DOI https://doi.org/10.30525/978-9934-26-261-6-62

NATURAL LANGUAGE PROCESSING AS AN ASPECT OF MODERN TECHNOLOGIES DEVELOPMENT

ОБРОБКА ПРИРОДНОЇ МОВИ В АСПЕКТІ РОЗВИТКУ СУЧАСНИХ ТЕХНОЛОГІЙ

Yemelianova O. V.
Candidate of Philological Sciences, Associate Professor, Associate Professor at the Department of Germanic Philology, Sumy State University

Ємельянова О. В.
кандидат філологічних наук,доцент, доцент кафедри германської філології, Сумський державний університет

Kuksenko O. O.
Master's student at the Department of Germanic Philology, Sumy State University Sumy, Ukraine

Куксенко О. О.
магістр кафедри германської філології, Сумський державний університет м. Суми, Україна

Комп'ютер було створено як апарат обчислювальної техніки, що допомагає з розрахунком. Пізніше функціональні можливості комп'ютера вийшли за межі вирішення математичних рівнянь, і сьогодні він є майже у кожному домі та великих, чи малих компаніях, виконуючи роль персонального помічника, але, як відомо, для кращої співпраці потрібна стабільна та якісна комунікація. З метою покращення зв'язку науковці зосередилися на складних питаннях обробкою природної мови.

Обробка природної мови (Natural Language Processing – NLP) – це міждисциплінарна галузь, яка стоїть на перетині комп'ютерних наук, штучного інтелекту та обчислювальної лінгвістики, основним проблемним полем якої є забезпечення прямої взаємодії між комп'ютером та людиною за допомогою природної мови. NLP та машинне навчання взаємопов'язані та є невід'ємною частиною розвитку штучного інтелекту. Обидві сфери поділяють методи, алгоритми, теорію та терміни [1]. Простіше кажучи NLP досліджує тексти, створені людиною, і перетворює їх у дані зрозумілі для комп'ютера, або навпаки.

Завдяки розвитку NLP сьогодні наші ноутбуки та телефони можуть розпізнавати голосові команди та рукописний текст, а це − всі веб-пошукові системи, що надають доступ до інформації світу, та машинний переклад, що дозволяє нам читати тексти, написані іноземною мовою.

В руслі надання більш природних людино-машинних інтерфейсів та більш складного доступу до зберігання інформації, мовна обробка стала

відігравати центральну роль у багатомовному інформаційному суспільстві [2 с. 9]. Задачі NLP – написання, переклад, покращення тексту, визначення теми тексту, дійових осіб тощо. В. Дьомкін вважає, що будь-яка інтелектуальна обробка тексту в тій чи іншій формі належить до сфери NLP [4].

Перші дослідження у царині машинного перекладу датуються 1954, коли у Сполучених Штатах Америки була проведена перша публічна демонстрація системи машинного перекладу, виконаного на машині IBM-701, що мала на меті вирішення практичних завдань [5].

Приблизно в той же час виокремилися дві ключові групи вивчення NLP: знакова, або заснована на правилах, та стохастична. Перша група базується на формальних мовах Чомського та синтаксисі; ця група складалася з багатьох лінгвістів і комп'ютерних інженерів, які вважали, що ця галузь покладе початок розвитку штучного інтелекту. Стохастичні дослідники більше цікавилися статистичними та імовірнісними методами NLP, працюючи над проблемами оптичного розпізнавання символів та розпізнавання образів між текстами.

У 70-х роках дослідники почали розширювати сфери NLP відповідно до появи нових технологій та знань. Однією з нових галузей були логічні парадигми, мови, які зосереджені на правилах кодування, та мова в математичній логіці. Згодом це сприяло розробці однієї з мов програмування – Prolog. Іншою сферою розвитку NLP дослідження стали самі природні мови через комп'ютерні програми. Професор комп'ютерних наук Террі Виноград написав програму SHRDLU, як наукову дисертацію. Ця програма помістила комп'ютер у світ блоків. Подібно до дитячої гри у кубики, програма дозволяла комп'ютеру маніпулювати блоками та відповідати на запитання відповідно до інструкцій природною мовою користувача. Але головним досягненням системи була здатність вивчати та розуміти людські мови. Це довело, що комп'ютер здатний з найвищою точністю будувати зв'язки між об'єктами та розуміти певні неоднозначності у мові [6].

До числа найбільш відомих прикладних задач машинного аналізу текстів природною мовою відносяться:

- машинний переклад (machine translation);
- інформаційний пошук (information retrieval);
- автоматична класифікація та кластеризація текстів (automatic text classification and clustering);
- автоматичне реферування та анотування текстів (automatic text summarization and annotation);
- автоматичне вилучення фактів інформації з текстів (information extraction, knowledge discovery);
- розробка автоматичних питально-відповідних систем (question-answering systems development).

Головними методами NLP є:

• Токенізація, аналіз тлумачення головної думки речення, яку можна розбити на менші компоненти: «блоки», слова, числа чи знаки.

• Лематизація. Перетворення частин мови у їх словникову форму для кращого і швидшого зрозуміння.

• Тегування частини мови. Допомагає визначати у реченні частини мови.

• Автоматизований емоційний аналіз тексту. Використовується для ретельного вивчення тексту, щоб визначити в текстах емоційно забарвлену лексику та емоційну оцінку авторів (думок) по відношенню до об'єктів, мова про які ведеться в тексті.

• Розпізнавання імен та власних назв. Визначення та класифікація власних імен, назв організацій, географічних назв, подій та дат.

• Виведення головних ідей тексту. Ця техніка NLP може коротко резюмувати текст [6].

В руслі традиційного підходу, текст в NLP розглядається як набір дискретних символів, подібно до формальних мов, чи мов програмування, які в подальшому представлені у вигляді одноразових векторів. Векторне представлення – це метод представлення строк у вигляді векторів зі значеннями. Створюється плоский (щільний) вектор для кожного слова так, щоб слова у схожих контекстах мали подібні вектори. Простіше кажучи, використання векторів, це застосування математики до слів. Застосовуючи до слів базову арифметику, можна обчислити вкладення слова *Королева*: «*Король − Чоловік + Жінка = Королева*» Це пояснюється тим, що вектори представляють саме значення слова, а не лише саме слово [7]. Вперше використання векторів було запропоновано у 1960-1970-х рр. Дослідницька група під керівництвом Дж. Салтон, розробила основні принципи інформаційного пошуку, векторну модель пошуку (vector space model) [3, с. 618].

Недолік такого способу представлення слів полягає у відсутності визначення схожості для одноразових векторів, оскільки більшість слів можуть мати семантично подібне значення. Вирішити дану проблему можна за допомогою кодування схожості у самих векторах [1, с. 187]. Такий спосіб представлення враховує стартову точку для більшості завдань NLP та робить глибинне навчання ефективним.

Можливість комп'ютера спілкуватися за допомогою програм мовою людини сприймається нами у 21 столітті як звична річ, але й досі залишається однією із найскладніших для втілення завдань. Завдяки синергетичному поєднанню, здавалося б двох різних наук − інформатики та лінгвістики − для нас відкриваються нові можливості для досліджень та наукових відкриттів.

Література*:*

1. Онищенко К., Данієль Я., Каменєв Р. Аналіз Методів Обробки Природної Мови : наукова стаття. Харків : Харківський національний університет радіоелектроніки, 2020. С 186−190.
2. Bird S., Klein E., Loper E., Natural Language Processing with Python First edition: Printed in the United States of America, 2009. 504p.
3. Salton G., Wong A., Yang C. S., A vector space model for automatic: Indexing Communications of the ACM. № 11. 1975. P. 613−620.
4. Дьомкін В. Громадське радіо. Незабаром на нас чекає digital humanity. URL: https://hromadske.radio/podcasts/einstein/nezabarom-na-nas-chekaye-digital-humanity-vsevolod-domkin (дата звернення 15.09.2016)
5. Предмет, мета і завдання обробки природної мови. URL: https://stud.com.ua/140008/informatika/predmet_meta_zavdannya_obrobki_prirodnoyi_movi (дата звернення 15.09.2016)
6. Madill W. Exploring Natural Language Processing (NLP) in Translation. URL: https://localizejs.com/articles/natural-language-processing-nlp/ (дата звернення 15.09.2016)
7. Natural Language Processing. URL: https://cs.stanford.edu/people/eroberts/courses/soco/projects/2004-05/nlp/overview_history.html (дата звернення 15.09.2016)