УДК 004.932

Б.С. Білецький

АДАПТАЦІЯ МЕТОДУ ДИСТИЛЯЦІЇ ЗНАНЬ ПРИРОДНОЮ МОВОЮ ДЛЯ КЛАСИФІКАЦІЇ ТЕМПОРАЛЬНИХ ФРАЗ

Вінницький національний технічний університет, Україна

Анотація

Дослідження представляє вдосконалення процесів автоматичного розпізнавання та класифікації темпоральних фраз у природномовних текстах за допомогою методу дистиляції знань. Підкреслюючи зростаючу потребу в автоматичному розумінні темпоральної інформації, дослідження зосереджено на створенні масштабного датасету з 1 078 862 записами та навчанні рекурентної нейромережі Bi-LSTM. Результати показали високу точність моделі, зокрема у розрізненні темпоральних фраз, відкриваючи нові перспективи для аналізу текстової інформації в різноманітних областях.

Ключові слова: інтелектуальна технологія, темпоральні фрази, дистиляція знань, Bi- LSTM, класифікація тексту, природномовні тексти, автоматичне розпізнавання, ChatGPT, навчання нейромережі, аналіз даних, обробка природної мови, штучний інтелект, машинне навчання.

Abstract

The research presents an improvement of automatic recognition and classification of temporal phrases in natural language texts using a knowledge distillation method. Emphasizing the growing need for automatic understanding of temporal information, the study focuses on creating a large-scale dataset with 1,078,862 records and training a recurrent neural network Bi-LSTM. The results showed high accuracy of the model, in particular in distinguishing temporal phrases, opening up new perspectives for analyzing textual information in various fields.

Keywords: intelligent technology, temporal phrases, knowledge distillation, Bi-LSTM, text classification, natural language texts, automatic recognition, ChatGPT, neural network training, data analysis, natural language processing, artificial intelligence, machine learning.

Вступ

Класифікація темпоральних фраз у природномовних текстах є ключовим компонентом для розуміння та аналізу даних, що забезпечує точне відтворення хронології подій. Це дослідження розкриває потенціал використання сучасних методів дистиляції знань для вдосконалення процесів автоматичного розпізнавання та класифікації фраз, що вказують на час.

Застосування цих технологій може суттєво покращити аналіз текстової інформації в різноманітних областях, включаючи історичні дослідження, журналістику, моніторинг соціальних медіа та багато інших, де ключовим елементом є розуміння та відтворення часових параметрів подій.

Метод дистиляції знань відіграє ключову роль у покращенні ефективності та швидкості обробки даних у дослідженнях обробки природної мови, спрямований на зменшення вуглецевого сліду [1]. Систематичне дослідження методів дистиляції показує, як різні компоненти впливають на результати, виявляючи ключові фактори для оптимізації [2]. Досвід створення DistilBERT ілюструє можливість зменшення розміру моделі BERT без значної втрати її можливостей, що відкриває шлях для ширшого застосування моделей у обмежених обчислювальних умовах [3].

Результати дослідження

Метод дистиляції знань. Цей метод дистиляції знань дозволяє компактній нейромережі ефективно навчатися, використовуючи знання, отримані від великих мовних моделей, забезпечуючи високу точність при менших вимогах до обчислювальних ресурсів. В цьому дослідженні було використано дві генеративні моделі для створення прикладів темпоральних фраз для датасету. Це ChatGPT 4.0 від компанії OpenAI та Gemini від Google.

Створення датасету. Було створено датасет з 1 078 862 записами, де використання ChatGPT сприяло ефективній анотації темпоральних фраз. Це забезпечило масштабність даних, необхідну для точного навчання нейромережі.

Рисунок 1 — Схема поповнення датасету прикладами темпоральних фраз рокам

Навчання нейронної мережі. Обрана архітектура Bi-LSTM, з огляду на її ефективність у роботі з послідовними даними, є ідеальною для нашої задачі. Ця модель забезпечує глибше розуміння контексту темпоральних фраз у тексті.

Спроба	Датасет, к-сть записів	Вага нейромережевої моделі	Точність моделі
1	1 500	08.2 МБ	0.99
2	500 000	28 МБ	0.95
3	1 078 862	108 МБ	0.98

Рисунок 2 — Порівняння моделей

Демонстрація результатів класифікації. Модель тестувалася на датасеті WISEST-SBB, що містить екологічні звіти про стан води в річці Південний Буг. На прикладах з датасету мережа показала точність 0.988 у розрізненні темпоральних фраз у природномовному тексті. Ці результати підкреслюють потенціал моделі для застосування в реальних умовах.

Комунальники курортного Хмільника Вінницької області другий місяць б'ють на сполох, адже місцева лабораторія фіксує понаднормове забруднення Південного Бугу у місці водозабору. Вперше у КП «Хмільникводоканал» звернулися до екологів наприкінці грудня. Досліди фахівців інспекцій Вінницької та Хмельницької областей у січні не виявили грубих порушень. Як зазначають у ДЕІ, по факту забруднення річки Південний Буг, встановленого 28 січня 2021 року на межі Вінницької та Хмельницької областей, Державною екологічною інспекцією у Вінницькій області 24 лютого 2021 року повторно проведено спостереження за якістю поверхневої води в річці від Хмельницької області до м. Вінниці. Тобто лютневі досліди проводили виключно в межах одного регіону. Як вказують у вінницькій інспекції, «спеціалістами відділу інструментально-лабораторного контролю був здійснений виїзд пересувною лабораторією в район «Кармелюкового крісла», яке знаходиться на межі Вінницької та Хмельницької областей. На місці були відібрані проби з річки Південний Буг та проведений експрес-аналіз на вміст азоту амонійого та фосфат-іону. За результатами експрес аналізу концентрація азоту амонійного становить 3,6 мг/дм3 при нормативі 2,0 мг/дм3, фосфати 1,85 мг/дм3 при нормі 3,5 мг/дм3».

Рисунок 3 — Приклад розмітки тексту. Червоний колір кодує ймовірність того що слово є темпоральною фразою

Оцінка результатів. На розміченому тексті видно, що модель класифікує як темпоральні вирази окрім дат ще й фізичні величини. Це зумовлено тим що тренувальний датасет не містив достатньої кількості виразів, що містять цифри, але не є датами. Поповнення датасету та покращення результатів класифікації є в планах дослідження. Як видно з прикладу модель добре розрізняє не тільки дати, але й неточні вираження про час такі як «наприкінці грудня», «другий місяць» тощо.

Висновки

Розглянуто застосування адаптованого методу дистиляції знань природною мовою для класифікації темпоральних фраз. Під час дослідження було створено та опубліковано датасет на навчена рекурентна нейромережа для класифікації темпоральних фраз з природномовних текстів. Здобуті результати

мають великий потенціал для розвитку та покращення. Збільшення точності вилучення темпоральних фраз та їх переведення в точний час є предметами подальших досліджень.

Суміжні дослідження пропонують наступні ідеї для розвитку. Динамічний вибір "вчителів" оптимізує процес дистиляції, покращуючи передання знань [5]. Багатоетапна дистиляція демонструє переваги у компресії великих багатомовних моделей, підкреслюючи важливість гнучких стратегій [6]. Подальший розвиток може включати використання знань зі зворотного проходу для генерації додаткових тренувальних зразків, що може значно покращити процес дистиляції [7]. Інноваційні методи, які сприяють покращенню передачі знань і компресії моделей, відкривають нові перспективи для подальших досліджень у галузі обробки природної мови [8].

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Anderson, M., & Gómez-Rodríguez, C. (2020). Distilling Neural Networks for Greener and Faster Dependency Parsing. ArXiv, abs/2006.00844. https://doi.org/10.18653/v1/2020.iwpt-1.2.
2. He, H., Shi, X., Mueller, J., Zha, S., Li, M., & Karypis, G. (2021). Distiller: A Systematic Study of Model Distillation Methods in Natural Language Processing. ArXiv, abs/2109.11105. https://doi.org/10.18653/v1/2021.sustainlp-1.13.
3. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. ArXiv, abs/1910.01108.
4. Sun, S., Cheng, Y., Gan, Z., & Liu, J. (2019). Patient Knowledge Distillation for BERT Model Compression. , 4322-4331. https://doi.org/10.18653/v1/D19-1441.
5. Yuan, F., Shou, L., Pei, J., Lin, W., Gong, M., Fu, Y., & Jiang, D. (2020). Reinforced Multi-Teacher Selection for Knowledge Distillation., 14284-14291. https://doi.org/10.1609/aaai.v35i16.17680.
6. Mukherjee, S., & Awadallah, A. (2020). XtremeDistil: Multi-stage Distillation for Massive Multilingual Models., 2221-2234. https://doi.org/10.18653/v1/2020.acl-main.202.
7. Jafari, A., Rezagholizadeh, M., & Ghodsi, A. (2023). Improved knowledge distillation by utilizing backward pass knowledge in neural networks. ArXiv, abs/2301.12006. https://doi.org/10.48550/arXiv.2301.12006.
8. Lin, Y., Chen, K., & Kao, H. (2023). LAD: Layer-Wise Adaptive Distillation for BERT Model Compression. Sensors (Basel, Switzerland), 23. https://doi.org/10.3390/s23031483.

Білецький Богдан Сергійович — аспірант кафедри системного аналізу та інформаційних технологій, e-mail: bohdanbeletskyi@gmail.com.

Biletskyi Bohdan S. — Post-Graduate Student of the Department of System Analysis and Information Technologies, e-mail: bohdanbeletskyi@gmail.com.