Є. В. Муковнін
Житомирський державний університет імені І. Франка
Науковий керівник: канд. філол. наук., доцент
Гирин Олег Володимирович

ТРУДНОЩІ В РОБОТІ СИСТЕМ ОБРОБКИ ПРИРОДНОЇ МОВИ ТА ОСНОВНІ МЕТОДИ ЇХ ВИРІШЕННЯ

У статті досліджуються проблеми, які виникають під час роботи систем автоматичної обробки природної мови та методи, які застосовуються для їх подолання, проведено аналіз їх ефективності та доцільності використання на сучасному етапі

Ключові слова: обробка природної мови, токенізація, частиномовна розмітка, вирішення лексичної багатозначності, машинне навчання, парсинг.

Постановка проблеми. За два останні століття людство успішно впоралося з автоматизацією багатьох завдань використовуючи механічні та електричні прилади. У другій половині XX століття увага людини звертається і до автоматизації обробки природної мови. Тепер людині потрібна допомога не лише з механічною роботою, а й з інтелектуальними завданнями. Ми хочемо, щоб машина була здатною читати непідготований текст, перевіряти його на помилки, виконувати завдання, поставлені в тексті і навіть розуміти текст настільки добре, щоб дати відповідь ґрунтуючись на значенні тексту.

Проблему обробки природної мови не можна назвати простою. Труднощі виникають через низку об’єктивних причин, як от існування сотень природних мов, у кожній з яких діють свої синтаксичні правила. У межах однієї мови існують слова, які можуть мати різний зміст залежно від контексту вживання. Навіть на рівні окремих символів зустрічаються певні труднощі.

Аналіз останніх досліджень і публікацій. Дослідженню обробки природної мови присвячено багато досліджень в зарубіжній науці, зокрема дослідження таких науковців, як D. Jurafsky [7], J. H. Martin [7], R. M. Reese [10], Y. Goldberg [6]. В українській комп’ютерній лінгвістиці питанням обробки природної мови та автоматичного синтаксичного аналізу присвячені дослідження В. Ю. Таранухи [4] та Н. П. Дарчук [3].

Мета нашого дослідження є аналіз проблем в роботі систем обробки природної мови, які не є вирішеними на сучасному етапі розвитку галузі. На основі цього аналізу пропонуються методи, які покликані подолати дані проблеми.

Виклад основного матеріалу. У процесі обробки природної мови завжди слід враховувати **кодування**, яке використовується в конкретному документі. Текст може зберігатися в різних кодуваннях: ASCII, UTF-8, UTF-16 або Latin-1. Особливі види обробки можуть знадобитися для знаків пунктуації та для чисел. Іноді доводиться окремо обробляти використання знаків, які відображають емоції (комбінації символів або спеціальні символи), гіперпосилань, розділових

знаків що повторюються (... або ---), розширень файлів та імен користувачів, що містять крапки [10].

Під розподілом тексту на фрагменти або елементи зазвичай мається на увазі представлення тексту у вигляді послідовності слів. У цьому випадку слова позначаються терміном ”лексичний елемент”, ”лексема”, або просто ”токен” (token), а процес поділу тексту – “**токенізація**” (tokenization). Цей процес не викликає особливих труднощів у мовах, що використовують пробільні символи для розділення слів, але в мовах, подібних до китайської, це зробити набагато важче, оскільки ієрогліфи можуть позначати як склади, так і цілі слова. Та і в англійській мові з процесом токенізації можуть виникнути певні труднощі, адже існує велика кількість альтернативних варіантів, коли одне слово може писатися разом, окремо або через дефіс [10].

Слова об’єднуються в словосполучення і речення. **Визначення меж речень** теж може бути пов’язано з певними труднощами, хоча на перший погляд здається, що достатньо лише знайти крапку, що позначає кінець речення. Але крапки можуть зустрічатися і всередині речень, наприклад після скорочених слів [10].

При граматичному розборі все ще виникають серйозні проблеми з точністю. По-перше, багато що тут залежить від якості **морфологічної (частиномовної) розмітки** (part-of-speech tagging), яка повинна бути дуже високою (97-98%), проте в довгих реченнях дуже часто можна зустріти неправильно розпізнану певну частину мови, що призводить до подальших помилок розбору. По-друге, сучасний **автоматичний синтаксичний розбір** дає точність приблизно 90-93% а це, в свою чергу, означає, що в довгому реченні практично завжди будуть помилки розбору. Наприклад, при точності розбору 90%, ймовірність розбору речення довжиною 10 слів без жодної помилки складе всього 35%. Сучасний стан досліджень дає надію на покращення якості розбору, проте часто правильний синтаксичний розбір включає також розуміння семантики речення, але, наприклад, в англійській мові це нерідко викликає труднощі. Так, у реченні “*He saw a man with a hammer*” може бути два різних варіанти синтаксичного розбору залежно від того, чи вважаємо ми, що людину побачили за допомогою молотка або побачили людину з молотком. Звичайно, якщо потрібно отримати максимально точний синтаксичний розбір, то має сенс залишати кілька найбільш ймовірних варіантів, а потім визначати правильний за сукупністю різних факторів, в тому числі семантичних [2].

Іноді доводиться визначати зв’язки між словами. Наприклад, **встановлення кореферентності** (coreference resolution) визначає зв’язки між конкретними словами які позначають один і той же об’єкт, тобто мають один і той же референт в одному або в декількох реченнях. Наприклад в реченнях *”The city is large but beautiful. It fills the entire valley.”* слово “*it*” кореферує, тобто є референційно тотожним слову “*city*”. Явища кореферентності обумовлені фундаментальними закономірностями організації тексту. Оскільки текст має лінійну будову, а ситуація, яку він описує, як правило, нелінійна, у тексті майже неминуче повинні міститися повторні згадки елементів ситуації, що описується. При кожній новій згадці того ж об’єкта проводиться нова номінація

цього об'єкта, яка базується на тому, що вже було сказано про цей об'єкт, і на тих знаннях, які в тексті не вербалізовані (екстралінгвальні знання мовця про контекст предметної області). Хоча проблема кореферентності в лінгвістиці досить докладно досліджена, втілення цих теоретичних знань на практиці на сьогоднішній день є досить складним [1: 41].

Якщо слово може мати кілька смислових значень, для визначення його сенсу в даному конкретному випадку може знадобитися виконання операції **вирішення лексичної багатозначності** (word sense disambiguation, WSD). Іноді це пов'язано з певними труднощами. Наприклад в реченні "*John went back home*." слово "*home*" може означати "*housing that someone is living in*" або "t*he country or state or city where someone lives*" [10].

Однією з найбільших відкритих проблем при обробці природномовних текстів є **неоднозначність (багатозначність)** її одиниць, що виявляється на всіх її рівнях та виражається в явищах полісемії, омонімії та синонімії. Говорячи про неоднозначність, можна зазначити лексичну (існування більш ніж одного значення слова, наприклад, "*bank*"); синтаксичну, або структурну (коли одне речення має декілька імовірних варіантів граматичної структури і, відповідно має інше значення, наприклад проблема приєднання (attachment ambiguity), коли PP може приєднуватися як до VP так і до NP в межах одного речення із зміною значення: "*The police shot the rioters with guns*"); семантична неоднозначність (коли одне і те ж речення можна по-різному розуміти в різних контекстах, хоча лексична чи структурна багатозначність відсутня: "*all linguists prefer a theory*"); прагматична неоднозначність (коли одне речення можна по-різному розуміти в контексті, в якому воно існує "*every student thinks he is a genius*") [9].

Сучасні системи вирішення лексичної багатозначності мають точність в діапазоні 60-70% і частіше за все, представлені як самостійні методи. Вирішення проблеми зняття неоднозначності вимагатиме інтеграції декількох джерел інформації та методів [9].

Незважаючи на всі перераховані труднощі, технологія обробки природної мови в більшості випадків здатна досить успішно впоратися зі своїми завданнями, тому дуже корисна в багатьох галузях.

Приблизно в половині випадків має місце будь-яка форма омонімії, і набір морфологічних ознак виявляється недостатнім для її вирішення. Зменшити неоднозначність можна за допомогою **синтаксичного і семантичного аналізу із використанням статистичних методів**, які дають змогу відкинути вкрай малоймовірні варіанти. Природна мова хоч і є за своєю природою символічною, обробити її за допомогою символічних, базованих на логіці, правилах та об'єктивних моделях є досить складним процесом. Природна мова є вкрай неоднозначною та мінливою, тому для її обробки необхідно застосовувати і статистичні алгоритми, тому домінантними підходами до сучасної ОПМ є підходи, базовані на **статистичному машинному навчанні** (statistical machine learning) [6].

На початку 90-х років стали розвиватися методи машинного навчання і одночасно було проведено ряд досліджень з статистичної лінгвістики. В

машинному навчанні чудово показали алгоритми класифікації для різних задач, пов'язаних з обробкою текстів: визначення спаму, сортування документів за тематикою, виділення іменованих сутностей. Використання в комп'ютерній лінгвістиці статистичних методів, зокрема прихованих ланцюгів Маркова та моделі максимальної ентропії дало змогу зробити визначення частин мови високоточним. З'явилися парсери на базі стохастичних контекстно-вільних граматик, створюються проекти з статистичного машинного перекладу. Також були закладені основи глибокого навчання, яке лише недавно дало перші результати, зумовлені прогресом в галузі високопродуктивних систем і появою великих обсягів даних, що використовуються для навчання. [2].

У 2010 році була запропонована **модель лексикалізованої ймовірнісної (стохастичної) граматики**, яка дозволила підвищити точність граматичного розбору до 93%, що, звичайно, далеко від ідеалу. Точність розбору – це відсоток правильно визначених граматичних зв'язків, а також вірогідність того, що довге речення буде розібрано правильно, яка зазвичай дуже низька. Одночасно, завдяки новим алгоритмам і підходам, включаючи глибоке навчання, збільшилася швидкість граматичного розбору. Крім того, практично всі провідні алгоритми і моделі стали доступні широким масам дослідників, і, напевно, найвідомішою роботою в області глибокого навчання для ОПМ став алгоритм Томаса Міколова [8]

Завдяки новим **методам глибокого навчання**, сьогодні можна отримати якісні семантичні уявлення для слів, фраз і речень, причому навіть без навчальної вибірки. Все менше зусиль зараз потрібно для створення власних семантичних словників і баз знань, тому розробляти системи автоматичної обробки текстів стало простіше. Однак ми все ще дуже далекі від адекватного вирішення завдання розуміння взаємопов'язаних подій, представлених у вигляді послідовності речень або образів, а також діалогів. Всі відомі сьогодні методи успішно працюють або при вирішенні завдань "поверхневого" розуміння мови, або при істотному обмеженні предметної області [2].

Методи глибокого навчання є більш точними ніж поверхневі методи, які не намагаються осмислити текст, однак на практиці бази знань, необхідні для їх роботи, існують для вкрай обмежених предметних сфер і тому на сучасному етапі досить часто використовуються і поверхневі методи. Такі методи беруть до уваги найближчі слова, використовуючи подібну інформацію, досліджуючи сполучуваність слів. Правила можуть бути автоматично отримані за допомогою комп'ютера, при використанні навчальної текстової бази даних слів, доданих з їх смисловими значеннями. Цей метод, теоретично, не такий дієвий, як глибокі методи, хоча на практиці він дає кращі результати, внаслідок обмеженості знань комп'ютера про світ [5].

Процес розуміння та генерації природної мови з використанням комп'ютерних технологій є надзвичайно складним. Найбільш відомими методами роботи із даними мови є методи з використанням алгоритмів **машинного навчання з учителем**, коли система намагається виділити мовні структури та правила із анотованих даних корпусу. Наприклад, завдання класифікації документів за категоріями: спорт, політика, економіка, розваги, є

досить простим, адже слова, які використовуються в документах таких тематик є підказкою. Читач може з легкістю віднести текст до однієї із тем, базуючись на власному досвіді, але навряд чи зможе назвати конкретні правила, якими він послуговувався. Написання групи правил для автоматичного категоризування тексту є складним та трудоємким. Використовуючи алгоритми машинного навчання з учителем можна дозволити машині визначити мовні структури, які дадуть змогу категоризувати документи. Методи машинного навчання є відмінним рішенням для галузей, де важко визначити конкретні правила, а створити анотований корпус досить просто [2].

Висновки. Обробка природної мови активно досліджується в наш час; існує багато методів, за допомогою яких теоретично можна подолати труднощі в роботі систем обробки природної мови, але на практиці застосування таких методів є ще досить далеким від ідеалу та потребує подальшого покращення та вдосконалення.

Список використаних джерел та літератури

1. Боярский К. К. Введение в компьютерную лингвистику. Учебное пособие / К. К. Боярский. – СПб: НИУ ИТМО, 2013. – 72 с.
2. Велихов П. Машинное обучение для понимания естественного языка / П. Велихов [Електронний ресурс] – Режим доступу: https://www.osp.ru/os/2016/01/13048649/
3. Дарчук Н. П. Комп'ютерна лінгвістика (автоматичне опрацювання тексту): підручник / Н. П. Дарчук. – К.: Видавничо-поліграфічний центр "Київський університет". – 2008. – 351 с.
4. Тарануха В. Ю. Інтелектуальна обробка текстів: [навчальний посібник] / В. Ю. Тарануха. – К.: електронна публікація на сайті факультету, 2014. – 80 с.
5. Chen P. 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. – 2009. – P. 28 – 36.
6. Goldberg Y. Neural Network Methods for Natural Language Processing / Y. Goldberg. – Morgan & Claypool Publishers. – 2017. – 309p.
7. Jurafsky D. Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics. 2nd edition, / Jurafsky D., Martin J. – 2009 [Електронний ресурс] – Режим доступу: http://www.cse.iitk.ac.in/users/mohit/Speech-and-Language-Processing.pdf
8. Mikolov T. Efficient Estimation of Word Representations in Vector Space / T. Mikolov, K. Chen, G. Corrado, J. Dean. – 2013. – 12p.
9. Mohd S. H. Word Sense Ambiguity: A Survey / S. H. Mohd, R. B. Mohd // International Journal of Computer and Information Technology Vol. 02– Issue 06. – 2013. – P. 1161-1168

10. Reese R. M. Natural Language Processing with Java / R. M. Reese. – Packt Publishing, 2015. – 262 p.