УДК 811.111'367.7:659.3
DOI: 10.28925/2311-2425.2020.145

ОСНОВНІ ПРОБЛЕМИ СИСТЕМ ОБРОБКИ ПРИРОДНОЇ МОВИ

Гирин О.В.,
Житомирський державний університет імені Івана Франка,
вул. Велика Бердичівська, 40, м. Житомир, 10008
oleg_hyryn@ukr.net
ORCID iD 0000-0002-3641-2440

Стаття присвячена обробці природної мови, а саме автоматичній синтаксичній обробці англійських речень. Висвітлено проблеми, спричинені цим процесом, що пов'язані з графічною, семантичною та синтаксичною неоднозначністю. Визначено шляхи вирішення цих проблеми, зумовлених застосуванням автоматичного синтаксичного аналізу, та яким чином такі методи аналізу можуть бути корисні для розробки його нових алгоритмів аналізу. Дослідження зосереджене на питаннях, які унеможливлюють основу обробки природної мови — парсинг. Це процес аналізу речень за їх структурою, змістом і значенням, метою якого є вивчення граматичної структури речення, розподіл речень на складові компоненти і визначення зв'язків між ними.

Ключові слова: *синтаксичний аналіз, обробка природної мови, статистичне машинне навчання, неоднозначність.*

O. Hyryn

Basic challenges in natural language processing systems

The article proceeds from the intended use of parsing for the purposes of automatic information search, question answering, logical conclusions, authorship verification, text authenticity verification, grammar check, natural language synthesis and other related tasks, such as ungrammatical speech analysis, morphological class definition, anaphora resolution etc. The study covers natural language processing challenges, namely of an English sentence. The article describes formal and linguistic problems, which might arise during the process and which are connected with graphic, semantic, and syntactic ambiguity. The article provides the description of how the problems had been solved before the automatic syntactic analysis was applied and the way, such analysis methods could be helpful in developing new analysis algorithms today. The analysis focuses on the issues, blocking the basis for the natural language processing — parsing — the process of sentence analysis according to their structure, content and meaning, which aims to examine the grammatical structure of the sentence, the division of sentences into constituent components and defining links between them. The analysis identifies a number of linguistic issues that will contribute to the development of an improved model of automatic syntactic analysis: lexical and grammatical synonymy and homonymy, hypo- and hyperonymy, lexical and semantic fields, anaphora resolution, ellipsis, inversion etc. The scope of natural language processing reveals obvious directions for the improvement of parsing models. The improvement will consequently expand the scope and improve the results in areas that already employ automatic parsing. Indispensable achievements in vocabulary and morphology processing shall not be neglected while improving automatic syntactic analysis mechanisms for natural languages.

Key words: *parsing, natural language processing, statistical machine learning, ambiguity.*

Вступ. Використання цифрових технологій стало невід'ємною частиною нашого життя, що викликало нагальну потребу замінити роботу, що виконується людьми, на автоматичну. Обробка природної мови (далі — ОПМ) (NLP — natural language processing) — одне із завдань, яке може здійснюватися автоматично. Метою ОПМ є вивчення механізмів природної мови (як внутрішніх, так і зовнішніх) та використання цих знань у застосунках та програмах, що допоможуть полегшити повсякденне спілкування з використанням машин.

Теоретичні передумови дослідження. Обробка природної мови вивчається в численних працях зарубіжної лінгвістики з 1967 р. Питання, пов'язані з автоматичним аналізом мовлення, знайшли своє відображення у працях Флейс Дж.Л. [8], Голлінгсворт Ч. [10], Ковар В. [11] та ін.

Хоча в Україні дослідження аналізу англійської мови до цього часу мало теоретичний характер, проте досвід та теоретичні результати в галузі англійської граматики, зокрема з генеративної точки зору (Буніятова І.Р. [2], Полховська М.В. [5; 6]), може створити основу для їх прикладного застосування.

Поточне використання та перспективи ОПМ були вказані в [4]. Дослідження визначає застосування синтаксичного аналізу для цілей автоматич-

ного пошуку інформації, генерування автоматичних відповідей на запитання, логічних висновків, перевірки авторства, перевірки справжності тексту, перевірки граматичності тексту, синтезу природної мови та інших супутніх завдань, таких як аналіз неграматичних речень, визначення морфологічного класу слів, прив'язка відділених анафор тощо [4].

Мета статті — висвітлити стан вирішення проблем, які неминуче виникають під час ОПМ.

Методика дослідження. Це дослідження пропонує аналіз деяких лінгвістичних проблем, які слід враховувати при розробці моделей синтаксичного аналізу; застосування наукових методів аналізу, синтезу, опису та порівняння, лінгвістичних методів заміщення і трансформації для вирішення основних проблеми, зумовлених застосуванням автоматичного синтаксичного аналізу.

Результати й обговорення. ОПМ жодним чином не можна назвати чітко визначеним процесом. Численні труднощі виникають через низку об'єктивних причин, таких як існування сотень природних мов, кожна з яких має синтаксичні правила, а також їх варіації в мові. У межах однієї мови є слова, які можуть мати різне значення залежно від контексту вживання. Навіть графічний рівень наводить на думку про деякі технічні труднощі. Таким чином, ОПМ має враховувати тип кодування, який використовується в конкретному документі. Текст може зберігатися в різних кодуваннях: ASCII, UTF-8, UTF-16 або Latin-1 [14, 74]. Для пунктуації та цифр можуть знадобитися спеціальні типи обробки. Іноді доводиться обробляти використання символів, що представляють емоції (комбінації символів або спеціальних символів), гіперпосилання, повторювані розділові знаки (... або —), розширення файлів та імена користувачів, що містять крапки.

Розбити текст на фрагменти або елементи зазвичай означає подання його у вигляді послідовності слів. У цьому разі останні називаються «лексичним елементом», «лексемою» або «токеном», а сам процес — «токенізацією». Розбиття тексту не викликає особливих труднощів у тих мовах, які використовують пробіли для розділення слів. Проте в мовах, подібних до китайської, це зробити набагато складніше, оскільки символи можуть позначати як склади, так і цілі слова. Більше того, власне англійська мова може представляти певні труднощі під час процесу лексемізації, оскільки їй властива велика кількість альтернативних способів офіційного подання того самого слова: його можна писати разом, окремо або через дефіс.

Слова природним чином поєднуються у фрази та речення. Визначення меж речень також може бути пов'язане з певними труднощами. Здавалося б перший очевидний логічний спосіб — знайти крапки, що вказують на його закінчення, проте вони можуть траплятися і всередині речення, наприклад після скорочених слів тощо.

Однак більш серйозну проблему, що стосується точності аналізу, являє граматичний аналіз. По-перше, багато залежить від якості позначення частини мови, яка має бути дуже високою (97–98 %) [3]. Проте в довгих реченнях можна натрапити на неправильно розпізнану частину мови, що може призвести до дальшого хибного аналізу. По-друге, автоматичний синтаксичний аналіз, який існує на сьогодні, дає точність приблизно 90–93 % [3], а це означає, що в довгому реченні майже завжди будуть помилки синтаксичного аналізу. Наприклад, з точністю 90 % ймовірність позначення мовної частини без помилок для речення довжиною 10 слів становитиме 35 % [3].

Сучасний стан досліджень дає надію на покращення якості синтаксичного аналізу, та часто правильність останнього передбачає розуміння семантики речення. Адже в природній мові існують речення, які можна дослідити лише «людським оком». Так, речення "*I hit a man with a camera*" може передбачати два різних варіанти синтаксичного розбору, що залежатиме від того, чи вважаємо ми, що у нападника був фотоапарат, чи останній використовувався як інструмент для удару. Звичайно, щоб отримати найбільш точний синтаксичний аналіз, має сенс залишити деякі найбільш вірогідні варіанти, а потім визначити правильний за допомогою комбінації різних факторів, включаючи семантичні.

Іноді під час ОПМ важливо встановити взаємозв'язки між словами в різних синтаксичних групах. Така роздільна здатність визначає співвідношення між конкретними словами, що позначають той самий об'єкт, тобто вони мають однаковий референт в одному або кількох реченнях. Наприклад, у реченнях "*The town is small but beautiful. It is located at the foot of the mountain*". Слово *it* співвідноситься, тобто є референтно ідентичним до *town*. Явища ко-референтності походять від основних закономірностей організації тексту, оскільки останній має лінійну структуру, а ситуація, яка в ньому описується, зазвичай нелінійна. Текст майже неминуче має містити багаторазові номінації елементів в описаній ситуації. При кожному новому посиланні на один і той самий об'єкт виникає нова номінація цього об'єкта на основі того, що вже було сказано про нього й тих знань, які не виражені у тексті. Хоча проблема узгодженості в лінгвістиці була ретельно вивчена, практична реалізація цих теоретичних завдань залишається проблемною [1, 41].

Якщо слово має кілька семантичних інтерпретацій, то для з'ясування його значення в кожному конкретному випадку може знадобитися використання окремого інструмента, завданням якого буде усунути неоднозначність слова [14, 77]. Це

допоможе подолати певні труднощі. Наприклад, у реченні "*Mary returned home*" слово *home* може означати як будівлю, так і країну або населений пункт.

Однією з найбільш відкритих проблем ОПМ є двозначність її одиниць, яка може виникати на всіх мовних рівнях, що включає явища полісемії, омонімії та синонімії. Неоднозначність може бути: лексична (існування більше одного значення слова, наприклад *bank*); синтаксична або структурна (коли одне речення має декілька можливих граматичних варіантів і, отже, різне значення, таке як неоднозначність конструкцій, коли PP (прийменникова фраза) може займати позицію у реченні як після VP, так і NP у межах одного речення із відповідною зміною значення: *"The police shot the burglars with guns"*); семантична багатозначність (коли одне і те саме речення можна розуміти по-різному в різних контекстах, хоча лексична чи структурна багатозначність при цьому відсутня: "*All philologists stick to a theory"*); прагматична (коли одне і те саме речення можна розуміти по-різному в різних контекстах, у яких воно може вживатися *"My brother thinks he is a genius"*).

Системи усунення лексичної багатозначності, які на сьогодні розроблені, мають точність в межах 60–70 % [13, 1165] і швидше за все будуть представлені як окремі методи. Розв'язання питання однозначності потребує інтеграції кількох джерел інформації та методів.

Таким чином, основне завдання синтаксичного аналізу полягає у з'ясуванні, чи є речення граматично правильним з точки зору загальновизнаних правил побудови фраз на певній мові. Завдання розуміння тексту машиною — розпізнати граматичну структуру речення, що дає змогу формалізовано подати значення тексту. Синтаксична структура може являти або проміжний результат, який є вхідним матеріалом для дальшого семантичного аналізу, або зручне представлення тексту природною мовою для розв'язання прикладних завдань, наприклад, в інформаційно-аналітичних системах або системах машинного перекладу.

Незважаючи на всі перелічені труднощі, технологія обробки природної мови здебільшого здатна успішно справлятися зі своїми завданнями, а отже, може бути застосована в багатьох галузях діяльності.

Природна мова, хоч і структурована та систематизована, видається досить проблематичною для символічних алгоритмів, спрямованих на її обробку, тому домінуючими підходами до сучасної ОПМ є ті, що засновані на статистичному машинному навчанні [9, 49]. Слід зазначити, що приблизно в половині випадків омонімії набору морфологічних ознак недостатньо для визначення синтаксичних класів одиниць. Хоч двозначність і можна зменшити, здійснюючи синтаксичний та семантичний аналізи за допомогою статистичних методів, які дають змогу відхиляти вкрай малоймовірні варіанти. Природна мова, хоч і є символічною за своєю природою, проте її обробка за допомогою символічних, заснованих на логіці правилах та об'єктивних моделях є досить складним процесом.

На початку 1990-х років почали розвиватися методи машинного навчання і паралельно з цим було здійснено низку досліджень із статистичної лінгвістики. У машинному навчанні алгоритми класифікації для різних завдань виявилися ефективними, зокрема, для обробки текстів: виявлення спаму, сортування документів за темою, виділення названих сутностей. Застосування статистичних методів у комп'ютерній лінгвістиці дало змогу з високою точністю визначати частини мови, з'явилися парсери на основі стохастичних безконтекстних граматик, було розроблено проєкти із статистичного машинного перекладу. Також закладено основи глибокого машинного навчання, які завдяки прогресу у високопродуктивних системах та появі великих обсягів даних, що використовуються з цією метою, лише нещодавно дали перші результати [3].

У 2010 р. була запропонована модель лексичної імовірнісної (стохастичної) граматики, яка дала змогу збільшити точність граматичного синтаксичного аналізу до 93 %, що, звичайно, далеко не ідеально. Точність синтаксичного аналізу — це відсоток правильно визначених граматичних зв'язків, а також ймовірність (яка зазвичай дуже низька) того, що довге речення буде належним чином проаналізоване. Водночас завдяки новим алгоритмам та підходам, включаючи глибоке навчання, швидкість граматичного розбору зросла. Більше того, всі провідні алгоритми та моделі стали доступними для широкого кола дослідників і, мабуть, найвідомішим у цій галузі став алгоритм Томаса Міколова [12].

Після появи нових методів глибокого навчання стало можливе отримання чітких семантичних описів слів, фраз та речень, навіть без аналізу безпосереднього оточення мовленнєвих одиниць. Створення власних семантичних словників та баз даних зараз вимагає менших зусиль, а отже простіше розробити автоматичні системи обробки тексту. Однак ОПМ ще залишається далекою від адекватного аналізу взаємопов'язаних явищ, представлених у формі послідовності речень чи зображень, а також діалогів. Усі відомі методи на сьогодні успішно працюють або при вирішенні проблем «поверхневого» розуміння мови, або з істотним обмеженням ділянки аналізу [3].

Слід зазначити, що методи глибокого навчання є більш точними, ніж поверхневі, адже останні не намагаються «зрозуміти» текст. Вони враховують найближче безпосереднє ото-

чення слова, використовуючи інформацію щодо валентності слів. Правила можна також отримати автоматично, за допомогою комп'ютера, послуговуючись текстовою базою даних для слів, доданих разом з їх лексичною семантикою. Теоретично цей метод не такий ефективний, як метод глибокого аналізу, хоча на практиці він дає кращі результати [7].

Висновки. Процес розуміння та породження природної мови за допомогою комп'ютерних технологій надзвичайно складний. На сьогодні найефективнішими методами роботи з мовними даними є методи алгоритму машинного навчання з оператором-«учителем», що допомагає системі відрізнити мовні структури та правила від анотованих корпусних даних. Наприклад, завдання класифікації текстової інформації за такими категоріями, як спорт, політика, економіка та розваги видається досить простим, оскільки слова, що вживаються в документах цих предметних галузей, слугують маркером. Спираючись на власний досвід, читач може легко визначити тематику тексту, але навряд чи назве конкретні правила, за якими це можна зробити. Створення правила або набору правил для автоматичної категоризації тексту є складним і копітким. Застосовуючи алгоритми машинного навчання з оператором-«учителем», який вводить потрібні дані, машина може визначити мовні структури, що дадуть змогу категоризувати текстову інформацію. Цей підхід є ефективним для лише обмежених сфер, а саме спорту, права чи економіки. Однак для більш широких галузей, таких як історія, політика, соціологія тощо, він нерезультативний, оскільки трудомісткий за своїм характером.

Перспективи. Враховуючи вищесказане, необхідність ефективного синтаксичного аналізу видається очевидною. Вивчення класичних та сучасних синтаксичних аналізів є перспективою для подальших досліджень.

ДЖЕРЕЛА

1. Боярский К.К. Введение в компьютерную лингвистику: учеб. пособ. СПб.: НИУ ИТМО, 2013. 72 с.
2. Буніятова І.Р. Еволюція гіпотаксису в германських мовах (IV–XIII ст.): монографія. К.: Вид. центр КНЛУ, 2003. 327 с.
3. Велихов П. Машинное обучение для понимания естественного языка. *Открытые системы. СУБД.* 2016. № 01. URL: https://www.osp.ru/os/2016/01/13048649/ (дата звернення: 26.09.20).
4. Гирин О.В. Автоматичний синтаксичний аналіз англійської мови: застосування та перспективи. *Вісник Житомирського державного університету імені Івана Франка*. Житомир: Вид-во ЖДУ ім. І. Франка, 2017. Вип. 1 (85). С. 26–30.
5. Полховська М.В. Аналіз англійських медіальних конструкцій з позиції генеративної граматики. *Studia philologica*. Київ, 2013. Вип. 2. С. 32–36.
6. Полховська М.В. Критерії розрізнення медіальних та ергативних конструкцій в англійській мові. *Наукові записки [Національного університету «Острозька академія»]. Серія: Філологічна.* Острог, 2012. Вип. 26. С. 277–280.
7. Chen P. A Fully Unsupervised Word Sense Disambiguation Method Using Dependency Knowledge. *Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. 2009. P. 28–36.
8. Fleiss J. L. Statistical Methods for Rates and Proportions. John Wiley & Sons, 2013. 800 p.
9. Goldberg Y. Neural Network Methods for Natural Language Processing. Morgan & Claypool Publishers, 2017. 309p.
10. Hollingsworth Ch. Using Dependency-based Annotations for Authorship Identification. *Proceedings of Text, Speech and Dialogue, 15th International Conference*. Berlin. v. 7499 . 2012. P 314–319.
11. Kovar V. Information Extraction for Czech Based on Syntactic Analysis. *Proceedings of the 5th Language & Technology Conference*. Poznan : Fundacja Universytetu im. A. Mickiewicza, 2011. P. 466–470.
12. Mikolov T. Efficient Estimation of Word Representations in Vector Space. 2013. 12p.
13. Mohd S. H. Word Sense Ambiguity: A Survey. *International Journal of Computer and Information Technology.* 2013. Vol. 02. Issue 06. P. 1161–1168.
14. Reese R. M. Natural Language Processing with Java. Packt Publishing, 2015. 262 p.

REFERENCES

1. Boiarskii, K. K. (2013). Vvedenie v kompiuternuiu lingvistiku. Uchebnoe posobie [Introduction to Computational Linguistics]. SPb: NIU ITMO, 72 p. (in Russian).
2. Buniiatova, I. R. (2003). Evoliutziia hipotaksysu v hermanskykh movakh (IV–XIII st.) [Evolution of Hypotaxis in Germanic Languages (4th-13th c.)]. K.: Vyd. tzentr KNLU, 327 p. (in Ukrainian).

3. Velikhov, P. (2016*)*. Mashinnoie obuchenie dlia ponimaniia estestvennogo yazyka [Machine Learning for Understanding Natural Language]. *Otkrytyie sistemy. SUBD, № 01* (in Russian). https://www.osp.ru/os/2016/01/13048649/
4. Hyryn, O. V. (2017). Avtomatychnyi syntaksychnyi analiz anhliiskoi movy: zastosuvannia ta perspektyvy [Automatic Syntactic Analysis of English Language. Application and Perspectives]. *Visnyk Zhytomyrskoho derzhavnoho universytetu imeni Ivana Franka*, Zhytomyr: Vyd-vo ZhDU im. I. Franka, *Vypusk 1 (85)*, 26–30 (in Ukrainian).
5. Polkhovska, M. V. (2013). Analiz anhliiskykh medialnykh konstruktsii z pozytsii heneratyvnoi hramatyky [Generative Perpective to the Analysis of English Media Constructions]. *Studia Philologica, Vyp. 2*, 32–36 (in Ukrainian).
6. Polkhovska, M. V. (2012). Kryterii rozriznennia medialnykh ta erhatyvnykh konstruktsii v anhliiskii movi [Defining Criteria of Media and Ergative Constructions in English]. *Naukovi zapysky Natsionalnoho universytetu "Ostrozka akademia", Ser.: Filolohichna*, *Vyp. 26*, 277–280 (in Ukrainian).
7. Chen, P. (2009). A Fully Unsupervised Word Sense Disambiguation Method Using Dependency Knowledge. P. Chen, W. Ding, C. Bowes, D. Brown, *Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 28–36.
8. Fleiss, J. L. (2013). Statistical Methods for Rates and Proportions. John Wiley & Sons, 800 p.
9. Goldberg, Y. (2017). Neural Network Methods for Natural Language Processing. Morgan & Claypool Publishers, 309p.
10. Hollingsworth, Ch. (2012). Using Dependency-based Annotations for Authorship Identification. *Proceedings of Text, Speech and Dialogue, 15th International Conference,* Berlin, V. 7499, 314–319.
11. Kovar, V. (2011). Information Extraction for Czech Based on Syntactic Analysis. *Proceedings of the 5th Language & Technology Conference*, Poznan: Fundacja Universytetu im. A. Mickiewicza, 466–470.
12. Mikolov, T. (2013). Efficient Estimation of Word Representations in Vector Space. 12 p.
13. Mohd, S. H. (2013). Word Sense Ambiguity: A Survey. *International Journal of Computer and Information Technology, Vol. 02*, Issue 06, 1161–1168.
14. Reese, R. M. (2015) Natural Language Processing with Java. Packt Publishing, 262 p.

Дата надходження статті до редакції: 23.09.2020 р.

Прийнято до друку: 23.10.2020 р.