І.С. Кілімінська
Житомирський державний університет
імені І. Франка
Науковий керівник:
к.ф.н., доц. О.В. Гирин

АВТОМАТИЧНИЙ СИНТАКСИЧНИЙ АНАЛІЗ АНГЛІЙСЬКИХ ПОШУКОВИХ ЗАПИТІВ

Сьогодні опрацювання текстової інформації залишається одним із провідних завдань лінгвістики. Як наука, лінгвістика прагне до постійного розвитку та прогресивних змін. Водночас, цифрові технології давно стали частиною нашого повсякденного життя, тому стає очевидним той факт, що автоматична обробка текстової інформації та використання отриманих знань в різноманітних корисних програмах або ж додатках може полегшити спілкування та принести користь в наукових розробках та дослідженнях.

Проблемі інтелектуалізації інформаційних технологій приділяється значна увага в провідних країнах Європи та, зокрема, в США. На розвиток лінгвістичних інформаційних систем спрямовується велика кількість науково-дослідницьких програм. У зв'язку з невпинним розвитком комп'ютерно-комунікаційних технологій ця проблема набуває ще більшої значущості.

Ще в минулому столітті науковці витрачали величезні зусилля на розробку спеціальних алгоритмів та комп'ютерних програм обробки текстів. Аби мати змогу аналізувати та синтезувати тексти автоматизовано, науковці створювали різноманітні моделі процесів обробки тексту, алгоритми та структури представлення даних. Аналіз природномовних текстів складався з певної послідовності процесів: від морфологічного аналізу, синтаксичного аналізу до семантичного аналізу. Тобто для кожного етапу створювався власний алгоритм та моделі [4: 4].

При обробці природної мови (ОПМ) найбільшими труднощами є обробка таких явищ як полісемія, омонімія, обробка фразеологізмів і т.д. Такі явища додають неоднозначності та досить ускладнюють можливість встановлення коректного відображення семантично-синтаксичної структури в його формальне логічне представлення.

Значний внесок у вивчення автоматичного синтаксичного аналізу зробили Н.Г. Чейлитко [2], І. Р. Буніятова [3], Дарчук Н.П. [4]. Також автоматичний

аналіз мовлення висвітлюють такі зарубіжні дослідники як Fleiss J. L. [9], Hollingsworth Ch. [10], Kovar V. [11].

Мета статті полягає у висвітленні результатів дослідження у галузі автоматичного синтаксичного аналізу та аналізу англійських пошукових запитів.

Обробка природної мови (Natural Language Processing, NLP) – це напрямок дослідження штучного інтелекту і математичної лінгвістики. ОПМ вивчає проблеми комп'ютерного аналізу і синтезу природних мов. Аналіз у цьому контексті означає розуміння мови, а синтез – генерацію тексту. Вирішення цих проблем буде означати створення більш зручної форми взаємодії комп'ютера та людини. До основних напрямів обробки природної мови відносять такі як витяг фактів, аналіз тональності тексту, відповіді на питання, інформаційний пошук, генерація тексту, переклад і т.д. [1: 138].

Розглянемо основні задачі обробки природної мови, на основі яких і здійснюються пошукові запити [8]:

1. Пошук інформації – це пошук документів або ж сайтів, веб-сторінок, які мають відношення до поданного запиту.
2. Отримання відповіді на запитання – постановка питання природною мовою, а не запиту з використанням ключових слів. Пошукова система у такому випадку «витягує» ключові слова з питання.
3. Розпізнавання авторства – атоматичне визначення авторства твору невідомого автора.
4. Перевірка граматики є однією з найважливіших задач обробки мови. Насьогодні автоматична перевірка грамматика не в змозі вмправити всі помилки в тексті.
5. Синтез природної мови. Сьогодні ми маємо змогу чути синтезоване мовлення. У громадському транспорті генератор мовлення відтворює заздалегідь введені в программу мовні одиниці. Однак насьогодні важливим завданням є власне незалежне генерування мовлення.
6. Крім вищезазначених сфер застосування аналізу існують також супутні завдання, пов'язані з обробкою мови. Ці процесі приховані від звичайних користувачів, але надзвичайно важливі при обробці природної мови . Одним з таких завдань є вибір правильного морфологічного маркера для певного слова. Наприклад, визначення до якої частини мови належить слово answer (іменник чи дієслово).
7. Формулювання логічних висновків – створення нових формул, тобто генерування ряду нових речень природною мовою відповідно до інших речень природною мовою, які називають базою знань.

Під витягом інформації або фактів мається на увазі пошук в неструктурованому або слабо структурованому документі окремих фактів, які вас цікавлять. Наприклад, є певна кількість статей, в яких фігурує велика кількість різних особистостей, і ви хочете скласти базу даних, яка буде зберігати дані про те, хто з цих людей є братом та сестрою. Аналіз тональності тексту означає автоматичне визначення емоційного забарвлення тексту і виявлення ставлення людини, яка написала текст, до об'єкта обговорення. Також одним з найбільш відомих і часто використовуваних напрямків обробки природного тексту є його переклад з однієї природної мови на іншу [5: 29].

Всі тексти на природній мові мають велику кількість слів, які не несуть інформації про текст. В англійській мові такими є артиклі, в українській до них можна віднести прийменники, сполучники, частки. Ці слова називають шумовими або «стоп-словами». Для досягнення кращої якості класифікації на першому етапі попередньої обробки текстів зазвичай необхідно видаляти такі слова. Другий етап попередньої обробки текстів – це приєднання кожного слова до основи, яка є однаковою для всіх його граматичних форм. Це необхідно, адже слова мають однаковий сенс і можуть бути записані в різній формі. Наприклад, одне й те ж саме слово може відмінятись по-різному, мати різні префікси та закінчення. Варто зазначити, що для вирішення вищезазначених завдань, дослідники користуються величезною кількістю інструментів і технік аналізу природної мови [5: 30].

Сучасні інформаційно-пошукові системи, які працюють з запитами користувача у вигляді ключових слів, характеризуються непоганим рівнем повноти і точності результатів пошуку. До процесу пошуку лінгвістичний процесор повинен перекодувати запит користувача з природної мови на інформаційно-пошукову мову. У зв'язку з цим вчені намагаються створити пошукові машини, які б володіли достатнім інтелектом, який дозволив би і людині, і комп'ютеру працювати на природній мові. Створений інтелект повинен мати можливості для вирішення ряду проблем, підвищення продуктивності, а також покращення якості роботи пошукової системи [6: 57]. Наразі інтелектуальні пошукові системи можуть вирішувати наступні завдання:

1. Автоматично визначати мову запиту. Ця функція дозволяє обмежити мовний сегмент Мережі, в якому буде проводитись пошук запитуваної інформації, що з урахуванням кількості інформації у віртуальному світі, позитивно позначається на процесі пошуку.
2. Виключати неінформативні слова або «стоп-слова». Це слова службових частин мови, які не мають ніякого змістового навантаження,

або ж деякі найбільш загальновживані слова. Їх видалення значно скорочує обсяг індексу і збільшує релевантність результатів пошуку. Варто зазначити, що універсального переліку неінформативних слів не існує, тому що він постійно оновлюється за рахунок додавання нових і виключення старих слів. Якщо користувач вводить запит, що складається тільки з стоп-слів, про релевантність результатів пошуку не може бути й мови. Тому при побудові запиту користувачеві потрібно по можливості виключати такі слова (Наприклад: and, or, not).

3. Проводити лінгвістичний аналіз вихідного запиту користувача, а також знайдених текстових документів. Лінгвістичний аналіз включає наступні процедури: 1) лексичний аналіз; 2) морфологічний аналіз; 3) синтаксичний аналіз; 4) семантичний аналіз [3: 25].

Як відомо, інтернет-користувачі під час пошуку потрібної їм інформації в інтернеті формулюють запити «ключовими словами» (keyword queries), і, як наслідок, обсяг оброблюваних пошукових запитів значно збільшується щороку. В результаті накопичуються великі за обсягом журнали, які містять пошукові запити користувачів (search query logs). Однак, будь-які колекції даних марні, якщо не існує методики для їх аналізу. Запити користувачів найважливіша інформація для власників інтернет-ресурсів. Висновки, отримані шляхом аналізу пошукових запитів, потенційно можуть поліпшити якість пошуку, так як вони допомагають краще зрозуміти інтереси користувачів. І з урахуванням отриманих знань пошукові машини (search engine) будуть показувати найбільш релевантні користувачеві документи [7: 3].

Таким чином, в сучасному світі, коли постійно обсяги інформації зростають, аналіз текстових даних має великий потенціал і широке застосування. Інтелектуалізація інформаційних технологій відбувається постійно, тому основні відмінності інтелектуальної пошукової системи від класичної полягають в умінні системи обробляти запити користувача, вони сформульовані в довільній формі на природній мові, а також представляють результати пошуку не у вигляді адрес сайтів, а у вигляді конкретного текстового фрагменту, що містить потрібну користувачу інформацію. Варто зазначити, що у всіх вищезгаданих процесах величезну роль грають надбання та дослідження у царині лексики та морфології.

Список використаної літератури

1. Когаловський М.Р. Перспективные технологии информационных систем / / М.Р. Когаловский. - М:Компания АйТи, 2003 – 288 с.

2. 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

3. Буніятова І. Р. Еволюція гіпотаксису в германських мовах (IV – XIII ст.) : монографія / І.Р. Буніятова. – К. : Вид. центр КНЛУ, 2003. – 327 с.
4. Тарануха В.Ю Інтелектуальна обробка текстів: навчальний посібник / В.Ю. Тарануха. – К.: Вид-во. КНУ ім.Тараса Шевченка, 2014. – 80 с.
5. Дарчук Н. П. Автоматичний синтаксичний аналіз текстів корпусу української мови / Н. П. Дарчук // Українське мовознавство. – КНУ ім. Т. Шевченка, 2013. – № 43. – С. 11–19
6. Диковицкий В.В. Обработка текстов естественного языка в моделях поисковых систем / В.В. Диковицкий, М.Г. Шишаев / / Труды Кольского научного центра РАН, 2010. – Вып. № 3. - С. 29–34.
7. Толубко В.Б. Розробка моделі автоматичного синтаксичного аналізу і синтезу тексту в системі машинного перекладу / В.Б. Толубко // Військово-спеціальні науки. – КНУ ім. Т.Шевченка, 2013. – 2(31). – С.57-60.
8. Киселёва Ю. Е. Методы группировки и структуризации поисковых запросов и их реализация: автореф. дис. на соискание уч. степ. кандидата физико-матем. наук: спец. 05.13.11 «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей» / Ю.Е.Киселёва – С-П.2011-14 с.
9. 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:
10. Fleiss J. L., Statistical methods for rates and proportions / J. L. Fleiss, B. Levin, Ch. P. Myunghee. – John Wiley & Sons, 2013. – 800 p.
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. Kovar V. Information extraction for Czech based on syntactic analysis / V. Baisa, V. Kovar // Proceedings of the 5th Language & Technology Conference. – Poznan : Funcacja Universytetu im. A. Mickiewicza, 2011. – P. 466–470.