граматик природніх мов часто призводить до появи дуже великих автоматів із дуже великою кількістю епсілон-переходів. У цій статті визначаються і порівнюються кілька алгоритмів створення підмножин, які застосовують епсілон-переходи. Проведено експерименти, які свідчать, що алгоритми значно відрізняються на практиці, як за розміром отриманого детермінованого автомата, так і за ефективностю. Крім того, експерименти наводять на думку, що середня кількість епсілон-переходів у стані може допомогти передбачити, який алгоритм, скоріше за все, буде найшвидшим для конкретного вхідного автомата.

Переклад Д. Попової

Kiraz, G.A. Multitiered Nonlinear Morphology Using Multitape Finite Automata: A Case Study on Syriac and Arabic [Багаторівнева нелінійна морфологія з використанням багатострічкових скінченних автоматів:

тематичне дослідження на основі сирійської та арабської мов] / George Anton Kiraz // Computational linguistics. – 2000. – Vol. 26. – No. 1. – Pages 77–105. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120100561647#.WIUlq33sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120100561647

У статті представлено обчислювальну модель нелінійної морфології з прикладами з сирійської та арабської мов. Модель є багаторівневою, оскільки допускає численні лексичні репрезентації, що відповідають численним рівням автосегментної фонології. Модель складається з трьох основних компонентів: 1) лексикону, який є сукупністю підлексиконів, де кожний підлексикон представляє матеріал певного рівня, 2) компонента правил виводу, який відображає декілька лексичних репрезентацій у одній поверхневій формі і навпаки, та (iii) морфотактичного компонента, який використовує автоматні граматики. Система є скінченною, позаяк лексикони і правила можуть бути представленими багатострічковими скінченними автоматами.

Переклад Д. Попової

Morrill, G. Incremental Processing and Acceptability [Поетапна обробка та відповідність вимогам] / Glyn Morrill // Computational linguistics. – 2000. – Vol. 26. – No. 3. – Pages 319–338. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120100561728#.WIUmDH3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120100561728

У статті описано процедуру поетапної, зліва направо, обробки категоріальної граматики Ламбека шляхом створення захисної сітки. Простий показник складності, характеристика в момент великої кількості невирішених валентностей, точно прогнозує різноманітні проблеми обробки, зокрема неоднозначність (garden pathing), неприпустимість вбудовування всередину, очікування пізнього закриття, вибір сфери дії квантифікаторів зліва-направо і зсув іменної групи.

Переклад О. Мартинюк

Rambow, O. D-Tree Substitution Grammars [Граматики заміщення d-дерев] / Owen Rambow, K. Vijay-Shanker, David Weir // Computational linguistics. – 2001. – Vol. 27. – No. 1. – Pages 87–121. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120101300346813#.WIUmdn3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120101300346813

Багато комп'ютерних лінгвістів цікавляться лексикалізованими

граматичними моделями. Одним із добре відомих прикладів є лексикалізована граматика складання дерев (lexicalized tree adjoining grammar, скор. LTAG). У статті пропонується розглядати виводи у LTAG не як маніпуляції над деревами, а як маніпуляції над описами дерев. Новий погляд на лексикалізовану модель піднімає питання про доцільність деяких її аспектів. Описано нову модель — граматику заміщення d-дерев (DSG). Виводи в DSG включають структуру d-дерев, спеціальні види опису дерев. Дерева зчитуються з похідних d-дерев. Показано, як можна використати граматику DSG, яка успадкувала багато характеристик LTAG, для здійснення різних лінгвістичних досліджень, які неможливі з LTAG.

Переклад М. Драчової

Wintner, S. A Note on Typing Feature Structures [Нотатки про типізацію ознакових структур] / Shuly Wintner, Anoop Sarkar // Computational linguistics. – 2002. – Vol. 28. – No. 3. – Pages 389–397. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120102760276027#.WIXJrX3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120102760276027

Ознакові структури використовуються для передачі лінгвістичної інформації у різноманітних лінгвістичних формалізмах. Існують різні визначення ознакових структур; однією з площин варіації є типізація: на відміну від структур нетипізованих ознак, структури типізованих ознак асоціюють кожну структуру з типом і обмежують появу ознак і значення, яке вони набирають, вимогами відповідності. У статті продемонстровано переваги, які типізація дає навіть тим лінгвістичним формалізмам, які використовують структури нетипізованих ознак. Описано метод валідації узгодженості вимог до структури нетипізованих ознак шляхом дотримання порядку обслуговування типів. Цей метод спрощує велику кількість перевірок на стадії компілювання: багато можливих помилок можна виявити до застосування формалізму для здійснення синтаксичного розбору. Розроблено сигнатуру типу для існуючої граматики англійської мови з широким діапазоном можливих застосувань і запроваджено алгоритм виведення типів на основі специфікації ознакової структури у граматиці, який повідомляє про несумісності з сигнатурою. Виявлено велику кількість помилок у граматиці, деякі з яких описано у статті.

Переклад В. Коломієць

Poesio, M. Centering: A Parametric Theory and Its Instantiations [Центрування: параметрична теорія і її трактування] / Massimo Poesio , Rosemary Stevenson , Barbara Di Eugenio , Janet Hitzeman // Computational linguistics. – 2004. – Vol. 30. – No. 3. – Pages 309–363. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120104185091 1#.WIXNp

33sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/0891201041850911

Теорія центрування є найвідомішою платформою для узагальнень про локальні зв'язність і виділеність, проте її положення сформульовані мовою понять, які конкретизуються лише частково, таких як «висловлення», «реалізація» або «ранжування». Детальнішої специфікації цих параметрів теорії намагалася досягти велика кількість дослідників, внаслідок чого положення центрування можуть трактуватися багатьма різними способами. У дослідженні систематично проаналізовано вплив цих різних способів налаштування вказаних параметрів на положення теорії. Для цього потрібно було, в першу чергу, уточнити, які положення містить теорія (одним із висновків є те, що так зване «Обмеження №1» є власне основним положенням теорії). По-друге, потрібно було чітко визначити ці параметричні аспекти. Наприклад, ми стверджуємо, що поняття «займенник», яке використовується у Правилі 1, повинне уважатися параметром. По-третє, потрібно було знайти відповідні методи для оцінки цих положень. З'ясовано, що хоча основне положення теорії про виділеність і прономіналізацію, Правило 1 – перевага прономіналізації ретроспективного центру (ЦР), мало залежить від налаштувань, Обмеження 1 – положення про зв'язність (об'єкту) і унікальність ЦР – значно більше залежить від налаштувань. Воно не підтверджується, якщо параметри налаштовані відповідно загальноприйнятим поглядам («стандартне налаштування»), воно підтверджується, лише якщо дозволена непряма реалізація, і навіть при найсприятливіших налаштуваннях його порушують від 20% до 25% висловлювань у нашому корпусі. Також встановлене оптимальне співвідношення між Правилом 1 з одного боку і Обмеженням 1 і Правилом 2 з другого боку. Налаштування параметрів для зведення порушень локальної зв'язності до мінімуму призводить до зростання порушень виділеності і навпаки. Отримані результати свідчать, що зв'язність «об'єкта» – постійне посилання на ті ж самі об'єкти – повинна бути доповнена принаймні повідомленням про споріднену зв'язність.

Переклад В. Коломієць

Nederhof, M. A General Technique to Train Language Models on Language Models [Загальний метод навчання мовних моделей на мовних моделях] / Mark-Jan Nederhof // Computational linguistics. – 2005. – Vol. 31. – No. 2. – Pages 173–185. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/0891201054223986#.WIXRIX3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/0891201054223986

Показано, що за певних умов мовну модель можна тренувати на основі іншої мовної моделі. Головним прикладом цього методу є навчання кінцевого автомату на основі імовірнісної контекстно-вільної граматики,

завдяки чому відстань Кульбака-Лейблера між граматикою і навченим автоматом є очевидно мінімальною. Навчання n-грамної моделі на основі імовірнісної контекстно-вільної граматики є суттєвим узагальненням існуючого алгоритму.

Переклад В. Коломієць

Malouf, R. Maximal Consistent Subsets [Максимальні упорядковані підмножини] / Robert Malouf // Computational linguistics. – 2007. – Vol. 33. – No. 2. – Pages 153–160. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2007.33.2.153#.WIXRZX3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2007.33.2.153

Стандартні операції уніфікації комбінують вивірену інформацію з інформацією від однієї або більше заперечних ознакових структур. Багато таких операцій включають знаходження максимальних підмножин набору елементарних обмежень, узгоджених між собою і з строгою ознаковою структурою, у якій підмножина максимально упорядкована з точки зору структури класифікації, оскільки до неї не можна додати жодного обмеження, не порушивши упорядкованості. Хоча вказана проблема є НП-повною, існує багато евристичних методів оптимізації, за допомогою яких можна значно зменшити обсяг пошукового простору. У статті пропонується новий метод оптимізації, **обрізка листових верхівок**, який у деяких випадках на декілька порядків прискорює час виконання завдання у порівнянні з описаними раніше алгоритмами. Завдяки цьому стандартні операції уніфікації є достатньо ефективними для застосування до широкого кола проблем і додатків.

Переклад В. Коломієць

Smith, N. Weighted and Probabilistic Context-Free Grammars Are Equally Expressive [Зважені та імовірнісні контекстно-вільні граматики є однаково точними] / Noah A. Smith, Mark Johnson // Computational linguistics. – 2007. – Vol. 33. – No. 4. – Pp. 477–491. – Режим доступу до анотації http://www.mitpressjournals.org/doi/abs/10.1162/coli.2007.33.4.477#.WH6BNX3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2007.33.4.477

У статті аналізується співвідношення між зваженими контекстно-вільними граматиками, у яких кожному правилу виводу відповідає позитивне дійсне число, та імовірнісними контекстно-вільними граматиками, у яких ваги правил виводу, яким відповідає нетермінал, повинні в сумі дорівнювати одиниці. Оскільки клас зважених контекстно-вільних граматик по суті включає в себе імовірнісні контекстно-вільні граматики, можна припустити, що зважені контекстно-вільні граматики здатні описати розподіли, які не

можна описати за допомогою імовірнісних контекстно-вільних граматик. Однак З. Чі (Комп'ютерна лінгвістика. – 1999. – V. 25. – Issue 1. – P. 131-160) і С. П. Ебні, Д. А. МакАлестер і П. Перейра (Матеріали 37-ї щорічної конференції Асоціації комп'ютерної лінгвістики – Коледж Парк, Меріленд, 1999. – С. 542-549) довели, що будь-який розподіл, описаний зваженою контекстно-вільною граматикою, є еквівалентним певному розподілу, описаному імовірнісною контекстно-вільною граматикою. Ми застосували їхні висновки до умовних розподілів і показали, що будь-який визначений зваженою контекстно-вільною граматикою умовний розподіл синтаксичних дерев за наявності ланцюжків є також умовним розподілом, визначеним певною імовірнісною контекстно-вільною граматикою, навіть коли функції розподілу граматик не співпадають. Це свідчить про те, що будь-яке поліпшення точності синтаксичного аналізу або анотування від умовного визначення зваженими контекстно-вільними граматиками або умовними довільними полями до комбінованого визначення імовірнісними контекстно-вільними граматиками або прихованими моделями Маркова пояснюється процедурою визначення, а не зміною типу моделі, оскільки імовірнісні контекстно-вільні граматики і приховані моделі Маркова настільки ж точні, як і, відповідно, зважені контекстно-вільні граматики і ланцюгові умовні довільні поля.

Переклад А. Бобкової

Miyao, Y. Feature Forest Models for Probabilistic HPSG Parsing [Моделі лісу ознак для вірогіднісного синтаксичного аналізу на основі HPSG] / Yusuke Miyao, Jun'ichi Tsujii // Computational linguistics. – 2008. – Vol. 34. – No. 1. – Pages 35–80. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2008.34.1.35#.WIXSaX3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2008.34.1.35

Вірогіднісне моделювання лексикалізованих граматик є складним завданням, оскільки ці граматики використовують складні структури даних, такі як структури типізованих ознак. Це дозволяє уникнути застосування звичайних методів вірогіднісного моделювання, у яких уся структура ділиться на підструктури, виходячи з припущення про статистичну незалежність підструктур. Наприклад, частиномовна розмітка речення розкладається на розмітку кожного слова, а автоматичний синтаксичний аналіз на основі контекстно-вільної граматики (КВГ) розкладається на застосування правил КВГ. Ці методи спираються на структуру поставленого завдання, тобто решітки і дерева, і не можуть застосовуватися до структур графів, які включають структури типізованих ознак.

У статті пропонується вирішити проблему вірогіднісного моделювання складних структур даних, зокрема структур типізованих ознак, за допомогою моделі лісу ознак. Модель лісу ознак допускає спосіб вірогіднісного моделювання без припущення про незалежність, якщо вірогіднісні події

представлені лісами ознак. Ліси ознак є універсальними структурами даних, які представляють омонімічні дерева у спакованій структурі лісу. Моделі лісу ознак є моделями максимальної ентропії, визначеними на основі лісів ознак. Для оцінки по методу максимальної ентропії без розпакування лісів ознак запропоновано алгоритм динамічного програмування. Отже, вірогіднісне моделювання будь-яких структур даних є можливим, якщо вони представлені лісами ознак.

У статті також описано методи представлення за допомогою лісів ознак синтаксичних структур граматики HPSG і предикатно-аргументних структур. Отже, у статті подано повний опис сратегії розробки вірогіднісних моделей для синтаксичного аналізу на основі граматики HPSG. Ефективність запропонованих методів емпірично оцінена за допомогою експериментів із синтаксичним аналізом на основі корпусу Penn Treebank, проаналізовано можливість їх застосування для синтаксичного аналізу реальних речень.

Переклад В. Коломієць

Riggle, J. The Complexity of Ranking Hypotheses in Optimality Theory [Складність ранжування гіпотез у теорії оптимальності] / Jason Riggle // Computational linguistics. – 2009. – Vol. 35. – No. 1. – Pages 47–59. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.07-031-R2-06-98#.WIXTcH3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.07-031-R2-06-98

Якщо маємо задану обмеженнями множину з k обмеженнями в рамках теорії оптимальності (*англ.* Optimality Theory, *скор.* OT), яка її здатність стати класифікаційною схемою для лінгвістичних даних? Одна корисна міра цієї здатності – обсяг найбільшого набору даних, у якому кожна підвибірка узгоджується з унікальною граматичною гіпотезою. Ця міра відома як розмірність Вапника-Червоненкіса (*англ.* Vapnik-Chervonenkis dimension, *скор.* VCD) і є стандартною мірою складності для класів понять у теорії складності обчислень. У статті використовується тризначна логіка базових умов ранжування для того, щоб показати, що VCD теорії оптимальності з k обмеженнями становить k-1. Аналіз OT з точки зору VCD свідчить, що складність OT є регулярною функцією k і що «складність» обчислень в OT має лінійний вигляд у k для великої кількості теорій, які використовують вірогіднісні визначення складності.

Переклад В. Коломієць

Huang, L. Binarization of Synchronous Context-Free Grammars [Бінаризація синхронних контекстно-вільних граматик] / Liang Huang, Hao Zhang, Daniel Gildea, Kevin Knight // Computational linguistics. – 2009. – Vol. 35. – No. 4. – Pages 559–595. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2009.35.4.35406#.WIXUzH3sSGA – Режим доступу до повнотекстової статті:

http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2009.35.4.35406

Системи на основі синхронних граматик і перетворювачів дерев обіцяють поліпшити якість статистичного машинного перекладу, але вони часто потребують величезних обчислювальних потужностей. Через довільне переупорядкування між двома мовами складність в обсязі окремих граматичних правил стрімко зростає. Нами розроблена теорія бінаризації для синхронних контекстно-вільних граматик і описано лінійний за часом алгоритм для бінаризаціїї синхронних правил, якщо це можливо. Проведені нами широкомасштабні експерименти виявили, що майже всі правила бінаризуються і отриманий набір бінаризованих правил значно покращує швидкість і точність сучасної системи машинного перекладу на основі синтаксису. Також проаналізована загальніша, і складніша в обчислювальному плані, проблема знаходження ефективних стратегій синтаксичного аналізу для правил, які неможливо бінаризувати, і описано приблизний алгоритм поліноміального часу для цієї проблеми.

Переклад В. Коломієць

Nesson, R. Complexity, Parsing, and Factorization of Tree-Local Multi-Component Tree-Adjoining Grammar [Складність, синтаксичний аналіз і факторизація багатокомпонентної граматики з'єднання дерев у часткові дерева] / Rebecca Nesson, Giorgio Satta, Stuart M. Shieber // Computational linguistics. – 2010. – Vol. 36. – No. 3. – Pp. 443–480. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli_a_00005#.WH6Ecn3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli_a_00005

Багатокомпонентна граматика з'єднання дерев у часткові дерева (Tree-Local Multi-Component Tree-Adjoining Grammar, скор. TL-MCTAG) є привабливим формалізмом для репрезентації природної мови, бо уважається, що вона уможливлює інкапсуляцію правильної області розташування всередині своїх базових структур. Її багатокомпонентна структура дозволяє моделювати лексичні одиниці, елементи яких можуть знаходитися у реченні на великій відстані один від одного, такі як квантифікатори і питальні слова. Коли вона використовується як базовий формалізм для синхронної граматики, її гнучкість дозволяє їй виражати як тісні зв'язки, так і неоднорідну структуру, потрібні для визначення зв'язків між синтаксисом і семантикою однієї мови або синтаксисом двох різних мов. Її помірна виразність обмежує відхилення і, на нашу думку, можливо надала їй додаткової популярності, яка грунтується на неправильному уявленні про складність її синтаксичного аналізу.

Хоча під час першого преставлення TL-MCTAG було показано, що за експресивністю вона еквівалентна граматиці з'єднання дерев, складність TL-MCTAG все ще недостатньо вивчена. У статті детально описано дослідження

проблеми розпізнавання TL-MCTAG, яке свідчить, що навіть найбільш обмежені форми TL-MCTAG є НП-повними для розпізнавання. Проте незважаючи на довідну складність проблеми розпізнавання, ми запропонували кілька алгоритмів, які можуть суттєво поліпшити ефективність обробки. По-перше, ми описали алгоритм синтаксичного аналізу, який удосконалює базовий метод синтаксичного аналізу і здійснює обробку за поліноміальний час, коли у вхідній граматиці обмежені як максимальна кількість дерев у наборі дерев, так і максимальна кількість дерев, які можна з'єднати у задане дерево. По-друге, ми запропонували оптимальний, ефективний алгоритм факторизації граматики для отримання високоеквівалентної TL-MCTAG із мінімальною кількістю дерев у дереві.

Переклад В. Коломієць

Erk, K. A Flexible, Corpus-Driven Model of Regular and Inverse Selectional Preferences [Гнучка корпусно-керована модель регулярної і зворотної вірогідної сполучуваності] / Katrin Erk, Sebastian Padó, Ulrike Padó // Computational linguistics. – 2010. – Vol. 36. – No. 4. – Pages 723–763. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli_a_00017#.WIXWVX3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli_a_00017

У статті представлена модель вірогідної сполучуваності на основі векторного простору, яка вираховує показник імовірності для стрижневих слів аргументів. Модель не потребує жодних лексичних ресурсів (таких як WordNet). Її можна навчати як на одному синтаксично анотованому корпусі, так і поєднуючи невеликий за обсягом вихідний корпус із семантичною розміткою та великий узагальнюючий корпус із синтаксичною розміткою. Наша модель здатна передбачити зворотну вірогідну сполучуваність, тобто показники імовірності для предикатів з урахуванням стрижневих слів аргументів.

Для оцінки розробленої моделі було використано одне завдання з обробки природної мови (псевдо-зняття неоднозначностей) та одне когнітивне завдання (прогнозування експертних оцінок достовірності), визначено вплив різних параметрів і здійснено порівняння розробленої моделі з іншими класами моделей. Використання вирішення лексичної неоднозначності та інформації про сематичні ролі, яка міститься у семантично анотованому вихідному корпусі, забезпечило постійні переваги. Відносно параметрів, визначено налаштування, які забезпечують високу продуктивність у різних експериментальних умовах. Проте, основним чинником, який впливає на якість прогнозування, залишається частота. Також визначено більш детальні налаштування параметрів, потрібних для завдань із великою кількістю низькочастотних одиниць .

Переклад М. Погребної

Sygal, Y. Towards Modular Development of Typed Unification Grammars [На шляху до модульної розбудови типізованих уніфікаційних граматик] / Yael Sygal, Shuly Wintner // Computational linguistics. – 2011. – Vol. 37. – No. 1. – Pages 29–74. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli_a_00035#.WIXW-n3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli_a_00035

Створення великомасштабних граматик природних мов є складним завданням. Граматики створюються колективно командами лінгвістів, комп'ютерних лінгвістів і програмістів так само, як створюється великомасштабне програмне забезпечення. Граматики пишуться за допомогою граматичних формалізмів, які нагадують мови програмування дуже високого рівня, і тому є дуже схожими на комп'ютерні програми. Проте розробка граматик все ще знаходиться на початковій стадії розвитку. Дуже мало середовищ розробки граматики підтримують створення складних модульних граматик шляхом розподілу завдань по створенню граматики, комбінування підграматик, нарізної компіляції і автоматичної компоновки, інкапсуляції даних тощо.

Ця праця заклала базовий фундамент для модульної розбудови типізованих уніфікаційних граматик природніх мов. Переважна частина даних у таких формалізмах шифрується за допомогою сигнатури типу, отже в даному дослідженні проблема розв'язується шляхом розподілу сигнатур між різними модулями. Наведено визначення сигнатурного модуля і запропоновано оператори комбінування модулів. Модулі можуть визначати лише частину інформації про компоненти сигнатури і можуть спілкуватися через параметри, так само як виклики функцій у мовах програмування. В основу запропонованих визначень покладено методи і прийоми теорії мов програмування і розробки програмного забезпечення, а також реальні потреби розробників граматик, визначені шляхом ретельного аналізу існуючих граматик. Показано, що наведені визначення відповідають цим потребам, оскільки задовольняють детальний набір побажань. Користь запропонованих визначень продемонстрована шляхом наведення модульної конструкції HPSG-граматики К. Полларда та І. Сага.

Переклад В. Коломієць

Greenhill, J. S. Levenshtein Distances Fail to Identify Language Relationships Accurately [Відстані Левенштейна нездатні точно визначати ступені спорідненості мов] / Simon J. Greenhill // Computational linguistics. – 2011. – Vol. 37. – No. 4. – Pages 689–698. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00073#.WIUU8n3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00073

Відстань Левенштейна – це проста міра відстані, що дорівнює кількості

операцій редагування, необхідних для перетворення одного рядка в інший. Останнім часом цією мірою цікавляться як засобом автоматичної класифікації мов на генеалогічні підгрупи. У статті ефективність використання відстані Левенштейна для класифікації мов протестована шляхом субдискретизації трьох мовніх підгруп з великої бази даних австронезійських мов. Порівняння класифікації, отриманої за допомогою відстані Левенштейна, з класифікацією, отриманою за допомогою порівняльного методу, свідчить про те, що точність класифікації на основі відстані Левенштейна становить 40%. Стандартизація орфографії поліпшує результати продуктивності, але не більше ніж до 65% точності всередині мовних підгруп. Точність класифікації на основі відстані Левенштейна різко зменшується з філогенетичною відстанню, унеможливлюючи розрізнення гомологічності і випадкової схожості віддалено споріднених мов. Така низька продуктивність свідчить про необхідність лінгвістично чутливіших методів для автоматичної класифікації мов.

Переклад В. Коломієць

Schütze, H. Half-Context Language Models [Напівконтекстні моделі мови] / Hinrich Schütze, Michael Walsh // Computational linguistics. – 2011. – Vol. 37. – No. 4. – Pages 843–865. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00078#.WIXXe33sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00078

У статті досліджується вплив різних ступенів деталізації контексту на продуктивність моделі мови. Описано нову модель мови, яка поєднує в собі кластеризацию і часткову контекстуалізацію, новий спосіб представлення контекстів. Основою часткової контекстуалізації є гіпотеза про частковий контекст, згідно з якою найкращого представлення дистрибутивних характеристик слова або біграма можна досягти, аналізуючи окремо його дистрибуцію у правому і лівому контекстах і беручи до уваги лише найголовнішу інформацію про дистрибуцію. Кластеризація виконується за допомогою нового алгоритму кластеризації для мовних моделей на основі класів, який вигідно відрізняється від алгоритму обміну. Показано, що у поєданні з моделлю Кнезера-Нея напівконтекстні моделі досягають вищого показника невизначеності, ніж широко використовувані інтерпольовані моделі на основі n-грамів та традиційні підходи на основі класів. Новий, детальний, контекстнозалежний аналіз виділяє ті контексти, у яких модель досягає високої ефективності, і ті, які краще аналізувати за допомогою існуючих моделей, які не базуються на класах.

Переклад М. Погребної

Cohen, B. S. Empirical Risk Minimization for Probabilistic Grammars: Sample Complexity and Hardness of Learning [Емпірична мінімізація ризику для імовірнісних граматик: кількість прикладів і складність

навчання] / Shay B. Cohen, Noah A. Smith // Computational linguistics. – 2012. – Vol. 38. – No. 3. – Pages 479–526. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00092#.WIXaWn3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00092

Імовірнісні граматики – це генеративні статистичні моделі, корисні для композиційних і послідовних структур. Вони широко застосовуються у комп'ютерній лінгвістиці. У статті описано схожу на структурну мінімізацію ризику концепцію емпіричної мінімізації ризику імовірнісних граматик за допомогою логарифмічного декремента. У цій концепції визначено кількість прикладів, яка потрібна як для навчання з учителем, так і для навчання без учителя. Висуваючи відповідні описам природних мов припущення про вихідний розподіл, можна на основі розподілу визначити потрібну кількість прикладів для імовірнісних граматик. Також наведено прості алгоритми для здійснення емпіричної мінімізації ризику, використовуючи цю концепцію як із залученням учителя, так і без учителя. Показано, що у навчанні без учителя проблема мінімізації емпіричного ризику є НП-складною. Тому для мінімізації емпіричного ризику запропоновано приблизний алгоритм, схожий на максимізацію очікувань.

Переклад В. Коломієць

Kuhlmann, M. Tree-Adjoining Grammars Are Not Closed Under Strong Lexicalization [Граматики з'єднання дерев не закриваються під впливом сильної лексикалізації] / Marco Kuhlmann, Giorgio Satta // Computational linguistics. – 2012. – Vol. 38. – No. 3. – Pp. 617–629. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00090#.WH6G-33sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00090

Лексикалізована граматика з'єднання дерев – це граматика з'єднання дерев, у якій кожне атомарне дерево містить якусь очевидну лексичну одиницю. Такі граматики використовуються для надання лексичних пояснень синтаксичних явищ, у яких початкове дерево визначає домен розташування синтаксичних і семантичних залежностей його лексичних одиниць. У літературі стверджувалося, що для кожної граматики з'єднання дерев можна сконструювати абсолютно еквівалентну лексикалізовану версію. Ми показали, що подібної процедури не існує. Граматики з'єднання дерев не закриваються під впливом сильної лексикалізації.

Переклад В. Коломієць

Tan, M. A Scalable Distributed Syntactic, Semantic, and Lexical Language Model [Широкомасштабна розподілена синтаксична, семантична і лексична модель мови] / Ming Tan, Wenli Zhou, Lei Zheng, Shaojun Wang

// Computational linguistics. – 2012. – Vol. 38. – No. 3. – Pages 631–671. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00107#.WIXa5H3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00107

У статті зроблено спробу побудувати широкомасштабну розподілену складену мовну модель, створену органічним об'єднанням N-грамної моделі, структурної моделі мови та імовірнісного латентно-семантичного аналізу під спрямованою парадигмою випадкових полів Маркова для одночасного пояснення лексичного значення локального слова, синтаксичної структури середньомасштабного речення і семантичного змісту довгого документу. Складену мовну модель навчали шляхом виконання наближеного EM-алгоритму з конвергентним списком N-кращих гіпотез і додаткового EM-алгоритму з метою покращення ефективності передбачення слів на основі корпусів обсягом до мільярда слів. Модель зберігали у суперкомп'ютері. Широкомасштабна розподілена складена мовна модель дає різке зменшення перплексивності N-грамів і досягає значно кращої якості перекладу за метрикою Bleu і "читабельності" перекладів при повторному ранжуванні списку N-кращих гіпотез із сучасної системи машинного перекладу на основі синтаксичного аналізу.

Переклад Т. Павлущенко і М. Погребної

Lembersky, G. Language Models for Machine Translation: Original vs. Translated Texts [Мовні моделі для машинного перекладу: порівняння оригінальних і перекладених текстів] / Gennadi Lembersky, Noam Ordan, Shuly Wintner // Computational linguistics. – 2012. – Vol. 38. – No. 4. – Pages 799–825. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00111#.WIdNqH3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00111

У статті досліджуються відмінності між мовними моделями, створеними на основі оригінальних текстів мовою перекладу, і моделями, створеними на основі текстів, вручну перекладених на мову перекладу. На підтвердження загальновідомих спостережень перекладознавців продемонстровано, що останні є значно кращими індикаторами перекладених речень, ніж перші, а тому краще підходять для набору зразків. Більше того, перекладені тексти дозволяють отримати кращі мовні моделі для статистичного машинного перекладу, ніж оригінальні тексти.

Переклад В. Коломієць

Wedekind, J. LFG Generation by Grammar Specialization [Генерування ЛФГ шляхом спеціалізації граматики] / Jürgen Wedekind, Ronald M. Kaplan // Computational linguistics. – 2012. – Vol. 38. – No. 4. – Pages 867–

915. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00113#.WIXbln3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00113

У статті описано підхід до генерування лексико-функціональної граматики (ЛФГ), яка базується на тому факті, що набір ланцюгів, які ЛФГ пов'язує з певною ациклічною f-структурою, є контекстно-вільною мовою. У статті описано алгоритм створення для довільної ЛФГ і довільної вихідної ациклічної f-структури контекстно-вільної граматики, яка описує саме той набір ланцюгів, які вказана ЛФГ асоціює з цією f-структурою. Конкретні речення потім подаються через стандартний контекстно-вільний генератор, який працює на основі цієї граматики. Контекстно-вільна граматика будується шляхом адаптації контекстно-вільної основи ЛФГ для конкретної f-структури і є компактним представленням усіх результатів генерації, які ЛФГ ставить у відповідність із уведенням. Вказаний підхід розповсюджується на інші граматичні формалізми із очевидними контекстно-вільними основами, такі як PATR, а також на формалізми, які дозволяють видобути контекстно-вільну основу із складніших специфікацій. Він забезпечує загальну математичну концептуальну схему для розуміння і удосконалення функціонування серії алгоритмів генерування на основі блок-схем.

Переклад В. Коломієць

Kuhlmann, M. Mildly Non-Projective Dependency Grammar [Помірно непроективна граматика залежностей] / Marco Kuhlmann // Computational linguistics. – 2013. – Vol. 39. – No. 2. – Pp. 355–387. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00125#.WH6Km33sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00125

Синтаксичні представлення на основі залежностей між словами мають давню традицію у дескриптивній лінгвістиці і активно використовуються у численних прикладних програмах. Проте з формальної точки зору граматика залежностей залишається до деякої міри автономією. Більше того, більшість наявних формалізмів граматики залежностей використовуються лише у проективному аналізі і через це не можуть забезпечити зрозумілого відображення таких явищ як переміщення питального слова або перехресні залежності.

У статті представлено формалізм непроективної граматики залежностей у контексті лінійних контекстно-вільних систем переписування. Характерною особливістю нашого формалізму є тісна відповідність між непроективними деревами залежностей, які допускає граматика, з одного боку і граматичною складністю синтаксичного розбору з другого боку. Ми показуємо, що

синтаксичний аналіз на основі необмеженої граматики є важкоконтрольованим. Через це ми аналізуємо два обмеження непроективності: рівень блокування і високий рівень вкладеності. Разом ці два обмеження визначають клас помірно непроективних граматик залежностей, які можна аналізувати за поліноміальний час. Тестування з використанням п'яти банків дерев залежностей показало, що ці граматики ефективно обробляють емпіричні дані.

Переклад Д. Попової

Crabbé, B. XMG: eXtensible MetaGrammar [XMG: розширювана метаграматика] / Benoît Crabbé, Denys Duchier, Claire Gardent, Joseph Le Roux, Yannick Parmentier // Computational linguistics. – 2013. – Vol. 39. – No. 3 – Pages 591–629. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00144#.WIXcGn3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00144

Стаття знайомить із eXtensible MetaGrammar (XMG), концепцією для опису граматик на основі дерев, таких як Feature-Based Lexicalized Tree-Adjoining Grammars (FB-LTAG) та Interaction Grammars (IG). Стверджується, що XMG притаманні три характеристики, які полегшують як написання граматик, так і швидке моделювання граматик на основі дерев. По-перше, XMG є повністю декларативною. Наприклад, вона допускає декларативне використання діатези, яка суттєво відійшла від процедурних лексичних правил, які часто використовуються для детального опису граматик на основі дерев. По-друге, мова XMG має високі виражальні можливості, оскільки вона підтримує численні лінгвістичні виміри, успадкування і ретельну обробку індентифікаторів. По-третє, XMG може бути розширена, оскільки її обчислювальна структура дозволяє розширення у інші лінгвістичні формалізми. У статті пояснюється, як ця структура природно підтримує розробку трьох лінгвістичних формалізмів, а саме FB-LTAG, IG і багатокомпонентної граматики складання дерев (*англ.* Multi-Component Tree-Adjoining Grammar, *скор.* MC-TAG). Також показано, як вона уможливлює пряму інтеграцію додаткових механізмів, таких як лінгвістичні і формальні принципи. Щоб докладніше проілюструвати декларативність, виражальні можливості і розширюваність XMG, у статті описано методику, яка використовувалась для докладного опису FB-LTAG для французької мови, розширеної композиційною семантикою на основі уніфікації. Це ілюстрація того, як XMG спрощує і моделювання ієрархій фрагментів дерев, необхідне для опису граматик на основі дерев, і синтаксичний/семантичний інтерфейс між семантичними репрезентаціями і синтаксичними деревами. Нарешті, у статті коротко повідомляється про декілька граматик для французької, англійської та німецької мов, які були застосовані на практиці за допомогою XMG і здійснюється порівняння XMG з іншими існуючими концептуальними схемами опису граматик для граматик на основі дерев.

Переклад В. Коломієць

Kuhn, T. A Survey and Classification of Controlled Natural Languages [Аналіз і класифікація контрольованих природних мов] / Tobias Kuhn // Computational linguistics. – 2014. – Vol. 40. – No. 1. – Pages 121–170. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00168#.WIXcsH3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00168

Те, що у статті називається контрольованою природною мовою (*англ.* controlled natural language, *скор.* CNL), відоме під різними назвами. Велика кількість таких мов створена переважно протягом останніх чотирьох десятиліть. Вони використовуються для удосконалення спілкування між людьми, для удосконалення перекладу або для природних і інтуїтивно зрозумілих представлень формальних позначень. Незважаючи на очевидні відмінності, варто об'єднати всі ці мови в одну групу. Для того щоб упорядкувати різні мови, у статті запропонована загальна схема класифікації. Представлено комплексне дослідження існуючих CNL на основі англійської мови, у якому описано 100 мов, створених після 1930 року. Класифікація цих мов свідчить, що вони утворюють єдину розірвану хмару, яка заповнює концептуальний простір між природними мовами, такими як англійська, з одного боку і формальними мовами, такими як пропозиціональна логіка, з іншого боку. Мета статті – розробити спільну термінологію і спільну модель для CNL, сприяти розумінню їх загальної природи, створити вихідний пункт для дослідників у цій галузі і допомогти розробникам у прийнятті проектних рішень.

Переклад В. Коломієць

Chung, T. Sampling Tree Fragments from Forests [Відбір фрагментів дерев із лісів] / Tagyoung Chung, Licheng Fang, Daniel Gildea, Daniel Štefankovič // Computational linguistics. – 2014. – Vol. 40. – No. 1. – Pages 203–229. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00170#.WH6MG33sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00170

У статті досліджується проблема відбору дерев із лісів за умови, що вірогідності для кожного дерева можуть бути функцією довільно великих фрагментів дерева. За цієї умови сучасні проекти по формуванню вибірки для навчання граматик заміщення дерев повинні включати випадки, коли структура дерева (дерево утворене на основі граматики заміщення дерев) не є фіксованою. Розроблено алгоритм Монте Карло з ланцюгами Маркова, який виправляє спотворення, спричинені незбалансованими лісами, і описано експерименти з використанням цього алгоритму для навчання правилам

синхронної контекстно-незалежної граматики для машинного перекладу. У цьому додатку відібрані ліси представляють набір правил граматики Hiero, які узгоджуються із фіксованими вихідними вирювнюваннями на рівні слів. Якість машиного перекладу не відрізняється від стандартних методик, але досягається за допомогою значно менших граматик.

Переклад В. Коломієць

Wedekind, J. On the Universal Generation Problem for Unification Grammars [Про універсальну проблему генерації для уніфікаційних граматик] / Jürgen Wedekind // Computational linguistics. – 2014. – Vol. 40. – No. 3. – Pages 533–538. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00191#.WIXdJ33sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00191

Універсальною проблемою генерації для уніфікаційних граматик є проблема визначення, чи породжує певна граматика якийсь термінальний ланцюг із певною ознаковою структурою. Відомо, що для формалізмів LFG і PATR цю проблему можна вирішити, якщо до уваги беруться лише нециклічні ознакові структури. У статті показано, що для циклічних структур вказана проблема є нерозв'язною. Це стосується навіть граматик, які аналізуються автономно.

Переклад В. Коломієць

Sproat, R. Applications of Lexicographic Semirings to Problems in Speech and Language Processing [Застосування лексикографічних напівкілець до розв'язання проблем обробки природної мови] / Richard Sproat, Mahsa Yarmohammadi, Izhak Shafran, Brian Roark // Computational linguistics. – 2014. – Vol. 40. – No. 4. – Pages 733–761. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00198#.WIXdbX3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00198

У статті досліджуються *лексикографічні півкільця* та їх застосування у вирішенні проблем обробки природної мови. Зокрема, описано два приклади бінарних лексикографічних півкілець, одне з яких включало пару тропічних вагів, а друге – тропічну вагу разом із півкільцем у новому рядку, названим нами *категоріальним півкільцем*. Перше з них використовується для отримання точного коду відтермінованих моделей з іпсилон-переходами. Таке *лексикографічне півкільце мовної моделі* дозволяє оптимізувати в автономному режимі точні моделі, представлені як великі зважені кінцеві перетворювачі на відміну від розмитих (онлайн) моделей зі збоями при переходах. Представлені емпіричні результати свідчать, що навіть у простих переходах, де можуть використовуватися невдалі переходи, використання потужнішого лексикографічного півкільця є доцільним з точки зору часу

перетину. Друге з цих лексикографічних півкілець застосовується для вирішення проблеми вилучення з решітки послідовностей слів з частиномовною розміткою лише найкращої частиномовної розмітки для кожної послідовності слів. Здійснюється це шляхом додавання міток в якості категоріальної ваги у другому компоненті ⟨Тропічного, Категоріального⟩ лексикографічного півкільця, детермінізації користувача отриманою решіткою слів у тому півкільці, а потім використання тегів в якості вихідних міток перетворювача словесних решіток. Цей метод порівнюється з конкуруючим методом Поуві та інших [Povey et al., 2012].

Переклад А. Синящик

Kuhlmann M. Lexicalization and Generative Power in CCG [Лексикалізація та генеративна ефективність в ККГ] / Marco Kuhlmann, Alexander Koller, Giorgio Satta // Computational linguistics. – 2015. – Vol. 41. – No. 2. – Pages 215–247. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00219 – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00219

Слабка еквівалентність між комбінаторною категоріальною граматикою (ККГ) і граматикою з'єднання дерев (ГЗД) є основним результатом у літературі про граматичні формалізми з помірною залежністю від контексту. Проте, категорійний формалізм, для якого була встановлена ця еквівалентність, суттєво відрізняється від сучасних версій ККГ. Зокрема, він дозволяє обмежувати комбінаторні правила на основі граматики, у той час як сучасні ККГ передбачають універсальний набір правил, що виключає будь-які міжмовні варіації у лексиконі. У статті досліджується формальна значимість цієї розбіжності. Основний висновок полягає в тому, що лексикалізовані версії класичного формалізму ККГ явно менш ефективні, ніж ГЗД.

Переклад А. Шульги

Tanaka-Ishii, K. Computational Constancy Measures of Texts—Yule's K and Rényi's Entropy [Міри обчислювальної стійкості текстів – показник Юла (К) та ентропія Реньї] / Kumiko Tanaka-Ishii, Shunsuke Aihara // Computational linguistics. – 2015. – Vol. 41. – No. 3. – Pages 481–502. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00228 – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00228

У статті описно математичну й емпіричну перевірку мір обчислюваної стійкості текстів природною мовою. Міра стійкості характеризує певний текст, приписуючи інваріантну величину будь-якому обсягу, що перевищує певну кількість. Вивчення таких мір проводиться вже 70 років, починаючи з

показника Юла (K), який спочатку призначався для встановлення авторства. У статті розглянуто різні міри, запропоновані після Юла, і перевірено зроблені дотепер висновки, тобто зроблено огляд досліджень мір стійкості. Також, у статті пояснено, чому K є по суті еквівалентом апроксимації ентропії Реньї другого порядку, тобто визначено значимість цього показника в лінгвістиці. Крім того, емпірично досліджено кандидатів у міри стійкості в цьому новому, ширшому контексті. Наближена ентропія вищого порядку демонструє стабільне зближення між різними мовами та видами текстів. Втім, також з'ясовано, що, всупереч очікуванням Юла, вона не може встановлювати авторство. Насамкінець, K застосовано до двох невідомих рукописів – манускрипту Войніча та ронго-ронго і продемонстровано, що результати підтверджують попередні гіпотези щодо цих рукописів.

Переклад М. Дубка

Karimi, S. Evaluation Methods for Statistically Dependent Text [Методи оцінювання статистично залежних текстів] / Sarvnaz Karimi, Jie Yin, Jiri Baum // Computational linguistics. – 2015. – Vol. 41. – No. 3. – Pages 539–548. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00230 – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00230

Протягом останніх років опубліковано багато досліджень даних, зібраних у соціальних мережах, зокрема в мікроблогах на кшталт Twitter. Проте, лише в кількох з них розглядалися методи оцінювання, які враховують статистично залежний характер таких даних, що порушує теоретичні умови використання перехресної перевірки. Незважаючи на питання, які піднімались у минулому щодо застовання перехресної перевірки до даних зі схожими характеристиками, наприклад, динамічних рядів, деякі з цих досліджень оцінюють свої результати за допомогою стандартної k-кратної перехресної перевірки. Завдяки експериментам на основі даних Twitter, зібраних протягом двохрічного періоду, який включає катастрофічні події, було показано, що через ігнорування статистичної залежності опублікованих у соціальних мережах текстових повідомлень стандартна перехресна перевірка може призвести до помилкових висновків у завданні з машинного навчання. Проаналізовано альтернативні методи оцінювання, які напряму використовують статистичну залежність у тексті. Отримані результати також викликають питання до будь-яких інших даних, до яких можна застосувати подібні умови.

Переклад М. Дубка

Paperno, D. When the Whole Is Less Than the Sum of Its Parts: How Composition Affects PMI Values in Distributional Semantic Vectors [Коли ціле є меншим, ніж сума його частин: як структура впливає на значення ПВІ в семантичних контекстних векторах] / Denis Paperno, Marco Baroni

// Computational linguistics. – 2016. – Vol. 42. – No. 2. – Pages 345–350. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00250 – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00250

Дистрибутивні семантичні моделі, які виводять представлення слів на основі векторів із прикладів вживання слів у корпусах, мають багато корисних застосувань (Turney and Pantel, 2010). Останнім часом зажили популярності композиційні дистрибутивні моделі, які виводять вектори для фраз з представлень слів, з яких вони складаються (Mitchell and Lapata, 2010). Значення контекстних векторів часто є оцінкою в балах поточкової взаємної інформації (ПВІ), отриманою з грубих показників одночасної появи слів. У статті проаналізовано зв'язок між координатами ПВІ вектора фрази і його компонентів, щоб з'ясувати, які операції повинна виконувати належна композиційна модель. Математично доведено, що величина різниці між координатою ПВІ вектора фрази і сумою показників ПВІ на відповідних координатах частин фрази інтерпретується незалежно, а саме шляхом квантифікації впливу контексту, пов'язаного з відповідною координатою на внутрішній когезії фрази, що також вимірюється за допомогою ПВІ. Потім цей показник досліджено емпірично, шляхом аналізу сполучень прикметник-іменник.

Переклад М. Дубка

Bos J. Expressive Power of Abstract Meaning Representations [Експресивна сила представлень абстрактних значень] / Johan Bos // Computational linguistics. – 2016. – Vol. 42. – No. 3. – Pages 527–535. – Режим доступу до анотації: https://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00257 – Режим доступу до повнотекстової статті: https://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00257

Синтаксис представлень абстрактних значень (ПАЗ) можна визначити рекурсивно, і можна визначити систематичний переклад на логіку першого порядку (ЛПП), зокрема правильне опрацювання заперечення. ПАЗ без повторюваних змінних знаходяться у розв'язному фрагменті ЛПП з двома змінними. Поточне визначення ПАЗ має обмежену експресивну силу для універсального кількісного визначення (до одного універсального квантифікатора на речення). Просте розширення синтаксису ПАЗ і переклад на логіку першого порядку уможливлює представлення проективності та області дії.

Переклад А. Шульги

Ionescu, R. T. String Kernels for Native Language Identification: Insights from Behind the Curtains [Рядкові ядра для визначення мови автора: із досвіду розробки й використання] / Radu Tudor Ionescu, Marius Popescu,

Aoife Cahill // Computational linguistics. – 2016. – Vol. 42. – No. 3. – Pages 491–525. – Режим доступу до анотації: https://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00256 – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00256

Найбільш поширеним підходом до задач класифікації в інтелектуальному аналізі текстових даних є використання таких категорій, як слова, частиномовні мітки, основи або інші лінгвістичні категорії вищого рівня. Нещодавно для завдання визначення рідної мови автора (ВРМ) було запропоновано підхід, який використовує в якості категорій виключно символьні p-грами. Шляхом об'єднання декількох стрічкових ядер за допомогою багатоядерних обчислень цей підхід дозволив отримати результати на рівні останніх досягнень галузі. Незважаючи на продуктивність підходу на основі стрічкових ядер, є кілька питань про цей метод, які чекають відповіді. По-перше, не зрозуміло, чому такий простий підхід може конкурувати з набагато складнішими підходами, які враховують слова, леми, синтаксичну інформацію чи навіть семантику. По-друге, хоча підхід створювався як незалежний від мови, всі експерименти досі проводились на англійській мові. Ця праця є детальним дослідженням, яке повинне дати системне уявлення про підхід на основі стрічкових ядер та відповісти на вищезгадані відкриті питання.

Щоб порівняти підхід на основі стрічкових ядер з іншими сучасними методами, проведено велику кількість експериментів із визначення рідної мови автора. Емпіричні результати, отримані в усіх експериментах, проведених у цьому дослідженні, вказують на те, що запропонований підхід відповідає останнім досягненням у ВРМ, досягаючи точності, що на 1,7% перевершує найкращі результати у змаганні систем ВРМ 2013 року. Крім того, результати, отримані на базі як арабського, так і норвезького корпусів, свідчать, що запропонований підхід є незалежним від мови. При визначенні носіїв арабської мови стрічкові ядра показали точність, що перевершує найкращі відомі на сьогодні показники на 17%. Результати стрічкових ядер при визначенні носіїв норвезької мови також значно перевершують найсучасніший підхід. Крім того, в експерименті з кількома корпусами запропонований підхід перевершив результати найсучаснішої системи на 32,3%, продемонструвавши, що він може також бути незалежним від тематики.

Щоб отримати додаткові уявлення про підхід на основі стрічкових ядер, в цій статті аналізуються категорії, виділені класифікатором як більш розрізнювальні. Аналіз також містить інформацію про наслідки переносу локалізованої мови, оскільки критерії, які використовуються запропонованою моделлю, є p-грамами різної довжини. Виділені моделлю категорії, як правило, включають основи, службові слова, а також префікси та суфікси, які можуть бути узагальнені на відміну від суто словесних ознак. Завдяки аналізу розрізнювальних ознак, стаття дає уявлення про два види ефекту

мовного перенесення, а саме вибір слів (лексичне перенесення) та морфологічні розбіжності. Мета цього дослідження полягає в тому, щоб дати повне уявлення про підхід на основі стрічкових ядер, а також пояснити, чому цей підхід працює так добре.

Переклад М. Дубка

Cohen S. Parsing Linear Context-Free Rewriting Systems with Fast Matrix Multiplication [Синтаксичний аналіз лінійних контекстно-незалежних систем переписування за допомогою швидкого множення матриць] / Shay B. Cohen, Daniel Gildea // Computational linguistics. – 2016. – Vol. 42. – No. 3. – Pages 421–455. – Режим доступу до анотації: https://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00254 – Режим доступу до повнотекстової статті: https://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00254

У статті описано алгоритм розпізнавання підмножини бінарних лінійних контекстно-незалежних систем переписування (ЛКНСП) з часом виконання O (nωd), де M (m) = O (mω) - це тривалість перемноження матриць m × m, a d –“контактний ранг” ЛКНСП, тобто максимальна кількість комбінаторних та некомбінаторних точок, які з’являються у правилах граматики. Показано також, що цей алгоритм можна також використовувати як підпрограму для отримання алгоритму розпізнавання загальної бінарної ЛКНСП з тривалістю виконання O (nωd + 1). Нині найбільш відомий ω є меншим, ніж 2,38. Отриманий результат є ще одним підтвердженням найбільш відомого результату автоматичного синтаксичного аналізу слабко контекстно-залежних формалізмів, таких як комбінаторні категоріальні граматики, вершинні граматики, лінійні індексовані граматики та граматики об’єднання дерев, аналіз яких триває O (n4.76). Він також свідчить, трансдукційні інвертовані граматики можна проаналізувати за O(n5.76). Крім того, бінарна ЛКНСП включає в себе багато інших формалізмів і типів граматик, для деяких з яких також вдосконалено асимптотичну складність автоматичного синтаксичного аналізу.

Переклад А. Шульги

Kuhlmann M. Towards a Catalogue of Linguistic Graph Banks [На шляху до каталогу банків лінгвістичних графів] / Marco Kuhlmann, Stephan Oepen // Computational linguistics. – 2016. – Vol. 42. – No. 4. – Pages 819–827. – Режим доступу до анотації: https://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00268 – Режим доступу до повнотекстової статті: https://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00268

Графи, які перевищують формальну складність кореневого дерева, стають все більш актуальними для багатьох прикладних лінгвістичних досліджень. Хоча ця проблема формально добре досліджена в теорії графів, існує значна

варіативність у типах лінгвістичних графів та інтерпретаціях різних структурних властивостей. Для забезпечення стандартної термінології та прозорих статистичних даних у різних наборах графів у опрацюванні природної мови, запропоновано створити загальнодоступний ресурс спільноти з відкритою еталонною реалізацією для отримання загальних даних.

Переклад А. Шульги

Sajjad H. Statistical Models for Unsupervised, Semi-Supervised, and Supervised Transliteration Mining [Статистичні моделі для видобування пар із транслітерованими словами без залучення, з частковим і з повним залученням учителя] / Hassan Sajjad, Helmut Schmid, Alexander Fraser, Hinrich Schütze // Computational linguistics. – 2017. – Vol. 43. – No. 2. – Pages 349–375. – Режим доступу до анотації: https://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00286 – Режим доступу до повнотекстової статті: https://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00286

У статті представлено генеративну модель, яка ефективно та послідовно видобуває пари з транслітерованими словами за трьох різних умов: без залучення, з частковим і з повним залученням учителя. Ця модель інтерполює дві підмоделі: одну для генерування пар із транслітерованими словами, а другу для генерування пар без транслітерованих слів (тобто шуму). Модель навчається на зашумлених немаркованих даних за допомогою ЕМ-алгоритму. Під час навчання підмодель транслітерації вчиться генерувати пари з транслітерованими словами, а фіксована модель пар без транслітерованих слів генерує зашумлені пари. Після навчання знімається омонімія немаркованих даних на основі апостеріорної ймовірності двох підмоделей. Систему для видобування пар із транслітерованими словами оцінено на даних із змагань систем видобування пар із транслітерованими словами та паралельних корпусів. Для трьох з чотирьох мовних пар описана система перевершила всі системи з частковим та повним залученням учителя, які приймали участь у змаганнях NEWS 2010. При використанні пар слів, видобутих із паралельних корпусів з менш ніж 2% пар із транслітерованими словами, запропонована система досягає F-показника 86,7% з точністю 77,9% і повнотою 97,8%.

Переклад А. Шульги

Nguyen, D. A Kernel Independence Test for Geographical Language Variation [Тест незалежності ядра для географічного варіювання мов] / Dong Nguyen, Jacob Eisenstein. – 2017. – Vol. 43. – No. 3. – Pages 567–592. – Режим доступу до анотації: https://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00293 – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00293

Кількісна оцінка ступеня просторової залежності лінгвістичних змінних є ключовим завданням для аналізу діалектного різноманіття. Проте існуючі підходи мають суттєві недоліки. По-перше, вони базуються на параметричних моделях залежності, що обмежує їхню ефективність у випадках, коли порушуються основні параметричні припущення. По-друге, їх не можна застосувати до всіх видів лінгвістичних даних. Одні підходи застосовуються лише до частот, інші – до булевих вказівників наявності мовної змінної. У статті представлено новий метод вимірювання географічного варіювання мови, який вирішує обидві ці проблеми. Запропонований підхід ґрунтується на представленнях Відтворюваного Ядра Гілбертового Простору (ВЯГП) для непараметричних статистичних даних і має форму тестової статистики, яка обчислюється з пар окремих геотегованих спостережень без агрегації до заданих географічних контейнерів. Здійснено порівняння цього тесту з попередньою роботою із використанням синтетичних даних, а також неоднорідного набору автентичних сукупностей даних: корпусу нідерландських твітів, голландського синтаксичного атласу та набору листів, адресованих редакторам північноамериканських газет. Продемонстровано, що запропонований тест підтверджує стійкі висновки в широкому діапазоні сценаріїв та типів даних.

Переклад М. Дубка

Constant, M. Multiword Expression Processing: A Survey [Опрацювання багатослівних виразів: огляд] / Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, Amalia Todirascu // Computational linguistics. – 2017. – Vol. 43. – No. 4. – Pages 837–892. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00302 – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00302

Багатослівні вирази (БВ) – це клас мовних форм, які збігаються із звичайними словами, і які є як унікальними, так і поширеними в різних мовах. Для врахування БВ потрібно переосмислити структуру опрацювання природної мови, яка залежить від чіткого розмежування слів і фраз. Питання опрацювання БВ має вирішальне значення для програм опрацювання природної мови, де воно викликає низку проблем. Цей огляд, обумовлений появою рішень за відсутності керівних принципів, має на меті не лише цілеспрямований аналіз опрацювання БВ, а й уточнення характеру взаємодії між опрацюванням БВ та цільовими програмами. Запропоновано концептуальну основу, в рамках якої можна розглядати складні проблеми та результати досліджень. Вона забезпечує спільне розуміння того, що мається на увазі під «опрацюванням БВ», відокремлюючи підзадачі знаходження та ідентифікації БВ. Вона також висвітлює взаємодію між опрацюванням БВ і

двома варіантами використання: автоматичним синтаксичним аналізом і машинним перекладом. Багато підходів у літературі можна диференціювати залежно від того, як вимірюється час опрацювання БВ відносно варіантів використання. Проаналізовано, яким чином такі механізми управління впливають на можливості систем, які опрацьовують БВ. Для кожної з двох підзадач опрацювання БВ і для кожного з двох варіантів використання зроблено висновок про невирішені питання та перспективи дослідження.

Переклад М. Дубка

Створення прикладних систем

Автоматичне реферування

Radev, R. D. Introduction to the Special Issue on Summarization [Вступ до спеціального випуску, присвяченого реферуванню] / Dragomir R. Radev, Eduard Hovy, Kathleen McKeown // Computational linguistics. – 2002. – Vol. 28. – No. 4. – Pages 399–408. – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120102762671927

У статті подається короткий огляд сучасного стану мистецтва реферування, який описує загальні напрями досліджень, зокрема однодокументне реферування на основі екстракції, перші спроби однодокументного реферування шляхом укладання анотацій і різноманітні підходи до багатодокументного реферування.

В. Коломієць

Teufel, S. Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status [Реферування наукових статей: експерименти з релевантністю і риторичним статусом] / Simone Teufel, Marc Moens // Computational linguistics. – 2002. – Vol. 28. – No. 4. – Pages 409–445. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120102762671936#.VREqUdyhGCA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120102762671936

У статті запропоновано метод реферування наукових статей на основі риторичного статусу речень у статті. Матеріал для рефератів відбирається таким чином, щоб реферати висвітлювали наукову новизну вихідної статті і визначали її значимість у контексті попередніх досліджень.

Для рефератів даного типу створено золотий стандарт, який складається з великого корпусу доповідей на конференціях з комп'ютерної лінгвістики, кожне речення яких вручну марковане висновками про його риторичний статус та релевантність. У статті описано декілька експериментів для визначення узгодженості висновків експертів стосовно цих маркувань.

У статті також представлено алгоритм, який на основі анотованого тренувального матеріалу відбирає контент з нових статей і класифікує його згідно заданого набору з семи риторичних категорій. Вихідні дані цієї системи пошуку і класифікації можуть розглядатися як самостійний реферат у вигляді єдиного документу або ж можуть використовуватися для генерації задачноорієнтованих, модифікованих відповідно вимог замовника рефератів для загального ознайомлення з науковою галуззю.

Переклад І. Снєгурова, М. Погребної

Zechner, K. Automatic Summarization of Open-Domain Multiparty Dialogues in Diverse Genres [Автоматичне реферування багатосторонніх діалогів різних жанрів із відкритою тематикою] / Klaus Zechner // Computational linguistics. – 2002. – Vol. 28. – No. 4. – Pages 447–485. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120102762671945#.VRHomdyhGCA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120102762671945

Автоматичне реферування усних діалогів із відкритою тематикою є відносно новим напрямом досліджень. Стаття знайомить із цим завданням та труднощами його виконання, а також містить обґрунтування та опис системи автоматичної екстракції конспектів протоколів багатосторонніх діалогів чотирьох різних жанрів без жодних тематичних обмежень.

Розглядаються наступні питання, які є невід'ємною складовою реферування усних діалогів, але зазвичай можуть бути проігноровані при реферуванні письмового тексту, наприклад повідомлень інформаційних агентств: (1) виявлення та видалення мовленнєвих збоїв; (2) знаходження та позначення меж речень; і (3) виявлення та об'єднання діалогічних єдностей (пар питання-відповідь).

Оцінка системи здійснюється за допомогою корпусу з 23 уривків діалогів середньою тривалістю близько 10 хвилин, що складається з 80 тематичних сегментів і близько 47 000 слів. Маркування відповідних сегментів тексту здійснили вручну шість анотаторів. Глобальна оцінка свідчить, що для двох жанрів більш розмовного стилю наша система реферування з опорою на притаманні діалогам компоненти значно перевершує два вихідні показники: (1) алгоритм розрахування мінімально допустимої відповідності за допомогою міри важливості термінів TF*IDF, та (2) вихідний показник LEAD, який відбирає перші n слів із тексту.

Переклад Т. Павлущенко

Silber, G. H. Efficiently Computed Lexical Chains as an Intermediate Representation for Automatic Text Summarization [Ефективно обчислені лексичні ланцюжки як проміжний етап автоматичного реферування тексту] / H. Gregory Silber, Kathleen F. McCoy // Computational linguistics. – 2002. – Vol. 28. – No. 4. – Pages 487–496. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120102762671954#.VRHqBNyhGCA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120102762671954

Не дивлячись не те, що автоматичне реферування тексту є галуззю, якій у сучасних дослідженнях приділяється багато уваги, питання про його ефективність піднімається рідко. Коли розглядається обсяг і кількість

документів, доступних в Інтернеті та з інших джерел, стає очевидною потреба у високоефективному інструменті для створення прийнятних рефератів. У статті описано лінійний алгоритм для обчислення лексичних ланцюжків. На проміжному етапі автоматичного реферування тексту алгоритм розраховує, які лексичні ланцюжки є ймовірним кандидатом. Також представлено та реалізовано метод оцінювання лексичних ланцюжків на проміжному етапі процесу реферування. Така оцінка була досі неможливою через складність обчислень лексичних ланцюжків попередніми алгоритмами.

Переклад Т. Павлущенко

Saggion, H. Generating Indicative-Informative Summaries with SumUM [Створення показово-інформативних оглядів за допомогою системи SumUM] / Horacio Saggion, Guy Lapalme // Computational linguistics. – 2002. – Vol. 28. – No. 4. – Pages 497–526. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120102762671963#.VRHrI9yhGCA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120102762671963

В статті описується й оцінюється система реферування текстів SumUM, яка використовує непідготовлений технічний текст в якості вхідних даних і створює індикативний інформативний реферат. У індикативній частині реферату визначаються теми документу, а інформативна частина містить розширену інформацію про ті теми, які цікавлять читача. Система SumUM обгрунтовує теми, описує елементи предметної області і визначає поняття. Вона є першим кроком до дослідження питань динамічного реферування, яке здійснюється шляхом поверхневого синтаксичного і семантичного аналізу, ідентифікації понять та переписування тексту. Запропонований метод було розроблено за допомогою аналізу корпусу анотацій, створених спеціалістами з написання анотацій. Скориставшись судженнями експертів, ми оцінили індикативність, інформативність та прийнятність текстів автоматично створених рефератів. Отримані результати свідчать про хорошу ефективність системи в порівнянні з іншими технологіями реферування.

Переклад О. Мартинюк

Jing, H. Using Hidden Markov Modeling to Decompose Human-Written Summaries [Використання прихованої марківської моделі для розбиття речень рефератів, написаних людиною] / Hongyan Jing // Computational linguistics. – 2002. – Vol. 28. – No. 4. – Pages 527–543. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120102762671972#.VRHr4tyhGCA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120102762671972

Люди, які професійно займаються реферуванням, часто використовують

вихідні документи для генерування рефератів. Мета розбиття речень реферату полягає в тому, щоб встановити, чи було використано первинний текст при побудові речення реферату та визначити використані словосполучення. Точніше, програма, що виконує розбиття речень, має відповісти на три питання стосовно певного речення реферату: 1) Чи використано текст вихідного документу при побудові цього речення реферату? 2) Якщо так, то які словосполучення у складі цього речення запозичені з вихідного документу? і 3) В якій частині вихідного документу вжито ці словосполучення? Вирішення проблеми розбиття речень сприятиме появі кращих способів генерування рефератів. Також, завдяки розбиттю речень можна створити великі корпуси для тренування і тестування систем реферування на основі екстракції. Для розбиття речень ми використовуємо приховану марківську модель. Оцінка запропонованого алгоритму свідчить про його ефективність.

Переклад І. Снєгурова

Barzilay, R. Sentence Fusion for Multidocument News Summarization [Злиття речень у процесі реферування декількох новинних повідомлень] / Regina Barzilay, Kathleen R. McKeown // Computational linguistics. – 2005. – Vol. 31. – No. 3. – Pages 297–328. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120105774321091#.VRHsxNyhGCA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120105774321091

Система, яка здатна створювати інформативні реферати, висвітлюючи інформацію, яка повторюється у великій кількості документів у Всесвітній Мережі, допоможе користувачам Всесвітньої Мережі швидко знаходити потрібну їм інформацію. У цій статті ми представляємо новітній метод генерування “текст-до-тексту” для синтезу загальної інформації, спільної для ряду документів. Злиття речень передбачає висхідне часткове паралельне вирівнювання для визначення словосполучень, які передають одну й ту саму інформацію, та статистичне генерування для об’єднання повторюваних словосполучень у речення. Завдяки методу злиття речень у галузі реферування здійснено перехід від використання виключно методів екстракції до генерування анотацій, що містять речення, яких немає в жодному з вихідних документів. Метод також дозволяє синтезувати інформацію з різних джерел.

Переклад І. Снєгурова

Daumé, H. III. Induction of Word and Phrase Alignments for Automatic Document Summarization [Використання вирівнювання слів і речень у автоматичному реферуванні документів] / Hal Daumé III, Daniel Marcu // Computational linguistics. – 2005. – Vol. 31. – No. 4. – Pages 505–530. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120105775299140#.VRH

uRdyhGCA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120105775299140

У сучасних дослідженнях автоматичного реферування окремого документа домінують два ефективні, але прості методи: реферування на основі екстракції речень і генерація заголовка на основі моделі «мішка слів». Хоча ці моделі дозволяють успішно вирішувати деякі завдання, жодна з них не здатна адекватно відтворити великий набір лінгвістичних засобів, які використовують при реферуванні люди. Однією з можливих причин широкого використання цих моделей є наявність ефективних методів екстракції підходящої інформації для їх тренування із існуючих корпусів документів/анотацій та документів/заголовків. Ми уважаємо, що подальший прогрес в автоматичному реферуванні буде пов'язаний як із створенням складніших, лінгвістично налаштованих моделей, так і з ефективнішим використанням корпусів документів/анотацій. Для одночасного досягнення обох цілей ми розробили методи автоматичного створення пар слів та фраз із документів та їх анотацій, написаних людиною. Ці пари виявляють відповідності, які існують між такими парами документів і анотацій, і створюють потенційно багату базу даних, яку можна використовувати для тренування складних алгоритмів реферування. У статті описано експерименти, які ми провели, щоб проаналізувати здатність людей робити таке вирівнювання. На основі результатів здійсненого аналізу ми описуємо експерименти для створення системи автоматичного вирівнювання. Наша модель вирівнювання базується на розширенні класичної прихованої моделі Маркова і вчиться створювати вирівнювання без учителя. Ми детально описуємо нашу модель та повідомляємо результати експериментів, які свідчать, що наша модель здатна навчитися надійно ідентифікувати вирівнювання на рівні слова та фрази у корпусі пар «документ, анотація».

Переклад Д. Попової

Kazantseva, A. Summarizing Short Stories [Реферування новел] / Anna Kazantseva, Stan Szpakowicz // Computational linguistics. – 2010. – Vol. 36. – No. 1. – Pages 71–109. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2010.36.1.36102#.VRHvttyhGCA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2010.36.1.36102

У статті представлено метод автоматичної генерації рефератів-екстрактів новел. Реферування здійснюється з конкретною метою: допомогти читачеві вирішити, чи хоче він прочитати всю новелу. З цією метою реферати дають читачеві необхідне уявлення про час і місце дії, не розкриваючи сюжету новели. У системі використовуються різні поверхневі показники предикативних одиниць у новелі, найважливішими з яких є ті, що пов'язані з аспектуальними характеристиками предикативних структур і з головними дійовими особами у новелі. Реферати були оцінені п'ятнадцятьма експертами

за допомогою низки зовнішніх та внутрішніх показників. Результати оцінювання дають підстави уважати, що отримані реферати відповідають поставленій меті.

Переклад Д. Попової

Clarke, J. Discourse Constraints for Document Compression [Дискурсна модель компресії тексту] / James Clarke, Mirella Lapata // Computational linguistics. – 2010. – Vol. 36. – No. 3. – Pages 411–441. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli_a_00004#.WITBzn3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli_a_00004

Компресія речень відкриває перспективи для багатьох прикладних програм: від автоматичного реферування до генерації підзаголовків. Вона звичайно виконується для ізольованих речень без урахування контексту, незважаючи на те, що більшість прикладних програм обробляють увесь текст. У статті представлено дискурсну модель, яка може створювати зв'язні і інформативні анотації текстів. Модель спирається на теорії локальної когерентності і формулюється в рамках цілочислового лінійного програмування. Експериментальні результати свідчать, що вона значно перевершує сучасний підхід, який не бере дискурс до уваги.

Переклад В. Коломієць

Conroy, J. M. Nouveau-ROUGE: A Novelty Metric for Update Summarization [Nouveau-ROUGE: нова метрика для генерації дайджесту оновлень] / John M. Conroy, Judith D. Schlesinger, Dianne P. O'Leary // Computational linguistics. – 2011. – Vol. 37. – No. 1. – Pages 1–8. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli_a_00033#.VRHwmdyhGCA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli_a_00033

Якщо читач уже переглянув попередні документи чи реферати, дайджест оновлень повинен містити стислий виклад нової інформації на тему, яка обговорюється протягом тривалого часу. У 2007 та 2008 роках щорічні змагання систем автоматичного реферування передбачали генерацію дайджесту оновлень. Оцінка за допомогою критерія ROUGE показала, що декілька систем-учасників генерували дайджести оновлень, які неможливо відрізнити від дайджестів, створених вручну. Проте жодна автоматична система не змогла зрівнятися з людиною у ручних оцінках, таких як пірамідальний показник та коефіцієнт загальної сили відгуку.

Ми представляємо метрику Nouveau-ROUGE, яка краще співвідноситься з показниками ручної оцінки і може бути використана для визначення як пірамідального показника, так і коефіцієнта загальної сили відгуку для

дайджестів оновлень. Nouveau-ROUGE може стати дешевшою заміною ручних оцінок при порівнянні існуючих систем і розробці нових.

Переклад В. Туз, М. Погребної

Louis, A. Automatically Assessing Machine Summary Content Without a Gold Standard [Автоматична оцінка змісту автоматично сформованих рефератів без золотого стандарту] / Annie Louis, Ani Nenkova // Computational linguistics. – 2013. – Vol. 39. – No. 2. – Pages 267–300. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00123#.WH3wp33sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00123

Найпопулярніші методи оцінки змісту рефератів використовують процедуру порівняння реферату з золотим стандартом (рефератами, створеними експертами), який традиційно називають еталонними рефератами. Така модель оцінювання не може бути застосована при відсутності еталонних рефератів і дає менш точні результати при наявності лише одного еталонного реферату. У статті запропоновано три нових методи оцінки. Два з них не використовують моделей і не потребують золотого стандарту для оцінювання. Третій метод удосконалює стандартні автоматичні оцінки шляхом додавання до набору наявних еталонних рефератів відібрані автоматично сформовані реферати.

У статті показано, що квантифікація схожості вихідного тексту і його реферату за допомогою правильно підібраних оцінок дозволяє отримати оцінку реферата в балах, яка точно відтворює експертну оцінку. Також досліджено шляхи підвищення якості оцінювання при наявності лише одного створеного експертом зразкового реферата, який використовується як золотий стандарт. Описано псевдомоделі, які є автоматично створеними рефератами, що отримали високі оцінки за зміст при автоматичному оцінюванні. Комбінування псевдомоделей із єдиним створеним експертом зразком для створення золотого стандарту дозволяє підвищити кореляцію з експертними оцінками у порівнянні з використанням лише однієї наявної моделі. Нарешті, досліджено пригодність ще однієї оцінки – схожості між автоматично створеним рефератом і фондом усіх інших автоматично створених рефератів на однакову тематику. Такий метод порівняння із консенсусом систем дає вражаюче точні оцінки автоматичних рефератів, досягаючи кореляції з експертними оцінками понад 0,9.

Переклад В. Коломієць

Діалогові системи

Radev, R. D. Introduction to the Special Issue on Summarization [Вступ до спеціального випуску, присвяченого реферуванню] / Dragomir R. Radev, Eduard Hovy, Kathleen McKeown // Computational linguistics. – 2002. – Vol. 28. – No. 4. – Pages 399–408. – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120102762671927

У статті подається короткий огляд сучасного стану мистецтва реферування, який описує загальні напрями досліджень, зокрема однодокументне реферування на основі екстракції, перші спроби однодокументного реферування шляхом укладання анотацій і різноманітні підходи до багатодокументного реферування.

В. Коломієць

Teufel, S. Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status [Реферування наукових статей: експерименти з релевантністю і риторичним статусом] / Simone Teufel, Marc Moens // Computational linguistics. – 2002. – Vol. 28. – No. 4. – Pages 409–445. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120102762671936#.VREqUdyhGCA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120102762671936

У статті запропоновано метод реферування наукових статей на основі риторичного статусу речень у статті. Матеріал для рефератів відбирається таким чином, щоб реферати висвітлювали наукову новизну вихідної статті і визначали її значимість у контексті попередніх досліджень.

Для рефератів даного типу створено золотий стандарт, який складається з великого корпусу доповідей на конференціях з комп'ютерної лінгвістики, кожне речення яких вручну марковане висновками про його риторичний статус та релевантність. У статті описано декілька експериментів для визначення узгодженості висновків експертів стосовно цих маркувань.

У статті також представлено алгоритм, який на основі анотованого тренувального матеріалу відбирає контент з нових статей і класифікує його згідно заданого набору з семи риторичних категорій. Вихідні дані цієї системи пошуку і класифікації можуть розглядатися як самостійний реферат у вигляді єдиного документу або ж можуть використовуватися для генерації задачноорієнтованих, модифікованих відповідно вимог замовника рефератів для загального ознайомлення з науковою галуззю.

Переклад І. Снєгурова, М. Погребної

Zechner, K. Automatic Summarization of Open-Domain Multiparty Dialogues in Diverse Genres [Автоматичне реферування багатосторонніх діалогів різних жанрів із відкритою тематикою] / Klaus Zechner // Computational linguistics. – 2002. – Vol. 28. – No. 4. – Pages 447–485. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120102762671945#.VRHomdyhGCA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120102762671945

Автоматичне реферування усних діалогів із відкритою тематикою є відносно новим напрямом досліджень. Стаття знайомить із цим завданням та труднощами його виконання, а також містить обґрунтування та опис системи автоматичної екстракції конспектів протоколів багатосторонніх діалогів чотирьох різних жанрів без жодних тематичних обмежень.

Розглядаються наступні питання, які є невід'ємною складовою реферування усних діалогів, але зазвичай можуть бути проігноровані при реферуванні письмового тексту, наприклад повідомлень інформаційних агентств: (1) виявлення та видалення мовленнєвих збоїв; (2) знаходження та позначення меж речень; і (3) виявлення та об'єднання діалогічних єдностей (пар питання-відповідь).

Оцінка системи здійснюється за допомогою корпусу з 23 уривків діалогів середньою тривалістю близько 10 хвилин, що складається з 80 тематичних сегментів і близько 47 000 слів. Маркування відповідних сегментів тексту здійснили вручну шість анотаторів. Глобальна оцінка свідчить, що для двох жанрів більш розмовного стилю наша система реферування з опорою на притаманні діалогам компоненти значно перевершує два вихідні показники: (1) алгоритм розрахування мінімально допустимої відповідності за допомогою міри важливості термінів TF*IDF, та (2) вихідний показник LEAD, який відбирає перші n слів із тексту.

Переклад Т. Павлущенко

Silber, G. H. Efficiently Computed Lexical Chains as an Intermediate Representation for Automatic Text Summarization [Ефективно обчислені лексичні ланцюжки як проміжний етап автоматичного реферування тексту] / H. Gregory Silber, Kathleen F. McCoy // Computational linguistics. – 2002. – Vol. 28. – No. 4. – Pages 487–496. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120102762671954#.VRHqBNyhGCA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120102762671954

Не дивлячись не те, що автоматичне реферування тексту є галуззю, якій у сучасних дослідженнях приділяється багато уваги, питання про його ефективність піднімається рідко. Коли розглядається обсяг і кількість документів, доступних в Інтернеті та з інших джерел, стає очевидною потреба у високоефективному інструменті для створення прийнятних

рефератів. У статті описано лінійний алгоритм для обчислення лексичних ланцюжків. На проміжному етапі автоматичного реферування тексту алгоритм розраховує, які лексичні ланцюжки є ймовірним кандидатом. Також представлено та реалізовано метод оцінювання лексичних ланцюжків на проміжному етапі процесу реферування. Така оцінка була досі неможливою через складність обчислень лексичних ланцюжків попередніми алгоритмами.

Переклад Т. Павлущенко

Saggion, H. Generating Indicative-Informative Summaries with SumUM [Створення показово-інформативних оглядів за допомогою системи SumUM] / Horacio Saggion, Guy Lapalme // Computational linguistics. – 2002. – Vol. 28. – No. 4. – Pages 497–526. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120102762671963#.VRHrI9yhGCA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120102762671963

В статті описується й оцінюється система реферування текстів SumUM, яка використовує непідготовлений технічний текст в якості вхідних даних і створює індикативний інформативний реферат. У індикативній частині реферату визначаються теми документу, а інформативна частина містить розширену інформацію про ті теми, які цікавлять читача. Система SumUM обгрунтовує теми, описує елементи предметної області і визначає поняття. Вона є першим кроком до дослідження питань динамічного реферування, яке здійснюється шляхом поверхневого синтаксичного і семантичного аналізу, ідентифікації понять та переписування тексту. Запропонований метод було розробено за допомогою аналізу корпусу анотацій, створених спеціалістами з написання анотацій. Скориставшись судженнями експертів, ми оцінили індикативність, інформативність та прийнятність текстів автоматично створених рефератів. Отримані результати свідчать про хорошу ефективність системи в порівнянні з іншими технологіями реферування.

Переклад О. Мартинюк

Jing, H. Using Hidden Markov Modeling to Decompose Human-Written Summaries [Використання прихованої марківської моделі для розбиття речень рефератів, написаних людиною] / Hongyan Jing // Computational linguistics. – 2002. – Vol. 28. – No. 4. – Pages 527–543. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120102762671972#.VRHr4tyhGCA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120102762671972

Люди, які професійно займаються реферуванням, часто використовують вихідні документи для генерування рефератів. Мета розбиття речень реферату полягає в тому, щоб встановити, чи було використано первинний

текст при побудові речення реферату та визначити використані словосполучення. Точніше, програма, що виконує розбиття речень, має відповісти на три питання стосовно певного речення реферату: 1) Чи використано текст вихідного документу при побудові цього речення реферату? 2) Якщо так, то які словосполучення у складі цього речення запозичені з вихідного документу? і 3) В якій частині вихідного документу вжито ці словосполучення? Вирішення проблеми розбиття речень сприятиме появі кращих способів генерування рефератів. Також, завдяки розбиттю речень можна створити великі корпуси для тренування і тестування систем реферування на основі екстракції. Для розбиття речень ми використовуємо приховану марківську модель. Оцінка запропонованого алгоритму свідчить про його ефективність.

Переклад І. Снєгурова

Barzilay, R. Sentence Fusion for Multidocument News Summarization [Злиття речень у процесі реферування декількох новинних повідомлень] / Regina Barzilay, Kathleen R. McKeown // Computational linguistics. – 2005. – Vol. 31. – No. 3. – Pages 297–328. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120105774321091#.VRHsxNyhGCA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120105774321091

Система, яка здатна створювати інформативні реферати, висвітлюючи інформацію, яка повторюється у великій кількості документів у Всесвітній Мережі, допоможе користувачам Всесвітньої Мережі швидко знаходити потрібну їм інформацію. У цій статті ми представляємо новітній метод генерування “текст-до-тексту” для синтезу загальної інформації, спільної для ряду документів. Злиття речень передбачає висхідне часткове паралельне вирівнювання для визначення словосполучень, які передають одну й ту саму інформацію, та статистичне генерування для об’єднання повторюваних словосполучень у речення. Завдяки методу злиття речень у галузі реферування здійснено перехід від використання виключно методів екстракції до генерування анотацій, що містять речення, яких немає в жодному з вихідних документів. Метод також дозволяє синтезувати інформацію з різних джерел.

Переклад І. Снєгурова

Daumé, H. III. Induction of Word and Phrase Alignments for Automatic Document Summarization [Використання вирівнювання слів і речень у автоматичному реферуванні документів] / Hal Daumé III, Daniel Marcu // Computational linguistics. – 2005. – Vol. 31. – No. 4. – Pages 505–530. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120105775299140#.VRHuRdyhGCA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120105775299140

У сучасних дослідженнях автоматичного реферування окремого документа домінують два ефективні, але прості методи: реферування на основі екстракції речень і генерація заголовка на основі моделі «мішка слів». Хоча ці моделі дозволяють успішно вирішувати деякі завдання, жодна з них не здатна адекватно відтворити великий набір лінгвістичних засобів, які використовують при реферуванні люди. Однією з можливих причин широкого використання цих моделей є наявність ефективних методів екстракції підходящої інформації для їх тренування із існуючих корпусів документів/анотацій та документів/заголовків. Ми уважаємо, що подальший прогрес в автоматичному реферуванні буде пов'язаний як із створенням складніших, лінгвістично налаштованих моделей, так і з ефективнішим використанням корпусів документів/анотацій. Для одночасного досягнення обох цілей ми розробили методи автоматичного створення пар слів та фраз із документів та їх анотацій, написаних людиною. Ці пари виявляють відповідності, які існують між такими парами документів і анотацій, і створюють потенційно багату базу даних, яку можна використовувати для тренування складних алгоритмів реферування. У статті описано експерименти, які ми провели, щоб проаналізувати здатність людей робити таке вирівнювання. На основі результатів здійсненого аналізу ми описуємо експерименти для створення системи автоматичного вирівнювання. Наша модель вирівнювання базується на розширенні класичної прихованої моделі Маркова і вчиться створювати вирівнювання без учителя. Ми детально описуємо нашу модель та повідомляємо результати експериментів, які свідчать, що наша модель здатна навчитися надійно ідентифікувати вирівнювання на рівні слова та фрази у корпусі пар «документ, анотація».

Переклад Д. Попової

Kazantseva, A. Summarizing Short Stories [Реферування новел] / Anna Kazantseva, Stan Szpakowicz // Computational linguistics. – 2010. – Vol. 36. – No. 1. – Pages 71–109. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2010.36.1.36102#.VRHvttyhGCA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2010.36.1.36102

У статті представлено метод автоматичної генерації рефератів-екстрактів новел. Реферування здійснюється з конкретною метою: допомогти читачеві вирішити, чи хоче він прочитати всю новелу. З цією метою реферати дають читачеві необхідне уявлення про час і місце дії, не розкриваючи сюжету новели. У системі використовуються різні поверхневі показники предикативних одиниць у новелі, найважливішими з яких є ті, що пов'язані з аспектуальними характеристиками предикативних структур і з головними дійовими особами у новелі. Реферати були оцінені п'ятнадцятьма експертами за допомогою низки зовнішніх та внутрішніх показників. Результати

оцінювання дають підстави уважати, що отримані реферати відповідають поставленій меті.

Переклад Д. Попової

Clarke, J. Discourse Constraints for Document Compression [Дискурсна модель компресії тексту] / James Clarke, Mirella Lapata // Computational linguistics. – 2010. – Vol. 36. – No. 3. – Pages 411–441. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli_a_00004#.WITBzn3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli_a_00004

Компресія речень відкриває перспективи для багатьох прикладних програм: від автоматичного реферування до генерації підзаголовків. Вона звичайно виконується для ізольованих речень без урахування контексту, незважаючи на те, що більшість прикладних програм обробляють увесь текст. У статті представлено дискурсну модель, яка може створювати зв'язні і інформативні анотації текстів. Модель спирається на теорії локальної когерентності і формулюється в рамках цілочислового лінійного програмування. Експериментальні результати свідчать, що вона значно перевершує сучасний підхід, який не бере дискурс до уваги.

Переклад В. Коломієць

Conroy, J. M. Nouveau-ROUGE: A Novelty Metric for Update Summarization [Nouveau-ROUGE: нова метрика для генерації дайджесту оновлень] / John M. Conroy, Judith D. Schlesinger, Dianne P. O'Leary // Computational linguistics. – 2011. – Vol. 37. – No. 1. – Pages 1–8. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli_a_00033#.VRHwmdyhGCA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli_a_00033

Якщо читач уже переглянув попередні документи чи реферати, дайджест оновлень повинен містити стислий виклад нової інформації на тему, яка обговорюється протягом тривалого часу. У 2007 та 2008 роках щорічні змагання систем автоматичного реферування передбачали генерацію дайджесту оновлень. Оцінка за допомогою критерія ROUGE показала, що декілька систем-учасників генерували дайджести оновлень, які неможливо відрізнити від дайджестів, створених вручну. Проте жодна автоматична система не змогла зрівнятися з людиною у ручних оцінках, таких як пірамідальний показник та коефіцієнт загальної сили відгуку.

Ми представляємо метрику Nouveau-ROUGE, яка краще співвідноситься з показниками ручної оцінки і може бути використана для визначення як пірамідального показника, так і коефіцієнта загальної сили відгуку для

дайджестів оновлень. Nouveau-ROUGE може стати дешевшою заміною ручних оцінок при порівнянні існуючих систем і розробці нових.

Переклад В. Туз, М. Погребної

Louis, A. Automatically Assessing Machine Summary Content Without a Gold Standard [Автоматична оцінка змісту автоматично сформованих рефератів без золотого стандарту] / Annie Louis, Ani Nenkova // Computational linguistics. – 2013. – Vol. 39. – No. 2. – Pages 267–300. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00123#.WH3wp33sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00123

Найпопулярніші методи оцінки змісту рефератів використовують процедуру порівняння реферату з золотим стандартом (рефератами, створеними експертами), який традиційно називають еталонними рефератами. Така модель оцінювання не може бути застосована при відсутності еталонних рефератів і дає менш точні результати при наявності лише одного еталонного реферату. У статті запропоновано три нових методи оцінки. Два з них не використовують моделей і не потребують золотого стандарту для оцінювання. Третій метод удосконалює стандартні автоматичні оцінки шляхом додавання до набору наявних еталонних рефератів відібрані автоматично сформовані реферати.

У статті показано, що квантифікація схожості вихідного тексту і його реферату за допомогою правильно підібраних оцінок дозволяє отримати оцінку реферата в балах, яка точно відтворює експертну оцінку. Також досліджено шляхи підвищення якості оцінювання при наявності лише одного створеного експертом зразкового реферата, який використовується як золотий стандарт. Описано псевдомоделі, які є автоматично створеними рефератами, що отримали високі оцінки за зміст при автоматичному оцінюванні. Комбінування псевдомоделей із єдиним створеним експертом зразком для створення золотого стандарту дозволяє підвищити кореляцію з експертними оцінками у порівнянні з використанням лише однієї наявної моделі. Нарешті, досліджено пригодність ще однієї оцінки – схожості між автоматично створеним рефератом і фондом усіх інших автоматично створених рефератів на однакову тематику. Такий метод порівняння із консенсусом систем дає вражаюче точні оцінки автоматичних рефератів, досягаючи кореляції з експертними оцінками понад 0,9.

Переклад В. Коломієць

Інформаційний пошук

Weeber, M. Extracting the Lowest-Frequency Words: Pitfalls and Possibilities [Виокремлення слів із найнижчою частотністю: труднощі та можливості] / Marc Weeber, Rein Vos, R. Harald Baayen // Computational linguistics. – 2000. – Vol. 26. – No. 3. – Pages 301–317. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120100561719#.WIEwmn3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120100561719

У системі видобування медичних даних нами використовуються стандартні методи асоціацій слів для виокремлення термінів на позначення побічних реакцій. Багато таких термінів мають частотність менше п’яти. Стандартні програми на основі словесних асоціацій ігнорують слова із найнижчою частотністю, ігноруючи таким чином корисну інформацію. Тому було розроблено систему видобування слів з усіма частотностями. Ця система вираховує значимість асоціацій за допомогою логарифмічного відношення правдоподібності та точного критерія Фішера. На виході програма демонструє повторювану, незалежну від корпусу тенденцію як у відносній, так і в абсолютній частоті значимих слів. Ці тенденції пояснюються статистичною поведінкою слів з найнижчою частотністю. Щоб показати універсальний характер виявлених закономірностей, було використано голландські фразові дієслова у якості другої і незалежної програми виокремлення колокацій. Зроблено наступні висновки: а) системи виокремлення слів на основі словесних асоціацій можна удосконалити шляхом урахування слів із найнижчою частотністю; б) рівні значущості не повинні бути фіксованими, а підлаштовуватися до оптимального розміру вікна; в) hapax legomena, слова, що зустрілися в тексті лише один раз, повинні апріорно ігноруватися у статистичному аналізі, та г) розподіл об’єктів для виокремленя слід розглядати разом із методом виокремлення.

Переклад О. Мартинюк

Stamatatos, E. Automatic Text Categorization in Terms of Genre and Author [Автоматична категоризація текстів за жанром і автором] / Efstathios Stamatatos, Nikos Fakotakis, George Kokkinakis // Computational linguistics. – 2000. – Vol. 26. – No. 4. – Pages 471–495. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120100750105920#.WIE1In3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120100750105920

Два основні фактори, які характеризують текст, - це його зміст і стиль, і обидва можуть бути використані як засіб категоризації. У статті описано

метод категоризації тексту за жанром і автором для сучасної грецької мови. На відміну від попередніх методів статистичної стилістики, зроблено спробу використовувати в повній мірі наявні інструменти обробки природної мови. Для цього розроблено набір показників стилю, зокрема аналітичні метрики, які показують, яким чином був проаналізований уведений текст і фіксують корисну стилістичну інформацію без додаткових витрат. Описано ряд невеликих, але достатніх експериментів із розпізнавання жанру тексту, встановлення особи автора та підтвердження авторства, і показано, що запропонований метод є ефективнішим, ніж надзвичайно популярні міри дистрибуції лексики, тобто функції багатства лексики і частоти вживання найчастотніших слів. У всіх описаних експериментах використовувався довільний текст, завантажений із Інтернету, без будь-якої ручної попередньої обробки або скорочення. Розглянуто різні проблеми використання методу, що стосуються обсягу навчального матеріалу і значущості запропонованих показників стилю. Створена система може бути використана в будь-якому додатку, де потрібна швидка категоризація тексту, яку можна легко адаптувати в плані стилістично однорідних категорій. Крім того, використовуючи існуючі інструменти обробки тексту, можна простежити процес визначення аналітичних метрик, щоб видобути корисну стилістичну інформацію.

Переклад К. Погорєлова

Cucchiarelli, A. Unsupervised Named Entity Recognition Using Syntactic and Semantic Contextual Evidence [Неконтрольоване розпізнавання власних назв з урахуванням синтаксичного і семантичного контексту] / Alessandro Cucchiarelli, Paola Velardi // Computational linguistics. – 2001. – Vol. 27. – No. 1. – Pages 123–131. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120101300346822#.WIE2H33sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120101300346822

Власні назви утворюють відкритий клас, тому неповнота правил класифікації, укладених вручну або автоматично, є очевидною проблемою. Стаття має дві цілі: по-перше, запропонувати використання додаткового "допоміжного" методу для підвищення надійності будь-якого маркувальника власних назв, створеного вручну або на основі машинного навчання, а по-друге, проаналізувати ефективність використання точніших даних − а саме, інформації про синтаксичний і семантичний контекст − для класифікації власних назв.

Переклад К. Погорєлова

Kehler, A. The Need for Accurate Alignment in Natural Language System Evaluation [Необхідність точної вивірки в оцінюванні системи обробки природної мови] / Andrew Kehler, John Bear, Douglas Appelt // Computational linguistics. – 2001. – Vol. 27. – No. 2. – Pages 231–248. –

Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120101750300517#.WIExsH3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120101750300517

Оскільки оцінки технологій комп'ютерної лінгвістики переміщаються до завдань вищого рівня складності, завдання встановлення відповідностей між відповідями системи та правильними відповідями може ускладнитись. У статті подано вичерпний аналіз процедури вивірки, яка використовувалась для оцінки технології видобування інформації на шостій конференції по розумінню повідомлень (Message Understanding Conference 6, *скор.* MUC-6). Виявлено причини, які заважають досягненню заявлених цілей аналізу. Показано, що ці причини настільки розповсюджені, що здатні негативно вплинути на процес розвитку технології. Отримані результати свідчать про необхідність використання точних критеріїв вивірки в оцінюванні природної мови та розмежування критеріїв вивірки і механізмів підрахунку оцінок.

Переклад А. Синящик

Weeds, J. Co-occurrence Retrieval: A Flexible Framework for Lexical Distributional Similarity [Виявлення одночасної появи слів: гнучкий підхід до лексичної дистрибутивної схожості] / Julie Weeds, David Weir // Computational linguistics. – 2005. – Vol. 31. – No. 4. – Pages 439–475. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120105775299122#.WIE3zH3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120105775299122

У багатьох галузях обробки природної мови пропонуються методи, які використовують інформацію про дистрибутивну схожість слів. Наприклад, у моделюванні мови можна зменшити проблему відсутності повних даних, спрогнозувавши вірогідність одночасної появи слів, яка не спостерігалась раніше, на основі одночасної появи подібних слів, яка спостерігалась раніше. У інших додатках дистрибутивна схожість уважається наближенням до семантичної схожості. Проте, завдяки широкому спектру потенційного застосування і через відсутність чіткого визначення поняття дистрибутивної схожості були запропоновані або запозичені багато методів обчислення дистрибутивної схожості.

У цій роботі пропонується гнучкий, параметризований підхід до обчислення дистрибутивної схожості. У рамках цього підходу проблема знаходження слів зі схожою дистрибуцією розглядається як різновид пошуку одночасної появи (ПОП), для якого можна визначити точність і повноту по аналогії з методами їх вимірювання у пошуці документів. Як буде продемонстровано, в рамках підходу ПОП використовувались налаштування параметрів для моделювання великої кількості популярних нині мірок дистрибутивної схожості. Після цього підхід ПОП був використаний у

дослідженні для систематичного дослідження трьох основних питань, які стосуються дистрибутивної схожості. По-перше, чи завжди відношення лексичної схожості є симетричними, і чи дає якісь переваги ставлення до них як до відношень асиметричних? По-друге, чи є деякі випадки одночасного вживання по своїй суті важливішими, ніж інші, у обчисленні дистрибутивної схожості? По-третє, чи потрібно брати до уваги різницю між кількістю появ кожного слова у кожному різновиді одночасної появи?

Оцінювання здійснювалось за допомогою двох завдань з використанням додатків: автоматичного створення тезаурусу і імітації розв'язання багатозначності. Можна значно поліпшити результати виконання обох вказаних завдань не шляхом використання інших існуючих критеріїв дистрибутивної схожості, а варіюючи параметри в рамках підходу ПОП. Також доведено, що будь-яка окрема непараметризована мірка навряд чи зможе показати вищу ефективність у обох завданнях. Це пояснюється притаманною лексичній замінності, а отже і лексичній дистрибутивній схожості, асиметрією.

Переклад В. Коломієць

Tanaka-Ishii, T. Sorting Texts by Readability [Сортування текстів за складністю] / Kumiko Tanaka-Ishii, Satoshi Tezuka, Hiroshi Terada // Computational linguistics. – 2010. – Vol. 36. – No. 2. – Pages 203-227. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.09-036-R2-08-050#.WIE5wn3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.09-036-R2-08-050

У статті описано новаторський метод оцінювання складності текстів шляхом сортування. За допомогою машинного навчання створюється компаратор, який порівнює відносну складність пари текстів, потім цей компаратор сортує заданий набір текстів. Корисність розробленого методу в тому, що він вирішує проблему відсутності навчального набору даних, адже для створення компаратора потрібен лише набір даних, розсортованих за двома рівнями складності. Розроблений метод порівнюється з методами регресії і новітнім класифікаційним методом. Крім того, описано розроблену нами програму під назвою Terrace, яка знаходить тексти, співставні за рівнем складності із заданим вхідним текстом.

Переклад В. Коломієць

Verberne, S. What Is Not in the Bag of Words for Why-QA? [Чого немає у мішку слів для системи питання «Чому…»-відповідь?] / Suzan Verberne, Lou Boves, Nelleke Oostdijk, Peter-Arno Coppen // Computational linguistics. – 2010. – Vol. 36. – No. 2. – Pages 229–245. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.09-032-R1-08-034#.WIEyS33sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.09-032-R1-08-034

У процесі розробки моделі ПИТАННЯ «Чому…»-відповідь до системи пошуку фрагментів, яка використовує готові технології інформаційного пошуку, було додано модуль переранжування, який містить синтаксичну інформацію. Було отримано значно вищі показники середнього оберненого рангу MRR@150 (від 0.25 до 0.34) і success@10. Досягнуте поліпшення на 23% показників середнього оберненого рангу співставне з досягненнями інших дослідників у цій області у вирішенні різних задач, пов'язаних з питально-відповідальними системами, хоча у запропонованому методі переранжування використовуються порівняно спрощені і частково дубльовані міри, які включають синтаксичні складники, сигнальні слова і структуру документа.

Переклад В. Коломієць

Petrenz, P. Stable Classification of Text Genres [Стабільна жанрова класифікація текстів] / Philipp Petrenz, Bonnie Webber // Computational linguistics. – 2011. – Vol. 37. – No. 2. – Pages 385–395. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00052#.WIEMCX3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00052

Кожен текст має прнаймні одну тему і належить принаймні до одного жанру. Про тему і жанр тексту частково свідчать його лексичні та синтаксичні характеристики – характеристики, які використовуються як для автоматичної тематичної класифікації, так і для автоматичної жанрової класифікації (АЖК). Оскільки ідеальна система АЖК не повинна залежати від змін у розподілі тем, здійснено оцінку п'яти опублікованих методів АЖК як щодо їх ефективності на тій самій тематичній і жанровій дистрибуції, на якій вони навчалися, так і щодо стабільності цієї ефективності при змінах у тематичній і жанровій дистрибуції. Здійснені експерименти дозволили зробити висновок, що (1) до критеріїв оцінювання нових підходів до АЖК потрібно додати стабільність в умовах зміни тематичної дистрибуції і (2) що при розробці високопродуктивної, стабільної системи АЖК для конкретного, можливо динамічного, корпусу ознаки частин мови потрібно враховувати окремо.

Переклад І. Снєгурова

Pan, F. Annotating and Learning Event Durations in Text [Анотування і автоматичне визначення тривалості подій у текстах] / Feng Pan, Rutu Mulkar-Mehta, Jerry R. Hobbs // Computational linguistics. – 2011. – Vol. 37. – No. 4. – Pages 727–752. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00075#.WIHw_X3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00075

У статті описується розробка корпусу публікацій новин з розміткою приблизної тривалості подій і машинне навчання на основі цього корпусу. Описано правила анотування, розроблену з метою зменшення серйозних розходжень між судженнями анотаторів класифікацію подій, а також використання нормального розподілу для моделювання неконкретної і імпліцитної інформації про тривалість подій і визначення міри узгодженості між анотаторами щодо розподілів тривалості подій. Потім показано, що застосувавши до цих даних методи машинного навчання, можна автоматично отримати приблизну інформацію про тривалість подій, що значно перевершує базові дані продуктивності і наближається до людських оцінок. Описані у статті методи можна застосовувати до інших видів неконкретних, але суттєвих даних у тексті.

Переклад В. Коломієць

Chen, Y. A Joint Model to Identify and Align Bilingual Named Entities [Комбінована модель для розпізнавання і вирівнювання власних назв двома мовами] / Yufeng Chen, Chengqing Zong, Keh-Yih Su // Computational linguistics. – 2013. – Vol. 39. – No. 2. – Pages 229–266. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00122#.WII0BX3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00122

У статті теоретично виведена інтегрована модель, яка і визначає, і вирівнює власні назви (ВН) двома мовами – китайскою і англійською. Модель підказана такими спостереженнями: 1) вибір семантичного або фонетичного перекладу ВН великою мірою залежить від їх різновиду, 2) власні назві у вирівняній парі повинні належати до одного типу і 3) ВН, визначені першими, можуть виступати в ролі якорів і надавати додаткову інформацію при відборі кандидатів у ВН. На основі цих спостережень у статті пропонується характеристика співвідношення способів перекладу (яка визначається як відсоток усіх ВН, перекладених семантичним способом), уводиться обмеження коректності типів об'єктів і використовуються додаткові нові можливі ВН (на основі визначених на початку якорів ВН).

Експерименти свідчать, що цей новітній метод значно перевершує стандартні методи. У вирівнюванні китайських і англійських ВН показник F-score, незалежний від типу розпізнаних пар ВН, зріс із 78,4% до 88,0% (відносне покращення на 12,2%), а показник F-score, залежний від типу розпізнаних пар, зріс із 68,4% до 83% (відносне покращення на 21,3%). Крім того, запропонована модель показала свою надійність при тестуванні у різних предметних областях. Нарешті, при застосуванні навчання із частковим залученням учителя для тренування розробленої моделі розпізнавання англійських ВН запропонована модель також значно поліпшує залежний від типу розпізнаних англійських ВН показник F-score.

Переклад В. Коломієць

Di Marco, A. Clustering and Diversifying Web Search Results with Graph-Based Word Sense Induction [Кластеризація і диверсифікація результатів інформаційного пошуку за допомогою встановлення значення слів на основі графів] / Antonio Di Marco, Roberto Navigli // Computational linguistics. – 2013. – Vol. 39. – No. 3. – Pages 709–754. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00148#.WIE0BH3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00148

Метою кластеризації результатів інформаційного пошуку є полегшення пошуку інформації у Інтернеті. Замість представлення результатів запиту у вигляді плаского списку, вони групуються на основі схожості і пізніше пред'являються користувачеві як список кластерів. Призначення кожного кластера – представляти різні значення пошукового запиту, враховуючи таким чином проблему лексичної неоднозначності (або полісемії). Проте існуючі методи кластеризації Всесвітньої мережі звичайно базуються на якомусь поверховому уявленні про текстову схожість фрагментів пошукових результатів. В результаті, текстові фрагменти, які не містять однакових слів, потрапляють у окремі кластери, навіть якщо вони схожі за змістом, а текстові фрагменти, які містять однакові слова, групуються разом, навіть якщо вони відносяться до різних значень запиту.

У статті представлено новий підхід до кластеризації результатів інформаційного пошуку, який базується на автоматичному встановленні значень слів із сирого тексту, завданні, яке називається індукцією значення слова. Ключом до нашого підходу є встановлення різних смислів (тобто значень) неоднозначного запиту і наступна кластеризація результатів пошуку на основі їх семантичної схожості із встановленими смислами слів. Експерименти, проведені на наборах даних, які складалися з неоднозначних запитів, свідчать, що наш підхід перевершує як мережеву кластеризацію, так і інформаційно-пошукові системи.

Переклад В. Коломієць

D'hondt, E. Text Representations for Patent Classification [Репрезентації текстів для класифікації патентів] / Eva D'hondt, Suzan Verberne, Cornelis Koster, Lou Boves // Computational linguistics. – 2013. – Vol. 39. – No. 3. – Pages 755–775. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00149#.WIE62n3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00149

У зв'язку із збільшенням кількості заявок на видачу патентів, збільшується економічна важливість автоматичної класифікації патентів. У статті

досліджується, як можна поліпшити класифікацію патентів, використовуючи різні представлення патентної документації. За допомогою системи лінгвістичної класифікації (*англ.* Linguistic Classification System, *скор.* LCS) порівнюється вплив додавання статистичних словосполучень (у формі біграмів) і лінгвістичних словосполучень (з двома різними видами залежностей) до стандартної репрезентації тексту у вигляді мішка слів на виборці з 532 264 англійських аннотацій з корпусу CLEF-IP 2010. На відміну від попередніх досліджень класифікації за допомогою словосполучень із бази даних Reuters-21578, у класифікації патентів додавання словосполучень призводить до значного підвищення якості у порівнянні зі стандартними показниками уніграму. Найкращі показники були отримані шляхом об'єднання усіх чотирьох репрезентацій, на другому місці знаходяться показники, отримані шляхом комбінування уніграмів і лематизованих біграмів. У статті здійснено ретельний аналіз моделей класів (або опис класів), створених класифікаторами у рамках LCS, для визначення типу словосполучень, які є найбільш інформативними для класифікації патентів. З'ясовано, що підвищення точності класифікації залежить в першу чергу від біграмів. Щоб визначити ступінь можливого узагальнення отриманих результатів подібні експерименти були проведені на підмножинах уривків з патентів французькою і німецькою мовами.

Переклад В. Коломієць

Barrón-Cedeño, A. Plagiarism Meets Paraphrasing: Insights for the Next Generation in Automatic Plagiarism Detection [Плагіат і парафраза: ідеї для наступного покоління систем атоматичного виявлення плагіату] / Alberto Barrón-Cedeño, Marta Vila, M. Antònia Martí, Paolo Rosso // Computational linguistics. – 2013. – Vol. 39. – No. 4. – Pages 917–947. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00153#.WIE7lH3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00153

Хоча парафразування є лінгвістичним механізмом, який лежить в основі багатьох випадків плагіату, його аналізу в рамках автоматичного розпізнавання плагіату приділялось мало уваги. Саме тому сучасним детекторам плагіату складно виявити випадки парафразового плагіату. У статті здійснено аналіз зв'язку парафрази та плагіату з метою виділення різновидів парафрази, які є характерними для плагіату, і тих із них, які можна виявити за допомогою систем розпізнавання плагіату. Для досягнення поставленої мети було створено корпус P4P, новий ресурс, у якому використано типологію парафрази, для анотування частини корпусу PAN-PC-10 для автоматичного розпізнавання плагіату. З точки зору цього анотування проаналізовано результати другого міжнародного змагання із визначення плагіату.

Описані експерименти свідчать, що (1) складніші парафрази та висока

щільність парафразових конструкцій ускладнюють розпізнавання плагіату, (2) лексичні заміни є парафразовими конструкціями, які найчастіше використовуються у процесі списування, і (3) парафразові конструкції, як правило, скорочують списаний текст. Це перше дослідження парафразових конструкцій, які використовуються у процесі плагіату, у якому висловлено ідеї, важливі для вдосконалення автоматичних систем розпізнавання плагіату.

Переклад В. Туз

Seroussi, Y. Authorship Attribution with Topic Models [Встановлення авторства за допомогою тематичних моделей] / Yanir Seroussi, Ingrid Zukerman, Fabian Bohnert // Computational linguistics. – 2014. – Vol. 40. – No. 2. – Pages 269–310. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00173#.WIE-AH3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00173

Автоматичне визначення авторства полягає у встановленні авторства анонімних текстів. Раніше дослідження у цій галузі стосувалися переважно офіційних документів, таких як есе і романи, але останнім часом більше уваги приділяється текстам, створеним користувачами мережі Інтернет, таким як електронні листи та блоги. Встановити авторство таких текстів значно важче, ніж встановити авторство офіційних документів, оскільки обсяг тексту є меншим, а кількість претендентів на авторство – більшою. Ми вирішуємо цю проблему, отримуючи репрезентації авторів за допомогою тематичних моделей. Окрім вивчення нових способів застосування двох відомих тематичних моделей для розв'язання цієї задачі, протестовано нашу нову модель, яка проектує авторів та документи на два окремі тематичні простори. Використання нашої моделі при встановленні авторства текстів продемонструвало її високу ефективність у кількох наборах даних, які містили або офіційні документи, написані кількома авторами, або неофіційні документи, створені десятками тисяч користувачів Інтернету. Також описано результати експериментів, які засвідчили можливість застосування тематичних репрезентацій авторів при розв'язанні ще двох проблем: визначенні тональності текстів та прогнозуванні можливих оцінок користувачами таких продуктів, як кінофільми.

Переклад М. Погребної

Chang, C. Practical Linguistic Steganography using Contextual Synonym Substitution and a Novel Vertex Coding Method [Практична лінгвістична стенографія на основі підстановки контекстних синонімів і нового методу кодування вершин] / Ching-Yun Chang, Stephen Clark // Computational linguistics. – 2014. – Vol. 40. – No. 2. – Pages 403–448. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00176#.WIE-

Z33sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00176

Мета лінгвістичної стенографії – сховати інформацію в тексті природною мовою. Однією з основних трансформацій, які використовуються у лінгвістичній стенографії, є підстановка синонімів. Проте досліджень практичного застосування цього підходу мало. У статті запропоновано два вдосконалення до застосування підстановки синонімів для кодування прихованих бітів інформації. По-перше, для перевірки застосовності синоніма в контексті використано корпус Google n-grams, а оцінювання методу здійснено за допомогою даних із завдання на лексичну підстановку з конференції SemEval і даних, анотованих вручну. По-друге, розглянуто спричинену багатозначними словами проблему створення потенційної неоднозначності: які біти представлені конкретним словом. Розроблено новий метод, у якому слова є вершинами графа, синоніми з'єднані ребрами, а приписані слову біти визначаються за допомогою алгоритму кодування вершини. Вказаний метод гарантує, що кожне слово представляє унікальну послідовність бітів без виключення великої кількості синонімів і таким чином зберігає достатню шифрувальну здатність.

Переклад В. Коломієць

Shaalan, K. A Survey of Arabic Named Entity Recognition and Classification [Огляд розпізнавання і класифікації власних назв арабською мовою] / Khaled Shaalan // Computational linguistics. – 2014. – Vol. 40. – No. 2. – Pages 469–510. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00178#.WIE-3H3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00178

Оскільки завдяки службам Інтернету і Інтранету приватні і корпоративні користувачі отримують у Всесвітній мережі доступ до зростаючої кількості текстів арабською мовою, існує нагальна потреба у технологіях і інструментах для обробки потрібної інформації. Розпізнавання власних назв (*англ.* Named Entity Recognition, *скор.* NER) – це завдання видобування інформації, яке стало невід'ємною частиною багатьох інших завдань обробки природної мови, таких як машинний переклад та інформаційний пошук. NER арабською мовою стало привертати увагу протягом останніх років. Через характерні особливості арабської мови, яка належить до семітської групи мов, розпізнавання власних назв є складним завданням. Продуктивність компонента NER арабською мовою позитивно впливає на загальну продуктивність системи обробки природної мови. У статті робиться спроба детально описати зростання за останній час інтересу і наявні досягнення у дослідженнях NER арабською мовою. Обгрунтовано важливість виконання NER, висвітлено основні характеристики арабської мови, проілюстровано особливості стандартизації в анотуванні власних назв. Крім того, описано

різні лінгвістичні ресурси арабською мовою і розглянуто підходи, які використовуються в області NER арабською мовою. Описано особливості типових інструментів, які виористовуються у NER арабською мовою і проілюстровано стандартні оціночні показники. Крім того, проаналізовано огляд сучасних досліджень NER арабською мовою. Нарешті, представлено висновки автора. Для ясності виклад матеріалу проілюстровано прикладами.

Переклад В. Коломієць

Clercq O. All Mixed Up? Finding the Optimal Feature Set for General Readability Prediction and Its Application to English and Dutch [Все змішалося? Пошук оптимального набору параметрів для прогнозування загальної легкочитаності та його застосування до англійської і нідерландської мов] / Orphée De Clercq, Véronique Hoste // Computational linguistics. – 2016. – Vol. 42. – No. 3. – Pages 457–490. – Режим доступу до анотації: https://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00255 – Режим доступу до повнотекстової статті: https://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00255

Дослідження легкочитності має довгу і багату історію, але досі майже не приділялося уваги прогнозуванню загальної легкочитності без орієнтації на конкретну аудиторію чи жанр тексту. Крім того, хоча прикладні лінгістичні дослідження зосереджуються на додаванні більш складних параметрів легкочитності, досі не існує єдиної думки щодо того, які параметри грають найважливішу роль у прогнозуванні. У статті докладно досліджено можливість побудови системи прогнозування легкочитності текстів загального змісту англійською та нідерландською мовами за допомогою навчання з учителем. На основі експертної та краудсорсингової оцінок легкочитності застосовано різні типи характеристик тексту, від легкообчислюваних поверхневих до характеристик, які потребують глибокої лінгвістичної обробки. Всього виділено десять груп характеристик. Досліджено як регресійні, так і класифікаційні моделі, що відображають два можливі завдання прогнозування легкочитності: оцінювання окремих текстів або порівняння двох текстів. У статті показано, що вихід за межі обчислень кореляції для оптимізації легкочитності за допомогою методу оптимізації генетичного алгоритму на основі обкладинки, є перспективним завданням, яке дає чимало інформації про те, які комбінації характеристик дозволяють прогнозувати загальну легкочитаність. Оскільки для тих функцій, які потребують глибинної обробки, існує золотий стандарт, можна дослідити справжню верхню межу системи нідерландської мови. Зауважимо, що видається цікавим той факт, що продуктивність описаної системи автоматичного прогнозування легкочитаності, співставна з продуктивністю системи на основі золотостандартної повної синтаксичної і семантичної інформації.

Переклад А. Шульги

Машинний переклад

Dan Melamed, I. Models of Translational Equivalence among Words [Моделі перекладацької еквівалентності серед слів] / I. Dan Melamed // Computational linguistics. – 2000. – Vol. 26. – No. 2. – Pages 221–249. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120100561683#.WIH2An3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120100561683

Паралельні тексти (бітексти) мають характеристики, які відрізняють їх від інших видів паралельних даних. По-перше, більшість слів перекладаються лише одним словом. По-друге, бітекстова відповідність зазвичай є частковою, тобто багато слів у кожному тексті не мають чітких відповідників у другому тексті. У статті описано методи налаштування моделей статистичного перекладу для відображення цих властивостей. Оцінка на основі суджень незалежних експертів підтвердила, що налаштовані таким чином моделі перекладу значно точніші, ніж базова модель без застосування знань. У статті також показано, як статистична модель перекладу може використовувати вже існуючі знання, наявні для певних мовних пар. Продемонстровано, що навіть елементарні знання про конкретну мову, такі як відмінність між самостійними і службовими частинами мови, забезпечують значне підвищення продуктивності моделі перекладу при виконанні деяких завдань. Статистичні моделі, які відображають знання про предметну галузь, поєднують у собі найкращі риси раціоналістичного та емпіричного підходів.

Переклад Д. Попової

Och, F.J. A Systematic Comparison of Various Statistical Alignment Models [Систематичне порівняння різних статистичних моделей вирівнювання] / Franz Josef Och, Hermann Ney // Computational linguistics. – 2003. – Vol. 29. – No. 1. – Pages 19–51. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120103321337421#.WIIHDn3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120103321337421

У статті описуються і порівнюються різні методи для обчислення вирівнювання слів за допомогою статистичних і евристичних моделей. Розглянуто п’ять моделей вирівнювання, описаних у праці П. Брауна та ін. (Brown, P., Della Pietra, S. A., Della Pietra, V., J., and Mercer, R. L., 1993), приховану марківську модель вирівнювання, методи згладжування, а також уточнення. Ці статистичні моделі порівнюються з двома евристичними моделями на основі коефіцієнта Дайса. Описано різні методи комбінування

вирівнювання слів для створення відображення моделей спрямованого статистичного вирівнювання. В якості критерія оцінювання використано якість отриманого вирівнювання Вітербі у порівнянні з створеним вручну вирівнюванням референцій. Для оцінки моделей використовувались німецько-англійський перекладач Verbmobil і французько-англійський корпус Hansards. Здійснено ретельний аналіз різних проектів системи статистичного вирівнювання і їх оцінка за допомогою тренувальних корпусів різних розмірів. Важливим результатом є те, що вдосконалені моделі вирівнювання з залежністю першого порядку і модель родючості дають кращі результати, ніж прості евристичні моделі. У додатку вміщено ефективний тренувальний алгоритм для моделей вирівнювання.

Переклад В. Коломієць

Tillmann, C. Word Reordering and a Dynamic Programming Beam Search Algorithm for Statistical Machine Translation [Зміна порядку слів і алгоритм променевого пошуку на основі динамічного програмування для статистичного машинного перекладу] / Christoph Tillmann, Hermann Ney // Computational linguistics. – 2003. – Vol. 29. – No. 1. – Pages 97–133. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120103321337458#.WIIIgn3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120103321337458

В даній статті описується ефективний алгоритм променевого пошуку на основі динамічного програмування (ДП) для статистичного машинного перекладу. Алгоритм пошуку використовує модель перекладу, представлену в праці Брауна та ін. (Brown et al., 1993). Починаючи з вирішення завдання комівояжера на основі ДП, ми представляємо новий спосіб обмеження можливих змін порядку слів у процесі перекладу, щоб створити ефективний алгоритм пошуку. Визначено обмеження змін порядку слів, потрібні при перекладі з німецької мови на англійську. Обмеження узагальнено і запропоновано сукупність чотирьох параметрів для контролю змін порядку слів, котру можна легко адаптувати для перекладу інших мовних пар. Процедура променевого пошуку була успішно протестована у системі Verbmobil (німецька – англійська, у словнику 8000 слів) та у корпусі Canadian Hansards (французька – англійська, у словнику 100 000 слів). Під час виконання середнього за розміром завдання Verbmobil речення може бути перекладене за декілька секунд, кількість помилок невелика, а погіршення результатів, яке вимірюється за критерієм помилок у вживанні слів, який використовується в даній статті, не зафіксовано.

Переклад М. Драчової

Kraaij, W. Embedding Web-Based Statistical Translation Models in Cross-Language Information Retrieval [Вбудовування моделей статистичного перекладу на основі інтернет-технологій у пошук інформації різними

мовами] / Wessel Kraaij, Jian-Yun Nie, Michel Simard // Computational linguistics. – 2003. – Vol. 29. – No. 3. – Pages 381–419. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120103322711587#.WIIKZ33sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120103322711587

Хоча кількість мовних пар, які обслуговують системи машинного перекладу, зростає, залишається ще багато пар, для яких немає інструментів перекладу. Одним із практичних завдань, яке потребує перекладацького забезпечення порівняно невисокої якості, є пошук інформації різними мовами (ПІРМ), адже основою діючих моделей інформаційного пошуку (ІП) і досі залишається мішок слів. Інтернет є величезним ресурсом для автоматичного створення паралельних корпусів, які можуть використовуватися для автоматичного тренування статистичних моделей перекладу. Отримані таким чином моделі перекладу можна вбудувати різними способами у модель інформаційного пошуку. У статті розглядається проблема автоматичного пошуку в паралельних текстах з Інтернету і різні способи вбудовування моделей перекладу в процес інформаційного пошуку. Експерименти на основі стандартних наборів текстів для ПІРМ свідчать, що перекладацькі моделі на основі Інтернет-технологій можуть перевершити комерційні системи машинного перекладу у завданнях ПІРМ. Ці результати відкривають можливість створення при дуже низьких затратах повністю автоматичної системи перекладу запитів для ПІРМ.

Переклад В. Коломієць

Way, A. *wEBMT*: Developing and Validating an Example-Based Machine Translation System Using the World Wide Web [*Машинний переклад із використанням Всесвітньої мережі*: розробка і оцінка ефективності системи машинного перекладу на основі прецедентів, що використовує Всесвітнє павутиння] / Andy Way, Nano Gough // Computational linguistics. – 2003. – Vol. 29. – No. 3. – Pages 421–457. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120103322711596#.WIIKrH3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120103322711596

Ми розробили систему машинного перекладу на основі прецедентів (Example-Based Machine Translation, *скор.* EBMT), що використовує Всесвітню мережу з двома різними цілями: по-перше, ми заповнюємо пам'ять системи перекладами, отриманими із розташованих у мережі систем машинного перекладу (МП) на основі правил. Вихідні ланцюжки, уведені в ці системи, було автоматично вилучено з дуже маленької підгрупи типів правил в банку дерев Penn-II. На наступних етапах отримані пари типу «оригінал – переклад» автоматично перетворюються на низку ресурсів, що підвищують ефективність процесу перекладу. Хоча результат роботи онлайн-

систем МП часто містить помилки, ми продемонстрували у численних експериментах, що насправді вони можуть бути корисними при створенні високоякісних перекладів, якщо використовуються для заповнення пам'яті системи EBMT. Крім того, ми показуємо переваги систем EBMT над онлайн-системами. По-друге, незважаючи на те, що якість наявних у мережі документів сумнівна, ми доводимо ефективність використання таких ресурсів у процесі автоматичного постредагування варіантів перекладу, запропонованих нашою системою.

Переклад М. Драчової

Li, H. Word Translation Disambiguation Using Bilingual Bootstrapping [Використання двомовного самоналаштування для вирішення багатозначності при перекладі слів] / Hang Li, Cong Li // Computational linguistics. – 2004. – Vol. 30. – No. 1. – Pages 1–22. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120104773633367#.WIILk33sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120104773633367

В статті запропоновано новий метод вирішення багатозначності слів при перекладі за допомогою способу машинного навчання під назвою двомовне самоналаштування. У процесі навчання усуненню неоднозначності слів, які потрібно перекласти, двомовне самоналаштування використовує невеликий обсяг класифікованих даних і великий обсяг некласифікованих даних у мові оригіналу і у мові перекладу. Він багатократно будує класифікатори одночасно на двох мовах і підвищує їх продуктивність за допомогою класифікації некласифікованих даних двома мовами та шляхом обміну інформацією щодо класифікованих даних між двома мовами. Результати експериментів свідчать, що вирішення багатозначності слів при перекладі з допомогою двомовного самоналаштування дозволяє отримати значно кращі результати, ніж існуючі методи, в яких використовується одномовне самоналаштування.

Переклад О. Мартинюк, М. Погребної

Nießen, S. Statistical Machine Translation with Scarce Resources Using Morpho-syntactic Information [Використання морфо-синтаксичної інформації у статистичному машинному перекладі з недостатньо великим корпусом для тренування] / Sonja Nießen, Hermann Ney // Computational linguistics. – 2004. – Vol. 30. – No. 2. – Pages 181–204. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120104323093285#.WIIMTn3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120104323093285

У статистичному машинному перекладі відповідності між словами у мові оригіналу та мові перекладу встановлюються автоматично за допомогою паралельних корпусів, а лінгвістичні знання при формуванні базових моделей зазвичай використовуються мало або не використовуються взагалі. Зокрема, існуючі статистичні системи машинного перекладу часто розглядають різні похідні форми однієї і тієї ж леми так, ніби вони незалежні одна від одної. Двомовні корпуси для тренування можуть використовуватися ефективніше за умови детального врахування взаємозалежностей між спорідненими похідними формами. Ми пропонуємо створювати ієрархічні моделі лексиконів на основі еквівалентних класів слів. Крім цього, ми пропонуємо трансформації реструктурування на рівні речень, мета яких полягає в уподібненні порядку слів у споріднених реченнях. Ми ретельно визначили обсяг двомовних даних для тренування, необхідних для підтримання прийнятної якості машинного перекладу. Тестування сукупності запропонованих методів покращення якості перекладу в умовах обмежених ресурсів виявилось успішним. Нам вдалося зменшити кількість двомовних даних для тренування до менш ніж 10% вихідного корпусу, при цьому якість перекладу знизилась лише на 1.6%. Покращення результатів перекладу продемонстровано на двох німецько-англійських корпусах з проектів Verbmobil та Nespole!

Переклад О. Мартинюк

Casacuberta, F. Machine Translation with Inferred Stochastic Finite-State Transducers [Машинний переклад з допомогою автоматично побудованих стохастичних скінченних перетворювачів] / Francisco Casacuberta, Enrique Vidal // Computational linguistics. – 2004. – Vol. 30. – No. 2. – Pages 205–225. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/089120104323093294#.WIIMi33sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120104323093294

Скінченні перетворювачі – це моделі, які використовуються в різних галузях розпізнавання образів та комп'ютерної лінгвістики. Однією з таких галузей є машинний переклад, у якому набувають популярності підходи з використанням автоматичної побудови моделей на основі навчальної вибірки. Скінченні перетворювачі доцільно використовувати при виконанні обмежених завдань за наявності навчальних виборок пар речень. У цій статті представлено метод автоматичної побудови скінченних перетворювачів. Даний метод спирається на формальні зв'язки між скінченними перетворювачами та граматичними моделями. Запропонований підхід передбачає застосування методів статистичного вирівнювання до корпусу для тренування, який складається із речень з їх перекладами, з метою створення набору стандартних ланцюжків, з якого виводиться ймовірна модель (наприклад, n-грам). Ця модель зрештою трансформується у

скінченний перетворювач. Запропоновані методи протестовано в процесі виконання серії експериментів з машинного перекладу в рамках проекту E u Trans.

Переклад М. Погребної

Och, F. J. The Alignment Template Approach to Statistical Machine Translation [Статистичний машинний переклад з використанням алгоритмів вирівнювання] / Franz Josef Och, Hermann Ney // Computational linguistics. – 2004. – Vol. 30. – No. 4. – Pages 417–449. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/0891201042544884#.WIINLn3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/0891201042544884

У цій статті описується статистичний машинний переклад на основі словосполучень – переклад з використанням алгоритмів вирівнювання. Цей підхід до перекладу дозволяє встановлювати загальні відносини між словами типу «багато до багатьох». Таким чином, ця модель перекладу враховує контексти слів, також можна точно визначити зміни в порядку слів при переході від мови оригіналу до мови перекладу. Модель описується за допомогою логарифмічно-лінійного підходу до моделювання, який є узагальненням популярного методу на основі вихідного каналу. Отже, цю модель розширити легше, ніж традиційні системи статистичного машинного перекладу. Ми детально описуємо процес навчання перекладу по словосполученням, застосовані функції та алгоритм пошуку. Оцінювання цього підходу здійснюється за допомогою трьох різних проектів. За допомогою системи усного перекладу з німецької мови на англійську Verbmobil проаналізовано роль різних компонентів системи. За допомогою французько-англійського корпусу Canadian Hansards продемонстровано, що модель перекладу з використанням алгоритмів вирівнювання дає значно кращі результати, ніж модель перекладу по окремих словах. У здійсненому в 2002 році Національним інститутом стандартів і технологій (National Institute of Standards and Technology, скор. NIST) оцінюванні машинних перекладів з китайської мови на англійську зазначена система досягла статистично істотно вищого показника NIST, ніж усі інші конкуруючі дослідницькі й комерційні системи перекладу.

Переклад М. Погребної

Munteanu, D. S. Improving Machine Translation Performance by Exploiting Non-Parallel Corpora [Удосконалення машинного перекладу шляхом використання непаралельних корпусів] / Dragos Stefan Munteanu, Daniel Marcu // Computational linguistics. – 2005. – Vol. 31. – No. 4. – Pages 477–504. – Режим доступу до анотації:

http://www.mitpressjournals.org/doi/abs/10.1162/089120105775299168#.WIIN0X3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/089120105775299168

Ми пропонуємо новий метод для виявлення паралельних речень у порівняльних, непаралельних корпусах. Ми навчаємо класифікатор максимальної ентропії достовірно визначати, чи є пара речень перекладами один одного. Використовуючи цей підхід, ми отримуємо паралельні дані з великих непаралельних корпусів газет китайською, арабською і англійською мовами. Ми здійснюємо оцінку якості отриманих даних, демонструючи, що вони підвищують продуктивність сучасної статистичної системи машинного перекладу. Ми також показуємо, що можна створити якісну систему машинного перекладу з нуля, маючи дуже малий за обсягом паралельний корпус (100 000 слів) та використовуючи великі непаралельні корпуси. Отже, наш метод можна ефективно застосовувати для мовних пар, для яких наявна дуже обмежена кількість ресурсів.

Переклад Т. Павлущенко, М. Погребної

Mariño, J. B. N-gram-based Machine Translation [Машинний переклад на основі N-грамів] / José B. Mariño, Rafael E. Banchs, Josep M. Crego, Adrià de Gispert, Patrik Lambert, José A. R. Fonollosa, Marta R. Costa-jussà // Computational linguistics. – 2006. – Vol. 32. – No. 4. – Pages 527–549. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2006.32.4.527#.WIIOM33sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2006.32.4.527

У цій статті детально описано статистичний машинний переклад із застосуванням n-грамів. Цей підхід полягає у логлінійній комбінації моделі перекладу на основі n-грамів двомовних одиниць, які називають кортежами, з чотирма особливими функціями. Якість перекладу, яка є однією з найкращих сьогодні, продемонстровано за допомогою перекладів пленарних засідань Європейського парламенту (European Parliament Plenary Sessions, скор. EPPS) з іспанської на англійську та з англійської на іспанську.

Переклад Т. Павлущенко, М. Погребної

Ueffing, N. Word-Level Confidence Estimation for Machine Translation [Оцінка достовірності машинного перекладу на рівні слів] / Nicola Ueffing, Hermann Ney // Computational linguistics. – 2007. – Vol. 33. – No. 1. – Pages 9–40. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2007.33.1.9#.WIIOqn3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2007.33.1.9

В цій статті представлено та оцінено декілька різних показників

достовірності машинного перекладу на рівні слів. Ці показники використовуються для маркування кожного слова у автоматично створеному тексті перекладу як правильного чи неправильного. Всі підходи до оцінювання достовірності, представлені в цій роботі, базуються на ймовірності наступного слова. Ми пропонуємо та порівнюємо різні концепції ймовірності наступного слова, а також різні способи їх розрахунку. Їх можна розділити на дві категорії: системні методи, які досліджують дані, надані системою перекладу, що генерує переклади, та прямі методи, які не залежать від системи перекладу. Системні методи використовують вихідні дані системи, такі як графи слів або списки N-кращих гіпотез. Ймовірність наступного слова визначаєтья як сума ймовірностей речень у можливому варіанті перекладу, що містить дане слово. Прямі показники достовірності спираються на інші джерела інформації, такі як словники слів або словосполучень. Їх можна також застосовувати до перекладів, виконаних нестатистичними системами машинного перекладу.

У статті представлено результати експериментального оцінювання різних показників достовірності у різних перекладацьких завданнях та для декількох мовних пар. Крім того, досліджується застосування показників достовірності для перевірки гіпотез перекладу.

Переклад Т. Павлущенко, М. Погребної

Chiang, D. Hierarchical Phrase-Based Translation [Переклад на основі складних словосполучень] / David Chiang // Computational linguistics. – 2007. – Vol. 33. – No. 2. – Pages 201–228. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2007.33.2.201#.WIIO6n3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2007.33.2.201

Ми представляємо статистичну модель машинного перекладу, яка використовує *складні словосполучення*, що складаються з простих словосполучень. Формально вона є синхронною контекстно-вільною моделлю, але її отримують із паралельного тексту без синтаксичної розмітки. Тому її можна розглядати як сукупність основних ідей як перекладу на основі синтаксису, так і перекладу на основі словосполучень. Ми детально описуємо методи тренування та декодування нашої системи та оцінюємо її за критеріями швидкості й точності перекладу. Застосувавши алгоритм BLEU для визначення точності перекладу, ми виявили, що наша система працює значно краще, ніж «Система на основі алгоритму вирівнювання», найновіша система перекладу на основі словосполучень.

1. *Переклад К. Погорєлова, М. Погребної*

Fraser, A. Measuring Word Alignment Quality for Statistical Machine Translation [Визначення якості вирівнювання слів у статистичному

машинному перекладі] / Alexander Fraser, Daniel Marcu // Computational linguistics. – 2007. – Vol. 33. – No. 3. – Pages 293–303. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2007.33.3.293#.WIIPsH3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2007.33.3.293

Автоматичне вирівнювання слів відіграє важливу роль у статистичному машинному перекладі. На жаль, зв'язок між якістю вирівнювання та якістю статистичного машинного перекладу поки ще не є повністю вивченим. В останніх дослідженнях проблема вирівнювання часто розглядалась окремо від перекладацького завдання і висловлені в них припущення щодо визначення якості вирівнювання для машинного перекладу, як виявилось, не підтвердились. Зокрема, у жодній з десяти статей, опублікованих за останні 5 років, не було сказано, що значне зменшення частоти помилок вирівнювання (alignment error rate, скор. AER) призводить до суттєвого покращення якості перекладу. У даній статті подано огляд досліджень та запропоновано такі способи визначення якості вирівнювання, які дозволяють передбачити якість статистичного машинного перекладу.

Переклад К. Погорєлова, М. Погребної

Barrachina, S. Statistical Approaches to Computer-Assisted Translation [Статистичні підходи до автоматизованого перекладу] / Sergio Barrachina, Oliver Bender, Francisco Casacuberta, Jorge Civera, Elsa Cubel, Shahram Khadivi, Antonio Lagarda, Hermann Ney, Jesús Tomás, Enrique Vidal, Juan-Miguel Vilar // Computational linguistics. – 2009. – Vol. 35. – No. 1. – Pages 3–28. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2008.07-055-R2-06-29#.WIIP_33sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2008.07-055-R2-06-29

Сучасні системи машинного перекладу (МП) все ще є недосконалими. На практиці переклади, виконані такими системами, потрібно редагувати. Для підвищення продуктивності процесу перекладу (машинний переклад плюс ручне редагування) можна включити у процес перекладу роботу людини-редактора, таким чином переходячи від машинного до автоматизованого перекладу. Така модель передбачає ітеративний процес, у якому робота людини-перекладача є частиною циклу: при кожній ітерації людина перевіряє префікс перекладу (приймає або редагує його), а система вираховує найкращий (чи *n*-найкращий) гіпотетичний суфікс перекладу для доповнення цього префіксу. Успішною моделлю машинного перекладу є так званий статистичний машинний переклад (або розпізнавання паттернів). Цікаво, що при цьому підході адаптація систем машинного перекладу до інтерактивного сценарію впливає здебільшого на процес пошуку, дозволяючи активно використовувати ефективні методи і моделі. У цій статті обговорюється

використання алгоритмів вирівнювання, моделей на основі словосполучень та стохастичних кінцевих перетворювачів для розробки систем автоматизованого перекладу. Ці системи було використано у Європейському проекті (TransType2) для виконання двох реальних завдань: перекладу інструкцій до принтерів; інструкцій і перекладу *Бюлетеня Європейського Союзу*. У кожному завданні здійснювався двосторонній переклад між трьома парами мов: англійська-іспанська, англійська-німецька та англійська-французька.

Переклад М. Погребної

Wang, W. Re-structuring, Re-labeling, and Re-aligning for Syntax-Based Machine Translation [Повторні структуризація, розмітка та вирівнювання у машинному перекладі на основі синтаксису] / Wei Wang, Jonathan May, Kevin Knight, Daniel Marcu // Computational linguistics. – 2010. – Vol. 36. – No. 2. – Pages 247–277. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2010.36.2.09054#.WIIQbH3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2010.36.2.09054

У даній статті показано, що структура двомовного матеріалу, опрацьованого за допомогою стандартного програмного забезпечення для синтаксичного аналізу та вирівнювання, не є оптимальною для тренування систем статистичного машинного перекладу (СМП) на основі синтаксису. Ми представляємо три модифікації даних для тренування МП з метою підвищення точності сучасних систем МП на основі синтаксису: повторна структуризація змінює синтаксичну структуру навчальних дерев залежності і робить можливим повторне використання простих структур, повторне анотування вносить зміни до поміт для розширення умов використання правил, а повторне вирівнювання уніфікує вирівнювання слів у реченнях, видаляє неправильні вирівнювання та уточнює правильні. За допомогою EM-алгоритму удосконалюються структури, поміти та вирівнювання слів. Ми показуємо, що кожний окремий метод сприяє підвищенню ефективності за оцінкою BLEU, але ми також демонструємо, що шляхом поєднання цих методів можна досягти значно більшого підвищення ефективності. Ми повідомляємо про підвищення на 1.48 показника BLEU у наборі еталонів NIST08 порівняно із загальним рівнем китайсько-англійського перекладу.

Переклад К. Погорєлова, М. Погребної

Ravi, S. Does GIZA++ Make Search Errors? [Чи робить GIZA++ пошукові помилки?] / Sujith Ravi, Kevin Knight // Computational linguistics. – 2010. – Vol. 36. – No. 3. – Pages 295–302. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli_a_00008#.WIIRvn3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli_a_00008

Вирівнювання на рівні слів є важливою процедурою у статистичному машинному перекладі (СМП). Найпопулярніший на сьогодні алгоритм вирівнювання на рівні слів, який був використаний у додатках GIZA [Al-Onaizan et al., 1999] і GIZA++ [Och and Ney 2003] і застосований майже в кожному проекті СМП, було запропоновано у статті П. Брауна та ін. [Brown et al., 1993]. У цій статті досліджується, чи робить вказаний алгоритм помилки при обчисленні вирівнювань Вітебрі, тобто чи обчислює він вирівнювання, які згідно навченої моделі є напівоптимальними.

Переклад В. Коломієць

Liu, Y. Discriminative Word Alignment by Linear Modeling [Дискримінативне вирівнювання слів на основі лінійного моделювання] / Yang Liu, Qun Liu, Shouxun Lin // Computational linguistics. – 2010. – Vol. 36. – No. 3. – Pages 303–339. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli_a_00001#.WIISO33sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli_a_00001

Важливу роль у багатьох задачах обробки природної мови відіграє вирівнювання на рівні слів, оскільки воно показує відповідність між словами у паралельних текстах. Хоча для вирівнювання великих двомовних корпусів широко використовуються породжувальні моделі, їх важко розширити для включення додаткової корисної лінгвістичної інформації. У статті представлено дискримінативний підхід до вирівнювання на рівні слів на основі лінійної моделі. В рамках цього підходу всі джерела інформації розглядаються як функції-ознаки, які залежать від речення вихідною мовою, від речення цільовою мовою і вирівнювання між ними. Описано багато функцій, які могли б забезпечити симетричні вирівнювання. Запропоновану модель можна легко розширити і оптимізувати в плані безпосередньо показників оцінювання. Модель забезпечила високу якість вирівнювання на трьох конкурсних завданнях вирівнювання для п'яти мовних пар з різним ступенем схожості і доступності ресурсів. Крім того, показано, що наш підхід підвищує якість перекладу різних статистичних систем машинного перекладу.

Переклад О. Мартинюк, М. Драчової

Graça, V. J. Learning Tractable Word Alignment Models with Complex Constraints [Навчання гнучких моделей вирівнювання слів із комплексними обмеженнями] / João V. Graça, Kuzman Ganchev, Ben Taskar // Computational linguistics. – 2010. – Vol. 36. – No. 3. – Pages 481–504. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli_a_00007#.WIISZn3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli_a_00007

Вирівнювання двомовних текстів на рівні слів є важливим ресурсом для зростаючої кількості завдань. Імовірнісні моделі вирівнювання на рівні слів забезпечують основоположний компроміс між різноманітністю визначених обмежень і кореляцій та ефективністю і гнучкістю логічних виведень. Для того щоб включити до імовірнісних моделей на етапі машинного навчання комплексні обмеження, не міняючи ефективності базової моделі, у статті використано методологію апостеріорної регуляризації (J. V. Graça, K. Ganchev, and B. Taskar, 2007). Велику увагу приділено простій і гнучкій прихованій марківській моделі, а також описано ефективний алгоритм навчання для включення наближеної бієктивності і обмежень симетричності. Моделі, обчислені з цими обмеженнями, дають значне підвищення продуктивності, про що свідчать показники і повноти, і точності анотованих вручну вирівнювань для шести пар мов. Також описано експерименти з двома різними завданнями, які потребують вирівнювання на рівні слів: машинним перекладом на основі словосполучень і з перенесенням синтаксису, і показано обнадійливе покращення результатів у порівнянні з традиційними методами.

Переклад О. Мартинюк, М. Драчової

de Gispert, A. Hierarchical Phrase-Based Translation with Weighted Finite-State Transducers and Shallow-n Grammars [Ієрархічна модель статистичного машинного перекладу зі зваженими перетворювачами із скінченним числом станів і поверховими-n граматиками] / Adrià de Gispert, Gonzalo Iglesias, Graeme Blackwood, Eduardo R. Banga, William Byrne // Computational linguistics. – 2010. – Vol. 36. – No. 3. – Pages 505-533. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli_a_00006#.WIIS-n3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli_a_00006

У статті описано HiFST, декодер на решітковій базі для ієрархічного перекладу і вирівнювання на основі словосполучень. Декодер використовується у стандартних операціях зваженого скінченного перетворювача (*англ.* Weighted Finite-State Transducer, *скор.* WFST) як альтернатива добре відомій процедурі скорочення куба. З'ясовано, що використання WFST замість списків k-кращих зменшує обрізку в пошуках перекладу, результатом чого є зменшення кількості пошукових помилок, краща оптимізація параметрів і поліпшення якості перекладу. Пряма генерація решіток перенесень мовою перекладу може покращити наступні процедури переоцінки, даючи додаткові переваги, якщо застосовуються універсальні мовні моделі і розкодування з мінімальним байесовським ризиком. У статті також описано, як контролювати величину зони пошуку, задану правилами ієрархії. Показано, що поверхові-n граматики, конкатенація нижнього порядку та інші пошукові обмеження можуть

допомогти налаштувати потужність системи перекладу для конкретних пар мов.

Переклад В. Коломієць

Xiong, D. Linguistically Annotated Reordering: Evaluation and Analysis Grammars [Лінгвістично анотоване переупорядкування: граматика оцінки і аналізу] / Deyi Xiong, Min Zhang, Aiti Aw, Haizhou Li // Computational linguistics. – 2010. – Vol. 36. – No. 3. – Pages 535-568. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli_a_00009#.WIIksH3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli_a_00009

Важлива роль у статистичному машинному перекладі на основі словосполучень належить лінгвістичним знанням. Для ефективного застосування лінгвістичних знань у переупорядкуванні словосполучень запропоновано новий підхід: лінгвістично анотоване переупорядкування (Linguistically Annotated Reordering, *скор.* LAR). У LAR було створено апаратні ієрархічні скелети, вузли яких наповнювались під час перекладу програмно-сумісними лінгвістичними знаннями з вихідних синтаксичних дерев. Результати експерименту із застосуванням широкомасштабних навчальних даних свідчать, що LAR співставна з переупорядкуванням на основі межових слів (boundary word-based reordering, *скор.* BWR) (D. Xiong, Q. Liu and S. Lin, 2006), яке є дуже ефективним лексикалізованим методом переупорядкування. Комбінація BWR і LAR дозволяє отримати додаткові дані для переупорядкування словосполучень, які разом значно покращують показники BLEU.

Для того щоб глибше зрозуміти роль лінгвістичних знань LAR у переупорядкуванні словосполучень, для автоматичного визначення руху складників у еталонному і машинному перекладах застосовано метод аналізу на основі синтаксису і узагальнено синтаксичні закономірності переупорядкування, визначені моделями переупорядкування. За допомогою запропонованого методу здійснено порівняльний аналіз, який не тільки проливає світло на роль лінгвістичних знань у переупорядкуванні словосполучень, але також виявляє нові проблеми у їх переупорядкуванні.

Переклад В. Коломієць

Riezler, S. Query Rewriting Using Monolingual Statistical Machine Translation [Переписування пошукового запиту з використанням статистичного машинного перекладу на основі одномовного корпусу] / Stefan Riezler, Yi Liu // Computational linguistics. – 2010. – Vol. 36. – No. 3. – Pages 569–582. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli_a_00010#.WIITnH3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli_a_00010

Довгі пошукові запити часто призводять до низької повноти веб-пошуку через одночасне знаходження відповідних пар термінів. Ймовірність знаходження відповідників у релевантних документах можна збільшити, замінивши терміни у запиті новими термінами зі схожими статистичними характеристиками. Ми порівнюємо методи, які передбачають використання журналів запитів користувача для того, щоб навчитися переписувати пошукові запити, використовуючи терміни з текстів веб-документів. Ми демонструємо, що найкращих результатів можна досягти, застосувавши підхід, який полягає у заповненні "лексичної прогалини" між запитами та веб-документами шляхом перекладу запитів із вхідної мови створених користувачем запитів на вихідну мову веб-документів. Ми тренуємо новітню модель статистичного машинного перекладу на парах запит-текстовий фрагмент з журналів запитів користувача і отримуємо нові терміни з переписаних запитів, перекладених одномовною системою перекладу. За допомогою зовнішньої оцінки результатів реального пошуку інформації у Всесвітній Мережі ми показуємо, що поєднання перекладу мови запиту на мову текстового фрагменту із мовою пошукових запитів підвищує ефективність контекстно-залежного розширення пошукового запиту в порівнянні з новітньою моделлю розширення запитів, яку тренують на одних і тих даних з журналу запитів.

Переклад І. Снєгурова

Shen, L. String-to-Dependency Statistical Machine Translation [Статистичний машинний переклад «від ланцюжка до залежності»] / Libin Shen, Jinxi Xu, Ralph Weischedel // Computational linguistics. – 2010. – Vol. 36. – No. 4. – Pages 649–671. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli_a_00015#.WIIt633sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli_a_00015

Ми пропонуємо новий алгоритм статистичного машинного перекладу - «від ланцюжка до залежності». Під час декодування цей алгоритм використовує модель залежностей вихідної мови, щоб скористатись дистантними зв'язками між словами, які не можна змоделювати за допомогою традиційної мовної моделі на основі n-грамів,. Експерименти з використанням наборів еталонів NIST MT06 та MT08 свідчать, що даний алгоритм значно підвищує ефективність сучасної системи машинного перекладу на основі складних словосполучень, яка працює за принципом «від ланцюжка до ланцюжка».

Переклад В. Туз, М. Погребної

Popović, M. Towards Automatic Error Analysis of Machine Translation Output [На шляху до автоматичного аналізу помилок у машинному перекладі] / Maja Popović, Hermann Ney // Computational linguistics. –

2011. – Vol. 37. – No. 4. – Pages 657–688. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00072#.WIIuU33sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00072

Оцінка результатів та аналіз помилок машинного перекладу є важливими, проте складними завданнями. У цій статті пропонується модель автоматичного аналізу і класифікації помилок на основі ідентифікації помилкових слів за допомогою алгоритмів для підрахунку Коефіцієнту помилкових слів (Word Error Rate, скор. WER) та Позиційно-незалежного коефіцієнту помилкових слів (Position-independent word Error Rate, скор. PER), що є найпершою спробою розробити способи автоматичної оцінки, які дозволяють отримати детальнішу інформацію про певні проблеми перекладу. Запропонований підхід дозволяє використовувати різні типи лінгвістичних знань для здійснення класифікації перекладацьких помилок багатьма різними способами. У цій статті йдеться про одну з можливих типологій, яка включає п’ять категорій помилок: неправильна форма слова, неправильний порядок слів, пропущені слова, зайві слова та неправильний добір лексики. Для кожної з категорій ми з’ясовуємо відсоток різних частин мови. Ми порівняли результати автоматичного аналізу помилок з результатами аналізу помилок вручну, щоб дослідити два можливі способи застосування: вирахування відсотка кожного типу помилок у певному перекладі для виявлення основних причин помилок у даній системі перекладу, та порівняння різних перекладів, використовуючи запропоновані категорії помилок для отримання додаткової інформації про переваги та недоліки різних систем та можливості їх удосконалення, а також плюси і мінуси застосованих способів удосконалення. Ми використовували арабсько-англійські переклади онлайн- і радіоновин та китайсько-англійські переклади онлайн-новин, отримані у рамках проекту GALE, записи засідань Європарламенту іспанською та англійською, отримані в ході проекту TC-Star, та три німецько-англійські переклади, отримані під час четвертого Симпозіуму з машинного перекладу. Отримані нами результати добре корелюють із результатами аналізу помилок вручну і всі наші показники, за винятком зайвих слів, добре відображають як різницю між різними версіями однієї і тієї ж системи перекладу, так і між різними системами перекладу.

Переклад Д. Попової, М. Погребної

Baker, K. Use of Modality and Negation in Semantically-Informed Syntactic MT [Використання модальності і заперечення у синтаксичному машинному перекладі з семантичними можливостями] / Kathryn Baker, Michael Bloodgood, Bonnie J. Dorr, Chris Callison-Burch, Nathaniel W. Filardo, Christine Piatko, Lori Levin, Scott Miller // Computational linguistics. – 2012. – Vol. 38. – No. 2. – Pages 411-438. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00099#.WIIu233sS

GA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00099

У статті описано спроби створення ресурсів і систем для машинного перекладу з семантичними можливостями, здійснені у 8-тижневій літній школі з прикладної лінгвістики, організованій у 2009 році Центром передових досліджень лінгвістичних технологій при університеті Джонса Гопкінса. Описано нову схему анотування модальності/заперечення (МЗ), створення (загальнодоступного) лексикону МЗ і двох автоматизованих розмітників МЗ, побудованих за допомогою схеми анотування і лексикону. Наша схема анотування виокремлює три компоненти модальності і заперечення: пусковий елемент (слово, яке виражає модальність або заперечення), ціль (дія, яка асоціюється з модальністю або запереченням) і володільника (суб'єкт модальності). Описано, як було напівавтоматично створено лексикон МЗ, і продемонстровано, що структурований розмітник МЗ дозволяє досягти приблизно 86% точності (в залежності від жанру) анотування стандартного набору даних Консорціуму лінгвістичних даних.

Розроблена схема анотування МЗ застосована до статистичного машинного перекладу за допомогою синтаксичної моделі, яка підтримує додавання семантичних розміток. Синтаксичні мітки, збагачені семантичними помітами, присвоюються деревам розбору у тренувальних текстах на мові перекладу шляхом щеплення дерев. Хоча стаття присвячена модальнності і запереченню, процедура щеплення дерев є загальною і допускає інші типи семантичної інформації. Цю можливість використано шляхом включення міток, створених уже існуючим розмітником, на додаток до елементів МЗ, створених розмітниками, описаними у статті. Створена система значно перевершила лінгвістично примітивну базову модель (Hiero) і досягла найвищих показників, які до цього часу досягалися на тестових даних NIST 2009 на урду і англійській. Цей результат підтверджує гіпотезу про те, що і синтаксична, і семантична інформація можуть поліпшити якість перекладу.

Переклад В.Коломієць

Gildea, D. On the String Translations Produced by Multi Bottom–Up Tree Transducers [Про переклади ланцюжків, виконані висхідними мультидеревовидними перетворювачами] / Daniel Gildea // Computational linguistics. – 2012. – Vol. 38. – No. 3. – Pages 673–693. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00108#.WIIxVn3s SGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00108#.VRF6NvmsU5E

Зазвичай деревовидні перетворювачі визначають як відношення між деревами, але у машинному перекладі на основі синтаксису нас насамперед цікавлять відношення між вхідним і вихідним ланцюжками на верхівках

дерев розбору. Саме з цієї точки зору ми досліджуємо формальну продуктивність висхідних мультидеревовидних перетворювачів.

Переклад Д. Попової, М. Погребної

Lembersky, G. Improving Statistical Machine Translation by Adapting Translation Models to Translationese [Вдосконалення статистичного машинного перекладу шляхом адаптації моделей перекладу до перекладизмів] / Gennadi Lembersky, Noam Ordan, Shuly Wintner // Computational linguistics. – 2013. – Vol. 39. – No. 4. – Pages 999–1023. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00159#.WII01X3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00159#.VRF8U_msU5E

Моделі перекладу для статистичного машинного перекладу створюються на матеріалі паралельних корпусів, які перекладають вручну. Зазвичай уважається, що паралельні тексти є симетричними: напрям перекладу вважається неістотним і тому ігнорується. Однак велика кількість досліджень у галузі перекладознавства свідчить, що напрям перекладу має значення, оскільки мова перекладу (перекладизми) має багато специфічних властивостей. Вже було продемонстровано, що таблиці словосполучень, укладені на основі паралельних корпусів, перекладених в одному напрямку із перекладацьким завданням, перевершують таблиці на основі корпусів, перекладених у зворотному напрямку.

Ми підтверджуємо, що це дійсно так, але одночасно наголошуємо на важливості використання текстів, перекладених у “неправильному” напрямку. При складанні таблиць словосполучень ми використовуємо інформацію про напрям перекладу, адаптуючи модель перекладу до специфіки перекладизмів. Ми досліджуємо два способи адаптації. По-перше, ми створюємо змішану модель шляхом інтерполяції таблиць словосполучень на основі текстів, перекладених у прямому і зворотному напрямах. Ваги для інтерполяції визначаються з допомогою зменшення показника зв’язності. По-друге, ми визначаємо критерії на основі ентропії, що оцінюють відповідність словосполучень мови перекладу перекладизмам, тим самим виключаючи необхідність додавання до паралельного корпусу інформації про напрям перекладу. Ми демонструємо, що використання цих критеріїв у таблицях словосполучень систем статистичного машинного перекладу призводить до стійкого, статистично істотного підвищення якості перекладу.

Переклад І. Снєгурова, М. Погребної

Stymne, S. Generation of Compound Words in Statistical Machine Translation into Compounding Languages [Генерування складених слів у статистичному машинному перекладі на мови, схильні до утворення складених слів] / Sara Stymne, Nicola Cancedda, Lars Ahrenberg //

Computational linguistics. – 2013. – Vol. 39. – No. 4. – Pages 1067–1108. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00162#.WII1p33sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00162#.VSKjk_BR08k

У цій статті ми досліджуємо статистичний машинний переклад (statistical machine translation, скор. SMT) у германських мовах, зосереджуючись на обробці складених слів. Наша основна мета – уможливити генерування нових складених слів, які не зустрічались у тренувальному корпусі. Ми використовуємо метод розділення-об'єднання, при якому складені слова розділяються перед тренуванням системи статистичного машинного перекладу і об'єднуються після виконання перекладу. Такий підхід компенсує нестачу тренувальніх даних, але ризикує розмістити переклади частин складеного слова у неправильній послідовності. Він також вимагає об'єднання частин складеного слова після обробки для відтворення складених слів у вихідному перекладі. Ми пропонуємо спосіб збільшення шансів розміщення перекладених компонентів майбутнього складеного слова у суміжних позиціях та у правильному порядку, і показуємо, що це може привести до підвищення ефективності як при безпосередній перевірці, так і за стандартною системою показників якості перекладу. Ми також пропонуємо кілька нових способів об'єднання частин складеного слова на основі евристики і машинного навчання, які перевершують запропоновані раніше алгоритми. Ці способи генерують нові складені слова і їх переклади з таким же або кращим рівнем якості, що й стандартні системи. Для всіх проміжних завдань ми показуємо, що для перекладу складених слів варто для всіх проміжних завдань включати до процесу перекладу інформацію з урахуванням частин мови.

Переклад М. Погребної

Gimpel, K. Phrase Dependency Machine Translation with Quasi-Synchronous Tree-to-Tree Features [Машинний переклад на основі синтаксису словосполучень з використанням квазісинхронних характеристик «дерево до дерева»] / Kevin Gimpel, Noah A. Smith // Computational linguistics. – 2014. – Vol. 40. – No. 2. – Pages 349–401. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00175#.WII2RH3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00175#.VRF9F_msU5E

Останні дослідження свідчать про істотне підвищення якості перекладу завдяки використанню лінгвістичного синтаксису для мови оригіналу або мови перекладу. Однак з'ясовано, що при застосуванні синтаксису для обох

мов (переклад за принципом "дерево до дерева") розбіжність у синтаксичній будові може завадити отриманню корисних правил (Ding and Palmer, 2005). Сміт і Айснер (Smith and Eisner, 2006) запропонували квазісинхронну граматику – формалізм, який аналізує неізоморфну структуру поступово, використовуючи характеристики замість жорстких обмежень. Хоча ця граматика створена для моделювання перекладу, виявилось, що її гнучкість створює проблеми при розробці реальних систем. У цій статті представлено систему машинного перекладу на основі квазі-синхронної граматики, яка працює за принципом «дерево до дерева». Основа нашого підходу – нова модель, яка поєднує словосполучення і синтаксис залежності, інтегруючи в собі переваги перекладу на основі словосполучень і на основі синтаксису. Ми повідомляємо про статистично істотне підвищення ефективності в порівнянні зі стандартними системами на основі словосполучень у п'ятьох із семи тестових наборів для чотирьох мовних пар. Ми також представляємо обнадійливі попередні результати застосування неконтрольованого синтаксичного аналізу на основі граматики залежності для машинного перекладу на основі синтаксису.

Переклад І. Снєгурова, М. Погребної

Allauzen, C. Pushdown Automata in Statistical Machine Translation [Автомати з магазинною пам’яттю у статистичному машинному перекладі] / Cyril Allauzen, Bill Byrne, Adrià de Gispert, Gonzalo Iglesias, Michael Riley // Computational linguistics. – 2014. – Vol. 40. – No. 3. – Pages 687–723. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00197#.WII2cn3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00197#.VRF-OfmsU5E

У даній статті описується використання автомату з магазинною пам'яттю (МП) у контексті статистичного машинного перекладу та вирівнювання за синхронною контекстно-вільною граматикою. Ми використовуємо МП автомати для компактного представлення набору кандидатів перекладу, згенерованих граматикою для вхідного речення. Ми представляємо алгоритми заміни, склеювання, знаходження найкоротшого шляху та розширення для МП автоматів загального спрямування. Ми описуємо HiPDT – ієрархічний фразовий декодер, який використовує МП автомати та вищезгадані алгоритми. Ми порівнюємо складність цього декодера із декодером на основі кінцевих автоматів і показуємо, що МП автомати забезпечують кращі умови для точного декодування більших синхронних контекстно-вільних граматик і менших мовних моделей. Це підтверджується експериментально, шляхом вирівнювання та перекладу великої кількості текстів з китайської мови на англійську. Для перекладу ми пропонуємо декодування у два кроки, що передбачає залучення на першому кроці простішої мовної моделі для використання результатів аналізу складності

МП автомату. Ми детально вивчаємо умови експерименту та компроміси, при яких HiPDT може досягти сучасного рівня продуктивності у масштабному статистичному машинному перекладі.

Переклад В. Туз, М. Погребної

Durrani, N. The Operation Sequence Model—Combining N-Gram-Based and Phrase-Based Statistical Machine Translation [Модель послідовності операцій — поєднання машинного перекладу на основі N-грамів та фразового статистичного машинного перекладу] / Nadir Durrani, Helmut Schmid, Alexander Fraser, Philipp Koehn, Hinrich Schütze // Computational linguistics. – 2015. – Vol. 41. – No. 2. – Pages 185–214. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_002168 – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00218

У статті запропоновано інноваційну модель машинного перекладу – модель послідовності операцій (МПО), яка поєднує переваги фразового статистичного машинного перекладу і статистичного машинного перекладу (СМП) на основі N-грамів, а також усуває їхні недоліки. Послідовність включає не лише операції перекладу, а й операції зміни порядку слів. Подібно до СМП на основі N-грамів, модель (i) базується на мінімальних одиницях перекладу; (ii) враховує як вихідну, так і цільову інформацію; (iii) не виходить з фразової незалежності; (iv) уникає проблеми, пов’язаної з помилковою сегментацією фраз. Подібно до фразового СМП, модель (i) здатна запам'ятовувати чинники зміни порядку слів, (ii) динамічно будує пошуковий графік, (iii) декодує під час пошуку за допомогою великих одиниць перекладу. Унікальними властивостями моделі є (i) тісне поєднання зміни порядку слів і перекладу, причому рішення щодо перекладу та зміни порядку слів залежать від n попередніх відповідних рішень, (ii) здатність послідовно моделювати локальну та дистантну зміну порядку слів. Використавши BLEU в якості показника точності перекладу, було виявлено, що при виконанні стандартних завдань із перекладу створена система працює значно краще, ніж сучасні системи фразового СМП (Moses і Phrasal) та СМП на основі N-грамів (Ncode). Компонент зміни порядку слів МПО було порівняно з моделлю зміни порядку слів Moses шляхом його інтеграції МПО у систему Moses. Результати показали, що МПО перевершує зміну порядку слів при виконанні всіх перекладацьких завдань. Виявлено, що якість перекладу вдосконалюється далі завдяки машинному навчанню узагальнених представлень з використанням МПО на основі частин мови.

Переклад М. Дубка

Prud'hommeaux, E. Graph-Based Word Alignment for Clinical Language Evaluation [Вирівнювання за словами на основі графів для клінічної оцінки мовлення]/ Emily Prud'hommeaux, Brian Roark // Computational linguistics. – 2015. – Vol. 41. – No. 4. – Pages 549–578. – Режим доступу до

анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00232 – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00233

Одним із останніх застосувань алгоритмів обробки природної мови є аналіз усного мовлення для діагностичних та лікувальних цілей, викликаний потребою у простих, об'єктивних та безконтактних засобах діагностики неврологічних розладів, таких як деменція. Зокрема, компонентом такого діагностичного засобу може стати автоматичний аналіз переказів оповідань, оскільки в осіб з деменцією та її частим попередником, легким когнітивним порушенням, а також з іншими нейродегенеративними і нейроонтогенетичними розладами значно погіршується здатність створювати точні та змістовні розповіді. У статті представлено метод дуже точного автоматичного оцінювання повноти переказу на основі вирівнювання переказу і вихідної розповіді на рівні слів. Запропоновано вдосконалення вирівнювання за словами в існуючих системах машинного перекладу, зокрема інноваційний метод вирівнювання за словами на основі випадкових блукань по графу, який забезпечує кращу точність вирівнювання, ніж стандартні методи вирівнювання за словами на основі максимізації очікування, лише за крихту потрібного для максимізації очікувань часу. Окрім того, характеристики оцінок повноти переказу, отриманих на основі цього високоякісного вирівнювання за словами забезпечують точність діагностичної класифікації, яка не поступається за точністю оцінкам експертів і значно перевершує точність, досягнуту за допомогою показників подібності тексту на рівні реферату, які використовуються в інших галузях опрацювання природної мови. Ці методи можна легко адаптувати до зразків спонтанного мовлення, отриманих за допомогою немовних стимулів, що свідчить про гнучкість та узагальнюваність вказаних методів.

Переклад М. Дубка

Ortiz-Martínez D. Online Learning for Statistical Machine Translation [Онлайн-навчання систем статистичного машинного перекладу] / Daniel Ortiz-Martínez // Computational linguistics. – 2016. – Vol. 42. – No. 1. – Pages 121–161. – Режим доступу до анотації: https://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00244 – Режим доступу до повнотекстової статті: https://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00244

У статті представлено методи онлайн-навчання систем статистичного машинного перекладу (СМП). В галузі СМП зростає кількість великих наборів тренувальних даних, які безперервно збільшуються, наприклад, в контексті бюро перекладів або щоденних перекладів адміністративних розслідувань. Коли до моделей СМП потрібно включити нові знання, використання методів пакетного навчання вимагає дуже тривалого оцінювання усього набору навчання, виконання якого може розтягтися на

кілька днів або тижнів. За допомогою онлайн-навчання нові навчальні приклади можуть оброблятись індивідуально в режимі реального часу. З цією метою подано визначення сучасної моделі СМП, яка складається з набору підмоделей та оновленого і розширеного набору правил для кожної з них. Для перевірки розроблених методів було досліджено дві добре відомі програми СМП, які можна використовувати в бюро перекладів – постредагування та інтерактивного машинного перекладу. В обох випадках система СМП взаємодіє з користувачем для створення високоякісних перекладів. Ці перевірені користувачем переклади можуть бути використані для розширення моделей СМП за допомогою онлайн-навчання. Емпіричні результати в двох розглянутих випадках свідчать про великий вплив частих оновлень на продуктивність системи. Також, шляхом порівняння ефективності системи СМП на основі пакетного навчання і системи онлайн-навчання, було визначено затрати часу на такі оновлення і з'ясовано, що онлайн навчання можливе в режимі реального часу, тоді як пакетне перенавчання швидко стає неможливим через затрати часу. Емпіричні результати також показали, що ефективність онлайн-навчання співставна з результативністю пакетного навчання. Крім того, запропоновані методи можуть навчатися на основі вже оцінених методів або з нуля. У статті також запропоновано дві нові міри прогнозування ефективності онлайн-навчання в завданнях СМП. Представлена тут система перекладу з можливостями онлайн-навчання реалізована в Thot, програмному забезпеченні для СМП з відкритим вихідним кодом.

Переклад А. Шульги

Neubig, G. Optimization for Statistical Machine Translation: A Survey [Оптимізація статистичного машинного перекладу: огляд] / Graham Neubig, Taro Watanabe // Computational linguistics. – 2016. – Vol. 42. – No. 1. – Pages 1–54. – Режим доступу до анотації: https://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00241 – Режим доступу до повнотекстової статті: https://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00241

Важливою складовою практично всіх сучасних систем статистичного машинного перекладу (СМП) є оптимізація параметрів систем з метою підвищення точності перекладу. У статті здійснено огляд досліджень оптимізації статистичного машинного перекладу протягом 12 років: від плідних праць, присвячених розрізнювальним моделям (F. J. Och and H. Ney, 2002) і навчанню з мінімальною вірогідністю помилок (F. J. Och, 2003) до останніх досягнень. Після короткого вступного огляду основ систем статистичного машинного перекладу у статті розглянуто багато різних алгоритмів як для пакетної, так і для онлайн-оптимізації. Зокрема розглянуто збитки, спричинені а прямою мінімізацією помилок, максимальною правдоподібністю, максимальною різницею, мінімізацію ризиків, ранжування тощо, а також прийнятні способи мінімізації цих збитків. Також

проаналізовано останні досягнення, наприклад широкомасштабна оптимізація, нелінійні моделі, предметно-залежна оптимізація, а також вплив на оптимізацію критеріїв оцінки машинного перекладу або пошуку. Нарешті, розглянуто поточний стан оптимізації машинного перекладу і виділено деякі невирішені проблеми, які, ймовірно, стануть об'єктом подальших досліджень в області оптимізації машинного перекладу.

Переклад А. Шульги

Wang P. Source Language Adaptation Approaches for Resource-Poor Machine Translation [Адаптування вихідної мови для машинного перекладу мов з недостатньою кількістю ресурсів] / Pidong Wang, Preslav Nakov, Hwee Tou Ng // Computational linguistics. – 2016. – Vol. 42. – No. 2. – Pages 277–306. – Режим доступу до анотації: https://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00248 – Режим доступу до повнотекстової статті: https://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00248

Більшість мов світу мають недостатньо ресурсів для статистичного машинного перекладу. Проте, багато з них фактично споріднені з якоюсь мовою, яка має велику кількість ресурсів. Отже, у статті запропоновано три нові, незалежні від мови підходи до адаптування вихідної мови до статистичного машинного перекладу з недостатньою кількістю ресурсів. Зокрема, створено вдосконалені статистичні моделі машинного перекладу з бідної на ресурси мови (POOR) на цільову мову (TGT) шляхом адаптування і використання великого паралельного тексту спорідненою мовою, багатою на ресурси (RICH) та тією ж самою цільовою мовою TGT. За основу взято невеликий паралельний текст (POOR-TGT), з якого автоматично видобуто парафрази на рівні слова та фрази, а також різномовні морфологічні варіанти, наявні у багатій та бідній на ресурси мовах. Це дослідження має важливе значення для машинного перекладу з недостатньою кількістю ресурсів, оскільки воно може служити корисним орієнтиром для тих, хто створює системи машинного перекладу для мов з недостатньою кількістю ресурсів.

Експерименти з перекладом з індонезійської/малайської на англійську свідчать, що використання великого, адаптованого, багатого на ресурси паралельного тексту дозволило поліпшити показник BLEU на 7.26 бала у порівнянні з неадаптованим паралельним текстом і на 3.09 бала у порівнянні з вихідним невеликим паралельним текстом. Крім того, поєднання невеликого паралельного тексту (POOR–TGT) з адаптованим паралельним текстом перевершує аналогічні комбінації з неадаптованим паралельним текстом на 1,93-3,25 BLEU балів. У статті також продемонстровано можливість застосування запропонованих підходів до інших мов та інших предметних областей.

Переклад А. Шульги

Bisazza, A. A Survey of Word Reordering in Statistical Machine Translation: Computational Models and Language Phenomena [Дослідження зміни порядку слів у статистичному машинному перекладі: обчислювальні моделі та мовні явища] / Arianna Bisazza, Marcello Federico // Computational linguistics. – 2016. – Vol. 42. – No. 2. – Pages 163–205. – Режим доступу до анотації: https://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00245 – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00245

Зміна порядку слів є одним із найскладніших аспектів статистичного машинного перекладу (СМП) і важливим фактором його якості та ефективності. Незважаючи на велику кількість досліджень, опублікованих до сьогодні, інтерес дослідників до цієї проблеми не зменшився, і жоден окремий метод не видається сильно домінуючим у всіх мовних парах. Натомість, вибір оптимального підходу до нового перекладу все ще, здається, диктується переважно емпіричними випробуваннями.

Для того, щоб зорієнтувати читача в цій великій і складній галузі досліджень, у статті представлено докладне дослідження зміни порядку слів, яке розглядається як задача статистичного моделювання та як явище природної мови. У дослідженні детально описано, як моделюється зміна порядку слів в рамках різноманітних методів СМП на основі стрічок і дерев і в якості окремого завдання, включаючи системні огляди літератури в галузі передового моделювання зміни порядку слів.

Також досліджено, чому одні підходи є більш ефективними, ніж інші, для різних мовних пар. У статті стверджується, що окрім вимірювання глибини зміни порядку слів, важливо розуміти, які види зміни порядку слів відбуваються в певній мовній парі. З цією метою проведено якісний аналіз явищ зміни порядку слів у різних зразках мовних пар на основі великої колекції мовних знань. Показано, що емпіричні результати в літературі по СМП підтверджують гіпотезу про те, що декілька лінгвістичних фактів можуть бути дуже корисними для прогнозування характеристик зміни порядку слів для мовної пари та для вибору методу СМП, який найкраще підходить для них.

Переклад М. Дубка

Deng, D. Translation Divergences in Chinese–English Machine Translation: An Empirical Investigation [Перекладацькі розбіжності в китайсько-англійському машинному перекладі: емпіричне дослідження] / Dun Deng, Nianwen Xue // Computational linguistics. – 2017. – Vol. 43. – No. 3. – Pages 521–565. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00292 – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00292

У статті здійснено емпіричне дослідження перекладацьких розбіжностей між китайською та англійською мовами на базі паралельного банку дерев. З цією метою спочатку було розроблено схему ієрархічного вирівнювання, в якій завдяки вирівнюванню китайських та англійських синтаксичних дерев усуваються надмірності та конфлікти між вирівнюванням слів і синтаксичних дерев, щоб запобігти появі випадкових перекладацьких розбіжностей. Використання цього ієрархічно вирівняного китайсько-англійського паралельного банку дерев HACEPT уможливило напівавтоматичну ідентифікацію та категоризацію перекладацьких розбіжностей між двома мовами і квантифікацію кожного типу перекладацьких розбіжностей. Результати дослідження свідчать, що перекладацькі розбіжності є значно ширшими, ніж описано в попередніх дослідженнях, які переважно базуються на фрагментарних даних та лінгвістичних знаннях. Дистрибуція перекладацьких розбіжностей також показує, що деякі відомі перекладацькі розбіжності, які мотивували попередні дослідження, насправді дуже рідко трапляються серед даних цього дослідження, тоді як інші перекладацькі розбіжності, яким раніше приділялося мало уваги, насправді наявні у великих кількостях. Також показано, що банк дерев HACEPT дозволяє виводити правила перекладу на основі синтаксису, більшість з яких є достатньо конкретними, щоб відобразити перекладацькі розбіжності, і зазначено, що синтаксична розмітка в існуючих банках дерев не є оптимальною для виведення таких правил перекладу. Також описано наслідки цього дослідження для спроб подолати перекладацькі розбіжності шляхом розробки спільних семантичних представлень у різних мовах. Отримані кількісні результати ще раз підтверджують зауваження про те, що хоча деякі перекладацькі розбіжності можна подолати за допомогою семантичних представлень, інші перекладацькі розбіжності допускають кілька трактувань, тому побудова семантичного представлення, яке враховує всі можливі перекладацькі розбіжності, може бути недоцільною.

Переклад М. Дубка

Joty S. Discourse Structure in Machine Translation Evaluation [Структура дискурсу в оцінюванні машинного перекладу] / Shafiq Joty, Francisco Guzmán, Lluís Màrquez, Preslav Nakov // Computational linguistics. – 2017. – Vol. 43. – No. 4. – Pages 683–722. – Режим доступу до анотації: https://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00298 – Режим доступу до повнотекстової статті: https://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00298

У статті досліджуються можливості застосування структури дискурсу на рівні речення для оцінки машинного перекладу. Спочатку визначено критерії подібності з урахуванням структури дискурсу, які, відповідно до теорії риторичної структури (TPC), використовують ядра усіх піддерев для порівняння синтаксичного анотування дискурсу. Далі продемонстровано, що проста лінійна комбінація з цими критеріями може допомогти поліпшити

існуючі метрики оцінки машинного перекладу з точки зору кореляції з судженнями експертів і на сегментному, і на системному рівнях. Це свідчить про те, що інформація про дискурс доповнює інформацію, що використовується багатьма існуючими метриками оцінювання, і через це може братися до уваги при розробці детальніших метрик оцінювання, таких як комбінована метрика DiscoTKparty, переможниця міжнародного семінару з статистичного машинного перекладу WMT-14. Також надано детальний аналіз значущості різних елементів дискурсу та відношень із синтаксичних дерев на основі TPC для оцінювання машинного перекладу. Зокрема, доведено, що (i) всі аспекти синтаксичного маркування на основі TPC дискурсу є значущими, (ii) ядерність є важливішою, ніж тип відношення, і (iii) подібність синтаксичного анотування перекладу в перекладі до еталонного синтаксичного анотування дискурсу позитивно корелює з якістю перекладу.

Переклад А. Шульги

Мультимодальні системи

Bangalore, S. Robust Understanding in Multimodal Interfaces [Робастне розуміння у мультимодальних інтерфейсах] / Srinivas Bangalore, Michael Johnston // Computational linguistics. – 2009. – Vol. 35. – No. 3. – Pages 345–397. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.08-022-R2-06-26#.WIKJm33sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.08-022-R2-06-26

Мультимодальні граматики є ефективним механізмом для швидкої інтеграції і з'ясування можливостей інтерактивних систем, які підтримують одночасне використання методів багатоканального входу. Проте, як і інші підходи на основі створених вручну граматик, мультимодальні граматики можуть не впоратися з неочікуваним, неправильним або непослідовним введенням. У статті показано, як можна застосувати мультимодальну обробку мови за допомогою методу кінцевих станів для підтримки мультимодальних додатків, які об'єднують усне мовлення зі складним письмовим уведенням у вільній формі. Вказаний підхід оцінено за допомогою мультимодальної діалогової системи (MATCH). Досліджено декілька різних прийомів підвищення робастності мультимодальної інтеграції і розуміння, які включають прийоми створення ефективних мовних моделей для розпізнавання мови в умовах нестачі або відсутності мультимодальних тренувальних даних і прийоми робастного мультимодального розуміння на основі методів класифікації, машинного перекладу і редагування послідовностей. Також досліджено використання методів на основі редагування для подолання розбіжностей між потоком жестів і потоком мовлення.

Переклад В. Коломієць

Demir, S. Summarizing Information Graphics Textually [Реферування інформаційної графіки у вигляді тексту] / Seniz Demir, Sandra Carberry, Kathleen F. McCoy // Computational linguistics. – 2012. – Vol. 38. – No. 3. – Pages 527–574. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00091#.WITFO33sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00091

Інформаційна графіка (наприклад гістограми і діаграми) грає важливу роль у багатьох багатомодальних документах. Інформаційна графіка, яка з'являється у загальнодоступних засобах масової інформації, використовується здебільшого для передачі повідомлення, і дизайнер графіки використовує ретельно відібрані комунікативні сигнали, такі як виділення

кольором певних аспектів графіки, щоб виділити це повідомлення. Графіка, комунікативна мета (запланованe повідомлення) якої часто не згадується у супроводжуючому документ тексті, сприяє досягненню загальної мети документа і не може ігноруватися. У статті описано наш метод передачі основного змісту ненаукової інформаційної графіки у вигляді коротких текстових рефератів, які містять заплановане повідомлення і важливі характеристики графіки. У статті зібрано ідеї, запозичені з емпіричних досліджень, щоб визначити, що повинні включати такі реферати нетекстових вихідних даних і як можна видобути із візуального зображення і текстових компонентів графіки інформацію, потрібну для реалізації відібраного змісту. У статті також описано новий висхідний метод генерації для одночасного створення дискурсу і синтаксичних структур тестових рефератів шляхом використання різних характеристик дискурсу, таких як синтаксична складність створених речень і підпорядковані конструкції. Ефективність проведеного дослідження була підтверджена різними оціночними дослідженнями.

Переклад В. Коломієць

Питально-відповідні системи

Mollá, D. Question Answering in Restricted Domains: An Overview [Питально-відповідні системи для обмежених доменів: короткий огляд] / Diego Mollá, José Luis Vicedo // Computational linguistics. – 2007. – Vol. 33. – No. 1. – Pages 41–61. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2007.33.1.41#.WIS5kn3sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2007.33.1.41

Тема автоматизації відповідей на питання цікавила дослідників і розробників з моменту появи найперших програм із штучним інтелектом. З часу створення перших таких програм обчислювальна потужність збільшилась і загальна методологія еволюціонувала від створених вручну баз знань про прості домени до використання колекцій текстів в якості основного джерела знань про складніші домени. Проте залишається багато недосліджених проблем. Ця стаття присвячена використанню обмежених доменів у автоматизованих питально-відповідних системах. Стаття містить історичний огляд систем питання-відповідь для обмежених доменів та огляд сучасних методів та додатків, які використовуються в обмежених доменах. Основна особливість систем питання-відповідь у обмежених доменах – це інтеграція предметно-орієнтованої інформації, яка була розроблена або для системи питання-відповідь, або для інших цілей. У статті досліджуються основні методи застосування такої предметно-орієнтованої інформації.

Переклад А. Синящик

Demner-Fushman, D. Answering Clinical Questions with Knowledge-Based and Statistical Techniques [Відповіді на клінічні питання за допомогою методів на основі знань і статистичних методів] / Dina Demner-Fushman, Jimmy Lin // Computational linguistics. – 2007. – Vol. 33. – No. 1. – Pages 63–103. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2007.33.1.63#.WIS6G33sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2007.33.1.63

Сполучення останніх досягнень у дослідженні питально-відповідних систем і доступності унікальних ресурсів, розроблених спеціально для автоматичного семантичного аналізу медичних текстів надає унікальну можливість дослідження складних діалогових систем в області клінічної медицини. У статті описано систему, створену для задоволення інформаційних потреб лікарів, які практикують доказову медицину. Авторами розроблено низку систем автоматичного видобування знань на основі комбінованих, на основі знань і статистичних, методів для

автоматичного розпізнавання потрібної медичної інформації у анотаціях у MEDLINE. Видобуті фрагменти є вихідними даними для алгоритму, який оцінює значимість цитат відповідно до структурованих моделей інформаційних потреб згідно з принципами доказової медицини. Починаючи з попереднього списку фрагментів, видобутих за допомогою PubMed, розроблена система може перемістити значимі анотації ближче до початку списку і на їх основі згенерувати відповіді безпосередньо на питання лікарів. У статті описано три різні оцінювання: оцінювання точності систем автоматичного видобування знань, оцінювання завдання переранжування фрагментів і, нарешті, оцінювання відповідей двома лікарями. Експерименти, проведені з використанням набору реальних клінічних питань, свідчать, що наша система значно перевершує вже конкурентноздатні показники PubMed.

Переклад В. Коломієць

Hallett, C. Composing Questions through Conceptual Authoring [Створення питань за допомогою розробки концептів] / Catalina Hallett, Donia Scott, Richard Power // Computational linguistics. – 2007. – Vol. 33. – No. 1. – Pages 105–133. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2007.33.1.105#.WIS6233sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2007.33.1.105

У статті описується метод вільного утворення складних питань природною мовою, уникаючи типових помилок безкоштовних текстових запитів. Метод на основі розробки концептів призначений для систем питання-відповідь, у яких надійність і прозорість мають важливе значення, але немає можливості достатньо потренувати користувачів в укладанні питань. Цей сценарій зустрічається в більшості корпоративних доменів, особливо в додатках, які намагаються уникнути ризиків. У статті представлено розроблену авторами експериментальну систему: велика база історій хвороб з онкології з інтерфейсом «питання-відповідь». Продемонстровано, що запропонований метод дозволяє користувачам майже без підготовки успішно створювати складні запити.

Переклад А. Синящик

Surdeanu, M. Learning to Rank Answers to Non-Factoid Questions from Web Collections [Навчання ранжуванню відповідей на питання не про факти з веб-бібліотек] / Mihai Surdeanu, Massimiliano Ciaramita, Hugo Zaragoza // Computational linguistics. – 2011. – Vol. 37. – No. 2. – Pages 351–383. – Режим доступу до анотації: http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00051#.WIEzq33sSGA – Режим доступу до повнотекстової статті: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00051

У цій роботі досліджується використання лінгвістичних ознак для

покращення результативності пошуку, а саме для ранжування відповідей на питання, які не стосуються фактів. У статті показано, що для відбору таких ознак та навчання моделей ранжування, які їх ефективно поєднують, можна використовувати існуючі великі колекції пар «запитання-відповідь» (з соціальних діалогових вебсайтів). Досліджується широке коло типів ознак, деякі з яких вимагають обробки природної мови, а саме: поверхневого розв'язання лексичної омонімії, визначення іменованих сутностей, синтаксичного аналізу та розмітки семантичних ролей. Експерименти свідчать, що лінгвістичні ознаки, за умови їх комбінування, забезпечують значне підвищення точності. Залежно від налаштувань системи було досягнуто покращення від 14% до 21% за показником середнього оберненого рангу і Precision@1, що є одним з найпереконливіших доказів того, що складні лінгвістичні ознаки, такі як значення слова та семантичні ролі, можуть мати суттєвий вплив на широкомасштабний інформаційний пошук.

Переклад О. Мартинюк, М. Погребної

Jansen P. Framing QA as Building and Ranking Intersentence Answer Justifications [Розробка питально-відповідної системи як побудова та ранжування обґрунтування відповідей на рівні тексту] / Peter Jansen, Rebecca Sharp, Mihai Surdeanu, Peter Clark // Computational linguistics. – 2017. – Vol. 43. – No. 2. – Pages 407–449. – Режим доступу до анотації: https://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00287 – Режим доступу до повнотекстової статті: https://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00287

У статті описано питально-відповідний підхід до стандартизованого іспиту з природничих наук, який не тільки визначає правильні відповіді, але й надає переконливі, зрозумілі людині обґрунтування їх правильності. Спочатку цей метод визначає фактичну інформацію, необхідну для відповіді на запитання, використовуючи психолінгвістичні параметри конкретності. Потім ця інформаційна потреба використовується для побудови обґрунтувань відповідей шляхом об'єднання великої кількості речень з різних баз знань за допомогою синтаксичної і лексичної інформації. Після цього відповіді та їх обґрунтування спільно оцінюються за допомогою перцептрону, який розглядає якість обґрунтування як приховану змінну. Для оцінювання якості методу було використано 1000 завдань множинного вибору з екзамену з природничих наук у початковій школі. Було емпірично доведено, що описаний метод дає кращі результати, ніж кілька сильних контрольних систем, зокрема нейромережеві підходи. Найкраща конфігурація правильно відповідає на 44% питань, і найкращі обгрунтування 57% цих правильних відповідей містять переконливі, зрозумілі людині обґрунтування, які пояснюють умовиводи, потрібні для того, щоб дати правильну відповідь. У статті детально описано якість обґрунтувань як у запропонованому методі, так і в сильному контрольному варіанті, а також показано, що ключовим

компонентом у задоволенні інформаційної потреби у складних питаннях є об’єднання інформації.

Переклад А. Шульги

Покажчик назв статей журналу

Computational Linguistics

(2000-2017 pp.)

Автомати з магазинною пам’яттю у статистичному машинному перекладі [Pushdown Automata in Statistical Machine Translation] 40 (3)

Автоматична категоризація текстів за жанром і автором [Automatic Text Categorization in Terms of Genre and Author] 26 (4)

Автоматична класифікація дієслів на основі статистичного розподілу структури аргументів [Automatic Verb Classification Based on Statistical Distributions of Argument Structure] 27 (3)

Автоматична оцінка змісту автоматично сформованих рефератів без золотого стандарту [Automatically Assessing Machine Summary Content Without a Gold Standard] 39 (2)

Автоматична оцінка упорядкування інформації: тау Кендала [Automatic Evaluation of Information Ordering: Kendall's Tau] 32 (4)

Автоматична розмітка семантичних ролей [Automatic Labeling of Semantic Roles] 28 (3)

Автоматичне адаптування розмітки [Automatic Adaptation of Annotations] 41 (1)

Автоматичне визначення вихідних компонентів лексичних стягнень у англійській мові [Automatically Identifying the Source Words of Lexical Blends in English] 36 (1)

Автоматичне виявлення відношень частина-ціле [Automatic Discovery of Part-Whole Relations] 32 (1)

Автоматичне генерування референційних виразів: огляд [Computational Generation of Referring Expressions: A Survey] 38 (1)

Автоматичне зв’язування [Binding Machines] 28 (1)

Автоматичне зв’язування веб-каталогів зі значеннями слів [Automatic Association of Web Directories with Word Senses] 29 (3)

Автоматичне реферування багатосторонніх діалогів різних жанрів із відкритою тематикою [Automatic Summarization of Open-Domain Multiparty Dialogues in Diverse Genres] 28 (4)

Автоматичне створення двомовного словника емоційно-оціночної лексики шляхом використання двомовного маркування графів слів [Cross-lingual Sentiment Lexicon Learning With Bilingual Word Graph Label Propagation] 41 (1)

Автоматичний відбір синтаксичних дерев, створених аналізатором на основі граматики HPSG, для побудови банку дерев [Automatic Selection of HPSG-Parsed Sentences for Treebank Construction] 40 (3)

Автоматичний синтаксичний аналіз глибинних структур залежностей на основі переходів [Transition-Based Parsing for Deep Dependency Structures] 42 (3)

Автоматичний синтаксичний аналіз мов із розвиненою морфологією: передмова до спеціального випуску [Parsing Morphologically Rich Languages: Introduction to the Special Issue] 39 (1)
Автоматичний синтаксичний аналіз структур аргументації в есе-переконаннях [Parsing Argumentation Structures in Persuasive Essays] 43 (3)

Адаптивна генерація у діалогових системах за допомогою динамічного моделювання користувача [Adaptive Generation in Dialogue Systems Using Dynamic User Modeling] 40 (4)

Адаптування вихідної мови для машинного перекладу мов з недостатньою кількістю ресурсів [Source Language Adaptation Approaches for Resource-Poor Machine Translation] 42 (2)

Адаптування до помилок не носіїв мови з мінімальним залученням учителя [Adapting to Learner Errors with Minimal Supervision] 43 (4)

Алгоритм для сегментування китайських текстів на слова на основі стиснення [A Compression-based Algorithm for Chinese Word Segmentation] 26 (3)

Алгоритм неконтрольованого встановлення переважаючих значень слова [Unsupervised Acquisition of Predominant Word Senses] 33 (4)

Алгоритм розв’язання анафори у текстах іспанською мовою [An Algorithm for Anaphora Resolution in Spanish Texts] 27 (4)

Алгоритм-ескіз для оцінки двосторонніх і багатосторонніх асоціацій [A Sketch Algorithm for Estimating Two-Way and Multi-Way Associations] 33 (3)

Алгоритми детермінованого поетапного синтаксичного аналізу на основі дерев залежностей [Algorithms for Deterministic Incremental Dependency Parsing] 34 (4)

Алгоритми навчання перекладу на основі залежностей у вигляді наборів скінченних перетворювачів [Learning Dependency Translation Models as Collections of Finite-State Head Transducers] 26 (1)

Аналіз і інтеграція синтаксичних аналізаторів залежностей [Analyzing and Integrating Dependency Parsers] 37 (1)

Аналіз і класифікація контрольованих природних мов [A Survey and Classification of Controlled Natural Languages] 40 (1)

Аналіз і передбачення виправлень в усномовленнєвих діалогових системах [Characterizing and Predicting Corrections in Spoken Dialogue Systems] 32 (3)

Аналіз риторичної структури необмежених текстів: поверхневий підхід [The Rhetorical Parsing of Unrestricted Texts: A Surface-based Approach] 26 (3)

Анафора і структура дискурсу [Anaphora and Discourse Structure] 29 (4)

Анотування і автоматичне визначення тривалості подій у текстах [Annotating and Learning Event Durations in Text] 37 (4)

Анотування семантичних ролей: вступ до спеціального випуску [Semantic Role Labeling: An Introduction to the Special Issue] 34 (2)

Анотування семантичних ролей китайських присудків [Labeling Chinese Predicates with Semantic Roles] 34 (2)

Асимптотична модель співвідношення гапакс/вокабулярій у англійській мові [An Asymptotic Model for the English Hapax/Vocabulary Ratio] 36 (4)

Багатокомпонентні граматики об'єднання дерев із спільними вузлами на початковому дереві [Tree-Local Multicomponent Tree-Adjoining Grammars with Shared Nodes] 31 (2)

Багатомовне опрацювання метафор: експерименти з навчанням з мінімальним залученням учителя і без учителя [Multilingual Metaphor Processing: Experiments with Semi-Supervised and Unsupervised Learning] 43 (1)

Багатомовний об'єднаний синтаксичний аналіз синтаксичних і семантичних залежностей за допомогою моделі з прихованою змінною [Multilingual Joint Parsing of Syntactic and Semantic Dependencies with a Latent Variable Model] 39 (4)

Багаторівнева нелінійна морфологія з використанням багатострічкових скінченних автоматів: тематичне дослідження на основі сирійської та арабської мов [Multitiered Nonlinear Morphology Using Multitape Finite Automata: A Case Study on Syriac and Arabic] 26 (1)
Багатостратегійний підхід до покращення вимови за аналогією [A Multistrategy Approach to Improving Pronunciation by Analogy] 26 (2)
Банк пропозицій: анотований корпус семантичних ролей [The Proposition Bank: An Annotated Corpus of Semantic Roles] 31 (1)
Бінаризація синхронних контекстно-вільних граматик [Binarization of Synchronous Context-Free Grammars] 35 (4)

Вбудовування моделей статистичного перекладу на основі інтернет-технологій у пошук інформації різними мовами [Embedding Web-Based Statistical Translation Models in Cross-Language Information Retrieval] 29 (3)

Вдосконалення статистичного машинного перекладу шляхом адаптації моделей перекладу до перекладизмів [Improving Statistical Machine Translation by Adapting Translation Models to Translationese] 39 (4)

Вершинні статистичні моделі синтаксичного аналізу природної мови [Head-Driven Statistical Models for Natural Language Parsing] 29 (4)

Взаємодія баз даних у знятті лексичної неоднозначності [The Interaction of Knowledge Sources in Word Sense Disambiguation] 27 (3)

Ви впевнені, що це правда? Оцінка ступеня достовірності подій у тексті [Are You Sure That This Happened? Assessing the Factuality Degree of Events in Text] 38 (2)

Виведення семантичних ролей на основі схожості шляхом розбиття графа [Similarity-Driven Semantic Role Induction via Graph Partitioning] 40 (3)

Вивідні центри, центровані переходи і поняття когерентності [Inferable Centers, Centering Transitions, and the Notion of Coherence] 30 (2)

Видобування імпліцитних аргументів з порівняльних текстів: метод і його застосування [Inducing Implicit Arguments from Comparable Texts: A Framework and Its Applications] 41 (4)

Видобування онтологій предметних областей із сховищ документів і спеціалізованих веб-сайтів [Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites] 30 (2)

Видобування паралельних корпусів з мікроблогів Sina Weibo та Twitter [Mining Parallel Corpora from Sina Weibo and Twitter] 42 (2)

Визначення значення слова у контексті [Measuring Word Meaning in Context] 39 (3)

Визначення класу дієслова за допомогою інформативних пріоритетів [Verb Class Disambiguation Using Informative Priors] 30 (1)

Визначення коренів у семітських мовах: машинне навчання з використанням лінгвістичних правил [Identifying Semitic Roots: Machine Learning with Linguistic Constraints] 34 (3)

Визначення якості вирівнювання слів у статистичному машинному перекладі [Measuring Word Alignment Quality for Statistical Machine Translation] 33 (3)

Використання вирівнювання слів і речень у автоматичному реферуванні документів [Induction of Word and Phrase Alignments for Automatic Document Summarization] 31 (4)

Використання властивих людині обмежень обсягу пам’яті у синтаксичному аналізаторі з широким покриттям [Broad-Coverage Parsing Using Human-Like Memory Constraints] 36 (1)

Використання Всесвітньої мережі для отримання частот прихованих біграм [Using the Web to Obtain Frequencies for Unseen Bigrams] 29 (3)

Використання двомовного самоналаштування для вирішення багатозначності при перекладі слів [Word Translation Disambiguation Using Bilingual Bootstrapping] 30 (1)

Використання епсилон-переходів у створенні підмножин [Treatment of Epsilon Moves in Subset Construction] 26 (1)

Використання методу випадкових блукань у вирішенні проблеми лексичної багатозначності на основі знань [Random Walks for Knowledge-Based Word Sense Disambiguation] 40 (1)

Використання модальності і заперечення у синтаксичному машинному перекладі з семантичними можливостями [Use of Modality and Negation in Semantically-Informed Syntactic MT] 38 (2)

Використання морфо-синтаксичної інформації у статистичному машинному перекладі з недостатньо великим корпусом для тренування [Statistical Machine Translation with Scarce Resources Using Morpho-syntactic Information] 30 (2)

Використання обмежень у сполучуваності та субкатегорійних фреймів у синтаксичних аналізаторах на основі граматики залежностей [Integrating Selectional Constraints and Subcategorization Frames in a Dependency Parser] 42 (1)

Використання прихованої марківської моделі для розбиття речень рефератів, написаних людиною [Using Hidden Markov Modeling to Decompose Human-Written Summaries] 28 (4)

Використання розмітки семантичних ролей у знятті прийменникової омонімії [Exploiting Semantic Role Resources for Preposition Disambiguation] 35 (2)

Використання списків суфіксів для обчислення частоти термінів і частоти документів у всіх підрядках корпусу [Using Suffix Arrays to Compute Term Frequency and Document Frequency for All Substrings in a Corpus] 27 (1)

Використання теорії прив’язування і пристосування для розв’язання анафори і проекції пресупозиції [Implementing the Binding and Accommodation Theory for Anaphora Resolution and Presupposition Projection] 29 (2)

Виокремлення слів із найнижчою частотністю: труднощі та можливості [Extracting the Lowest-Frequency Words: Pitfalls and Possibilities] 26 (3)

Вирівнювання за словами на основі графів для клінічної оцінки мовлення [Graph-Based Word Alignment for Clinical Language Evaluation] 41 (3)

Вирівнювання упакованих дерев залежностей: теорія композиції для дистрибутивної семантики [Aligning Packed Dependency Trees: A Theory of Composition for Distributional Semantics] 42 (4)

Виявлення відношень логічного слідування шляхом оптимізації загальної структури графів [Learning Entailment Relations by Global Graph Structure Optimization] 38 (1)

Виявлення мовних показників суб’єктивності [Learning Subjective Language] 30 (3)

Виявлення одночасної появи слів: гнучкий підхід до лексичної дистрибутивної схожості [Co-occurrence Retrieval: A Flexible Framework for Lexical Distributional Similarity] 31 (4)

Від показника узгодженості між розмітниками до шумових моделей [From Annotator Agreement to Noise Models] 35 (4)

Відбір фрагментів дерев із лісів [Sampling Tree Fragments from Forests] 40 (1)

Відмінності між людьми і вибір лексики [Human Variation and Lexical Choice] 28 (4)

Відповіді на клінічні питання за допомогою методів на основі знань і статистичних методів [Answering Clinical Questions with Knowledge-Based and Statistical Techniques] 33 (1)

Відстані Левенштейна нездатні точно визначати ступені спорідненості мов [Levenshtein Distances Fail to Identify Language Relationships Accurately] 37 (4)

Відсутність єдиного підходу: пропозиція узгодити звітність про результати встановлення референції займенників [The Uncommon Denominator: A Proposal for Consistent Reporting of Pronoun Resolution Results] 27 (4)

Візки з фруктами: домен і корпус для досліджень у діалогових системах і психолінгвістиці [Fruit Carts: A Domain and Corpus for Research in Dialogue Systems and Psycholinguistics] 38 (3)

Вірогіднісна генерація діалогічних текстів за допомогою факторних моделей мови [Stochastic Language Generation in Dialogue using Factored Language Models] 40 (4)

Вірогіднісне пояснення логічної метонімії [A Probabilistic Account of Logical Metonymy] 29 (2)

Вплив когнітивного зусилля на розуміння надмірно конкретизованих описів [Effects of Cognitive Effort on the Resolution of Overspecified Descriptions] 43 (2)

Все змішалося? Пошук оптимального набору параметрів для прогнозування загальної легкочитаності та його застосування до англійської і нідерландської мов [All Mixed Up? Finding the Optimal Feature Set for General Readability Prediction and Its Application to English and Dutch] 42 (3)

Встановлення авторства за допомогою тематичних моделей [Authorship Attribution with Topic Models] 40 (2)

Встановлення кореферентності іменних груп на основі машинного навчання [A Machine Learning Approach to Coreference Resolution of Noun Phrases] 27 (4)

Вступ до спеціального випуску, присвяченого методам скінченних станів у обробці природної мови [Introduction to the Special Issue on Finite-State Methods in NLP] 26 (1)

Вступ до спеціального випуску, присвяченого реферуванню [Introduction to the Special Issue on Summarization] 28 (4)

Вступне слово до спеціального випуску, присвяченого Всесвітній мережі як корпусу текстів [Introduction to the Special Issue on the Web as Corpus] 29 (3)

Гамма (γ) уніфікованого й цілісного методу визначення та вирівнювання узгодженості між розмітниками [The Unified and Holistic Method Gamma (γ) for Inter-Annotator Agreement Measure and Alignment] 41 (3)

Генерування індивідуалізованих порівняльних описів із відповідною ситуації інтонацією [Generating Tailored, Comparative Descriptions with Contextually Appropriate Intonation] 36 (2)

Генерування ЛФГ шляхом спеціалізації граматики [LFG Generation by Grammar Specialization] 38 (4)

Генерування перефразувань словосполучень і речень: огляд методів, керованих даними [Generating Phrasal and Sentential Paraphrases: A Survey of Data-Driven Methods] 36 (3)

Генерування референційних виразів: булеві розширення покрокового алгоритму [Generating Referring Expressions: Boolean Extensions of the Incremental Algorithm] 28 (1)

Генерування референційних виразів з градуальними характеристиками [Generating Referring Expressions that Involve Gradable Properties] 32 (2)

Генерування референційних виразів на основі графів [Graph-Based Generation of Referring Expressions] 29 (1)

Генерування референційних виразів: спрощення ідентифікації референтів [Generating Referring Expressions: Making Referents Easy to Identify] 33 (2)

Генерування складених слів у статистичному машинному перекладі на мови, схильні до утворення складених слів [Generation of Compound Words in Statistical Machine Translation into Compounding Languages] 39 (4)

Генерування числових апроксимацій [Generating Numerical Approximations] 38 (1)

Гібридне, з підкріпленням і з учителем, навчання процедурам управління діалогом на основі фіксованих наборів даних [Hybrid Reinforcement/Supervised Learning of Dialogue Policies from Fixed Data Sets] 34 (4)

Гібридні граматики для синтаксичного аналізу перерваних фразових структур і непроективних структур залежностей [Hybrid Grammars for Parsing of Discontinuous Phrase Structures and Non-Projective Dependency Structures] 43 (3)

Глибинний аналіз аргументування у створеному користувачем веб-дискурсі [Argumentation Mining in User-Generated Web Discourse] 43 (1)

Глобальна об’єднана модель для розмітки семантичних ролей [A Global Joint Model for Semantic Role Labeling] 34 (2)

Гнучка корпусно-керована модель регулярної і зворотної вірогідної сполучуваності [A Flexible, Corpus-Driven Model of Regular and Inverse Selectional Preferences] 36 (4)

Граматики з’єднання дерев не закриваються під впливом сильної лексикалізації [Tree-Adjoining Grammars Are Not Closed Under Strong Lexicalization] 38 (3)

Граматики заміщення d-дерев [D-Tree Substitution Grammars] 27 (1)

Двостороннє розв’язання контексту [Bidirectional Contextual Resolution] 26 (4)

Декілька міркувань про корпус Penn Discourse TreeBank, порівняльні корпуси та додаткове анотування [Reflections on the Penn Discourse TreeBank, Comparable Corpora, and Complementary Annotation] 40 (4)

Детерміністичний підхід до встановлення кореференції на основі об’єктно-орієнтованих, ранжованих за точністю правил [Deterministic Coreference Resolution Based on Entity-Centric, Precision-Ranked Rules] 39 (4)

Деякі зауваження про типізацію ознакових структур [A Note on Typing Feature Structures] 28 (3)

Дискримінативне вирівнювання слів на основі лінійного моделювання [Discriminative Word Alignment by Linear Modeling] 36 (3)

Дискурсна модель компресії тексту [Discourse Constraints for Document Compression] 36 (3)

Дистрибутивна пам’ять: загальна методика корпусно-базованих досліджень семантики [Distributional Memory: A General Framework for Corpus-Based Semantics] 36 (4)

Диференціальне переранжування у синтаксичному аналізі природних мов [Discriminative Reranking for Natural Language Parsing] 31 (1)

Диференційоване впорядкування слів на основі синтаксису для генерування текстів [Discriminative Syntax-Based Word Ordering for Text Generation] 41 (3)

Ділимі системи переходів і мультипланарний синтаксичний аналіз на основі дерев залежностей [Divisible Transition Systems and Multiplanar Dependency Parsing] 39 (4)

Дослідження валідності деяких метрик автоматичного оцінювання систем генерування природної мови [An Investigation into the Validity of Some Metrics for Automatically Evaluating Natural Language Generation Systems] 35 (4)

Дослідження зміни порядку слів у статистичному машинному перекладі: обчислювальні моделі та мовні явища [A Survey of Word Reordering in Statistical Machine Translation: Computational Models and Language Phenomena] 42 (2)

Дослідження перебивання і відновлення у багатоцільових діалогах [An Investigation of Interruptions and Resumptions in Multi-Tasking Dialogues] 37 (1)

Дугоспрямований синтаксичний аналіз із обмеженням дерев [Arc-Eager Parsing with the Tree Constraint] 40 (2)

Емпіричне вивчення корпусних методів автоматизації відповіді для домену електронної служби технічної підтримки [An Empirical Study of Corpus-Based Response Automation Methods for an E-mail-Based Help-Desk Domain] 35 (4)

Експерименти з автоматичною індукцією семантичних класів німецьких дієслів [Experiments on the Automatic Induction of German Semantic Verb Classes] 32 (2)

Експресивна сила представлень абстрактних значень [Expressive Power of Abstract Meaning Representations] 42 (3)

Емпірична мінімізація ризику для імовірнісних граматик: кількість прикладів і складність навчання [Empirical Risk Minimization for Probabilistic Grammars: Sample Complexity and Hardness of Learning] 38 (3)

Емпіричні методи дослідження денотації у номіналізаціях [Empirical Methods for the Study of Denotation in Nominalizations in Spanish] 38 (4)

Ефективне комплексне автоматичне створення графів логічного слідування [Efficient Global Learning of Entailment Graphs] 41 (2)

Ефективно обчислені лексичні ланцюжки як проміжний етап автоматичного реферування тексту [Efficiently Computed Lexical Chains as an Intermediate Representation for Automatic Text Summarization] 28 (4)

Жадібний автоматичний синтаксичний аналіз залежностей на основі переходів за допомогою ТКПС [Greedy Transition-Based Dependency Parsing with Stack LSTMs] 43 (2)

Загальний метод навчання мовних моделей на мовних моделях [A General Technique to Train Language Models on Language Models] 31 (2)

ЗАДРА: новий диференційований підхід до риторичного аналізу [CODRA: A Novel Discriminative Framework for Rhetorical Analysis] 41 (3)

Застосування лексикографічних напівкілець до розв’язання проблем обробки природної мови [Applications of Lexicographic Semirings to Problems in Speech and Language Processing] 40 (4)

Застосування машинного навчання для моделювання налаштувань області дії [A Machine Learning Approach to Modeling Scope Preferences] 29 (1)

Застосування обчислювальних моделей просторових прийменників до візуально ситуаційних діалогів [Applying Computational Models of Spatial Prepositions to Visually Situated Dialog] 35 (2)

Збільшення швидкості й точності обробки природної мови за допомогою онлайн навчання, яке адаптується до частоти ознак [Feature-Frequency-Adaptive On-line Training for Fast and Accurate Natural Language Processing] 40 (3)

Зважений дедуктивний синтаксичний аналіз і алгоритм Нута [Weighted Deductive Parsing and Knuth's Algorithm] 29 (1)

Зважені та імовірнісні контекстно-вільні граматики є однаково точними [Weighted and Probabilistic Context-Free Grammars Are Equally Expressive] 33 (4)

Звернення до коренів синтаксичного аналізу на основі граматики залежностей [Going to the Roots of Dependency Parsing] 39 (1)

Злиття речень у процесі реферування декількох новинних повідомлень [Sentence Fusion for Multidocument News Summarization] 31 (3)

Зміна порядку слів і алгоритм променевого пошуку на основі динамічного програмування для статистичного машинного перекладу [Word Reordering and a Dynamic Programming Beam Search Algorithm for Statistical Machine Translation] 29 (1)

Знаходження багатослівних словосполучень шляхом комбінування різних джерел лінгвістичної інформації [Identification of Multiword Expressions by Combining Multiple Linguistic Information Sources] 40 (2)

Зняття неоднозначності номіналізацій [The Disambiguation of Nominalizations] 28 (3)

Зняття омонімії іменників, дієслів і прикметників за допомогою автоматично визначених селекційних преференцій [Disambiguating Nouns, Verbs, and Adjectives Using Automatically Acquired Selectional Preferences] 29 (4)

Ідентифікація та уникнення непорозумінь у діалогах із людьми з хворобою Альцгеймера [Identifying and Avoiding Confusion in Dialogue with People with Alzheimer's Disease] 43 (2)

Ієрархічна модель статистичного машинного перекладу зі зваженими перетворювачами із скінченним числом станів і поверховими-n граматиками [Hierarchical Phrase-Based Translation with Weighted Finite-State Transducers and Shallow-n Grammars] 36 (3)

Іменники у WordNet: класи і представники класів [WordNet Nouns: Classes and Instances] 32 (1)

Імовірнісна дистрибутивна семантика та моделі латентних змінних [Probabilistic Distributional Semantics with Latent Variable Models] 40 (3)

Імовірнісний низхідний синтаксичний аналіз і мовне моделювання [Probabilistic Top-Down Parsing and Language Modeling] 27 (2)

Інтеграція планування тексту та лінгвістичного вибору без відмови від модульності: генератор IGEN [Integrating Text Planning and Linguistic Choice Without Abandoning Modularity: The IGEN Generator] 26 (2)

Інтеграція просодичної і лексичної інформації для автоматичної тематичного сегментування [Integrating Prosodic and Lexical Cues for Automatic Topic Segmentation] 27 (1)

Інтегрування теорії типів і дистрибутивної семантики: дослідження прикладів сполучень прикметник-іменник [Integrating Type Theory and Distributional Semantics: A Case Study on Adjective–Noun Compositions] 42 (4)

Інтернет як паралельний корпус [The Web as a Parallel Corpus] 29 (3)

Керований даними синтаксичний аналіз за допомогою імовірнісних лінійних контекстно-незалежних систем переписування [Data-Driven Parsing using Probabilistic Linear Context-Free Rewriting Systems] 39 (1)

Кернфункції для анотування семантичних ролей [Tree Kernels for Semantic Role Labeling] 34 (2)

Класифікація діалогічних реплік, які не є реченнями: метод машинного навчання [Classifying Non-Sentential Utterances in Dialogue: A Machine Learning Approach] 33 (3)

Кластеризація значень слів і здатність до кластеризації [Word Sense Clustering and Clusterability] 42 (2)

Кластеризація значень хештегів за часовою подібністю [Hashtag Sense Clustering Based on Temporal Similarity] 43 (1)

Кластеризація і диверсифікація результатів інформаційного пошуку за допомогою встановлення значення слів на основі графів [Clustering and Diversifying Web Search Results with Graph-Based Word Sense Induction] 39 (3)

Кластеризація синтаксичних позицій із подібними семантичними вимогами [Clustering Syntactic Positions with Similar Semantic Requirements] 31 (1)

Коєфіцієнт Каппа: новий погляд [The Kappa Statistic: A Second Look] 30 (1)

Коли ціле є меншим, ніж сума його частин: як структура впливає на значення ПВІ в семантичних контекстних векторах [When the Whole Is Less Than the Sum of Its Parts: How Composition Affects PMI Values in Distributional Semantic Vectors] 42 (2)

Коли ціле не більше, ніж комбінація його частин: "Декомпозиційний" погляд на композиційну дистрибутивну семантику [When the Whole Is Not Greater Than the Combination of Its Parts: A “Decompositional” Look at Compositional Distributional Semantics] 41 (1)

Комбінована модель для розпізнавання і вирівнювання власних назв двома мовами [A Joint Model to Identify and Align Bilingual Named Entities] 39 (2)

Коментар до статті Р. Карраско і М. Форкади “Покрокова побудова і супроводження мінімальних скінченних автоматів” [Comments on “Incremental Construction and Maintenance of Minimal Finite-State Automata,” by Rafael C. Carrasco and Mikel L. Forcada] 30 (2)

Комплексний аналіз виведення двомовного словника [A Comprehensive Analysis of Bilingual Lexicon Induction] 43 (2)

Комп’ютерна соціолінгвістика: огляд [Computational Sociolinguistics: A Survey] 42 (3)

Конкретні моделі та емпіричні оцінки категоріальної композиційної дистрибутивної моделі значення [Concrete Models and Empirical Evaluations for the Categorical Compositional Distributional Model of Meaning] 41 (1)

Контекстно-теоретична концепція композиційності в дистрибутивній семантиці [A Context-Theoretic Framework for Compositionality in Distributional Semantics] 38 (1)

Контроль сприйняття користувачем мовного стилю: генерування сегментування характеристик особистості на основі машинного навчання [Controlling User Perceptions of Linguistic Style: Trainable Generation of Personality Traits Segmentation] 37 (3)

Корпус дериватів і структур залежностей, видобутих з корпусу Penn Treebank на основі ККГ [CCGbank: A Corpus of CCG Derivations and Dependency Structures Extracted from the Penn Treebank] 33 (3)

Корпусно-базоване оцінювання центрування і встановлення референції займенників [A Corpus-Based Evaluation of Centering and Pronoun Resolution] 27 (4)

Крапки, слова з великої літери тощо [Periods, Capitalized Words, etc.] 28 (3)

Критерії варіативності засобів доступу для встановлення меж китайських слів [Accessor Variety Criteria for Chinese Word Extraction] 30 (1)

Критичний аналіз і вдосконалення метрики оцінювання сегментування тексту [A Critique and Improvement of an Evaluation Metric for Text Segmentation] 28 (1)

Лексикалізація та генеративна ефективність в ККГ [Lexicalization and Generative Power in CCG] 41 (2)

Лінгвістично анотоване переупорядкування: граматика оцінки і аналізу [Linguistically Annotated Reordering: Evaluation and Analysis Grammars] 36 (3)

Максимальні упорядковані підмножини [Maximal Consistent Subsets] 33 (2)

Маркування семантичних ролей імпліцитних аргументів номінативних присудків [Semantic Role Labeling of Implicit Arguments for Nominal Predicates] 38 (4)

Математична модель історичної семантики і групування значень слова у концепти [A Mathematical Model of Historical Semantics and the Grouping of Word Meanings into Concepts] 31 (2)

Машинний переклад з допомогою автоматично побудованих стохастичних скінченних перетворювачів [Machine Translation with Inferred Stochastic Finite-State Transducers] 30 (2)

Машинний переклад із використанням Всесвітньої мережі: розробка і оцінка ефективності системи машинного перекладу на основі прецедентів, що використовує Всесвітнє павутиння [wEBMT: Developing and Validating an Example-Based Machine Translation System Using the World Wide Web] 29 (3)

Машинний переклад на основі N-грамів [N-gram-based Machine Translation] 32 (4)

Машинний переклад на основі синтаксису словосполучень з використанням квазісинхронних характеристик «дерево до дерева» [Phrase Dependency Machine Translation with Quasi-Synchronous Tree-to-Tree Features] 40 (2)

Метод встановлення кореференції шляхом розбиття гіперграфів на основі обмежень [A Constraint-Based Hypergraph Partitioning Approach to Coreference Resolution] 39 (4)

Метод зняття лексичної багатозначності на основі теорії ігор [A Game-Theoretic Approach to Word Sense Disambiguation] 43 (1)

Метод оцінювання синтаксичного аналізу, керованого даними, є необ’єктивним і непослідовним [The DOP Estimation Method Is Biased and Inconsistent] 28 (1)

Метод представлення інформації в усномовленнєвих діалогових системах [A Strategy for Information Presentation in Spoken Dialog Systems] 37 (3)

Метод розв’язання анафори з використанням машинного навчання з одним кандидатом [A Twin-Candidate Model for Learning-Based Anaphora Resolution] 34 (3)

Методи навчання для об'єднання лінгвістичних показників: поліпшення видової класифікації та виявлення лінгвістичної інформації [Learning Methods to Combine

Linguistic Indicators: Improving Aspectual Classification and Revealing Linguistic Insights] 26 (4)

Методи оцінювання статистично залежних текстів [Evaluation Methods for Statistically Dependent Text] 41 (3)

Міри обчислювальної стійкості текстів – показник Юла (K) та ентропія Реньї [Computational Constancy Measures of Texts—Yule's K and Rényi's Entropy] 41 (3)

Множинна ад'юнкція у категоріальній граматиці з'єднання дерев [Multiple Adjunction in Feature-Based Tree-Adjoining Grammar] 41 (1)

Мовні моделі для машинного перекладу: порівняння оригінальних і перекладених текстів [Language Models for Machine Translation: Original vs. Translated Texts] 38 (4)

Модальність і заперечення: вступ до спеціального видання [Modality and Negation: An Introduction to the Special Issue] 38 (2)

Моделі лісу ознак для вірогіднісного синтаксичного аналізу на основі HPSG [Feature Forest Models for Probabilistic HPSG Parsing] 34 (1)

Моделі оптимізації систем звуків на основі реферування генетичних алгоритмів [Optimization Models of Sound Systems Using Genetic Algorithms Summarization] 29 (1)

Моделі перекладацької еквівалентності серед слів [Models of Translational Equivalence among Words] 26 (2)

Моделі синтаксичного аналізу для розпізнавання багатослівних виразів [Parsing Models for Identifying Multiword Expressions] 39 (1)

Модель встановлення семантичної орієнтації на основі випадкового блукання [A Random Walk–Based Model for Identifying Semantic Orientation] 40 (3)

Модель каналу з перешкодами для неконтрольованого зняття лексичної багатозначності [The Noisy Channel Model for Unsupervised Word Sense Disambiguation] 36 (1)

Модель мультимодального встановлення референції [A Model for Multimodal Reference Resolution] 26 (2)

Модель послідовності операцій — поєднання машинного перекладу на основі N-грамів та фразового статистичного машинного перекладу [The Operation Sequence Model—Combining N-Gram-Based and Phrase-Based Statistical Machine Translation] 41 (1)

Модель співставлення семантичних карт різних мов (французька/англійська, англійська/французька) [A Model for Matching Semantic Maps between Languages (French/English, English/French)] 29 (2)

Модель швидкої покрокової інтерпретації під час розпізнавання мовлення [A Framework for Fast Incremental Interpretation during Speech Decoding] 35 (3)

Моделювання діалогічних актів для автоматичного анотування і розпізнавання розмовної мови [Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech] 26 (3)

Моделювання локальної когерентності на основі референтів [Modeling Local Coherence: An Entity-Based Approach] 34 (1)

Моделювання регулярної багатозначності: дослідження семантичної класифікації каталонських прикметників [Modeling Regular Polysemy: A Study on the Semantic Classification of Catalan Adjectives] 38 (3)

Морфологічні і синтаксичні відмінки у статистичному синтаксичному аналізі на основі дерев залежностей [Morphological and Syntactic Case in Statistical Dependency Parsing] 39 (1)

На шляху до автоматичного аналізу помилок у машинному перекладі [Towards Automatic Error Analysis of Machine Translation Output] 37 (4)

На шляху до автоматичного генерування питань на задану тему [Towards Topic-to-Question Generation] 41 (1)

На шляху до ефективної частиномовної розмітки китайської мови [Towards Accurate and Efficient Chinese Part-of-Speech Tagging] 42 (3)

На шляху до модульної розбудови типізованих уніфікаційних граматик [Towards Modular Development of Typed Unification Grammars] 37 (1)

На шляху до каталогу банків лінгвістичних графів [Towards a Catalogue of Linguistic Graph Banks] 42 (4)

Набір підрядних означальних речень для оцінки композиційної дистрибутивної семантики RELPRON [RELPRON: A Relative Clause Evaluation Data Set for Compositional Distributional Semantics] 42 (4)

Набори правил і процедур для заснованого на складниках синтаксичного аналізу німецької мови, морфологічно багатої мови з менш усталеним порядком слів [Knowledge Sources for Constituent Parsing of German, a Morphologically Rich and Less-Configurational Language] 39 (1)

Навчання гнучких моделей вирівнювання слів із комплексними обмеженнями [Learning Tractable Word Alignment Models with Complex Constraints] 36 (3)

Навчання і оцінювання стратегій діалогу для нових програм: емпіричні методи оптимізації на основі малих наборів даних [Learning and Evaluation of Dialogue Strategies for New Applications: Empirical Methods for Optimization from Small Data Sets] 37 (1)

Навчання композиційної семантики на основі залежностей [Learning Dependency-Based Compositional Semantics] 39 (2)

Навчання морфології без учителя [Unsupervised Learning of Morphology] 37 (2)

Навчання морфології природної мови без учителя [Unsupervised Learning of the Morphology of a Natural Language] 27 (2)

Навчання перетворювачів дерев [Training Tree Transducers] 34 (3)

Навчання представлень для завдань обробки природної мови з незначним залученням учителя [Learning Representations for Weakly Supervised Natural Language Processing Tasks] 40 (1)

Навчання ранжуванню відповідей на питання не про факти з веб-бібліотек [Learning to Rank Answers to Non-Factoid Questions from Web Collections] 37 (2)

Напівконтекстні моделі мови [Half-Context Language Models] 37 (4)

Напівконтрольоване анотування семантичних ролей шляхом структурного вирівнювання [Semi-Supervised Semantic Role Labeling via Structural Alignment] 38 (1)

Незалежне від жанру і тематичної області визначення семантичної невпевненості [Cross-Genre and Cross-Domain Detection of Semantic Uncertainty] 38 (2)

Неконтрольоване виявлення кореферентних номінацій події [Unsupervised Event Coreference Resolution] 40 (2)

Неконтрольоване розпізнавання власних назв з урахуванням синтаксичного і семантичного контексту [Unsupervised Named Entity Recognition Using Syntactic and Semantic Contextual Evidence] 27 (1)

Неконтрольоване розпізнавання ідіоматичних виразів на основі типів і вживань [Unsupervised Type and Token Identification of Idiomatic Expressions] 35 (1)

Немінімальні деривати у синтаксичному аналізі на основі уніфікаційних граматик [Nonminimal Derivations in Unification-based Parsing] 27 (2)

Необхідність точної вивірки в оцінюванні системи обробки природної мови [The Need for Accurate Alignment in Natural Language System Evaluation] 27 (2)

Неточна синонімія і лексичний вибір [Near-Synonymy and Lexical Choice] 28 (2)

Новий неконтрольований метод визначення меж слів [A New Unsupervised Approach to Word Segmentation] 37 (3)

Нотатки про типізацію ознакових структур [A Note on Typing Feature Structures] 28 (3)

Обмежена непроективність: охоплення чи ефективність [Restricted Non-Projectivity: Coverage vs. Efficiency] 42 (4)

Обмежений дугоспрямований синтаксичний аналіз на основі граматики залежностей [Constrained Arc-Eager Dependency Parsing] 40 (2)

Обмеження сполучуваності для класифікації семантичних ролей [Selectional Preferences for Semantic Role Classification] 39 (3)

Обчислення лексичного контрасту [Computing Lexical Contrast] 39 (3)

Огляд розпізнавання і класифікації власних назв арабською мовою [Survey of Arabic Named Entity Recognition and Classification] 40 (2)

Ознаки статусу інформації і референційні вирази: емпіричне дослідження посилань на людей у зведеннях новин [Information Status Distinctions and Referring Expressions: An Empirical Study of References to People in News Summaries] 37 (4)

Онлайн-навчання систем статистичного машинного перекладу [Online Learning for Statistical Machine Translation] 42 (1)

Опрацювання багатослівних виразів: огляд [Multiword Expression Processing: A Survey] 43 (4)

Оптимізація статистичного машинного перекладу: огляд [Optimization for Statistical Machine Translation: A Survey] 42 (1)

Орфографічні помилки на веб-сторінках: на шляху до зменшення кількості помилок у веб-корпусах [Orthographic Errors in Web Pages: Toward Cleaner Web Corpora] 32 (3)

Осмислення алгоритму Яровського [Understanding the Yarowsky Algorithm] 30 (3)

Оцінка достовірності машинного перекладу на рівні слів [Word-Level Confidence Estimation for Machine Translation] 33 (1)

Оцінка застосування центрування в упорядкуванні інформації за допомогою корпусів [Evaluating Centering for Information Ordering Using Corpora] 35 (1)

Оцінка узгодженості між анотаторами у комп'ютерній лінгвістиці [Inter-Coder Agreement for Computational Linguistics] 34 (4)

Оцінювальна лексика поза мішками слів: лінгвістична інформація і комп'ютерні програми [Evaluative Language Beyond Bags of Words: Linguistic Insights and Computational Applications] 43 (1)

Оцінювання ймовірності на основі класу з семантичної ієрархії [Class-Based Probability Estimation Using a Semantic Hierarchy] 28 (2)

Оцінювання метрик лексико-семантичної спорідненості на основі WordNet [Evaluating WordNet-based Measures of Lexical Semantic Relatedness] 32 (1)

Оцінювання надійності без обмежень [Reliability Measurement without Limits] 34 (3)

Оцінювання парних оціночних суджень експертів [Evaluating Human Pairwise Preference Judgments] 41 (2)

Оцінювання схем анотування дискурсу і діалогу [Evaluating Discourse and Dialogue Coding Schemes] 31 (3)

Передмова до спеціального випуску, присвяченого розв'язанню анафори [Introduction to the Special Issue on Computational Anaphora Resolution] 27 (4)

Переклад на основі складних словосполучень [Hierarchical Phrase-Based Translation] 33 (2)

Перекладацькі розбіжності в китайсько-англійському машинному перекладі: емпіричне дослідження [Translation Divergences in Chinese–English Machine Translation: An Empirical Investigation] 43 (3)

Переписування пошукового запиту з використанням статистичного машинного перекладу на основі одномовного корпусу [Query Rewriting Using Monolingual Statistical Machine Translation] 36 (3)

Переформулювання правила 2 теорії центрування [A Reformulation of Rule 2 of Centering Theory] 27 (4)

Питально-відповідні системи для обмежених доменів: короткий огляд [Question Answering in Restricted Domains: An Overview] 33 (1)

Підвищення референційної когерентності у генеруванні текстів [Optimizing Referential Coherence in Text Generation] 30 (4)

Підвищення точності частиномовної розмітки шляхом об’єднання систем машинного навчання [Improving Accuracy in Word Class Tagging through the Combination of Machine Learning Systems] 27 (2)

Підпорядкування прийменникових груп без оракулів [Prepositional Phrase Attachment without Oracles] 33 (4)

Плагіат і парафраза: ідеї для наступного покоління систем атоматичного виявлення плагіату [Plagiarism Meets Paraphrasing: Insights for the Next Generation in Automatic Plagiarism Detection] 39 (4)

Повторні структуризація, розмітка та вирівнювання у машинному перекладі на основі синтаксису [Re-structuring, Re-labeling, and Re-aligning for Syntax-Based Machine Translation] 36 (2)

Поетапна обробка та відповідність вимогам [Incremental Processing and Acceptability] 26 (3)

Поетапний, прогностичний синтаксичний аналіз на основі психолінгвістично обумовленої граматики з'єднання дерев [Incremental, Predictive Parsing with Psycholinguistically Motivated Tree-Adjoining Grammar] 39 (4)

Показник узгодженості γcat, додаток до γ, призначений для категоризації континууму [The Agreement Measure γcat a Complement to γ Focused on Categorization of a Continuum] 43 (3)

Покрокова побудова і супроводження мінімальних скінченних автоматів [Incremental Construction and Maintenance of Minimal Finite-State Automata] 28 (2)

Покрокова побудова мінімальних ациклічних скінченних автоматів [Incremental Construction of Minimal Acyclic Finite-State Automata] 26 (1)

Помірно непроективна граматика залежностей [Mildly Non-Projective Dependency Grammar] 39 (2)

Поняття аргумента у приєднанні прийменникової групи [The Notion of Argument in Prepositional Phrase Attachment] 32 (3)

Порівняльне дослідження морфологічного сегментування методом часткового навчання з учителем [A Comparative Study of Minimally Supervised Morphological Segmentation] 42 (1)

Порівняння джерел знань для розв'язання іменної анафори [Comparing Knowledge Sources for Nominal Anaphora Resolution] 31 (3)

Послідовна перевірка валідності ручного і автоматичного анотування значень за допомогою семантичних графів [Consistent Validation of Manual and Automatic Sense Annotations with the Aid of Semantic Graphs] 32 (2)

Пошук причин неузгодженості: теорія узагальнюваності у дослідженнях ручного анотування [Identifying Sources of Disagreement: Generalizability Theory in Manual Annotation Studies] 33 (1)

Практична лінгвістична стенографія на основі підстановки контекстних синонімів і нового методу кодування вершин [Practical Linguistic Steganography using Contextual Synonym Substitution and a Novel Vertex Coding Method] 40 (2)

Практичні експерименти з формального ототожнення безконтекстних мов [Practical Experiments with Regular Approximation of Context-Free Languages] 26 (1)

Представлення зв'язності дискурсу: корпусне дослідження [Representing Discourse Coherence: A Corpus-Based Study] 31 (2)

Представлення значення за допомогою поєднання логічних і дистрибутивних моделей [Representing Meaning with a Combination of Logical and Distributional Models] 42 (4)

Представлення інформації шляхом створення оригінального тексту, діаграм та зовнішнього вигляду сторінки [Towards Constructive Text, Diagram, and Layout Generation for Information Presentation] 27 (3)

Представлення мовної форми і функції в рекурентних нейронних мережах [Representation of Linguistic Form and Function in Recurrent Neural Networks] 43 (4)

Прийменники у прикладних програмах: огляд досліджень і вступ до спеціального випуску [Prepositions in Applications: A Survey and Introduction to the Special Issue] 35 (2)

Припущення і заперечення: правила, ранжувальники і роль синтаксису [Speculation and Negation: Rules, Rankers, and the Role of Syntax] 38 (2)

Приховані дерева для встановлення кореференції [Latent Trees for Coreference Resolution] 40 (4)

Про апосинтез тематичної цілісності і міжреченнєву анафору [Toward an Aposynthesis of Topic Continuity and Intrasentential Anaphora] 28 (3)

Про кореферентність: кореферентність у системі розмітки конференції з розуміння повідомлень і споріднених системах [On Coreferring: Coreference in MUC and Related Annotation Schemes] 26 (4)

Про переклади ланцюжків, виконані висхідними мультидеревовидними перетворювачами [On the String Translations Produced by Multi Bottom–Up Tree Transducers] 38 (3)

Про перифраз і кореференцію [On Paraphrase and Coreference] 36 (4)

Про універсальну проблему генерації уніфікаційних граматик [On the Universal Generation Problem for Unification Grammars] 40 (3)

Програмні конвеєри і обмеження обсягу [Pipelines and Size Constraints] 26 (2)

Проектування та оцінювання систем опрацювання метафор [Design and Evaluation of Metaphor Processing Systems] 41 (4)

Пунктуація як імпліцитна розмітка для сегментування китайських текстів [Punctuation as Implicit Annotations for Chinese Word Segmentation] 35 (4)

Реляційні характеристики у точному аналізі думок [Relational Features in Fine-Grained Opinion Analysis] 39 (3)

Репрезентації текстів для класифікації патентів [Text Representations for Patent Classification] 39 (3)

Реферування інформаційної графіки у вигляді тексту [Summarizing Information Graphics Textually] 38 (3)

Реферування наукових статей: експерименти з релевантністю і риторичним статусом [Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status] 28(4)

Реферування новел [Summarizing Short Stories] 36 (1)

Робастне розуміння у мультимодальних інтерфейсах [Robust Understanding in Multimodal Interfaces] 35 (3)

Розбір залежностей на основі концепції розширеного скінченного автомату [Dependency Parsing with an Extended Finite-State Approach] 29 (4)

Розпізнавання діалектів арабської мови [Arabic Dialect Identification] 40 (1)

Розпізнавання контекстуальної полярності: дослідження ознак для аналізу модальності на рівні словосполучення [Recognizing Contextual Polarity: An Exploration of Features for Phrase-Level Sentiment Analysis] 35 (3)

Розподіл орфографічних помилок у бразильському варіанті португальської мови [Spelling Error Patterns in Brazilian Portuguese] 41 (1)

Розпутування чату [Disentangling Chat] 36 (3)

Розробка і вдосконалене оцінювання робастного алгоритму розв’язання анафори [Design and Enhanced Evaluation of a Robust Anaphor Resolution Algorithm] 27 (4)

Розробка моделей семантичного простору на основі залежностей [Dependency-Based Construction of Semantic Space Models] 33 (2)

Розробка питально-відповідної системи як побудова та ранжування обґрунтування відповідей на рівні тексту [Framing QA as Building and Ranking Intersentence Answer Justifications] 43 (2)

Розуміння алгоритму Яровського [Understanding the Yarowsky Algorithm] 30 (3)

Розширення словника оціночної лексики та виявлення об’єкта оцінювання шляхом подвійного розповсюдження [Opinion Word Expansion and Target Extraction through Double Propagation] 37 (1)

Розщеплюваність білексичних контекстно-незалежних граматик є нерозв’язною [Splittability of Bilexical Context-Free Grammars is Undecidable] 37 (4)

Роль автоматичного синтаксичного аналізу і логічного виведення в анотуванні семантичних ролей [The Importance of Syntactic Parsing and Inference in Semantic Role Labeling] 34 (2)

Рядкові ядра для визначення мови автора: із досвіду розробки й використання [String Kernels for Native Language Identification: Insights from Behind the Curtains] 42 (3)

Самоналаштування якості дистрибутивного вектора ознак [Bootstrapping Distributional Feature Vector Quality] 35 (3)

Сегментування і оцінювання нерозпізнаних морфем на основі моделі складу для гібридного частиномовного анотування корейської мови [Syllable-Pattern-Based Unknown-Morpheme Segmentation and Estimation for Hybrid Part-of-Speech Tagging of Korean] 28 (1)

Сегментування китайських текстів на слова і розпізнавання власних назв: прагматичний підхід [Chinese Word Segmentation and Named Entity Recognition: A Pragmatic Approach] 31 (4)

Сегментування слів, розпізнавання незнайомих слів і морфологічне узгодження у синтаксичному аналізаторі івриту [Word Segmentation, Unknown-word Resolution, and Morphological Agreement in a Hebrew Parsing System] 39 (1)

Семантичний аналіз локативів з мінімальною рекурсією [A Minimal Recursion Semantic Analysis of Locatives] 35 (2)

Синтаксис і семантика прийменників у автоматичній інтерпретації іменних груп і складних слів: порівняльне дослідження [The Syntax and Semantics of Prepositions in the Task of Automatic Interpretation of Nominal Phrases and Compounds: A Cross-Linguistic Study] 35 (2)

Синтаксичний аналіз іменних груп у банку дерев Penn Treebank [Parsing Noun Phrases in the Penn Treebank] 37 (4)

Синтаксичний аналіз за допомогою універсального перцептрона і променевого пошуку [Syntactic Processing Using the Generalized Perceptron and Beam Search] 37 (1)

Синтаксичний аналіз лінійних контекстно-незалежних систем переписування за допомогою швидкого множення матриць [Parsing Linear Context-Free Rewriting Systems with Fast Matrix Multiplication] 42 (3)

Синтаксичний аналіз сучасної літературної арабської мови на основі граматики залежностей за допомогою лексичних і флексійних характеристик [Dependency Parsing of Modern Standard Arabic with Lexical and Inflectional Features] 39 (1)

Синтаксичний аналіз турецької мови на основі граматики залежностей [Dependency Parsing of Turkish] 34 (3)

Синхронні контекстно-вільні граматики та стратегії оптимального автоматичного синтаксичного аналізу [Synchronous Context-Free Grammars and Optimal Parsing Strategies] 42 (2)

Система обробки визначених дескрипцій на основі дослідних даних [An Empirically Based System for Processing Definite Descriptions] 26 (4)

Система оцінювання PARADISE: проблеми і результати [The PARADISE Evaluation Framework: Issues and Findings] 32 (2)

Систематичне порівняння різних статистичних моделей вирівнювання [A Systematic Comparison of Various Statistical Alignment Models] 29 (1)

Скінченні регістрові автомати для розпізнавання неконкатенативної морфології [Finite-State Registered Automata for Non-Concatenative Morphology] 32 (1)

Скінченні табличні обмеження для спрощення конвейерного контекстно-незалежного синтаксичного аналізу [Finite-State Chart Constraints for Reduced Complexity Context-Free Parsing Pipelines] 38 (4)

Складність ранжування гіпотез у теорії оптимальності [The Complexity of Ranking Hypotheses in Optimality Theory] 35 (1)

Складність, синтаксичний аналіз і факторизація багатокомпонентної граматики з’єднання дерев у часткові дерева [Complexity, Parsing, and Factorization of Tree-Local Multi-Component Tree-Adjoining Grammar] 36 (3)

Скорочення великих лінгвістичних корпусів за допомогою алгоритмів ЗМП [Large Linguistic Corpus Reduction with SCP Algorithms] 41 (3)

Словник сполучуваності морфем [The Combinatory Morphemic Lexicon] 28 (2)

Словникові методи аналізу тональності [Lexicon-Based Methods for Sentiment Analysis] 37 (2)

Сортування текстів за складністю [Sorting Texts by Readability] 36 (2)

Спільне двомовне навчання для класифікації тональності китайських відгуків на товари [Bilingual Co-Training for Sentiment Classification of Chinese Product Reviews] 37 (3)

Спонтанне визначення меж речень різними мовами [Unsupervised Multilingual Sentence Boundary Detection] 32 (4)

Справжнє і шаблонне генерування природної мови: хибне протиставлення? [Real versus Template-Based Natural Language Generation: A False Opposition?] 31 (1)

Стабільна жанрова класифікація текстів [Stable Classification of Text Genres] 37 (2)

Статистична модель встановлення меж слів у транскрибованому мовленні [A Statistical Model for Word Discovery in Transcribed Speech] 27 (3)

Статистична модель синтаксичного аналізу для класифікації тональності [A Statistical Parsing Framework for Sentiment Classification] 41 (2)

Статистична обробка метафор [Statistical Metaphor Processing] 39 (2)

Статистичний машинний переклад «від ланцюжка до залежності» [String-to-Dependency Statistical Machine Translation] 36 (4)

Статистичний машинний переклад з використанням алгоритмів вирівнювання [The Alignment Template Approach to Statistical Machine Translation] 30 (4)

Статистичний підхід до мікропланування на основі граматики [A Statistical, Grammar-Based Approach to Microplanning] 43 (1)

Статистичні моделі для видобування пар із транслітерованими словами без залучення, з частковим і з повним залученням учителя [Statistical Models for Unsupervised, Semi-Supervised, and Supervised Transliteration Mining] 43 (2)

Статистичні підходи до автоматизованого перекладу [Statistical Approaches to Computer-Assisted Translation] 35 (1)

Стверджувальні підкази у цілеспрямованих діалогах [Affirmative Cue Words in Task-Oriented Dialogue] 38 (1)

Створення і використання лексичної бази даних про розбіжності між неточними синонімами [Building and Using a Lexical Knowledge Base of Near-Synonym Differences] 32 (2)

Створення морфологічних аналізаторів шляхом поєднання опитування інформантів і машинного навчання [Bootstrapping Morphological Analyzers by Combining Human Elicitation and Machine Learning] 27 (1)

Створення питань за допомогою розробки концептів [Composing Questions through Conceptual Authoring] 33 (1)

Створення показово-інформативних оглядів за допомогою системи SumUM [Generating Indicative-Informative Summaries with SumUM] 28 (4)

Створення робастної системи анотування семантичних ролей [Towards Robust Semantic Role Labeling] 34 (2)

Структура дискурсу в оцінюванні машинного перекладу [Discourse Structure in Machine Translation Evaluation] 43 (4)

Структура документа [Document Structure] 29 (2)

Схеми синтаксичного аналізу на основі граматики залежностей і слабо непроективний синтаксичний аналіз на основі граматики залежностей [Dependency Parsing Schemata and Mildly Non-Projective Dependency Parsing] 37 (3)

Схожість семантичних відносин [Similarity of Semantic Relations] 32 (3)

Теоретико-графічна модель семантичної відстані [Graph-Theoretic Framework for Semantic Distance] 36 (1)

Тест незалежності ядра для географічного варіювання мов [A Kernel Independence Test for Geographical Language Variation] 43 (3)

Тонкощі моделі синтаксичного аналізу Коллінза [Intricacies of Collins' Parsing Model] 30 (4)

Тут немає логічного заперечення, але є альтернативи: моделювання усного заперечення за допомогою дистрибутивної семантики [There Is No Logical Negation Here, But There Are Alternatives: Modeling Conversational Negation with Distributional Semantics] 42 (4)

Удосконалена оцінка ентропії для оцінювання виведення значень слів [Improved Estimation of Entropy for Evaluation of Word Sense Induction] 40 (3)

Удосконалення сегментування тексту за допомогою латентного семантичного аналізу: повторний аналіз статті Ф. Чой, П. Вімер-Гастінгс і Д. Мур [Improving Text Segmentation Using Latent Semantic Analysis: A Reanalysis of Choi, Wiemer-Hastings, and Moore (2001)] 32 (1)

Удосконалення машинного перекладу шляхом використання непаралельних корпусів [Improving Machine Translation Performance by Exploiting Non-Parallel Corpora] 31 (4)

Укладання корпусів для розробки і оцінки систем перефразування [Constructing Corpora for the Development and Evaluation of Paraphrase Systems] 34 (4)

Факторизація грамматики шляхом роз’єднання дерев [Grammar Factorization by Tree Decomposition] 37 (1)

Формальна дистрибутивна семантика: передмова до спеціального випуску [Formal Distributional Semantics: Introduction to the Special Issue] 42 (4)

Формування вибірки для статистичного синтаксичного аналізу [Sample Selection for Statistical Parsing] 30 (3)

Фреймово-семантичний синтаксичний аналіз [Frame-Semantic Parsing] 40 (1)

Центрування: параметрична теорія і її трактування [Centering: A Parametric Theory and Its Instantiations] 30 (3)

Чи робить GIZA++ пошукові помилки? [Does GIZA++ Make Search Errors?] 36 (3)

Чи це правда? Прагматична складність оцінювання адекватності сприйняття [Did It Happen? The Pragmatic Complexity of Veridicality Assessment] 38 (2)

Чого немає у мішку слів для системи питання «Чому…»-відповідь? [What Is Not in the Bag of Words for Why-QA?] 36 (2)

Швидкий приблизний пошук у великих словниках [Fast Approximate Search in Large Dictionaries] 30 (4)

Широкомасштабна оцінка сучасних методів зняття лексичної багатозначності на основі псевдослів [A Large-Scale Pseudoword-Based Evaluation Framework for State-of-the-Art Word Sense Disambiguation] 40 (4)

Широкомасштабна розподілена синтаксична, семантична і лексична модель мови [A Scalable Distributed Syntactic, Semantic, and Lexical Language Model] 38 (3)

Широкомасштабне отримання і оцінювання лексичних ресурсів із банків дерев Penn-II і Penn-III [Large-Scale Induction and Evaluation of Lexical Resources from the Penn-II and Penn-III Treebanks] 31 (3)

Широкомасштабний високоефективний статистичний синтаксичний аналіз на основі комбінаторної категорійної граматики і логлінійних моделей [Wide-Coverage Efficient Statistical Parsing with CCG and Log-Linear Models] 33 (4)

Широкомасштабний глибокий статистичний синтаксичний аналіз із використанням автоматичної розмітки структури залежностей [Wide-Coverage Deep Statistical Parsing Using Automatic Dependency Structure Annotation] 34 (1)

Що впливає на узгодженість між розмітниками? Металінгвістичне дослідження [What Determines Inter-Coder Agreement in Manual Annotations? A Meta-Analytic Investigation] 37 (4)

Що таке перифраза? [What Is a Paraphrase?] 39 (3)

Ядерні методи для мінімально контрольованого зняття багатозначності [Kernel Methods for Minimally Supervised WSD] 35 (4)

AutoExtend: поєднання векторів представлення слів з семантичними ресурсами [AutoExtend: Combining Word Embeddings with Semantic Resources] 43 (3)

CorMet: Корпусно-базована система автоматичного видобування стертих метафор [A Computational, Corpus-Based Conventional Metaphor Extraction System] 30 (1)

HyperLex: великомасштабне оцінювання градуйованого лексичного логічного слідування [HyperLex: A Large-Scale Evaluation of Graded Lexical Entailment] 43 (4)

Nouveau-ROUGE: нова метрика для генерації дайджесту оновлень [Nouveau-ROUGE: A Novelty Metric for Update Summarization] 37 (1)

OntoLearn Reloaded: алгоритм на основі графа для генерації таксономії [OntoLearn Reloaded: A Graph-Based Algorithm for Taxonomy Induction] 39 (3)

SimLex-999: оцінювання семантичних методів шляхом оцінки (справжньої) схожості [SimLex-999: Evaluating Semantic Models With (Genuine) Similarity Estimation] 41 (4)

XMG: розширювана метаграматика [XMG: eXtensible MetaGrammar] 39 (3)

Покажчик авторів журналу

Computational Linguistics

(2000-2017 pp.)

Зміст