

Название:

АДАПТАЦИЯ БОЛЬШИХ МУЛЬТИЯЗЫЧНЫХ ЯЗЫКОВЫХ МОДЕЛЕЙ ПУТЕМ НАСТРОЙКИ ТОКЕНИЗАЦИИ НА РУССКИЙ ЯЗЫК

Автор:

Тихомиров М.М.

МГУ им. М.В. Ломоносова, Научно-исследовательский вычислительный центр, Лаборатория анализа информационных ресурсов. Кандидат физ.-мат. наук.

tikhomirov.mm@gmail.com

Индекс УДК:

УДК 004.8

Аннотация:

В докладе рассматривается проблема адаптации больших языковых моделей (LLM) с целью повышения вычислительной эффективности и качества при работе на русском языке. Большинство современных открытых LLM обучаются на преимущественно англоязычных корпусах текстов, из-за чего их использование на русском языке менее эффективно. Обучение подобных моделей с нуля стоит миллионы долларов, поэтому в работе исследуется метод адаптации мультязычных LLM путем обучения более подходящего под язык токенизатора с дальнейшей его заменой в языковой модели и дообучением на репрезентативном корпусе специальным образом.

Ключевые слова:

Большие языковые модели, токенизация, искусственный интеллект, обработка естественного языка.

Основной текст:

В последние годы в области обработки естественного языка (и всей отрасли искусственного интеллекта) происходят революционные изменения за счет развития больших нейросетевых языковых моделей, таких как GPT (генеративные модели на архитектуре «трансформер») (Radford A. et al., 2018). Большие языковые модели (LLM) представляют собой большую нейросеть (десятки и сотни миллиардов параметров), которые обучаются задаче языкового моделирования на огромных корпусах текстов (триллионы слов). За счет размеров модели и количества данных для обучения, современные языковые модели содержат в себе обширные знания о мире и позволяют решать широкий спектр задач обработки естественного языка.

Несмотря на бурное развитие данного направления, преимущественно оно происходит для английского языка, а для русского языка отсутствуют модели, которые были бы способны иметь сопоставимое качество с англоязычными моделями на англоязычных бенчмарках. Это в первую очередь связано с тем, что обучение подобных моделей с нуля требует миллионов долларов, что могут позволить себе только некоторые коммерческие компании. Во-вторых, морфология русского языка более сложная, чем в

английском языке, что добавляет моделям дополнительных сложностей во время обучения.

По этим причинам, одной из актуальных задач является развитие способов адаптации подобных мультязычных / англоязычных моделей на русский язык. Нашей группой был предложен метод адаптации большой языковой модели LLaMa 7B (Touvron H. et al., 2023) на русский язык путем замены оригинального токенизатора на новый (а также слоя эмбедингов и слоя предсказания следующего слова, lm head). Новый токенизатор обучался на русскоязычных текстовых данных, из-за чего любое слово представляется в среднем меньшим количеством токенов. Так, например, слово **интеллект** в новом токенизаторе представляется одним токеном, а в оригинальном разбивалось на три: ['инте', 'лле', 'кт']. Из-за этого исходной модели при работе на русском языке приходится оперировать не более-менее разумными токенами, но символьными n-граммами, что негативно влияет на вычислительную эффективность, а также усложняет задачу для языковой модели, так как от нее требуется агрегировать информацию практически на уровне символов.

В нашем исследовании было показано, что алгоритм unigram (Kudo T., 2018) для токенизации подходит лучше для русского языка, чем повсеместно применяемый BPE (Sennrich R., Haddow B., Birch A., 2016). А также было обнаружено, что для адаптации модели достаточно заменить токенизацию, слой эмбедингов и слой предсказаний и дообучить только новые слои на русскоязычных текстах.

Разработанный подход позволил повысить среднее качество на бенчмарке Russian Super Glue (Shavrina T. et al., 2020) по сравнению с исходной моделью с 0.68 до 0.7. Также эксперименты показали, что unigram подход к токенизации позволяет достичь более высокого качества на прикладных задачах, чем BPE.

Помимо этого, за счет перехода на новую токенизацию, существенно выросла вычислительная эффективность обработки русскоязычных текстов. Во время использования модели скорость генерации текстовой последовательности одинаковой длины выросла на 60% (чем длиннее текст, тем больше эффект). Во время обучения модель стала быстрее на 35%. Помимо этого, также снизилось количество необходимой памяти для работы модели.

1. Radford A. et al. Improving language understanding by generative pre-training. – 2018.
2. Touvron H. et al. Llama: Open and efficient foundation language models //arXiv preprint arXiv:2302.13971. – 2023.
3. Sennrich R., Haddow B., Birch A. Neural Machine Translation of Rare Words with Subword Units //54th Annual Meeting of the Association for Computational Linguistics. – Association for Computational Linguistics (ACL), 2016. – С. 1715-1725.

4. Kudo T. Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates //Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). – Association for Computational Linguistics, 2018.
5. Shavrina T. et al. RussianSuperGLUE: A Russian Language Understanding Evaluation Benchmark //Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). – 2020. – C. 4717-4726.