УДК 811.111'367.7:659.3

О. В. Гирин,
кандидат філологічних наук, доцент
(Житомирський державний університет імені Івана Франка)
oleg_hyryn@ukr.net
ORCID: 0000-0002-3641-2440

АВТОМАТИЧНИЙ СИНТАКСИЧНИЙ АНАЛІЗ АНГЛІЙСЬКОГО РЕЧЕННЯ: ЗАСТОСУВАННЯ ТА ПЕРСПЕКТИВИ

Стаття присвячена автоматичному синтаксичному аналізу англійської мови; проведено огляд основних сфер використання автоматичного синтаксичного аналізу та визначено завдання, які достатньою мірою не вирішені. У статті обґрунтовано, як автоматичний синтаксичний аналіз може покращити вирішення того чи іншого завдання в межах проаналізованих сфер використання. У результаті проведеного аналізу було визначено коло лінгвістичних питань, що сприятиме розробці більш досконалої моделі автоматичного синтаксичного аналізу.

Ключові слова: *парсинг, обробка природних мов, стиломерія, формальна логіка.*

Постановка проблеми. По мірі того, як використання цифрових технологій стає все більш невід'ємною частиною нашого життя, існує нагальна необхідність заміни частини роботи, яку може виконувати людина, автоматичною операцією. З-поміж інших завдань, які можуть виконуватися автоматично, є обробка природної мови (ОПМ). Метою ОПМ є дослідження механізмів (як внутрішніх, так і зовнішніх) природних мов і використання цих знань в додатках та програмах, які будуть служити для полегшення повсякденного спілкування із використанням машин.

Можливо, самі того не помічаючи, ми щодня використовуємо додатки, що здійснюють ОПМ: ми пишемо повідомлення зі словником T9, чуємо синтезовану мову, коли користуємося громадським транспортом (вокзали, метро тощо), використовуємо пошукову систему для миттєвого доступу до інформації, перевіряємо орфографію в текстовому процесорі, використовуємо служби автоматичного перекладу для іноземних мов. Робота всіх цих програм та додатків ґрунтується на певному рівні розуміння того, як працює мова.

З іншого боку, ми хотіли б використовувати програми, для яких розробникам ще не вистачає знань: ми не можемо, скажімо, спілкуватися з автоматом / програмою продажу квитків про найкоротший шлях до місця, адресу якого ми не пам'ятаємо; ми не можемо отримати від програми логічного висновку на основі обробленого масиву тексту; автоматичний переклад все ще містить серйозні помилки в орфографії, стилі, лексиці, фразеології та, особливо, граматиці.

Аналіз останніх досліджень і публікацій. Обробці природної мови в зарубіжній лінгвістиці з 1967 року присвячено численні роботи. Однак з розвитком можливостей технологій змінюються та вдосконалюються підходи до ОПМ Так, сьогодні питання, пов'язані з автоматичним аналізом мовлення знаходять висвітлення у роботах таких дослідників як: Fleiss J. L. [1], Hobbs J. R [2], Hollingsworth Ch. [3] та ін. Натомість у вітчизняній лінгвістиці об'єктом аналізу є обробка української мови (Дарчук Н. П. [4], Чейлитко Н. Г. [5]). Разом з тим, напрацювання в царині граматики англійської мови, зокрема генеративної граматики (Буніятова І. Р. [6], Полховська М. В. [7]), дозволяють перехід до їх застосування у прикладній галузі.

Метою статті є висвітлення результатів аналізу поточного стану в галузі ОПМ та конкретних проблем, наявних у ній.

На основі цього аналізу пропонуються певні напрямки, які слід врахувати при розробці моделі синтаксичного аналізу.

Виклад основного матеріалу. Синтаксичний аналіз природних мов (парсинг) часто називають "наріжним каменем" ОПМ, необхідною основою для поглибленого аналізу і розуміння будь-якої природної мови. Синтаксичний аналіз речень має на меті виявлення структури речення, мовних одиниць, що несуть смислове навантаження, а також відношення між ними. З іншого боку, в сучасних "інтелектуальних" додатках синтаксичний аналіз часто підміняється статистичними [1] або навіть стохастичними методами. В царині ОПМ існує думка, що синтаксичний аналіз не є необхідним для прикладного застосування.

Отже, проведемо огляд проблем в області ОПМ за сферами її використання та визначимо завдання, які достатньою мірою не вирішені (хоча існують часткові рішення в більшості випадків), а також визначимо, як автоматичний синтаксичний аналіз може покращити вирішення того чи іншого завдання в межах зазначених сфер використання.

Пошук інформації передбачає пошук документів / сайтів, веб-сторінок, що відносяться до поданого запиту. Зокрема, в системі пошуку Google (та деяких інших пошукових системах) це завдання

вирішується поєднанням показників оцінки релевантності документа певному запиту на основі слів запиту (їх деривативів і синоніми) в межах, зокрема, алгоритму PageRank [8] тощо.

Однак досі існують ситуації, в яких більш складна синтаксична обробка запитів допомогла б отримати більш точні результати, наприклад, у випадку запитів типу "*science fiction novel about Asimov*" прийменник *about* ігнорується і у результатах пошуку будуть представлені посилання на романи авторства А. Азімова (*novels by Asimov*).

Інший напрям для покращення пошуку необхідної інформації полягає в обробці питань, наприклад, "*Who revealed Snowden?*". Ми хочемо знайти інформацію, в якій Сноуден є об'єктом викриття. Натомість у результатах пошуку ми знайдемо ряд посилань на документи, де Сноуден є суб'єктом викриття. Звичайно, у пошукових системах є опція розширеного пошуку, зокрема використання запиту в лапках, але і в цьому разі такий запит не дасть посилання на інформацію, яка нам потрібна, виходячи із синтаксичної структури питання-запиту.

Відповіді на запитання, метою чого є отримання відповіді на основі даних бази знань (в т.ч. інтернету). У деяких випадках, цей напрям розуміють так само, як і пошук інформації з тією лише різницею, що запитом є питання природною мовою, а не на запит, що містить ключові слова. Окрім того, більшість нинішніх систем відповідають на питання на цьому принципі – вони "витягають" ключові слова з вхідного питання, а потім використовують пошукову систему і просто видають користувачу ту інформацію, яку видала пошукова система.

У нашій інтерпретації відповіді на питання завдань є більш складним процесом. Так, відповідь має містити саме ту інформацію, запит на яку подається у запитанні. При цьому відповідь має бути максимально лаконічною, без втрати семантики та без надлишкової інформації. Наприклад, у базі знань наявна послідовність таких речень:

(1) *Mary is a student. She is 20. She majors in English.*

Правильною відповіддю на питання "*What does Mary study?*" є

(1') *Mary studies English.*

Послідовність речень з прикладу (1) містить надлишкову інформацію, про яку у питанні-запиті не йдеться (*She is 20*). Речення, яке містить ім'я *Mary*, не дає відповіді на поставлене запитання. Натомість речення із *English* не містить імені та не має дієслова *study*. Отримання відповіді (1') передбачає заміну анафори *She* антецедентом *Mary* та заміну *majors* синонімом із запитання – *does study > studies.*

У пошукових системах знайдені документи / сторінки зазвичай містять ключові слова із запиту, однак, це не означає, що вони містять відповідь на питання, як показано у прикладі із Сноуденом або на прикладі аналогічних запитів у вигляді питань, таких як "*Who supports Bill Gates?*", де пошукова система наводить посилання на статті, документи та ін., у яких йдеться про те, кого підтримує Б. Гейтс, а не хто підтримує його.

Розробка алгоритму синтаксичного аналізу для отримання точних, лаконічних відповідей на запитання-запити є перспективним напрямом нашої діяльності в царині ОПМ. Очевидно, необхідною є попередня обробка вихідної інформації, наприклад, визначення смислових елементів речень та маркування їх відповідним чином (наприклад граматичний та семантичний суб'єкт та об'єкт, модифікатор часу, місця тощо). Повне вирішення кола завдань, що постають при розробці алгоритму відповідей на питання, звичайно, передбачає врахування великої кількості додаткових мовних явищ, включаючи, анафори, гіперо- та гіпонімію, синонімію, деривацію.

Формулювання логічних висновків у контексті інформатики та логіки зазвичай передбачає створення нових формул, зокрема логічного формалізму на основі певних припущень. Це означає генерування ряду нових речень природною мовою відповідно до інших речень природною мовою, які називають базою знань. Приклад можна представити наступним чином:

(2) база знань: *Joseph Conrad (1857 –1924) is a British writer born in Berdychiv*
можливі генеровані речення:

(2') *Joseph Conrad is no longer alive;*

(2") *Joseph Conrad wasn't born in England;*

(2"') *Joseph Conrad is famous in literature;*

тощо.

Крім того, таке використання автоматичного синтаксичного (а також лексичного) аналізу може уможливлювати перевірку правильності конкретної формули, тобто чи знаходиться ця формула у наборі допустимих формул, створених з на базі вхідної інформації. Наприклад:

(3) *civil war > unrest;*

(3') *unrest in Ukraine ⇏ civil war in Ukraine*

Такий аналіз формул може використовуватися для спростування так званих "фейкових" новин та повідомлень.

На сьогодні існують кілька експериментальних підходів до створення системи логічних висновків, наприклад: здійснюється автоматичний переклад речень природної мови в мову логіку предикатів, після

цього відбувається генерування умовиводів на базі логіки предикатів, результати перекладаються назад на природну мову [2: 117]; використовуються ручні або стохастичні правила перетворення безпосередньо в природній мові без проміжного використання формальної логіки [9: 126].

Всі ці підходи потребують гнучкого автоматичного синтаксичного аналізу. Однак це завдання залишається невирішеним, тому існує багато перспективних напрямків для дослідження у цьому руслі.

Розпізнавання авторства спрямоване на надійне автоматичне визначення автора анонімного фрагменту тексту (наприклад, у судовій лінгвістиці). Так, автоматична перевірка не лише лексики, а й синтаксису з високою ймовірністю може визначити, чи два анонімні фрагменти тексту були написані одним і тим же автором. Натомість, перевірка автентичності тексту дозволяє визначити, чи певний фрагмент тексту було написано конкретним автором.

Зрозуміло, що ці два завдання водночас і тотожні, і протилежні. Тотожні, тому що вони слугують одній цілі – встановлення авторства. Протилежні, тому що при розпізнаванні авторства наявна визначена кількість текстів (мінімум 2 тексти, максимум визначається опрацьованим творчим доробком автора), а наявність високого відсотка збігів є ознакою авторства При перевірці автентичності зазвичай один текст співставляється з необмеженою кількістю текстів, наявних в базах (зокрема в мережі Інтернет), а наявність високого відсотка збігів є ознакою порушення авторства.

Лексичний аналіз, який використовується при автоматичній перевірці автентичності тесту, не має високої надійності результатів через неврахування лексичної, а особливо граматичної синонімії, як то: активний-пасивний стан, інверсія, еліпсис, використання інфінітивних, дієприкметникових, герундіальних конструкцій тощо. Окрім того, лексичний аналіз не враховує можливості використання перекладного матеріалу при написанні тексту.

У межах цих завдань вартим уваги є так званий стилометричний напрям аналізу тексту природною мовою. У межах цього напряму замість визначення конкретного автора здійснюється аналіз, націлений на різні характеристики, наприклад, вік, рівень освіти чи стать автора. У більшості випадків, ті ж методи можуть використовуватися для всіх завдань цього типу.

В цій галузі було здійснено певні дослідження [10] для виокремлення у тексті стиломів (stylome) – сукупності ознак, що характеризують автора (чи обрану категорію характеристик). Особливості, як правило, лінгвістично мотивовані, наприклад: частота вживання коротких слів, використання конкретних слів або частин мови, граматичних конструкцій, тощо.

Перевірка граматики тексту є однією з найбільш важливих задач ОПМ. Перевірка орфографії та простих граматичних явищ в текстових процесорах вже стала звичною. Однак перевірка більш складних мовних явищ досі представляє проблему. Так, граматична перевірка, що здійснюється за допомогою пакетів програмного забезпечення, таких як Microsoft Office, може вирішувати лише обмежене коло граматичних помилок, наприклад: узгодження і керування. Але вони далекі від того, щоб знайти всі помилки.

Очевидно, у природних мовах помилки при суб'єктно-предикатному узгодженні належать до числа найбільш грубих. Окрім того в кожній мові є "власні" найбільш поширені помилки. На сьогодні автоматична перевірка граматики не може усунути значної кількості таких помилок. Складністю є той факт, що велика кількість граматичних правил включають в себе не лише морфо-синтаксичні, але і семантичні та прагматичні аспекти, що робить їх формалізацію для автоматичної перевірки проблематичною.

Так, відомий приклад (речення (4)) при автоматичній перевірці не міститиме помилок:

(4) *One morning I shot an elephant in my pajamas.*

Тому на сьогодні повний парсинг при граматичній перевірці використовується рідко, натомість використовується підхід, що передбачає врахування найбільш поширених помилок або полегшеної модифікації повних синтаксичних формалізмів.

Синтез природної мови. Незважаючи на те, що синтезоване мовлення вже використовується у сучасному світі, сфера його застосування обмежується репродуктивністю такого мовлення. На залізничних вокзалах, у метро та при озвучуванні заданого тексту генератор мовлення відтворює мовні одиниці, які були попередньо введені в програму. У цьому випадку не йдеться власне про генерування мовлення, а лише про відтворення. Перспективним напрямком в галузі ОПМ є розробка алгоритму для незалежного генерування мовлення. Це завдання включає в себе деякі попередні, зокрема пошук інформації, відповіді на запитання, логічні висновки.

Супутні завдання. Крім наведених сфер застосування синтаксичного аналізу існують завдання, пов'язані з обробкою мови, які зазвичай приховані від користувачів, однак вони необхідні для обробки природної мови і можуть застосовуватися для вирішення завдань відразу у кількох сферах використання.

Одним з них є визначення морфологічного класу – вибір правильного морфологічного маркера з набору всіх можливих маркерів для окремого слова (наприклад, щоб вирішити, чи "point" – дієслово чи іменник). Багато програм, що існують на сьогодні, залежать від цієї обробки. Іншими завданнями, від яких залежить синтаксичний аналіз природної мови – є вирішення проблеми анафор (пошук

антецедентів), еліпсованих частин речення. Важливим супутнім завданням, яке може виокремитися в самостійне завдання, є розпізнавання неграматичного мовлення.

Висновки. Таким чином, сфера ОПМ на сьогодні є широкою, однак очевидними залишаються напрямки для покращення моделі парсингу, що, у свою чергу, розширить сферу його застосування та покращить результати у галузях, у яких вже використовується автоматичний синтаксичний аналіз. При покращенні автоматичного синтаксичного аналізу природної мови незамінними залишаються надбання в обробці лексики та морфології.

Перспектива подальшого дослідження. Врахування наступного кола лінгвістичних питань сприятиме розробці більш досконалої моделі парсингу: лексична та граматична синонімія, лексико-граматична омонімія, гіперо-, гіпонімія, лексико-семантичні поля, анафори, еліпсис, інверсія. При цьому наведений перелік не є вичерпним і може бути доповнений.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ ТА ЛІТЕРАТУРИ

1. Fleiss J. L. Statistical Methods for Rates and Proportions / J. L. Fleiss, B. Levin, Ch. P. Myunghee. – John Wiley & Sons, 2013. – 800 p.
2. Hobbs J. R. Discourse and Inference : Magnum Opus in Progress / J. R. Hobbs. – Marina del Rey, 2014. – 168 p.
3. Hollingsworth Ch. Using Dependency-Based Annotations for Authorship Identification / Ch. Hollingsworth // Proceedings of Text, Speech and Dialogue, 15th International Conference. – Berlin. – V. 7499. – 2012. – P. 314–319.
4. Дарчук Н. П. Автоматичний синтаксичний аналіз текстів корпусу української мови / Н. П. Дарчук //Українське мовознавство. – КНУ ім. Т. Шевченка, 2013. – № 43. – С. 11–19.
5. Чейлитко Н. Г. Корпусне дослідження зон зв'язків словоформ в українському реченні / Н. Г. Чейлитко // Лінгвістичні студії : [зб. наук. праць]. – Донецьк, Вид-во ДонНУ, 2009. – Вип. 18. – С. 268–275.
6. Буніятова І. Р. Еволюція гіпотаксису в германських мовах (IV – XIII ст.) : [монографія] / І. Р. Буніятова. – К. : Вид. центр КНЛУ, 2003. – 327 с.
7. Полховська М. В. Аналіз англійських медіальних конструкцій з позиції генеративної граматики / М. В. Полховська // Studia Philologica. – 2013. – Вип. 2. – С. 32–36.
8. Page L. The PageRank Citation Ranking : Bringing Order to the Web. Technical Report // L. Page, S. Brin, R. Motwani, T. Winograd. – 1999-66, Stanford : InfoLab, 1999. – 128 p.
9. Stern A. The BIUTEE Research Platform for Transformation-Based Textual Entailment Recognition / A. Stern, I. Dagan // Linguistic Issues in Language Technology. – No 9. – 2013. – P. 120–146.
10. Koppel M. Computational Methods in Authorship Attribution / M. Koppel, J. Schler, Sh. Argamon // Journal of the American Society for Information Science and Technology. – No 60 (1). – 2009. – P. 9–26.

REFERENCES (TRANSLATED & TRANSLITERATED)

1. Fleiss J. L. Statistical Methods for Rates and Proportions / J. L. Fleiss, B. Levin, Ch. P. Myunghee. – John Wiley & Sons, 2013. – 800 p.
2. Hobbs J. R. Discourse and Inference : Magnum Opus in Progress / J. R. Hobbs. – Marina del Rey, 2014. – 168 p.
3. Hollingsworth Ch. Using Dependency-Based Annotations for Authorship Identification / Ch. Hollingsworth // Proceedings of Text, Speech and Dialogue, 15th International Conference. – Berlin. – V. 7499. – 2012. – P. 314–319.
4. Darchuk N. P. Avtomatychnyi syntaksychnyi analiz tekstiv korpusu ukrains'koi movy [Automatic Syntactic Aanalysis of the Ukrainian Text Corpus] / N. P. Darchuk // Ukrains'ke movoznavstvo [Ukrainian Linguistics]. – KNU im. T. Shevchenka, 2013. – № 43. – S. 11–19.
5. Cheilytko N. H. Korpusne doslidzennia zon zv'iazkiv slovoform v ukrains'komu rechenni [Corpus Study of the Connection Zones of Word Forms an a Ukrainian Sentence] / N. H. Cheilytko // Linhvistychni studii : [zb. nauk. pratz]. – Donets'k, Vyd-vo DonNU, 2009. – Vyp. 18. – S. 268–275.
6. Buniiatova I. R. Evoliutsiia hipotaksysu v hermans'kykh movakh (IV – XIII st.) [Evolution of Hypo-Taxis in Germanic Languages (4^{th}–13^{th} c.)] : [monohrafiia] / I. R. Buniiatova. – K. : Vyd. tzentr KNLU, 2003. – 327 s.
7. Polkhovska M. V. Analiz anhliis'kykh medial'nykh konstruktsii z pozytzii heneratyvnoi hramatyky [A Generative Perpective to the Analysis of English Medial Constructions] / M. V. Polkhovska // Studia philologica. – 2013. – Vyp. 2. – S. 32–36.
8. Page L. The PageRank Citation Ranking : Bringing Order to the Web. Technical Report // L. Page, S. Brin, R. Motwani, T. Winograd. – 1999-66, Stanford : InfoLab, 1999. – 128 p.
9. Stern A. The BIUTEE Research Platform for Transformation-Based Textual Entailment Recognition / A. Stern, I. Dagan // Linguistic Issues in Language Technology. – No 9. – 2013. – P. 120–146.
10. Koppel M. Computational Methods in Authorship Attribution / M. Koppel, J. Schler, Sh. Argamon // Journal of the American Society for Information Science and Technology. – No 60 (1). – 2009. – P. 9–26.

Гирин О. В. Автоматический синтаксический анализ английского предложения: применение и перспективы.

Статья посвящена автоматическому синтаксическому анализу английского предложения; проведен обзор основных сфер использования автоматического синтаксического анализа и определены задачи, которые в достаточной степени не решены. В статье обосновано, как автоматический

синтаксический анализ может улучшить решение того или иного задания в пределах проанализированных сфер использования. В результате проведенного анализа был определен круг лингвистических вопросов, решение которых будет способствовать разработке более совершенной модели автоматического синтаксического анализа.

Ключевые слова: *парсинг, обработка естественных языков, стилометрия, формальная логика.*

Hyryn O. V. Automatic Syntactic Analysis of an English Sentence: Application and Perspectives.

The article deals with the automatic syntactic analysis (parsing) of an English sentence. The existing spheres of its application and the natural language processing approaches serve as the research material of the study. Automatic syntactic analysis of an English sentence has not received due attention in Ukrainian linguistics, though the scientific results of formal grammar schools and scholars enable wide perspectives thereof. The scientific methods of analysis, synthesis, description and comparison as well as linguistic methods of substitution and transformation have been used in order to single out the main areas of application of automatic syntactic analysis today and to define problematic issues, which have not been sufficiently solved yet. The article describes the 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 analysis of ungrammatical sentences, morphological class definition, anaphora resolution etc. The article provides tips of how automatic syntactic analysis can improve the solution of a particular task within the analyzed application spheres. The analysis identified 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.

Key words: *parsing, natural language processing, stylometry, formal logics.*