УДК 004.8

ОБУЧЕНИЕ МОДЕЛИ КРОСС-ЯЗЫЧНОГО СЕМАНТИЧЕСКОГО ПОИСКА

Юхневич Е.Д.[1]
Научный руководитель – Маслюхин С.М.[2]
[1]Университет ИТМО, [2]ООО "ЦРТ-инновации"
katyauhn@ya.ru

Введение.

Модели семантического поиска находят применения во множестве различных задач. Наиболее актуальным направлением исследований в данной области являются мультиязычные модели, способные работать с моно-, мульти- и кросс-язычными данными. Однако при обучении таких моделей может проявляться проблема негативной интерференции языков, когда обучение на множестве языков негативно сказывается на финальных метриках качества работы модели для каждого языка отдельно. При использовании архитектур, которые «разграничивают» языки, например, с помощью специализированных моноязычных адаптеров, может наблюдаться улучшение метрик на моноязычных задачах, однако страдает качество работы модели с кросс-язычными данными [1]. В связи с этим, актуальным и важным направлением исследований является разработка новых подходов к обучению моделей семантического поиска, способных решать проблему негативной интерференции языков без ухудшения метрик на кросс-язычных задачах.

Основная часть.

В работе было проведено экспериментальное сравнение различных модульных подходов к дообучению мультиязычных моделей семантического поиска с использованием LoRA адаптеров.

В первой части работы были обучены моноязычные адаптеры для 6 языков (русский, арабский, фарси, английский, испанский, французский) для нескольких базовых моделей. Экспериментально было доказано, что улучшения метрик на моноязычных задачах с применением подхода с адаптерами возможно добиться, если базовая модель является моноязычной. При обучении моноязычных адаптеров для мультиязычной модели наблюдалось снижение метрик, из-за чего можно предположить, что разделение языков неэффективно использовать в методах доработки архитектуры после обучения, когда модель уже была обучена на большом объёме мультиязычных данных и между эмбеддингами разных языков уже сформировались устойчивые связи.

Вторая часть работы была посвящена обучению LoRA адаптеров для применения в кросс-язычных задачах. Исследовался вопрос поиска оптимального использование как моно-, так и кросс-язычных данных и их распределения при обучении адаптеров. Были рассмотрены такие подходы, как простое дообучение на кросс-язычных данных адаптеров, обученных на моноязычных данных, а также Stacked LoRA [2], использование последовательно моноязычных и кросс-язычных адаптеров. Наилучшего улучшения метрик на кросс-язычных задачах удалось достичь при многоэтапном обучении адаптеров сначала на моно, а потом на кросс-язычных данных. При этом для всех языков в кросс-язычных данных использовались пары на целевом и английском языке. Таким образом, эмбеддинги на разных языках «выравнивались» относительно английского.

Полученный набор адаптеров был оценён на двух бенчмарках для кросс-язычных задач: belebele и xquad. Анализ результатов показал, что увеличение метрик очень неравномерно для разных пар языков. Максимальный прирост наблюдался для документов на арабском, фарси и русском языках. Были сделаны выводы о том, что выравнивание относительно английского языка существенно улучшает работу для неевропейских языков,

так как изначально модель лучше всего работает на английском языке.

Выводы.

В работе было исследовано применение архитектуры модели с набором LoRA адаптеров как для моноязычных, так и для кросс-язычных задач. Экспериментально было доказано, что при работе с моделями, уже обученными на большом количестве мультиязычных данных, применение подхода с адаптерами более актуально и эффективно для кросс-язычных задач. Был обучен набор адаптеров для шести различных языков с применением как моно-, так и кросс-язычных данных. Улучшение работы модели было протестировано на наборах данных belebele и xquad. Наилучших приростов метрик качества работы модели удалось достичь для поиска среди документов на неевропейских языках.

Литература

1. Huang Y. et al. Modular sentence encoders: Separating language specialization from cross-lingual alignment //Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). – 2025. – С. 2167-2187.
2. Patil H. V., Sanam V., Atre M. P. Stacked LoRA: Isolated Low-Rank Adaptation for Lifelong Knowledge Management //The 14th International Joint Conference on Natural Language Processing and The 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics. – 2025. – С. 36-46.