Король В. Я., студент, гр. ПЛ-5
Окунькова О. О., ст. викладач кафедри комп'ютерної інженерії та кібербезпеки
Державний університет «Житомирська політехніка»

СПРОЩЕНИЙ ТА РОЗШИРЕНИЙ ПІДХОДИ ОБРОБКИ ПРИРОДНОЇ МОВИ

Обробка природної мови (ОПМ, англ. *Natural Language Processing, NLP*) дає змогу машинним технологіям сприймати та генерувати мови, якими ми, люди, спілкуємося: українська, англійська, польська, тощо. В своїй основі ОПМ, як і будь-яке програмування, базується на чітких алгоритмах, визначених структурах даних і твердій об'єктивності, яка дає змогу ефективно обробляти будь-які дані, які керуються чіткими законами та взаємовідносинами. Обробка таких даних базується на законах та методах математики, статистики, логіки з якими комп'ютерні технології справляються значно краще людини.

З іншого боку, процес сприйняття природної мови людською свідомістю не підлягає чіткій алгоритмізації. Наше мислення оперує як аналітичними здібностями, так і абстрактними зв'язками. Прикладом першого може бути формування речення з чітким порядком або відмінювання слів з допомогою підстановки необхідних суфіксів та закінчень до кореня.

Саме тому, в рамках цього дослідження виділено 2 основні парадигми ОПМ: *спрощену* (яка спрощує мову в обмежену модель) та *розширену* (яка працює з мовою як суцільним об'єктом).

Лінгвістична екосистема мови Python по-справжньому блискуче справляється з будь-якими завданнями, що можна описати в жорстких правилах і рамках. Аналіз різноманітних бібліотек обробки природної мови (*Pandas, Spacy, NLTK, Scikit-learn*) дає змогу прийти до висновку, що вони можуть виконувати різноманітні завдання, які дозволяють спростити мову до обмеженої системи, а саме: поділ на токени (*tokenisation*), класифікація частин мови (*part of speech tagging*), лематизація (*lemmisation*) побудова синтаксичного дерева, розпізнавання іменованих сутностей (*named entity recognition*), екстракція тексту за патернами (зокрема Spacy Matcher і регулярні вирази), порівняння схожості між словами (*similarity comparison*).

Однак, мова перед усім - це інструмент спілкування між людьми. Лінгвісти виділяють 7 видів значень, основними з них є: пряме значення (концептуальне), припущенне (те, що не говориться, але мається на увазі), конотативне (те, що мовець асоціює зі словом), колокативне (слово, яке часто вживається з іншим словом, утворюючи сталий вираз), соціальне (вибір слів від неформальних до формальних).

Метод машинного навчання було розроблено, аби технології мали змогу обробляти мову в її більш широкому аспекті та справлятися зі завданнями, які потребують аналізу та синтезу самих значень слова: моделювання теми (*topic modelling*), узагальнення тексту (*text summirisation*), машинний переклад (*machine translation*), фільтрація тексту (*text filtration*) та класифікація настрою (*sentimental analysis*). До певної міри, всі вони справляються з поставленими задачами з достатньо високою точністю, проте машинне навчання, попри свої переваги, має спиратися на величезну кількість відібраних даних, аби лише тренувати модель, і вимагає достатньо великих ресурсів, аби продукувати свої відповіді під час виконання.

В наслідок цього машинне навчання не може обробляти природну мову спонтанно, на прикладі малих даних, відслідковувати причинно-наслідкові та логічні зв'язки. Ці обмеження позбавляють машину можливості виконувати цілий ряд завдань, пов'язаних з інформативним аналізом і обробкою чистої інформації.

Обробка природної мови, - це область обчислень, мета якої - допомогти комп'ютерам зрозуміти людську або «природну» мову. На даний момент хвиля інтересу до NLP тільки зростає, і у найближчому майбутньому отримає більшого поширення майже у всіх сферах діяльності.

Хоча природні мови не здаються чимось дивовижним, комп'ютерам дуже важко правильно їх інтерпретувати і використовувати. Жорсткий, обмежений правилами формат електронних таблиць і баз даних ідеальний для машин, а випадкова природа людських мов, що залежить від контексту і не завжди пов'язана певними правилами призводить штучний інтелект в ступор.

Звісно, NLP ще далекий від того, щоб змінити світ професій прямо зараз, але цей напрямок існує вже близько 30 років і роботи по його вдосконаленню ведуться постійно. Експерти вважають, що наступний прорив у розвитку NLP буде колосальний, зумовить перехід від структурованих (бази даних) до неструктурованих (текст) даних і значно поліпшить здатність машин «розуміти» людей в звичайній розмові. На сьогодні програми, що здатні розуміти людську мову швидко розвиваються і вдосконалюються.