Языковая сложность в мультиязыковых моделях

Капелюшников Андрей Сергеевич

Студент

ФРКТ МФТИ, Москва, Россия

E-mail: kapeliushnikov.as@phystech.edu

Научный руководитель *— Шаров Сергей Александрович*

В настоящее время многие задачи, связанные с обработкой естественного языка, решаются при помощи моделей–трансформеров, в частности BERT. При этом внутренние процессы (закономерности в распространении активаций на скрытых слоях) изучены слабо. В данной работе рассматриваются корреляции внутренних состояний на задаче переноса предсказания языковой сложности между парами языков.

Задача предсказания языковой сложности в работе рассматривалась в двух формулировках:

1. Классификация текстов, размеченных экспертами на 6 классов сложности, согласно CEFR (ReadMe++).

2. Классификация бинарная классификация текстов статей взятых из Vikidia (класс простых текстов) и из Wikipedia (класс сложных текстов).

В обоих постановках эксперимент проводился в 3 этапа:

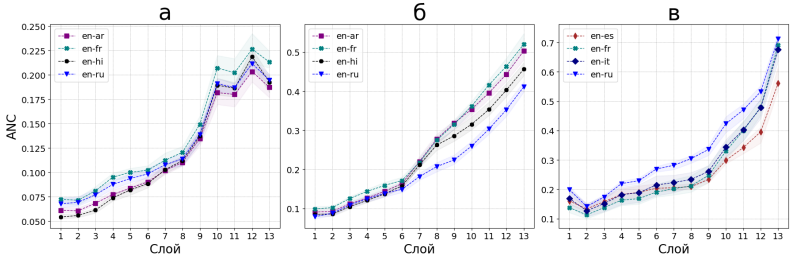
1. Дообучение (fine–tuning) модели на английских текстах;

2. Получение предсказаний и значений функций активации на скрытых слоях этой модели на других языках;

3. Расчёт метрики схожести активаций на скрытых слоях при предсказании для пар языков.

В качестве метрики схожести активаций использовали Average Neuron–Wise Correlation (ANC), предложенную Maksym D. и Mark F. В статье эта метрика использовалась для исследования переноса обучения между языками в задаче XNLI. В этой же работе показано преимущество ANC над другими ранее известными метриками (CKA и CCA).

Иллюстрации



ANC между активациями на слоях BERT в задачах: а) XNLI, б) Классификация текстов по CEFR, в) классификации Vikidia и Wikipedia. Заливка показывает область значений на других сэмплах из данных.

Выводы:

1. Метрика ANC устойчива к выбору данных для предсказания;
2. Характер поведения ANC значительно отличается на задачах XNLI и языковой сложности.

В дельнейшем планируется провести эксперименты по межязыковому переносу обучения для других задач классификации и добавить в сравнение поведение ANC на них.

Литература

1. Naous T. et al, ReadMe++: Benchmarking Multilingual Language Models for Multi–Domain Readability Assessment // arXiv preprint, 2023, arXiv:2305.14463.
2. Maksym D., Mark F., Cross–lingual Similarity of Multilingual Representations Revisited // In Proceedings of the 2nd Conference of the Asia–Pacific Chapter of the Association for Computational Linguistics, 2022, V. 1, P. 185—195.