Использование языковых моделей на основании архитектуры трансформеров для понимания поисковых запросов на электронных торговых площадках

Ф.В. Краснов

***Аннотация*— Определение намерения пользователя по тексту поискового запроса – один из этапов извлечения информации в системах интеллектуального поиска товаров на электронной торговой площадке. Рассматривая поисковые запросы как коллекцию коротких текстовых документов, а намерения пользователей как классы, автор продолжил исследование подходов к задаче многоклассовой классификации коротких текстов с помощью моделей на основании архитектуры трансформеров. В последнее время подход к обучению языковой модели на последовательностях токенов и дальнейшая тонкая настройка на предметную область хорошо себя зарекомендовали. В рамках этого подхода, автор рассмотрел вероятность появления метки класса, в качестве одного из токенов языковой модели на основании трансформера. Данный подход отличается от линейной суперпозиции токенов с применением функции активации для определения вероятности класса при тонком обучении. Одно из преимуществ такого подхода – классы приобретают компактные векторные представления (эмбеддинги). Автор экспериментально подтвердил преимущества и недостатки обоих подходов на текстовых данных поисковых запросов. При оптимальных гиперпараметрах точность предложенного подхода, полученная по метрике f1-score weighted, составила 96 %. Анализ небольших по размеру наборов данных позволил оценить характерные для языковых моделей недостатки, только усиливающиеся при масштабировании. А также вновь убедиться в том, что языковые модели – это вынужденное решение в условиях огромных наборов данных, а не безальтернативное преимущество.**

***Ключевые слова* — Language Models, LM, transformers, e-commerce, Information Retrieval, QU, IR.**

I. Введение

Большинство современных поисковых систем, рекламных платформ и рекомендательных систем используют схожую архитектуру многоуровневого поиска информации (IR), включающую этап отбора кандидатов или извлечения информации и этап упорядочивания кандидатов или ранжирования. Учитывая поисковый запрос и намерение пользователя, этап поиска сокращает количество возможных кандидатов с миллионов, иногда миллиардов, до сотен или меньше. Затем на этапе ранжирования настраивается порядок отбора кандидатов для представления пользователю. Такой подход является одновременно гибким и масштабируемым.

Понимание поисковых запросов (query understanding, QU) сопоставляет текст поискового запроса с категориями товаров, атрибутами или их комбинациями, позже используемыми в качестве ограничений поиска кандидатов. Такие методы часто используют понимание содержимого для извлечения метаданных из текста поискового запроса, чтобы выделить сущности и добавить структуру к тексту поискового запроса.

Понимание поисковых запросов – это только один из уровней их обработки. Другие уровни обработки поискового запроса схематически отображены на рис. 1.

Наличие нескольких уровней обработки поискового запроса служит, с одной стороны, для многокритериального уточнения списка товаров-кандидатов, а с другой – позволяет меньшими ресурсами обрабатывать большее количество поисковых запросов. Это достигается за счет того, что начальные уровни обработки поискового запроса построены на технологиях, требующих меньшее количество и большую производительность ресурсов, а дальнейшие уровни обработки поискового запроса могут использовать более сложные и ресурсоемкие модели машинного обучения.

1 Статья получена 31 мая 2023. Ф.В.Краснов, Исследовательский центр ООО "ВБ СК" на базе Инновационного Центра Сколково. krasnov.fedor2@wb.ru, http://orcid.org/0000-0002-9881-7371.

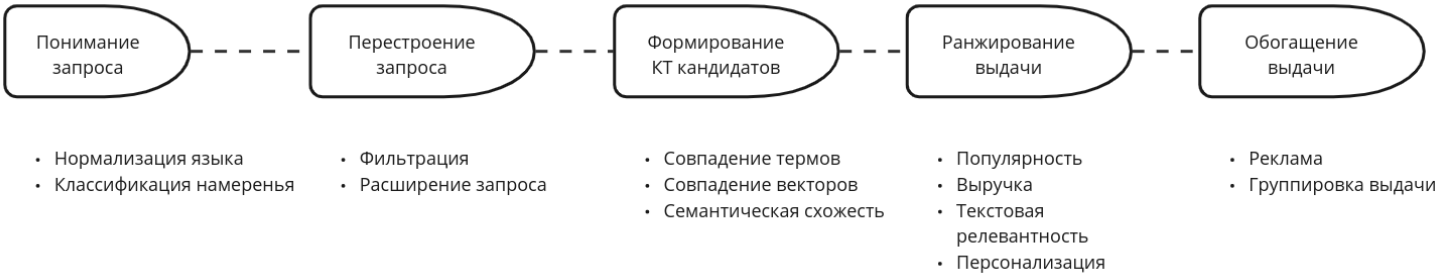

Рисунок 1 – Уровни обработки поискового запроса

II.Методика

Цель поискового запроса на электронной торговой площадке – выразить в виде текста намерение, интерес пользователя к определенному товару. При наличии в каталоге сотен миллионов товаров не всегда возможно определить товар точно по поисковому запросу. Кроме того, иногда необходимо представить ассортимент товаров: сопутствующие товары, товары-заменители. Поэтому намерение пользователя определяют через иерархию товаров – товарную категорию.Общая постановка задачи данного исследования соответствует задаче классификации, где намерения – это классы C_i , а тексты поисковых запросов – источник для получения признаков (feature extraction). Подходы к данной задаче разносторонне изучены. Например, в исследовании [1] рассматривается классификация на основе таксономии товаров, в работе [2] изучен вопрос неоднозначности текстового выражения намерений в поисковых запросах. В настоящее время исследователи поисковых запросов больше внимания уделяют использованию нейросетевых моделей глубокого обучения на основе трансформеров, например, таких как BERT. Они рассматривают переформулировки поисковых запросов для более точного выявления намерений [3] и построения семантического пространства с целью поиска наилучшего соответствия компактного векторного представления запроса и категории товара [4], [5].

Базовые значения для метрики f1-score исследованы в работах по распознаванию намерений в диалоговых системах [6], [7], [8]. Они составляют более 90 % при использовании BERT и LSTM [6].

В профессиональной литературе нет единого мнения о том, как рассматривать тексты поисковых запросов: в качестве упорядоченных последовательностей токенов или как набор неупорядоченных токенов. Априори обе точки зрения имеют свои преимущества и недостатки, которые влияют на метрики классификации. Также представляется важным проанализировать промежуточный, с точки зрения упорядоченности токенов, вариант: дистрибутивная семантика, в рамках которой окно из N токенов передвигается последовательно по тексту поискового запроса, а внутри окна токены рассматриваются безотносительно своего расположения, в роли контекста [9]. Механизм внимания (self attention), ставший фетишем для многих ученых после исследования [10], также не поддерживает упорядоченность токенов. В так называемых Causal language models, к которым относится модель GPT из работы [10], все предыдущие токены в рассматриваемом моделью фрагменте текста (блоке) связаны отношением каждый с каждым для предсказания следующего токена, то есть токены рассматриваются как не упорядоченные.

Содержание поисковых запросов не всегда является языковой конструкцией (например, утверждением или вопросом) в полном объеме. В поисковом запросе не всегда можно выделить синтаксически связанное сочетание слов, словесное построение, он не всегда обладает полноценными частями предложений, знаками препинания и пр. Поэтому представляется малоэффективным использование предобученных больших языковых моделей (LLM), так как их обучение произведено на корпусе текстов, составленном из веб-страниц и книг, содержащих полноценные языковые конструкции, в том числе с кодированием порядка слов [11]. В поисковом запросе порядок слов часто бывает не важен.

Отдельно стоит отметить такую специфику поисковых запросов, как опечатки и сленг. Согласно исследованиям [12], [13] доля опечаток в поисковых запросах составляет более 10 %. Чтобы определить намерение по тексту поискового запроса, высокий уровень низкочастотных токенов, создаваемый опечатками и сленгом, необходимо обрабатывать отдельно. Сленг, то есть лексика неформального регистра, так же как и опечатки увеличивает размер словаря, и делает еще более разряженными распределения токенов. Тем временем, сленг может оказаться наиболее сжатым выражением намерения. Например, такие сленговые слова, как «худи», «зипка» соответствуют различным моделям верхней одежды, и редкий пользователь в поисковом запросе развернуто описывает намерение.

Начиная с работы [14] для улучшения метрик модели стали применять подход, в котором сначала строится языковая модель, а затем производится тонкая настройка на решение определенной задачи. Так, в исследовании [10 перечисляются следующие варианты тонкой настройки языковой модели на основании архитектуры трансформеров (рис. 2):

1. Классификация документов;
2. Дополнение фрагмента текста;
3. Схожесть документов;
4. Выбор ответа на вопрос из списка.

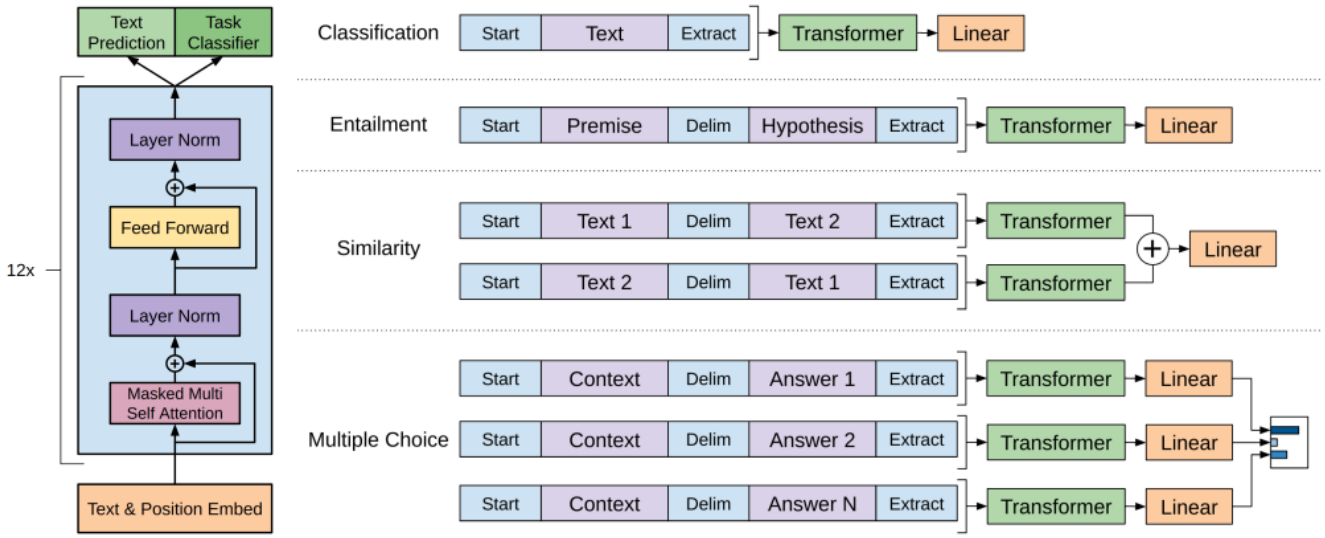

Рисунок 2 – Варианты тонкой настройки языковой модели на основании архитектуры трансформеров на предметный домен и задачу из [10]

В условиях задачи классификации предобученная языковая модель может быть использована без изменения или дообучена на примерах с разметкой классов.

В настоящее время исследователи наращивают количество параметров в моделях для улучшения метрик. Рациональное сравнение сложности и размера текстов в поисковых запросах и в корпусе текстов для обучения в исследовании [10] наводит на мысль о том, что, быть может, подобное количество параметров, используемое в современных больших языковых моделях, будет избыточным для настоящего исследования. Поэтому в качестве языковой модели целесообразно использовать уменьшенный вариант архитектуры нейронной сети из работы [10]. В качестве потенциальных возможностей для упрощения в первую очередь можно рассмотреть размерность компактных векторных представлений токенов (эмбеддингов), количество головок в механизме внимания и количество повторяющихся блоков-слоев в модели.

Внимание исследователей также обращено на увеличение размера рассматриваемого моделью фрагмента текста (block_size). В современных моделях архитектуры BERT используются фрагменты текста размером 512 токенов. Отметим, несмотря на то, что в недавней работе [15] рассмотрен вариант увеличения размера фрагмента текста до 1 млн токенов, в настоящем исследовании нет необходимости даже в стандартном для предобученных моделей размере в 512 токенов. Количество токенов в поисковом запросе зависит от выбранного способа токенизации, но в случае разбиения «через пробел» в тексте поискового запроса содержится порядка 5 токенов.

Распределение поисковых запросов по количеству токенов представлено на рис. 3.

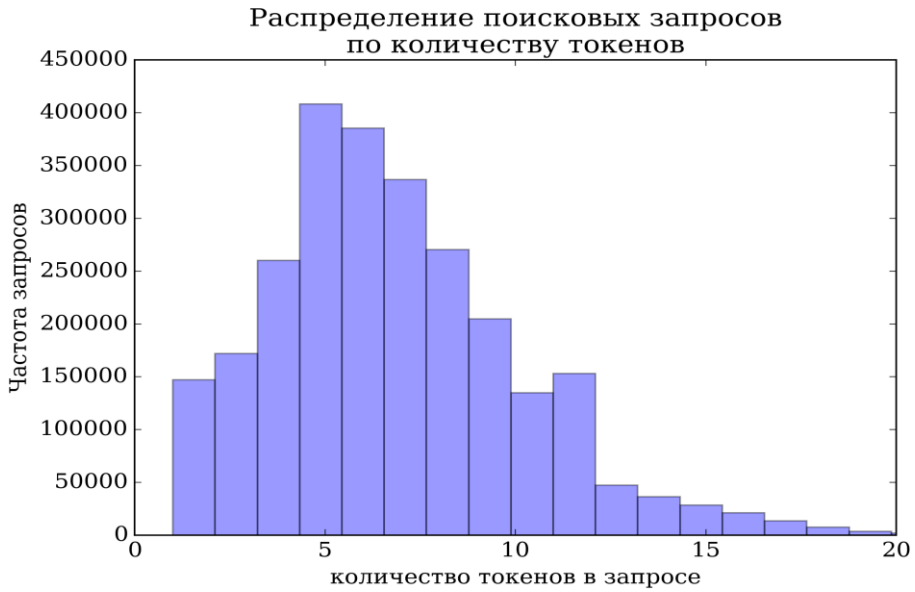

Рисунок 3 – Распределение поисковых запросов по количеству токенов

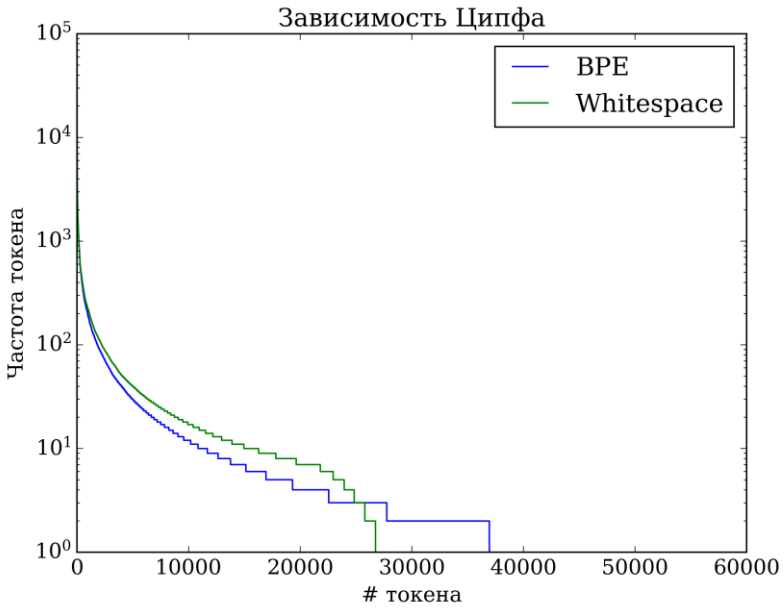

Рисунок 4 - Зависимость Ципфа для двух типов токенизации, «через пробел» (whitespace) и BPE с фиксированным размером словаря 30 тысяч токенов

Современные токенизаторы основываются на субсловарных N-граммах и используют такие алгоритмы сжатия словаря, как BPE [16]. Основным гиперпараметром для BPE является размер словаря. Для выбора настоящего параметра можно руководствоваться зависимостью Ципфа, представленной на рис.4 .

Из зависимости Ципфа на рис.4 можно отметить, что распределение по частот-встречаемости токенов, полученное с помощью whitespace-токенизации, мало отличается от сжатой с помощью алгоритма BPE версии словаря. Таким образом, потери информации от эффекта Out-of-Vocabulary будут незначительными.

Одно из преимуществ больших языковых моделей – это индуктивность обучения. Суть удобства индуктивности обучения состоит в том, что появление новых примеров может быть применено для дообучения существующей модели. В условиях огромных корпусов текстов и длительного времени обучения модели индуктивность незаменима. В рамках задачи настоящего исследования индуктивность не является обязательным требованием. Несмотря на то, что число примеров текстов поисковых запросов на электронной торговой площадке может достигать миллиарда в месяц, нет необходимости дообучаться, а не обучаться «с нуля». Эта особенность рассматриваемого предметного домена объясняется структурой распределения частот поисковых запросов, в котором менее 30 % всего количества поисковых запросов составляют высокочастотные поисковые запросы. Кроме того, появление новых категорий товаров также требует полного переобучения модели. Поэтому рассматриваемые архитектуры моделей могут быть как индуктивны, так и трансдуктивны. К примеру, такие модели «неглубоких» компактных векторных представлений токенов, как fastText [17], Glove [17] по своей сути являются трансдуктивными и используются

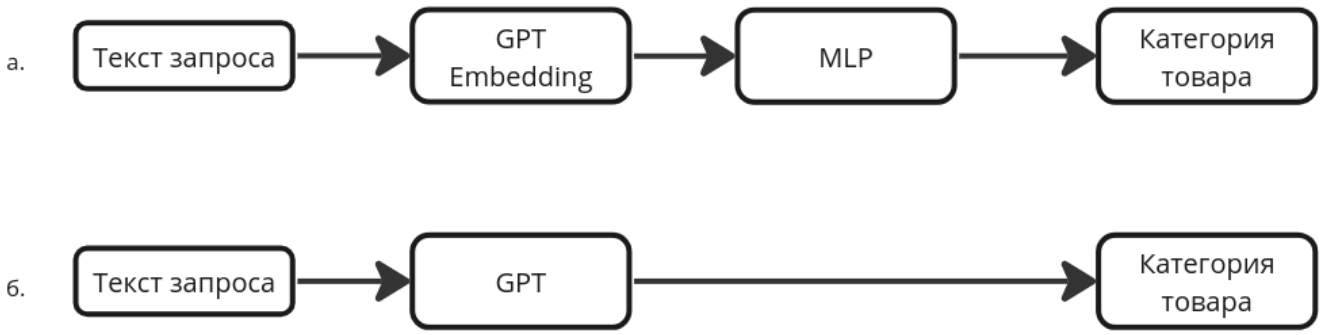

Рисунок 5 – Исследовательский фреймворк:
а. Использование эмбеддингов (компактных векторных представлений токенов) для классификации;
б. Использование языковой модели для классификации (вероятность метки класса как следующего токена в последовательности)

наравне с моделями глубокого обучения на основе архитектуры трансформеров.

Автор предлагает рассмотреть вариант использования языковых моделей на основании архитектуры трансформеров для понимания поисковых запросов, который отличается от изученных в работе [10]. Суть предлагаемой автором методики отображена на рис. 5 в виде схемы исследовательского фреймворка.

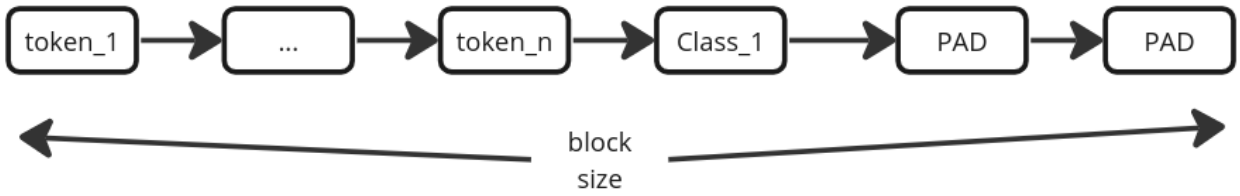

Рисунок 6 – Принцип выравнивания фрагмента текста с меткой класса

На рис. 5 для выделения основной идеи использована именно языковая модель на основании GPT, но для варианта 5а в эксперименте могут быть использованы и другие модели. Основное отличие варианта 5б в том, что категория товара становится частью словаря языковой модели, и как токен, следует за последним токеном поискового запроса.

Методически важно рассмотреть специфику механизма приведения входного для модели набора документов к единой длине (количеству токенов), которая возникает в случае б на рис. 5. После токенизации каждый документ представляется последовательностью ID токенов разной длины. Для применения тензорного исчисления все документы обрезают или дополняют пустыми токенами (PAD) до определенной длины блока. Основной критерий выбора длины блока – не смысловые границы фрагмента текста, а размер модели. В случае поисковых запросов выбор длины блока позволяет нам учитывать весь текст запроса целиком. Другой особенностью является то, что в случае добавления класса в качестве последнего токена запроса (рис. 5б) необходимо сделать выравнивание по длине блока уже после добавления (рис. 6).

Суммируя вышесказанное, автор сделал ряд эмпирических исследований на открытых наборах данных с целью проверки изложенной методики и выявления наиболее эффективного подхода к решению задачи использования языковых моделей на основании архитектуры трансформеров для понимания поисковых запросов на электронных торговых площадках.

III.Эксперимент

Эксперимент проведен на двух открытых наборах данных: dateset1, открытый набор данных, опубликованный исследователями лаборатории компании Amazon [19], dataset2, открытый набор данных, собранный экспертами из Leibniz Institute for the Social Sciences [20] для исследования целей потребителей электронных торговых площадок. Оба набора данных состоят из текстов поисковых запросов и категорий товаров, соответствующих этим запросам.

Набор данных dataset1 после обработки включает 97 тысяч поисковых запросов на английском языке. Так как набор данных dataset1 содержал категориальную оценку запроса из четырех суждений (Exact, Substitute, Complement, Irrelevant), то для настоящего исследования из набора данных была взята только информация о товаре, соответствующая суждению Exact – товар полностью соответствует запросу. Набор данных dataset2 состоит из 3.5 тысяч пар: текст поискового запроса и категория товара.

Для токенизации выбрана модель на основе BPE [16]. Размер словаря предпочтен согласно рекомендациям в статье [10] и равен 32 тысячам токенов. Однако в ходе обучения моделей токенизации размер словаря составил 16 тысяч токенов, что повлияло на уменьшение количества параметров моделей.

Упрощенная языковая модель на архитектуре трансформер содержит 8.63 млн параметров для dataset1 и 2.17 млн параметров для dataset2. По отношению к оригинальной модели из [14] уменьшено количество головок в механизме внимания до двух и количество повторяющихся блоков-слоев в модели до двух. Подробная информация обо всех слоях языковой модели приведена в приложении 1.

Для self-supervised обучения языковой модели был использован набор данных из поисковых запросов. Обучение проводилось в течение 600 эпох на процессорах CPU с использованием расширенных векторных инструкций AVX, без использования сопроцессоров на графических ускорителях GPU. Для оценки обучения использовалась отложенная выборка данных размером 10 % от основной выборки, стратифицированная по меткам классов. На рис. 7 приведен график значений функции потерь от эпох.

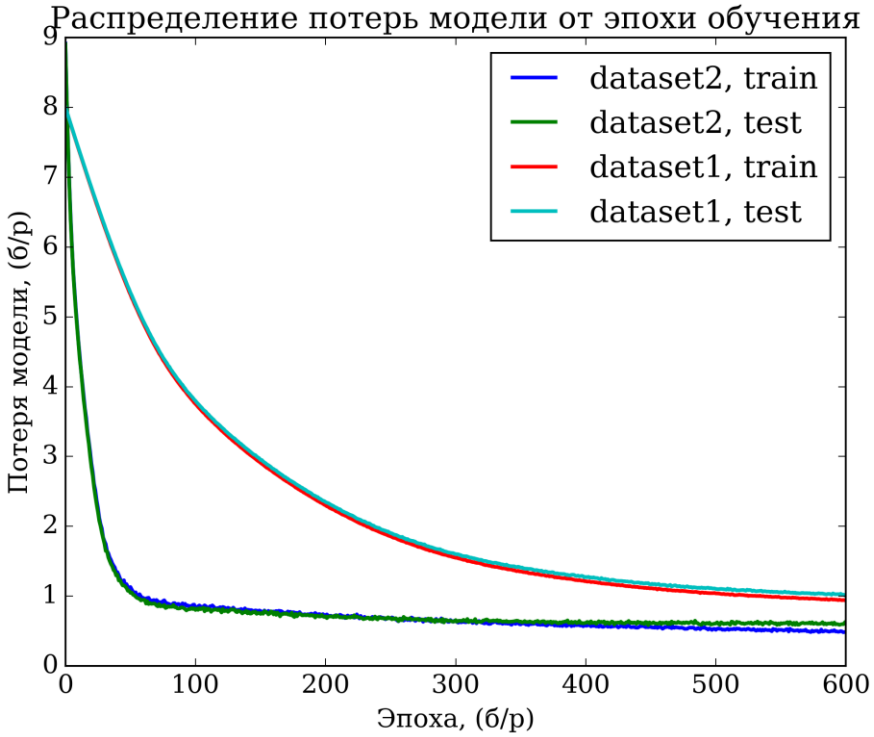

Рисунок 7 – Функции потерь языковой модели для dataset1 и dataset2 в зависимости от эпох обучения

Как видно из рисунка 7, потери уменьшаются неравномерно, и к 600 эпохе падение потерь замедляется, что свидетельствует, о минимизации потерь. Исходя из отношения потерь на учебной и отложенной выборке, эффект переобучения модели не наблюдался. В дальнейшем эксперименте полученные языковые модели использовались в двух вариантах, проиллюстрированных на рис. 5.

Таблица 1 – Значение метрики f1-score для моделей на отложенной выборке

Модель	F1-score (weighted)	
	dataset1	dataset2
Baseline (LR+BPE)	98%	99%
GPT + FC (5.а)	95%	96%
GPT (с классами, 5.б)	96%	97%
fastText (SkipGgram)	98%	99%

Результаты эксперимента, представленные в таблице 1, подтверждают гипотезы, которые изложены автором в разделе Методика. Использование моделей на архитектуре трансформеров для понимания поисковых запросов на электронных торговых площадках имеет ряд преимуществ и недостатков. Для рассмотренных автором небольших наборов данных метрика F1-score, полученная в результате применения моделей на архитектуре трансформеров немного ниже, чем для модели fastText на основе дистрибутивной семантики и для модели Logistic Regression с токенизацией BPE (LR+BPE). Модель fastText в отличие от GPT+FF, GPT (с классами) и LR не поддерживает онлайн-обучение, другими словами трансдуктивна, что, в свою очередь, не так важно для операционного использования в рамках задачи понимания поисковых запросов на электронных торговых площадках.

IV.ЗАКЛЮЧЕНИЕ

В данном исследовании автор рассмотрел различные методики постановки задачи определения намерения пользователя по тексту поискового запроса в системах интеллектуального поиска товаров на электронной торговой площадке.

Автор предложил и опробовал новый подход к обучению модели GPT с добавлением меток класса в словарь токенов. Предложенный подход показал точность по f1-score выше, чем подход, основанный на линейной суперпозиции токенов. Однако метрики модели на основе архитектуры трансформеров на небольших наборах данных уступают моделям на основе fastText (дистрибутивная семантика) и Logistic Regression с токенизацией BPE (LR+BPE).

Отдельно стоит отметить, что подход на основе нейросетевых моделей глубокого обучения с архитектурой трансформер не обязательно должен содержать миллиарды параметров, чтобы успешно справляться с задачами. Принцип рационального соответствия сложности модели и решаемой задачи позволил автору предложить и экспериментально опробовать наиболее современные нейросетевые архитектуры для небольшой прикладной задачи.

Библиография

[1] Skinner M., Kallumadi S. E-commerce Query Classification Using Product Taxonomy Mapping: A Transfer Learning Approach // eCOM@ SIGIR. – 2019.

[2] Papenmeier, A., Kern, D., Hienert, D., Sliwa, A., Aker, A., & Fuhr, N. (2021, March). Dataset of Natural Language Queries for E-Commerce. In Proceedings of the 2021 Conference on Human Information Interaction and Retrieval (pp. 307-311).

[3] Hirsch, S., Guy, I., Nus, A., Dagan, A., & Kurland, O. (2020, July). Query reformulation in E-commerce search. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 1319-1328).

[4] Kong, W., Khadanga, S., Li, C., Gupta, S. K., Zhang, M., Xu, W., & Bendersky, M. (2022, August). Multi-aspect dense retrieval. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (pp. 3178-3186).

[5] Zhang, Q., Yang, Z., Huang, Y., Chen, Z., Cai, Z., Wang, K., ... & Gao, J. (2022). A Semantic Alignment System for Multilingual Query-Product Retrieval. arXiv preprint arXiv:2208.02958.

[6] Gu Y. et al. Speech intention classification with multimodal deep learning // Advances in Artificial Intelligence: 30th Canadian Conference on Artificial Intelligence, Canadian AI 2017, Edmonton, AB, Canada, May 16-19, 2017, Proceedings 30. – Springer International Publishing, 2017. – С. 260-271.

[7] Chen Q., Zhuo Z., Wang W. Bert for joint intent classification and slot filling // arXiv preprint arXiv:1902.10909. – 2019.

[8] Gangadharaiah R., Narayanaswamy B. Joint multiple intent detection and slot labeling for goal-oriented dialog // 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). – 2019. – С. 564-569.

[9] Joulin, A., Grave, É., Bojanowski, P., & Mikolov, T. (2017, April). Bag of Tricks for Efficient Text Classification. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers (pp. 427-431).

[10] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[11] Ma, Y., Cao, Y., Hong, Y., & Sun, A. (2023). Large language model is not a good few-shot information extractor, but a good reranker for hard samples!. arXiv preprint arXiv:2303.08559.

[12] Dereza O. V., Kayutenko D. A., Marakasova A. A., Fenogenova A. S.A Complex Approach to Spellchecking and Autocorrection for Russian // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2016”. — 2016. –С. 1-11.

[13] Näther M. An in-depth comparison of 14 spelling correction tools on a common benchmark //Proceedings of the 12th Language Resources and Evaluation Conference. – 2020. – С. 1849-1857.

[14] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

[15] Bulatov, A., Kuratov, Y., & Burtsev, M. S. (2023). Scaling Transformer to 1M tokens and beyond with RMT. arXiv e-prints, arXiv-2304.

[16] Gage, P. (1994). A new algorithm for data compression. C Users Journal, 12(2), 23-38.

[17] Joulin, Armand, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. "Fasttext. zip: Compressing text classification models." arXiv preprint arXiv:1612.03651 (2016).

[18] Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532-1543. 2014.

[19] Reddy, C. K., Màrquez, L., Valero, F., Rao, N., Zaragoza, H., Bandyopadhyay, S., ... & Subbian, K. (2022). Shopping Queries Dataset: A Large-Scale ESCI Benchmark for Improving Product Search. arXiv preprint arXiv:2206.06588.

[20] Papenmeier, A., Kern, D., Hienert, D., Sliwa, A., Aker, A., & Fuhr, N. (2021, March). Dataset of Natural Language Queries for E-Commerce. In Proceedings of the 2021 Conference on Human Information Interaction and Retrieval (pp. 307-311).

ПРИЛОЖЕНИЕ 1. Схема слоев языковой модели

```
LanguageModel(
  (token_embedding_table): Embedding(32000, 128)
  (position_embedding_table): Embedding(780, 128)
  (blocks): Sequential(
    (0): Block(
      ...
    )
    (1): Block(
      (sa): MultiHeadAttention(
        (heads): ModuleList(
          (0): Head(
            ...
          )
          (1): Head(
            (key): Linear(in_features=128, out_features=64,
bias=False)
            (query): Linear(in_features=128, out_features=64,
bias=False)
            (value): Linear(in_features=128, out_features=64,
bias=False)
            (dropout): Dropout(p=0.1, inplace=False)
          )
        )
        (proj): Linear(in_features=128, out_features=128,
bias=True)
        (dropout): Dropout(p=0.1, inplace=False)
      )
      (ffwd): FeedFoward(
        (net): Sequential(
          (0): Linear(in_features=128, out_features=512,
bias=True)
          (1): ReLU()
          (2): Linear(in_features=512, out_features=128,
bias=True)
          (3): Dropout(p=0.1, inplace=False)
        )
      )
      (ln1): LayerNorm((128,), eps=1e-05,
elementwise_affine=True)
      (ln2): LayerNorm((128,), eps=1e-05,
elementwise_affine=True)
    )
  )
  (ln_f): LayerNorm((128,), eps=1e-05,
elementwise_affine=True)
  (lm_head): Linear(in_features=128, out_features=32000,
bias=True)
)
```

Query understanding via Language Models based on transformers for e-commerce

F.V. Krasnov

Abstract — Determining the user's intention by the text of the search query is one of the stages of extracting information in intelligent product search systems on an electronic trading platform. Considering search queries as a collection of short text documents, and user intentions as classes, the author continued to study approaches to the task of multi-class classification of short texts using models based on the architecture of transformers. The approach to teaching a language model based on token sequences and further fine-tuning to the subject area has proven itself well recently. Inspired by this approach, the author considered the probability of a class label appearing as one of the tokens of a language model based on a transformer. This approach differs from a linear superposition of tokens using an activation function to determine the probability of a class in fine learning. One of the advantages of this approach is that classes acquire compact vector representations (embeddings). The author experimentally confirmed the advantages and disadvantages of both approaches on the text data of search queries. With optimal hyper-parameters, the accuracy of the proposed approach obtained by the f1-score weighted metric was 96%. Consideration of small data sets allowed us to assess the disadvantages characteristic of language models, which will only increase with scaling, to make sure once again that language models are a forced solution in the conditions of huge data sets, and not an alternative advantage.

Key words — Language Models, LM, transformers, e-commerce, Information Retrieval, QU, IR

References

[1] Skinner M., Kallumadi S. E-commerce Query Classification Using Product Taxonomy Mapping: A Transfer Learning Approach // eCOM@ SIGIR. – 2019.
[2] Papenmeier, A., Kern, D., Hienert, D., Sliwa, A., Aker, A., & Fuhr, N. (2021, March). Dataset of Natural Language Queries for E-Commerce. In Proceedings of the 2021 Conference on Human Information Interaction and Retrieval (pp. 307-311).
[3] Hirsch, S., Guy, I., Nus, A., Dagan, A., & Kurland, O. (2020, July). Query reformulation in E-commerce search. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 1319-1328).
[4] Kong, W., Khadanga, S., Li, C., Gupta, S. K., Zhang, M., Xu, W., & Bendersky, M. (2022, August). Multi-aspect dense retrieval. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (pp. 3178-3186).
[5] Zhang, Q., Yang, Z., Huang, Y., Chen, Z., Cai, Z., Wang, K., ... & Gao, J. (2022). A Semantic Alignment System for Multilingual Query-Product Retrieval. arXiv preprint arXiv:2208.02958.
[6] Gu Y. et al. Speech intention classification with multimodal deep learning // Advances in Artificial Intelligence: 30th Canadian Conference on Artificial Intelligence, Canadian AI 2017, Edmonton, AB, Canada, May 16-19, 2017, Proceedings 30. – Springer International Publishing, 2017. – C. 260-271.
[7] Chen Q., Zhuo Z., Wang W. Bert for joint intent classification and slot filling // arXiv preprint arXiv:1902.10909. – 2019.
[8] Gangadharaiah R., Narayanaswamy B. Joint multiple intent detection and slot labeling for goal-oriented dialog // 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). – 2019. – C. 564-569.
[9] Joulin, A., Grave, É., Bojanowski, P., & Mikolov, T. (2017, April). Bag of Tricks for Efficient Text Classification. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers (pp. 427-431).
[10] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.
[11] Ma, Y., Cao, Y., Hong, Y., & Sun, A. (2023). Large language model is not a good few-shot information extractor, but a good reranker for hard samples!. arXiv preprint arXiv:2303.08559.
[12] Dereza O. V., Kayutenko D. A., Marakasova A. A., Fenogenova A. S.A Complex Approach to Spellchecking and Autocorrection for Russian // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2016”. — 2016. –C. 1-11.
[13] Näther M. An in-depth comparison of 14 spelling correction tools on a common benchmark //Proceedings of the 12th Language Resources and Evaluation Conference. – 2020. – C. 1849-1857.
[14] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.
[15] Bulatov, A., Kuratov, Y., & Burtsev, M. S. (2023). Scaling Transformer to 1M tokens and beyond with RMT. arXiv e-prints, arXiv-2304.
[16] Gage, P. (1994). A new algorithm for data compression. C Users Journal, 12(2), 23-38.
[17] Joulin, Armand, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. "Fasttext. zip: Compressing text classification models." arXiv preprint arXiv:1612.03651 (2016).
[18] Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532-1543. 2014.
[19] Reddy, C. K., Màrquez, L., Valero, F., Rao, N., Zaragoza, H., Bandyopadhyay, S., ... & Subbian, K. (2022). Shopping Queries Dataset: A Large-Scale ESCI Benchmark for Improving Product Search. arXiv preprint arXiv:2206.06588.
[20] Papenmeier, A., Kern, D., Hienert, D., Sliwa, A., Aker, A., & Fuhr, N. (2021, March). Dataset of Natural Language Queries for E-Commerce. In Proceedings of the 2021 Conference on Human Information Interaction and Retrieval (pp. 307-311).