НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
«КИЄВО-МОГИЛЯНСЬКА АКАДЕМІЯ»

12.06.2024

Порівняння моно- та багато-мовних моделей на основі BERT для вирішення задач обробки мови українською

Виконав: Ванін Данило Олегович,
Магістерська програма 122 Комп'ютерні науки

Науковий керівник: Крюкова Галина Віталіївна,
доцент, кандидат фіз.-мат. наук

Рецензент: Марченко Олександр Олександрович,
професор, доктор фіз.-мат. наук

Постановка проблеми

- Мовні моделі тренуються у два етапи: базове навчання на великих нерозмічених мовних корпусах та донавчання на менших наборах даних.
- Для деяких мов немає достатньої кількості якісних базових даних. Серед них наразі українська.
- Перспективне рішення - багатомовні моделі, натреновані на об'єднаних корпусах різних мов, які показують гарну ефективність для багатьох мов водночас.
- Вибір між одномовними та багатомовними моделями залежить від мови та завдання. Результати досліджень різняться в залежності від мови та завдання.

Вступ

Мета: оцінити ефективність різних типів моделей на наборі задач української мови.

Об’єкт дослідження: одно- та багатомовні моделі на основі BERT

Предмет дослідження: порівняння продуктивності таких моделей на завданнях ОПМ для української мови

Розділ 1. Основні поняття

1. Описано основні завдання обробки природньої мови (ОПМ)
2. Описано процес тренування моделей, детальніше описані конкретні випадки тренування моделей BERT
3. Описані особливості одно та багатомовних моделей
4. Огляд стану ОПМ для української мови
5. Огляд робіт з порівняння одно- та багатомовних моделей для інших мов

Розділ 2. Порівняння одно- та багатомовних моделей на завданнях української мови

1. Зібрано перелік бенчмарків для різних завдань
2. Зібрано перелік порівнянь моделей з різних досліджень
3. Для деяких завдань дотреновано моделі для перевірки ефективності
4. Для завдання заповнення пропусків створено окремий бенчмарк та проведено якісну оцінку та опитування для визначення ефективності моделей
5. Зроблено висновок з результатів порівняння

Результати дослідження

Завдання ОПМ для української мови	Порівняння одно- та багатомовних моделей
Розпізнавання іменованих сутностей (NER)	Одномовна модель тренована в основному на українських текстах показує найкращий результат
Розрізнення значень слів (Word Sense Disambiguation)	Багатомовна модель дотренована українською показує кращий результат
Класифікація текстів	Модель адаптована до української методом WECHSEL показує кращі результати, ніж українська LiBERTa
Eval-UA-tion (комплекс 3 завдань)	Модель Mistral-7B-Sherlock дотренована на українських даних показує кращі результати ніж багатомовна модель.
Заповнення пропусків	Модель адаптована до української методом WECHSEL показує кращі результати, ніж українська YouScan/RoBERTa
Розмічування частин мови (POS Tagging)	Незначна різниця у результатах одно- та багатомовних моделей

Висновки

- Моделі, натреновані “з нуля” на українських текстах або дотреновані на них, в загальному мають кращі результати, ніж багатомовні.
- Українські моделі показали кращі можливості захоплювати український контекст та враховувати культурні нюанси.
- На простих завданнях різниця між ефективністю різних типів моделей незначна.
- Більшість моделей не досягають рівня комерційної моделі ChatGPT4.
- Гостро відчутна нестача комплексного бенчмарку для оцінки (як GLUE).
- Українська мова сягнула моменту, коли даних достатньо, щоб одномовні моделі перевершували багатомовні.

Наступні кроки

- Публікація модифікованої роботи англійською (з врахуванням результатів UNLP 2024, проведеної 25 травня)
- Розробка комплексного бенчмарку КОРМ (Комплексна Оцінка Розуміння Мови) для української мови, який був би аналогом GLUE (англ.) та KLEJ (польск.)
- Розширення переліку наборів даних для інших завдань української мови

Список джерел

1. BERT: pre-training of deep bidirectional transformers for language understanding / J. Devlin et al. Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers) / ed. by J. Burstein, C. Doran, T. Solorio. Minneapolis, Minnesota, 2019. P. 4171–4186. URL: https://aclanthology.org/N19-1423 (дата звернення: 23.05.2024).
2. Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024, м. Torino / ред.: M. Romanyshyn та ін. 2024. URL: https://aclanthology.org/2024.unlp-1 (дата звернення: 05.06.2024).
3. Ruder S. The state of multilingual AI. ruder.io. URL: https://www.ruder.io/state-of-multilingual-ai/ (дата звернення: 23.05.2024).
4. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding / A. Wang et al. 2019.
5. Minixhofer B., Paischer F., Rekabsaz N. WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies / ed. by M. Carpuat, M.-C. de Marneffe, I. V. Meza Ruiz. P. 3992–4006. URL: https://aclanthology.org/2022.naacl-main.293 (дата звернення: 31.05.2024).

Дякую за увагу!

Презентував: Ванін Данило Олегович,
Магістерська програма 122 Комп’ютерні науки

НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
«КИЄВО-МОГИЛЯНСЬКА АКАДЕМІЯ»