DOI: https://doi.org/10.31392/UDU-MKfpgoe-2024.40
УДК 81'25:811.1/.8

Франчук Н. П.,
кандидат педагогічних наук, доцент,
доцент кафедри інформаційних технологій і програмування,
Український державний університет
імені Михайла Драгоманова,
м. Київ, Україна;
старший науковий співробітник відділу
відкритих освітньо-наукових інформаційних систем,
Інститут цифровізації освіти
Національної академії педагогічних наук України,
м. Київ, Україна

АВТОМАТИЗОВАНІ ТЕХНОЛОГІЇ ОПРАЦЮВАННЯ ПРИРОДНОЇ МОВИ: ВІД ТРАНСКРИПЦІЇ ДО МАШИННОГО ПЕРЕКЛАДУ

Вступ. Автоматизовані технології опрацювання природної мови (Natural Language Processing, NLP) є однією з ключових сфер штучного інтелекту, яка стрімко розвивається. Застосування цих технологій спрямовані на автоматизацію процесів аналізу, розуміння, генерації та перекладу текстових й мовленнєвих даних. Усе це робить NLP незамінним інструментом для розв'язання сучасних проблем у різних сферах: від освіти до комерційних застосувань [1, 2].

Сучасні технології опрацювання природної мови базуються на досягненнях у галузі машинного навчання, зокрема нейронних мереж, які дозволяють опрацьовувати великі обсяги текстових та аудіоданих з високою точністю. Розглянемо основні методи, алгоритми та моделі, що є актуальними в цій галузі.

До появи нейронних мереж опрацювання тексту значною мірою базувалося на статистичних моделях та методах на основі правил. Для прогнозування наступного слова в тексті на основі попередніх використовувалися статистичні N-грамні моделі. Хоча цей підхід є простим, він має обмеження у врахуванні довгострокових залежностей. Також використовувався алгоритм тематичного моделювання для класифікації текстів за темами.

Розглянемо автоматизовані технології опрацювання природної мови, що охоплюють повний спектр завдань (див. табл. 1).

Таблиця 1

Основні аспекти NLP

Аспект	*Характеристики*
Автоматична транскрипція мовлення	Автоматичне розпізнавання мовлення (Automatic Speech Recognition, ASR) дозволяє перетворювати аудіофайли в текст. Сучасні системи ASR, наприклад Whisper від OpenAI, використовують нейронні мережі для розпізнавання мови навіть у складних акустичних умовах. Їхнє використання сприяє створенню субтитрів, голосових помічників та аналізу дзвінків.
Морфологічний та синтаксичний аналіз	На цьому етапі текст розбивається на окремі слова або морфеми для аналізу їх граматичних властивостей. Синтаксичний аналіз передбачає побудову дерев залежностей, що дозволяє зрозуміти структуру речень. Ці завдання розв'язуються за допомогою інструментів, таких як SpaCy (бібліотека програмного забезпечення з відкритим вихідним кодом для опрацювання природної мови, написана на мовах програмування Python і Cython).
Семантичний аналіз і моделі трансформерів	Семантичний аналіз дозволяє машинам "розуміти" значення тексту. Тут використовуються трансформери, наприклад BERT або Generative Pre-trained Transformer (GPT), які вивчають контекст кожного слова в реченні. Завдяки цьому можна виконувати завдання контекстного пошуку, емоційного аналізу та автоматичного узагальнення тексту.
Машинний переклад	Системи машинного перекладу, такі як Google Translate чи DeepL, застосовують нейронні моделі для створення високоякісних перекладів. Технологія трансформерів значно підвищила якість перекладу, враховуючи як локальний, так і глобальний контекст речення.

Методи на основі правил залучають граматичні й семантичні конструкції, які задаються вручну. Прикладом є аналіз синтаксичних дерев та частиномовна розмітка. Хоча цей підхід забезпечує контрольованість, він погано масштабується для складних мов.

Тепер все частіше використовуються рекурентні нейронні мережі (RNN) та довгу короткочасну пам'ять Long Short-Term Memory (LSTM) для опрацювання послідовностей, таких як речення, але є обмеження у врахуванні довгих контекстів. Також використовується архітектура трансформерів (Transformers), що частково замінює RNN завдяки здатності паралельно опрацьовувати дані.

Двонаправлений шифратор репрезентацій з трансформерів – Bidirectional Encoder Representations from Transformers (BERT) використовується для завдань класифікації тексту, пошуку даних та семантичного аналізу.

Генеративний попередньо навчений трансформер – Generative Pre-trained Transformer (GPT) генерує текст із врахуванням контексту, та є основою чат-ботів та "творчих застосунків".

Трансформ передавання тексту в текст – Text-to-Text Transfer Transformer (T5) – це універсальна модель для багатьох задач NLP, включаючи переклад, класифікацію та генерацію тексту.

Сьогодні автоматизовані технології опрацювання природної мови застосовуються у всіх сферах, зокрема для: автоматичного оцінювання текстів, використання платформ для інтерактивного навчання, створення та використання чат-ботів, аналізу електронних записів, розпізнавання симптомів за описами, автоматизації опрацювання запитів, опрацювання великих обсягів текстових даних, автоматичного аналізу дослідницьких публікацій та ін. У попередніх публікаціях було досліджені деякі з цих питань [3, 4].

Проте є певні чинники, на які доречно звернути увагу та в подальшому розкрити їх, а саме:

– *малоресурсні мови*: створення моделей для мов із обмеженою кількістю текстових даних залишається складним завданням.

– *етичні аспекти*: необхідність запобігання дискримінаційним або упередженим результатам, які можуть виникати через обмеження навчальних даних.

– *продуктивність і швидкість*: оптимізація моделей для роботи на пристроях із обмеженими ресурсами.

Висновки. Автоматизовані технології опрацювання природної мови стали основою сучасних цифрових інновацій. Їх подальший розвиток дозволить значно підвищити ефективність роботи з даними, покращити міжмовну комунікацію та забезпечити доступ до знань для ширшої аудиторії. Інтеграція NLP у повсякденні процеси відкриває нові можливості для автоматизації, підвищення якості обслуговування та розв’язування соціальних викликів.

Використана література:

1. Kuzminov I. F., Bakhtin P. D., Timofeev A. A. *et al.* Modern Natural Language Processing Technologies for Strategic Analytics. *Sci. Tech. Inf. Proc.* 48, 467–475 (2021). DOI: https://doi.org/10.3103/S0147688221060071
2. The NLP Playbook: From Basics to Advanced Techniques and Algorithms – Let’s Data Science. URL : https://letsdatascience.com/natural-language-processing/
3. Франчук Н. П. Використання програм для автоматизованого синтаксичного аналізу тексту. *Теорія і практика використання інформаційних технологій в умовах цифрової трансформації освіти.* 29 червня 2023. м. Київ. С. 69-71. URL : https://lib.iitta.gov.ua/736682
4. Франчук Н. П. Використання штучного інтелекту в прикладній лінгвістиці. *Цифрова трансформація освіти: теоретико-методичні засади* : матеріали Міжнародної науково-практичної конференції з нагоди 70-річчя від дня народження доктора педагогічних наук, професора, заслуженого працівника освіти України, директора Навчально-наукового інституту перепідготовки та підвищення кваліфікації Сергієнка Володимира Петровича. м. Київ. 28 жовтня 2024 р. С. 408-409. DOI: https://doi.org/10.31392/UDU-MKNino-Sergiyenko70-2024