К.І. Марчук
Житомирський державний університет
імені Івана Франка
Науковий керівник:
кандидат філологічних наук, доцент
О. В. Гирин

Автоматичний синтаксичний аналіз англомовного тексту : застосування та перспективи

Сучасний світ лінгвістики потребує постійного розвитку тому опрацювання текстової інформації незмінно залишається провідним завданням лінгвістики. Наші знання про дійсність втілюються у певній вербалізованій формі. Дати можливість комп'ютеру "розуміти" текст означає, що комп'ютер матиме властивість виокремлювати з нього необхідну інформацію. Таке "розуміння" тексту полягає у вмінні аналізувати його на різних рівнях.

Ю.М.Марчук зазначає, що: "...У комп'ютерній лінгвістиці поняття морфологічного аналізу є поняттям операційним. Якщо у традиційній лінгвістиці до морфологічного аналізу належить те, що характеризує форму і відповідає на питання "що" класифікують, то в обчислювальній (прикладній) лінгвістиці важливо не "що", а "як" одержують ту чи іншу інформацію, тобто з форми слова у тексті" [9]

У повсякденному житті ми використовуємо безліч додатків, які допомагають нам аналізувати, редагувати, розуміти текст. Існує декілька підходів для аналізу текстів : автоматично-морфологічний аналіз тексту (АМА) та автоматичний синтаксичний аналіз АСА.

На рівні словосполучення АСА передбачає автоматичне виокремлення словосполучень, приписування їм типу синтаксичного зв'язку та автоматичне укладання словників словосполучень (дієслівних, іменних, ад'єктивних). На рівні речення має здійснюватися повний синтаксичний аналіз у вигляді дерев залежностей.

АСА спрямований на виявлення в тексті синтаксичних структур та їхнє формалізоване представлення.

Інтегральні системи як результат передбачають одержання всієї синтаксичної структури речення, локальні – лише якоїсь частини такої структури.

Необхідність вивчення сполучуваності лексичних одиниць зумовлена не розробленістю широкого кола як теоретичних, так і прикладних проблем. Теоретичні аспекти, які потребують вивчення, – це, зокрема, граматична і лексична валентність слів, типова сполучуваність, синонімія словосполучень різних структурних типів, лексична і граматична валентність як критерій синонімічності, закони комбінаторики словосполучень різних типів і розрядів, лексична валентність як критерій розмежування вільних і фразеологічних словосполучень, взаємодія стійкості й ідіоматичності тощо. До прикладних проблем можна віднести автоматизацію лінгвістичних досліджень, автоматичне визначення меж словосполучень, установлення критеріїв членування фрази на синтагми, автоматичний синтаксичний аналіз речення, автоматичне реферування й анотування тексту на основі сполучувальнісних критеріїв тощо.

Безліч робіт присвячено обробці природної мови в зарубіжній лінгвістиці. Fleiss J. L. [7], Hollingsworth Ch. [9], Kovar V. [11] та у вітчизняній лінгвістиці

об’єктом аналізу є обробка української мови (Дарчук Н. П. [1], Чейлитко Н. Г. [2] Вихованець І. Р[4],Загнітко А. П[4]).

Метою статті є пояснення проблематики обробки природної мови за допомогою цифрових та комп’ютерних ресурсів.

Автоматичний аналіз тексту є перспективним напрямом, адже він є основою для розуміння будь–якої мови. Синтаксичний аналіз має на меті виявлення зв’язків між членами речення, встановлення смислового значення, членування речення. Автоматичний синтаксичний аналіз текступредбачає використання комп’ютерного синтаксису. Він спрямований на виявлення в тексті синтаксичних структур та їхнє формалізоване представлення. Здійснення в таких системах процедури розкладу тексту на мінімальні синтагми – пари слів, пов’язані певним типом синтаксичного зв’язку: координацією (між членами предикативної пари – підметом та присудком), узгодженням, керуванням або приляганням. Локальні системи при цьому застосовують процедури методу безпосередніх складників, або аналізу контактних слів у реченні, розроблені представниками американської дескриптивної лінгвістики. Інтегральні системи використовують процедури граматики залежностей, розроблені представниками генеративної лінгвістики і спрямовані на виявлення в тексті головного й залежного слів безвідносно до їхньої позиції в реченні.

Методика послідовного аналізу тексту й виявлення синтаксичної структури представлених у ньому речень передбачає створення словника еталонів словосполук (синтагм), записаних у термінах граматичних класів слів. Методика передбачувального аналізу ґрунтується на наборах синтаксичних передбачень – гіпотетичних (ймовірних, можливих) синтаксичних структур та синтаксичних функцій окремих слів. Її розвитком є методика опорних точок, яка для слів з певними характеристиками визначає типові контексти, що діагностують вживання слова з тією чи іншою синтаксичною функцією в разі його багатофункціональності. Методика фільтрів дозволяє завдяки встановлюваним обмеженням на вживання, сполучуваність або переміщення слів у реченні з усього набору інформації про певні слова виявити інформацію, релевантну саме для тексту, що аналізується.

Зазвичай задаючи запит у систему пошуку, вона витягує інформацію за вказаними словами або символами. Але ми прагнемо отримати логічну та змістовну відповідь на поставлене запитання, без надлишкової інформації та втрати семантичних конструкцій.

Розробка алгоритму синтаксичного аналізу для отримання точних, лаконічних відповідей на запитання-запити є перспективним напрямом в 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[3].

На сьогодні існує кілька експериментальних підходів до створення системи логічних висновків, наприклад: здійснюється автоматичний переклад речень природної мови в мову логіки предикатів, після цього відбувається генерування

умовиводів на базі логіки предикатів, результати перекладаються назад на природну мову ; використовуються ручні або стохастичні правила перетворення безпосередньо в природній мові без проміжного використання формальної логіки [3].

Для автоматичного синтаксичного аналізу, в сучасному світі лінгвістики використовують парсинг. Парсинг - це процес зіставлення лінійної послідовності лексем (слів , токенів) мови з його формальної граматикою . Результатом зазвичай є дерево залежностей (синтаксичне дерево). Побудова автоматичних синтаксичних аналізаторів (парсерів) для великих корпусів є однією з найважливіших областей комп'ютерної лінгвістики. Більшість підходів об'єднує якісні та кількісні виміри . Поряд з різними статистичними підходами , які тренуються на забезпечених вручну позначками синтаксичних деревах (tree - banks) , багато синтаксичних аналізаторів використовують засновані на правилах або засновані на обмеженнях підходи , які прямо моделюють специфічні лінгвістичні теорії . Розробка цих синтаксичних аналізаторів тісно переплітається з розвитком цих теорій. Оскільки більшість пропозицій неоднозначні в будь-якої теорії , на основі правил (або переліку обмежень) має бути розроблена стратегія зняття неоднозначності. Багато стратегії зняття неоднозначності покладаються на кількісні дані - частоту даної структури в даному корпусі (тип) , обмеження на вибірку для даних лексичних одиниць , які були отримані або виділені з корпусних даних , і т.д.

Однак автоматичний аналіз природної мови багатозначний - він , як правило, дає кілька варіантів аналізу для однієї лексичної одиниці (слова , словосполучення, речення) . У цьому випадку говорять про граматичної омонімії . Зняття неоднозначності (морфологічної , синтаксичної) в цілому є однією з найважливіших і найскладніших задач комп'ютерної лінгвістики. При створенні корпусів для зняття неоднозначності використовуються автоматичні і ручні способи . Корпуси нового покоління включають сотні мільйонів слів, тому висуваються принципи розробки систем , які б мінімізували втручання людини. Можливе автоматичне усунення морфологічної або синтаксичної неоднозначності, як правило , ґрунтується на використанні інформації більш високого рівня (синтаксичного , семантичного) із застосуванням статистичних методів .

Для вирішення різних лінгвістичних завдань недостатньо мати масив текстів. Необхідно , щоб тексти містили в собі явним чином зазначену різного роду додаткову лінгвістичну та екстралінгвальну інформацію. Так , на матеріалі корпусу , наприклад браунівскому , можна легко виявити частотність слів – їх регулярне вживання в певних контекстах. Однак це буде частотність токенів (словоформ) . Для визначення частоти лексем кожному слову повинна бути приписана її лема.

На просторах зарубіжжя дане лінгвістичне та програмне забезпечення не є новим і швидко набирає свого розвитку. На теренах України такий проект розробляється на основі української мови в Київському Національному Університеті імені Т.Г.Шевченка. Саме тому нас цікавить дана галузь лінгвістики. Це єдиний лінгвістичний ресурс, що містить синтаксичне розмічування текстів мови, яке здійснюється тільки автоматично на базі

повного автоматичного морфологічного аналізу зі знятою омонімією (про перші спроби АСА див. [4]).

Київський національний університет імені Тараса Шевченка у теоретичному плані вирізнення словосполучення з реченнєвої структури на великих різностильових масивах текстів, які входять до Корпусу української мови, дає можливість дослідникам української мови визначити синтаксичну і семантичну ємність цієї синтаксичної одиниці[4]. Необхідність вивчення сполучуваності лексичних одиниць зумовлена не розробленістю широкого кола як теоретичних, так і прикладних проблем.

Для дослідження англомовного матеріалу в Україні існує доволі не велика кількість програм, яка дає змогу аналізувати англомовний матеріал на синтаксичному рівні у вільному доступі. Теоретичні аспекти, які потребують вивчення, – це, зокрема, граматична і лексична валентність слів, типова сполучуваність, синонімія словосполучень різних структурних типів, лексична і граматична валентність як критерій синонімічності, закони комбінаторики словосполучень різних типів і розрядів, лексична валентність як критерій розмежування вільних і фразеологічних словосполучень, взаємодія стійкості й ідіоматичності тощо. До прикладних проблем можна віднести автоматизацію лінгвістичних досліджень, автоматичне визначення меж словосполучень, установлення критеріїв членування фрази на синтагми, автоматичний синтаксичний аналіз речення, автоматичне реферування й анотування тексту на основі сполучувальнісних критеріїв тощо.

Не вдаючись до теоретичних дискусій щодо деяких питань синтаксису, зазначимо, що в основі АСА лежить формально-синтаксичний аспект вивчення речення. Ані семантико-синтаксичний і функціональний, ані комунікативний підхід до розгляду речення не можуть стати основою автоматизації. Тоді як дослідження формально-синтаксичної будови речення дає можливість створити словник синтаксем, для якого попередньо слід укласти таксономічну класифікацію лексики, що у майбутньому уможливить автоматичне визначення синтаксичних відношень між членами словосполучення. Формальна граматика, адаптована для потреб автоматизації, базуватиметься на гіпотаксисі як провідному аспекті синтаксичного ладу мови; а паратаксис буде додатковим аспектом, оскільки виокремлення сурядних словосполучень з погляду автоматизації не становить суттєвих труднощів. У ході автоматичного синтаксичного аналізу речення насамперед має здійснюватися автоматичний пошук зв’язків слів у реченні. Ознаки таких зв’язків наявні, зокрема, у словозмінних характеристиках слів. У реченні послідовно розгортається підпорядкування слів одне одному: одне слово (залежне) змінює форму, щоб адаптуватися до вимог іншого слова (головного). Таким чином, машина має виокремлювати пари слів, пов’язані граматичним зв’язком, позначаючи напрямок залежності.

Автоматизація лексико-граматичного аналізу досягла практично такий же точності. Синтаксичний аналіз на відміну від лексико - граматичного аналізу тексту, синтаксичний аналіз - розвивається в області прикладної лінгвістики. Мета синтаксичного аналізу – автоматична побудова функціонального дерева фрази, тобто знаходження взаємозалежностей між різнорівневими елементами пропозиції. Вважається, що маючи успішно побудоване

функціональне дерево фрази, можна виділити з пропозиції смислові елементи: логічний суб'єкт, логічний предикат, прямі і непрямі додатки і різні види обставин. Існує велика кількість різних підходів до синтаксичному аналізу текстів, наприклад система LTAG [8]. Головна особливість цієї системи полягає в побудові елементарних смислових дерев пропозиції. Кожне елементарне дерево містить в собі всю синтаксичну і семантичну інформацію про конкретний слові або групи слів. До цих деревах можуть бути застосовані операції примикання і підстановки. Підстановка є простою операцією - підстановкою дерева до висить вершині іншого дерева. Примикання є більш складною операцією - приєднання деякого дерева до внутрішніх вершин іншого дерева. Даний алгоритм докладно описаний у роботі [8]. Нижче розглянуто один із загальних підходів синтаксичного аналізу речення.

Синтаксичний розбір пропозиції відбувається шляхом набору послідовних перетворень:

- Пошук граматичних ідіом;
- Лексико-граматичний аналіз речення з усуненням неоднозначності у визначенні частин мови;
- Знаходження іменний групи об'єкта і суб'єкта;
- Знаходження дієслівної групи;
- Виділення головних і підрядних речень.

Наведемо приклад синтаксичного розбору пропозиції

[We] {have found} / that [subsequent addition] (of [the second inducer]) (of [either system]) <after {allowing} [single induction] {to proceed} +> (for [15 minutes]) (also) {results} (in [increased reproduction]) + \ + (of [both enzymes]).

Позначення:

[...] - Група іменника;
(...) - Група додатка;
{...} - Дієслівна група;
/ ... \ I <...> - головні і підрядні речення;
+ - Закінчення дієслівного оточення.

Сьогодні постійно зростає кількість програмних продуктів, пов'язаних з даною темою, підвищується їх якість. Але, незважаючи на це, ті системи перекладу, реферування та експертні системи, які на сьогодні вважаються кращими, далеко не ідеальні і вимагають серйозних доопрацювань. Все це говорить про необхідність продовження досліджень з питань, пов'язаних з обробкою природної мови в задачах ДН та розробкою нових підходів та алгоритмів, заснованих на методах штучного інтелекту.

Висновки. Синтаксичний аналіз природної мови потребує розвитку та покращення, особливо його граматична складова. Також існує нагальність розробки моделі для покращення парсингу, для поширення сфери його використання.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ ТА ЛІТЕРАТУРИ:

1. Буніятова І. Р. Еволюція гіпотаксису в германських мовах (IV – XIII ст.) : монографія / І.Р. Буніятова. – К. : Вид. центр КНЛУ, 2003. – 327 с
2. Вихованець І. Р. Граматика української мови. Синтаксис / І. Р. Вихова- нець. – К., 1993.
3. Гирин О.В. Автоматичний синтаксичний аналіз англійської мови: застосування та перспективи – [Електронний ресурс] / О.В.Гирин – Режим доступу: http://nniif.org.ua/File/17govasa.pdf

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. Загнітко А. П. Основи українського теоретичного синтаксису. Частина 1 / А. П. Загнітко. – Горлівка, 2004.
6. Кубрякова Е.С., Демьянков В.З., Панкрац Ю.Г., Лузина Л.Г. Краткий словарь когнитивных терминов. – М., 1996
7. Лингвистическая прагматика и общение с ЭВМ / Отв. ред. Ю.Н.Марчук. – М., 1989
8. Сучасна українська літературна мова. Морфологія. Синак- сис. – К., 2010. 6. Русская граматика: в 2 т. Т. 2. Синтаксис. – М., 1980. – С. 21.
9. 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.
10. Computer-Konkordanz zum Novum Testamentum Graece. – Berlin, New York, 1980.
11. 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.
12. Stern A. The BIUTEE research platform for transformationbased textual entailment recognition / A. Stern, I. Dagan // Linguistic Issues in Language Technology. – No 9. – 2013. – P. 120–146.