НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
ноябрь–декабрь 2024 Том 24 № 6 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
November–December 2024 Vol. 24 No 6 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2024-24-6-1016-1023

Prompt-based multi-task learning for robust text retrieval

Sergei M. Masliukhin[1], Pavel A. Posokhov[2], Stepan S. Skrylnikov[3], Olesia V. Makhnytkina[4]✉, Tatiana Yu. Ivanovskaya[5]

[1,2,4,5] ITMO University, Saint Petersburg, 197101, Russian Federation
[1,2,3] OOO "STC Innovations", Saint Petersburg, 194044, Russian Federation

[1] smmasliukhin@itmo.ru, https://orcid.org/0000-0002-9054-5252
[2] paposokhov@itmo.ru, https://orcid.org/0000-0001-9442-8021
[3] skrylnikov@speechpro.com, https://orcid.org/0009-0001-7557-7870
[4] makhnytkina@itmo.ru✉, https://orcid.org/0000-0002-8992-9654
[5] taturiva@mail.ru, https://orcid.org/0009-0006-8551-5100

Abstract

The exponential growth of digital information necessitates the development of robust text retrieval methods since most of the methods are domain or task-specific which limits their implementation. In this case multi-task learning is a promising alternative as it helps a model to have more meaningful embeddings; however such cases require usage of task separation methods. Many studies explore multi-task learning to improve generalization but tend to focus on large models. However, in real-world, speech analytics tasks that require searching through hundreds of millions of vectors in real-time, smaller models become more appropriate. This paper presents a novel approach to enhance the robustness of multi-task text retrieval models through the use of prompts. We use contrastive learning to train encoder models both in single-task and multi-task configurations and compare their performances as well as analyze the efficiency of different prompt usage strategies including hard prompts represented by explicit natural language instructions and soft prompts of varying lengths represented by model special tokens. Experiments are conducted by applying prompts to both the query and candidate document as well as to queries only keeping the candidate without prompts to reuse pre-encoded candidates in multi-task retrieval without significant quality loss. The obtained results are compared using R@1, R@5, and MRR metrics which are most applicable for evaluating in-domain and out-of-domain search. Single-task models show better performance on in-domain training data, while multi-task models demonstrate superior performance on out-of-domain data highlighting their increased robustness to domain shifts. Applying prompts to both elements–query and document–yields better performance than applying them solely to the query. Soft prompts are found to be preferable to hard as they better adapt the model to different domains. The findings of this study can be useful for improving text retrieval models, especially in scenarios involving multi-task systems where high adaptability and performance on new data are required. Trainable prompts could be an effective tool for enhancing the flexibility of models in various applications, such as information retrieval and question-answering systems.

Keywords

contrastive learning, text retrieval, question answering, multi-task learning, fine-tuning, persona, data collection methodology, dialog data, conversational agents, personalization, question and answer generation

Acknowledgements

This research was supported by a grant from the Russian Science Foundation (22-11-00128, https://www.rscf.ru/project/22-11-00128/).

For citation: Masliukhin S.M., Posokhov P.A., Skrylnikov S.S., Makhnytkina O.V., Ivanovskaya T.Yu. Prompt-based multi-task learning for robust text retrieval. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2024, vol. 24, no. 6, pp. 1016–1023. doi: 10.17586/2226-1494-2024-24-6-1016-1023

УДК 004.89

Многозадачное обучение на основе префиксов для устойчивого текстового поиска

Сергей Михайлович Маслюхин[1], Павел Александрович Посохов[2], Степан Сергеевич Скрыльников[3], Олеся Владимировна Махныткина[4]✉, Татьяна Юрьевна Ивановская[5]

[1,2,4,5] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
[1,2,3] ООО «ЦРТ-инновации», Санкт-Петербург, 194044, Российская Федерация

[1] smmasliukhin@itmo.ru, https://orcid.org/0000-0002-9054-5252
[2] paposokhov@itmo.ru, https://orcid.org/0000-0001-9442-8021
[3] skrylnikov@speechpro.com, https://orcid.org/0009-0001-7557-7870
[4] makhnytkina@itmo.ru✉, https://orcid.org/0000-0002-8992-9654
[5] taturiva@mail.ru, https://orcid.org/0009-0006-8551-5100

Аннотация

Введение. Экспоненциальный рост цифровой информации требует устойчивых методов текстового поиска, поскольку большинство методов направлено на решение конкретной задачи или домена, что ограничивает их использование. Решением в таком случае могут являться многозадачные модели, требующие использования методов разделения задач. Многие исследования изучают многозадачное обучение для улучшения обобщения и фокусируются на больших моделях. Вместе с тем в реальных задачах речевой аналитики, требующих поиска среди сотен миллионов векторов в реальном времени, более подходящими становятся модели меньшего размера. **Метод.** В работе представлен новый подход к повышению устойчивости многозадачных моделей текстового поиска на основе префиксов. Применяется контрастное обучение как для многозадачных, так и однозадачных моделей-энкодеров. Выполнено сравнение моделей на устойчивость и проанализирована эффективность различных стратегий использования подсказок, включая жесткие, представленные явными инструкциями на естественном языке (инструктивные префиксы), и мягкие подсказки разной длины, представленные специальными токенами модели (обучаемые префиксы) разной длины. Эксперименты выполнены с применением подсказок как к запросу и кандидату, так и отдельно к запросам, для повторного использования предварительно закодированных кандидатов в многозадачном поиске без значительной потери качества. **Основные результаты.** Проведено сравнение полученных результатов по метрикам R@1, R@5 и MRR, являющимися наиболее применимыми для оценки поисковых моделей внутри и вне домена обучения. Однозадачные модели показали себя лучше при работе с данными в пределах домена обучения. Многозадачные модели продемонстрировали лучшую применимость на данных вне домена обучения, что подчеркивает их повышенную устойчивость к его смене. Для сохранения этого свойства в данной работе рассмотрено применение префиксов к обоим элементам — запросу и документу, что обеспечивает лучшую устойчивость, чем их обособленное применение к запросу. Обучаемые префиксы оказались более предпочтительными по сравнению с инструктивными, поскольку они лучше адаптируют модель к различным доменам. **Обсуждение.** Результаты исследования могут быть полезны для улучшения моделей текстового поиска, особенно в сценариях, связанных с многозадачными системами, где требуется высокая адаптивность и производительность на новых данных. Обучаемые префиксы могут быть эффективным инструментом повышения устойчивости моделей в различных приложениях, таких как информационный поиск и системы вопросов-ответов.

Ключевые слова

контрастное обучение, текстовый поиск, многозадачное обучение, персона, методология сбора данных, диалоговые данные, разговорные агенты, персонализация, генерация вопросов и ответов

Благодарности

Исследование выполнено за счет гранта Российского научного фонда (22-11-00128, https://www.rscf.ru/project/22-11-00128/).

Ссылка для цитирования: Маслюхин С.М., Посохов П.А., Скрыльников С.С., Махныткина О.В., Ивановская Т.Ю. Многозадачное обучение на основе префиксов для устойчивого текстового поиска // Научно-технический вестник информационных технологий, механики и оптики. 2024. Т. 24, № 6. С. 1016–1023 (на англ. яз.). doi: 10.17586/2226-1494-2024-24-6-1016-1023

Introduction

The exponential growth of digital information requires the development of sophisticated text retrieval methods to find relevant documents since traditional keyword-based methods often fail to handle the semantic complexity of queries [1]. A promising alternative is semantic search which uses embeddings to analyze the relationships between words and concepts. This approach enables the retrieval of relevant documents without relying on keywords. This is especially important for queries with synonyms or paraphrases. However, semantic search models have difficulty adapting to new domains, which limits their generalizability. Many studies explore multi-task learning to improve generalization but focus on large models [2], however, in real-world, problems such as speech analytics, which require searching through hundreds of millions of vectors in real time, smaller models are more suitable. Besides, according to the MTEB [3] and BEIR [4] benchmarks, at the time of our experiments, state-of-the-art models in the field are predominantly represented by smaller models rather than Large Language Models (LLMs)

like ChatGPT and LLaMA. In addition, GritLM analysis [5] revealed that smaller models demonstrate a more substantial increase in performance compared to LLMs when using pre-trained models. Therefore, our research focuses on developing effective training approaches for smaller models.

Hard and soft prompts attached, respectively, as instructional and learnable prefixes to tokenized queries and candidate documents increase the adaptability of the model to new domains. Previous studies on creating efficient embeddings for semantic search [6] have shown significant improvement over keyword-based methods. Modern approaches use contrastive, self-supervised learning, achieving state-of-the-art results by automatically generating text pairs [7]. One of the unsolved problems in semantic search is ensuring the robustness of the model when working outside the domain [8]. This limitation reduces the applicability of such models in real-world settings, as they have difficulty generalizing to unknown information. Recent advances in multi-task learning offer a promising approach to addressing the robustness problem, as it allows the model to be trained on multiple related search tasks at once, potentially improving its generalization ability [9]. Research [10] shows the effectiveness of multi-task learning for tasks requiring extensive knowledge with significant performance improvement on out-of-domain data. Another method for enhancing the performance of search models is the use of instructions that demonstrate improvements in specific tasks [11]. On the other hand, less explicit instructions have a subtler effect on the model and can be used to improve multi-task models [12]. Based on multi-task learning and prompt strategies, we develop domain-robust and generalizable text retrieval models that effectively handle new information, examine the effectiveness of their application to both queries and candidate documents, and compare performance of hard and soft prompts of different lengths.

Data and Method

To enhance the model ability to grasp semantic connections in document retrieval, we leverage multi-tasking. This involves training the model on a variety of tasks that explore different text pair relationships. We use six main and two supportive tasks in the model. Main ones being: paraphrase, Question-Answering (QA), title-body, summary, dialog and persona, whilst the supportive ones are represented by inverse cloze task and next sentence prediction. By encountering paraphrases, question-answer pairs, titles and corresponding articles (bodies), conversation turn-history pairs (dialogues), summaries and full texts (summary), utterance-fact pairs (persona identification) [13–15], fragment-context pairs (Inverse Cloze Task, ICT), and previous-next sentence pairs (Next Sentence Prediction, NSP), the model is exposed to the diverse ways text interacts and conveys meaning.

The training data is obtained from a combination of automatically collected sources and existing datasets with the total size of training and test data for each task represented in Table 1.

Table 1. Size of training and test data for each task

Task	Train Size	In-Domain Test Size	Out-of-Domain Test Size
Paraphrase	455,624	5,689	3,499
QA	8,675,864	27,400	7,240
Title-Body	32,942,558	16,224	4,730
Summary	200,742	4,001	7,780
Dialog	190,367	6,768	—
Persona	64,767	2,024	—

During training, batches are formed from the same task and the same source dataset at a time. It is important to note that negative samples are taken from the same batch to leverage the effectiveness of in-batch negatives. This comprehensive training strengthens the model ability to understand the flow of information and identify semantic connections within documents.

Given a query q and a corpus $C = \{c_1, c_2, ..., c_m\}$ containing m candidate documents, the objective is to identify the k most relevant documents (where $k << m$). In this research, we use the contrastive learning approach. Its goal is to learn such an embedding space where similar sample pairs stay close to each other while dissimilar ones are far apart. The contrastive learning framework leverages the InfoNCE loss function with in-batch negatives detailed in [16]:

$$\min L = -\log \frac{\varphi(q_p^+, c_p^+)}{\varphi(q_p^+, c_p^+) + \sum\limits_{n_i \in N} \varphi(q_p^+, n_i)},$$

where q_p^+ = prompt:{query} and c_p^+ = prompt:{candidate}; N represents the set of negative examples; $\varphi(q_p, c_p)$ denotes a function that calculates the matching score between the query (q_p) and the document (c_p). For this purpose, we employ a temperature-scaled cosine similarity function:

$$\varphi(q_p, c_p) = \tau \cos(h_{qp}^+, h_{cp}^+).$$

Here, τ is a hyperparameter (set to 20 in the experiments) and h_{qp}^+ is an embedding of the query and h_{cp}^+ is an embedding of the candidate.

For each relevant query-candidate pair (q^+, c^+), task-specific prompts are incorporated into both the query (q^+p^+) and the candidate (c^+p^+) to generate enhanced representations. To improve the efficiency of an encoder model, we implemented the separation of queries and candidates. This can be achieved either by using distinct encoders for queries and candidates or by applying different task prompts. Multi-task model enables efficient query encoding for a wide range of tasks. However, the use of task-specific prompts prevents reuse of the same encoded candidate base across different tasks, which can be inefficient since candidate base encoding is one of the most resource-intensive operations. Therefore, we propose using prompts only on query encoding. This approach allows us to maintain candidate index compatibility with minimal loss in performance.

To facilitate multi-tasking, we explore two prompting techniques: hard and soft prompts. Both types of prompts proposed in this paper are used to guide the model towards interpreting the text for the particular task and are separated by a colon (":") from the actual text.

Hard Prompts: Explicit instructions specific to each task [11] written in Russian language. Their **translated** version can be found in Fig. 1.

Soft Prompts: Inspired by the Parameter-Efficient Fine-Tuning training method [17], we introduce special tokens during training. These tokens act as cues for the model to identify the relevant task without explicit instructions. They are represented in square brackets and include corresponding task and marker of input type (query or candidate). For example, special token "[qa-q]" represents query for question answering task. It is also worth mentioning that the paraphrase task is not represented by soft prompts as it is the only symmetrical task in a sense that both the query and the candidate are interchangeable. In addition, it also has only one hard prompt for both the query and the candidate. The length of the soft prompt has a direct impact on the model specialization for a given task and the separation of the query and candidate. The text embedding is computed as the average of all token vectors, while the soft prompt embeddings also modify the representations of other tokens in the text. The larger the soft prompt, the more the text embedding changes. Therefore, we consider two soft prompt configurations in this work.

Short: single special token for both query and candidate (Fig. 2).

Long: four different tokens are used to define task and query/candidate (Fig. 3).

Experiments and Results

The experiments in this research were conducted on multi-task model based on a bi-encoder ranking architecture with the RuBERT-tiny2 model having 29.4M parameters as the backbone. Besides, we evaluate ranking multi-task models compared to baseline single-task models for every specific task on in-domain data. For evaluation, we employ ranking metrics such as R@1 (Recall@1), R@5 (Recall@5), and Mean Reciprocal Rank (MRR) which are commonly used in retrieval and ranking tasks.

R@1 (Recall@1) and R@5 (Recall@5) metrics measure the proportion of times the correct answer (or relevant document) is found among the top-ranked candidates. R@1 evaluates whether the correct answer is ranked at the top position being rated as the most relevant, while R@5 extends this to the top 5 positions. These metrics provide insight into the model ability to rank relevant results at the

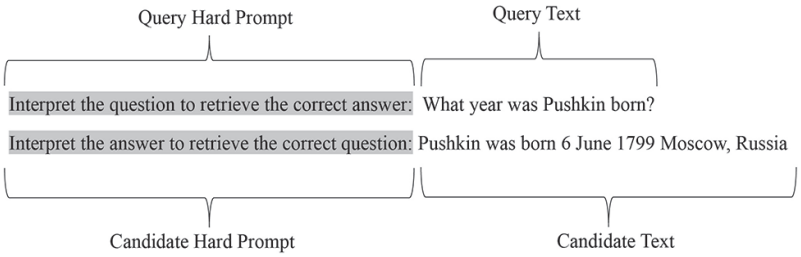

Fig. 1. Hard prompt example

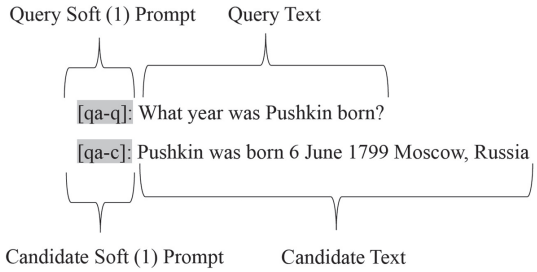

Fig. 2. Short soft prompt example

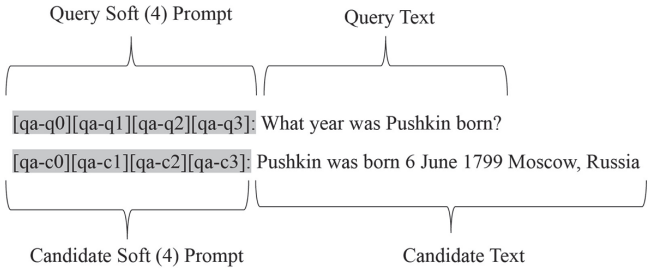

Fig. 3. Long soft prompt example

very top and within a broader set of retrieved candidates, offering a view on both the precision at the top rank and the scale of relevant candidates in the retrieval.

MRR computes the average reciprocal rank of the first relevant document across all queries. It considers the position of the first relevant result, offering a more nuanced evaluation by emphasizing early relevance. MRR is effective for understanding the ranking quality when there is a single correct answer or when early precision is critical.

These metrics are chosen because they effectively capture different aspects of ranking performance in information retrieval scenarios. R@1 and R@5 provide straightforward measures of retrieval success at different cutoffs which are essential for evaluating both precision at the top and the scale of relevance. MRR complements these metrics by focusing on the position of the first relevant result, offering a balanced evaluation that accounts for both precision and rank sensitivity.

By employing these ranking metrics (R@1, R@5, MRR), we ensure that our evaluation framework provides a comprehensive and informative assessment of the model performance across different scenarios, both in-domain and out-of-domain. This helps in selecting the optimal configuration for the multi-task model, ensuring it is both effective and robust. Furthermore, we estimate the robustness of the model by comparing retrieval performance on out-of-domain datasets not used in training. This helps in understanding how well the model generalizes to new, unseen data. In addition, we discuss the results that could be achieved with various prefix types for different task separation in multi-task model training and inference. We highlight the most effective prefix strategy we acquired which enhances the model ability to distinguish between tasks and improve retrieval accuracy. We evaluate the search quality of multi-task and single-task models on in-domain and out-of-domain datasets, the results are shown in Table 2.

The in-domain results shown in Table 2 demonstrate that multi-task models generally achieve quality comparable to single-task models. Multi-task models outperform single-task models in three tasks with smaller training datasets, suggesting that when task-specific data is limited, multi-task learning can effectively leverage shared representations and knowledge from other tasks. However, multi-task models are outperformed in three other tasks with larger training datasets, indicating the ability of single-task models to better utilize ample task-specific data to learn the nuances of that task.

However, multi-task model significantly outperforms the single-task models on out-of-domain data across all tasks. This indicates that multi-task learning significantly improves the robustness of models for all tasks used in training. The ability to learn more generalized patterns from diverse tasks enhances the model adaptability to new, unseen data, reducing the risk of over fitting to specific in-domain distributions.

The results demonstrate that multi-task learning provides significant advantages over single-task models, particularly in terms of generalization and robustness. While in-domain performance is comparable, with advantages varying based on dataset size, the multi-task model superior performance on out-of-domain data underscores its ability to generalize better to new contexts. This makes multi-task learning a more efficient and adaptable approach to develop versatile models capable of performing well across a range of tasks, reducing computational overhead and improving real-world applicability.

Table 2. Evaluation of multi-task and single-task models

Domain	Task	Model Type	R@1	R@5	MRR
In-Domain	Paraphrase	**Multi-task**	**0.769**	**0.930**	**0.902**
		Single-task	0.756	0.922	0.889
	QA	Multi-task	0.311	0.410	0.381
		Single-task	**0.327**	**0.425**	**0.402**
	Title-Body	Multi-task	0.423	0.535	0.590
		Single-task	**0.441**	**0.550**	**0.606**
	Dialog	Multi-task	0.027	0.108	0.074
		Single-task	**0.036**	**0.123**	**0.086**
	Summary	**Multi-task**	**0.515**	**0.697**	**0.599**
		Single-task	0.450	0.626	0.533
	Persona	**Multi-task**	**0.164**	**0.291**	**0.255**
		Single-task	0.103	0.198	0.171
Out-of-Domain	Paraphrase	**Multi-task**	**0.789**	**0.956**	**0.862**
		Single-task	0.756	0.941	0.834
	QA	**Multi-task**	**0.461**	**0.623**	**0.538**
		Single-task	0.408	0.552	0.479
	Title-Body	**Multi-task**	**0.094**	**0.151**	**0.125**
		Single-task	0.090	0.144	0.119
	Summary	**Multi-task**	**0.672**	**0.884**	**0.766**
		Single-task	0.365	0.588	0.469

Soft-prompts to hard-prompts comparison

In this work, we compare different prompting strategies for efficient task separation within multi-task training. Specifically, we consider hard prompts which are task-specific instructions in the Russian language for text interpreting. These prompts precede the meaningful part of the input sequence and are split by the ':' symbol. In comparison, we propose the use of soft-prompts deployed alongside full model fine-tuning. The use of soft-prompts involves a set of special tokens initialized for each specific task. We compare hard and soft prompts effectiveness within multi-task training; the results are shown in Table 3.

Our experiments determine that soft-prompts outperform hard-prompts. It could be attributed to the fact that instructions are less effectively interpreted by the small model used for vector encoding. Additionally, the relatively greater token length of instructions reduces the proportion of sequence informative content. Besides, we have identified that the optimal length of the soft-prompt prefix is one token. This length is sufficient for effective task separation in the model weight space.

Table 3. Prompt types effectiveness results on multi-task training

Task	Prefix Type	R@1	R@5	MRR
Paraphrase	Hard	0.768	0.928	0.899
	Soft (1)	**0.769**	**0.930**	**0.902**
	Soft (4)	0.768	0.925	0.898
QA	Hard	0.310	0.411	0.382
	Soft (1)	0.312	0.411	0.381
	Soft (4)	**0.312**	**0.412**	**0.384**
Title-Body	Hard	0.434	0.543	0.599
	Soft (1)	0.423	0.535	0.590
	Soft (4)	**0.435**	**0.545**	**0.603**
Dialog	Hard	0.026	0.108	0.072
	Soft (1)	**0.028**	**0.108**	**0.074**
	Soft (4)	**0.028**	**0.108**	**0.074**
Summary	Hard	0.482	0.657	0.566
	Soft (1)	**0.515**	**0.697**	**0.599**
	Soft (4)	0.485	0.670	0.573
Persona	Hard	0.279	0.533	0.446
	Soft (1)	**0.329**	**0.582**	**0.511**
	Soft (4)	0.300	0.563	0.478

Query and candidate encoding

We propose accompanying both queries and candidates from different tasks with corresponding prompts for their proper encoding in asymmetric tasks. However, this approach eliminates the compatibility of pre-encoded candidate bases between tasks. Therefore, we also propose second approach which involves using prefixes only for candidate encoding and allows us to maintain candidate index compatibility with minor quality loss. The results of different encoding settings are shown in Table 4.

The obtained results suggest that prefix-based separation of queries and candidates is sufficient for most of the tasks. Additionally, reducing the amount of training data for each individual encoder and doubling the total number of parameters when using separate encoders decreases the model performance. The lower quality of "Query Prefix" method compared to "Both Prefixes" indicates that in addition to query and candidate separation, prefix also contains information useful for more accurate candidate encoding.

Table 4. Prompt types effectiveness results on multi-task training

Task	Encoding Type	R@1	R@5	MRR
QA	**Both prefixes**	**0.312**	**0.411**	**0.381**
	Query Prefix	0.308	0.409	0.372
Title-Body	**Both prefixes**	**0.423**	**0.535**	**0.590**
	Query Prefix	0.411	0.530	0.583
Dialog	**Both prefixes**	**0.027**	**0.108**	**0.074**
	Query Prefix	0.024	0.094	0.065
Summary	**Both prefixes**	**0.515**	**0.0697**	**0.599**
	Query Prefix	0.507	0.0672	0.584
Persona	**Both prefixes**	**0.329**	**0.582**	**0.511**
	Query Prefix	0.310	0.565	0.498

Conclusion

Multi-task retrieval offers numerous advantages in practical applications. Therefore, obtaining a small-sized multi-task model with robustness comparable to single-task models can be a research goal in itself. However, this work demonstrates that we have not only achieved quality comparable to task-specific retrieval models but also improved the model quality on in-domain data and reduced the model size. Moreover, the proposed training method significantly enhances the model robustness to out-of-domain data. The key factor enabling us to achieve the desired results is the utilization of an optimal soft-prompting strategy. This strategy effectively separates the types of input sequences for multi-task model representation. Additionally, we propose an efficient candidate encoding strategy without prompts to reuse pre-encoded candidates for multi-task retrieval without significant quality loss.

Future work will focus on refining the soft-prompting strategy for better adaptability to specialized tasks and exploring its application in cross-lingual retrieval. We aim to optimize performance on larger, heterogeneous datasets while improving model efficiency. Additionally, enhancing explainability within the retrieval process will be a priority to ensure transparency and user trust. Lightweight model versions and techniques for reducing computational costs will also be explored for more practical deployment.

References

1. Hambarde K.A., Proença H. Information retrieval: recent advances and beyond. *IEEE Access*, 2023, vol. 11, pp. 76581–76604. https://doi.org/10.1109/access.2023.3295776
2. Zhang W., Xiong C., Stratos K., Overwijk A. Improving multitask retrieval by promoting task specialization. *Transactions of the Association for Computational Linguistics*, 2023, vol. 11, pp. 1201–1212. https://doi.org/10.1162/tacl_a_00597
3. Muennighoff N., Tazi N., Magne L., Reimers N. MTEB: Massive Text Embedding Benchmark. *Proc. of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2023, pp. 2014–2037. https://doi.org/10.18653/v1/2023.eacl-main.148
4. Thakur N., Reimers N., Rücklé A., Srivastava A., Gurevych I. BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models. *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021, pp. 105.
5. Muennighoff N., Su H., Wang L., Yang N., Wei F., Yu T., Singh A., Kiela D. Generative representational instruction tuning. *arXiv*, 2024, arXiv:2402.09906. https://doi.org/10.48550/arXiv.2402.09906
6. Reimers N., Gurevych I. Sentence-BERT: Sentence embeddings using siamese BERT-Networks. Sentence-BERT. *Proc. of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 3982–3992. https://doi.org/10.18653/v1/d19-1410
7. Wang L., Yang N., Huang X., Jiao B., Yang L., Jiang D., Majumder R., Wei F. Text embeddings by weakly-supervised contrastive pre-training. *arXiv*, 2024, arXiv:2212.03533. https://doi.org/10.48550/arXiv.2212.03533
8. Chen T., Zhang M., Lu J., Bendersky M., Najork M. Out-of-Domain semantics to the rescue! Zero-shot hybrid retrieval models. *Lecture Notes in Computer Science*, 2022, vol. 13185, pp. 95–110. https://doi.org/10.1007/978-3-030-99736-6_7
9. Ruder S. An overview of multi-task learning in deep neural networks. *arXiv*, 2017, arXiv:1706.05098. https://doi.org/10.48550/arXiv.1706.05098
10. Maillard J., Karpukhin V., Petroni F., Yih W., Oğuz B., Stoyanov V., Ghosh G. Multi-task retrieval for knowledge-intensive tasks. *Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Vol. 1*, 2021, pp. 1098–1111. https://doi.org/10.18653/v1/2021.acl-long.89
11. Su H., Shi W., Kasai J., Wang Y., Hu Y., Ostendorf M., Yih W., Smith N.A., Zettlemoyer L., Yu T. One embedder, any task: Instruction-finetuned text embeddings. *Findings of the Association for Computational Linguistics: ACL 2023*, 2023, pp. 1102–1121. https://doi.org/10.18653/v1/2023.findings-acl.71
12. Li X.L., Liang P. Prefix-tuning: Optimizing continuous prompts for generation. *Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Vol. 1*, 2021, pp. 4582–4597. https://doi.org/10.18653/v1/2021.acl-long.353
13. Matveev Y, Makhnytkina O, Posokhov P., Matveev A., Skrylnikov S. Personalizing hybrid-based dialogue agents. *Mathematics*, 2022, vol. 10, no. 24, pp. 4657. https://doi.org/10.3390/math10244657
14. Posokhov P., Apanasovich K., Matveeva A., Makhnytkina O., Matveev A. Personalizing dialogue agents for Russian: Retrieve and refine. *Proc. of the 31st Conference of Open Innovations Association (FRUCT)*, 2022, pp. 245–252. https://doi.org/10.23919/fruct54823.2022.9770895
15. Posokhov P., Matveeva A., Makhnytkina O., Matveev A., Matveev Y. Personalizing retrieval-based dialogue agents. *Lecture Notes in Computer Science*, 2022, vol. 13721, pp. 554–566. https://doi.org/10.1007/978-3-031-20980-2_47
16. Wang L., Yang N., Huang X., Yang L., Majumder R., Wei F. Improving text embeddings with large language models. *Proc. of the 62nd Annual Meeting of the Association for Computational Linguistics. Vol. 1*, 2024, pp. 11897–11916. https://doi.org/10.18653/v1/2024.acl-long.642
17. Xu L., Xie H., Qin S.-Z.J., Tao X., Wang F.L. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. *arXiv*, 2023, arXiv:2312.12148. https://doi.org/10.48550/arXiv.2312.12148

Литература

1. Hambarde K.A., Proença H. Information retrieval: recent advances and beyond // IEEE Access. 2023. V. 11. P. 76581–76604. https://doi.org/10.1109/access.2023.3295776
2. Zhang W., Xiong C., Stratos K., Overwijk A. Improving multitask retrieval by promoting task specialization // Transactions of the Association for Computational Linguistics. 2023. V. 11. P. 1201–1212. https://doi.org/10.1162/tacl_a_00597
3. Muennighoff N., Tazi N., Magne L., Reimers N. MTEB: Massive Text Embedding Benchmark // Proc. of the 17th Conference of the European Chapter of the Association for Computational Linguistics. 2023. P. 2014–2037. https://doi.org/10.18653/v1/2023.eacl-main.148
4. Thakur N., Reimers N., Rücklé A., Srivastava A., Gurevych I. BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models // Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2). 2021. P. 105.
5. Muennighoff N., Su H., Wang L., Yang N., Wei F., Yu T., Singh A., Kiela D. Generative representational instruction tuning // arXiv. 2024. arXiv:2402.09906. https://doi.org/10.48550/arXiv.2402.09906
6. Reimers N., Gurevych I. Sentence-BERT: Sentence embeddings using siamese BERT-Networks. Sentence-BERT // Proc. of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019. P. 3982–3992. https://doi.org/10.18653/v1/d19-1410
7. Wang L., Yang N., Huang X., Jiao B., Yang L., Jiang D., Majumder R., Wei F. Text embeddings by weakly-supervised contrastive pre-training // arXiv. 2024. arXiv:2212.03533. https://doi.org/10.48550/arXiv.2212.03533
8. Chen T., Zhang M., Lu J., Bendersky M., Najork M. Out-of-Domain semantics to the rescue! Zero-shot hybrid retrieval models // Lecture Notes in Computer Science. 2022. V. 13185. P. 95–110. https://doi.org/10.1007/978-3-030-99736-6_7
9. Ruder S. An overview of multi-task learning in deep neural networks // arXiv. 2017. arXiv:1706.05098. https://doi.org/10.48550/arXiv.1706.05098
10. Maillard J., Karpukhin V., Petroni F., Yih W., Oğuz B., Stoyanov V., Ghosh G. Multi-task retrieval for knowledge-intensive tasks // Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Vol. 1. 2021. P. 1098–1111. https://doi.org/10.18653/v1/2021.acl-long.89
11. Su H., Shi W., Kasai J., Wang Y., Hu Y., Ostendorf M., Yih W., Smith N.A., Zettlemoyer L., Yu T. One embedder, any task: Instruction-finetuned text embeddings // Findings of the Association for Computational Linguistics: ACL 2023. 2023. P. 1102–1121. https://doi.org/10.18653/v1/2023.findings-acl.71
12. Li X.L., Liang P. Prefix-tuning: Optimizing continuous prompts for generation // Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Vol. 1. 2021. P. 4582–4597. https://doi.org/10.18653/v1/2021.acl-long.353
13. Matveev Y., Makhnytkina O., Posokhov P., Matveev A., Skrylnikov S. Personalizing hybrid-based dialogue agents // Mathematics. 2022. V. 10. N 24. P. 4657. https://doi.org/10.3390/math10244657
14. Posokhov P., Apanasovich K., Matveeva A., Makhnytkina O., Matveev A. Personalizing dialogue agents for Russian: Retrieve and refine // Proc. of the 31st Conference of Open Innovations Association (FRUCT). 2022. P. 245–252. https://doi.org/10.23919/fruct54823.2022.9770895
15. Posokhov P., Matveeva A., Makhnytkina O., Matveev A., Matveev Y. Personalizing retrieval-based dialogue agents // Lecture Notes in Computer Science. 2022. V. 13721. P. 554–566. https://doi.org/10.1007/978-3-031-20980-2_47
16. Wang L., Yang N., Huang X., Yang L., Majumder R., Wei F. Improving text embeddings with large language models // Proc. of the 62nd Annual Meeting of the Association for Computational Linguistics. Vol. 1. 2024. P. 11897–11916. https://doi.org/10.18653/v1/2024.acl-long.642
17. Xu L., Xie H., Qin S.-Z.J., Tao X., Wang F.L. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment // arXiv. 2023. arXiv:2312.12148. https://doi.org/10.48550/arXiv.2312.12148

Authors

Sergei M. Masliukhin — Engineer, ITMO University, Saint Petersburg, 197101, Russian Federation; Leading Researcher, OOO "STC Innovations", Saint Petersburg, 194044, Russian Federation, https://orcid.org/0000-0002-9054-5252, smmasliukhin@itmo.ru

Pavel A. Posokhov — PhD Student, Software Developer, ITMO University, Saint Petersburg, 197101, Russian Federation; Scientific Researcher, OOO "STC Innovations", Saint Petersburg, 194044, Russian Federation, SC 57699334300, https://orcid.org/0000-0001-9442-8021, paposokhov@itmo.ru

Stepan S. Skrylnikov — Student, Junior Researcher, OOO "STC Innovations", Saint Petersburg, 194044, Russian Federation, SC 58029830000, https://orcid.org/0009-0001-7557-7870, skrylnikov@speechpro.com

Olesia V. Makhnytkina — PhD, Associate Professor, Associate Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 57208002090, https://orcid.org/0000-0002-8992-9654, makhnytkina@itmo.ru

Tatiana Yu. Ivanovskaya — Lecturer, ITMO University, Saint Petersburg, 197101, Russian Federation, https://orcid.org/0009-0006-8551-5100, taturiva@mail.ru

Авторы

Маслюхин Сергей Михайлович — инженер, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация; ведущий научный сотрудник, ООО «ЦРТ-инновации», Санкт-Петербург, 194044, Российская Федерация, https://orcid.org/0000-0002-9054-5252, smmasliukhin@itmo.ru

Посохов Павел Александрович — аспирант, программист, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация; научный сотрудник, ООО «ЦРТ-инновации», Санкт-Петербург, 194044, Российская Федерация, SC 57699334300, https://orcid.org/0000-0001-9442-8021, paposokhov@itmo.ru

Скрыльников Степан Сергеевич — магистр, младший научный сотрудник, ООО «ЦРТ-инновации», Санкт-Петербург, 194044, Российская Федерация, SC 58029830000, https://orcid.org/0009-0001-7557-7870, skrylnikov@speechpro.com

Махныткина Олеся Владимировна — кандидат технических наук, доцент, доцент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 57208002090, https://orcid.org/0000-0002-8992-9654, makhnytkina@itmo.ru

Ивановская Татьяна Юрьевна — преподаватель, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, https://orcid.org/0009-0006-8551-5100, taturiva@mail.ru

Received 07.10.2024
Approved after reviewing 24.10.2024
Accepted 24.11.2024

Статья поступила в редакцию 07.10.2024
Одобрена после рецензирования 24.10.2024
Принята к печати 24.11.2024