Typo-Robust Sentence Representation Learning for Dense Retrieval

Anonymous ACL submission

Abstract

Dense retrieval is a basic building block of information retrieval applications. One of the main challenges of dense retrieval in real-world settings is the handling of queries containing misspelled words. A popular approach to handling misspelled queries is minimizing the representations discrepancy between misspelled queries and their pristine ones. Unlike the existing approaches which only focus on the alignment between misspelled and pristine queries, our method also improves the contrast between each misspelled query and its surrounding queries. To assess the effectiveness of our proposed method, we compare it against the existing competitors using two benchmark datasets and two base encoders. Our method outperforms the competitors in all cases with misspelled queries.

1 Introduction

Dense retrieval is a fundamental component in many information retrieval applications, such as open-domain question answering and ad-hoc retrieval. The objective is to score and rank a large collection of candidate passages based on their similarity to a given query. The performance of dense retrieval relies on sentence representation learning. A popular approach is to finetune a pre-trained language model to create an embedding space that puts each query closer to its corresponding passages (Zhan et al., 2020; Karpukhin et al., 2020; Khattab and Zaharia, 2020; Xiong et al., 2021; Qu et al., 2021; Ren et al., 2021a,b).

One of the major challenges of dense retrieval is the handling of misspelled queries which induces representations of the misspelled queries to be closer to irrelevant passages than their corresponding passages. Several studies have demonstrated that misspellings in search queries can substantially degrade retrieval performance (Zhuang and Zuccon, 2021; Penha et al., 2022), specifically when informative terms, such as entity mentions, are misspelled (Sidiropoulos and Kanoulas, 2022).

To create a retrieval model that is capable of handling misspelled queries, researchers have proposed different training methods to align representations of misspelled queries with their pristine ones. Zhuang and Zuccon (2021, 2022) devise augmentation methods to generate misspelled queries and propose training methods, Typos-aware Training and Self-Teaching (ST), to encourage consistency between outputs of misspelled queries and their non-misspelled counterparts. Alternatively, Sidiropoulos and Kanoulas (2022) apply contrastive loss to enforce representations of misspelled queries to be closer to their corresponding non-misspelled queries. Although these methods can improve the performance of retrieval models for misspelled queries, there is still a substantial performance drop for misspelled queries.

In this paper, we propose a training method to improve dense retrieval for handling misspelled queries based on the following desired properties:

- ***Alignment***: the method should be able to align queries with their corresponding passages.
- ***Robustness***: the method should be able to align misspelled queries with their pristine queries.
- ***Contrast***: the method should be able to separate queries that refer to different passages and passages that correspond to different queries.

In contrast to the existing methods that only satisfy the *Alignment* and *Robustness* properties, our method also aims to satisfy the *Contrast* property. Increasing the distance between dissimilar queries should help distinguish misspelled queries from other distinct queries. We design the following components for our training method: (i) Dual Self-Teaching (DST) incorporates the ideas of Dual Learning (Xia et al., 2017; Li et al., 2021) and Self-Teaching (Zhuang and Zuccon, 2022) to train robust dense retrieval in a bidirectional manner: passage retrieval and query retrieval. (ii) Query

Expansion generates a large number of misspelling examples in a variety of misspelling variations for each query to supply our training objective.

Experimental studies were conducted to assess the efficiency of the proposed method in comparison to existing approaches. We conduct experiments based on two different pre-trained language models. We evaluate using two passage retrieval benchmark datasets, a standard one and a specialized one for misspellings robustness evaluation. For each dataset, we measure performance on both misspelled and non-misspelled queries, where the misspelled queries are both generated and real queries. The experimental results show that the proposed method outperforms the best existing methods for enhancing the robustness of dense retrieval against misspellings without sacrificing performance for non-misspelled queries.

We summarize our contributions as follow:

- We propose a novel training method to enhance the robustness of dense retrieval against misspellings by incorporating three desired properties: *Alignment*, *Robustness*, and *Contrast*.
- We introduce Dual Self-Teaching (DST) which adopts the idea of Dual Learning and Self-Teaching to learn robust sentence representations. In addition, we propose Query Expansion to generate multiple views of a particular query under different misspelling scenarios.
- We evaluate our method on misspelled and non-misspelled queries from two passage retrieval datasets. The results show that our method outperforms the previous state-of-the-art methods by a significant margin on misspelled queries.

2 Methodology

We propose a training pipeline to enhance the dense retrieval capability for handling spelling variations and mistakes in queries. As shown in Figure 1, the training pipeline comprises three steps. (i) *Query Expansion*: we augment each query in the training set into multiple misspelled queries using the typo generators provided by Zhuang and Zuccon (2021). (ii) *Similarity Score Calculation*: we compute similarity score distributions using in-batch negative queries and passages, with additional hard negative passages. (iii) *Dual Self-Teaching Loss Calculation*: we compute the DST loss using the similarity score distributions.

2.1 Query Expansion

The purpose of this step is to guide the learning with a broad array of possible misspelling patterns. Let $\mathbf{Q}$ denote a set $\{q_1, q_2, ..., q_N\}$ of N queries. From all queries in $\mathbf{Q}$, we generate a set of $K \times N$ misspelled queries $\boldsymbol{\mathcal{Q}}' = \{\langle q'_{1,k}, q'_{2,k}, ..., q'_{N,k}\rangle\}_{k=1}^{K}$, where K is the misspelling variations. We use five typo generators proposed by Zhuang and Zuccon (2021), including: RandInsert, RandDelete, RandSub, SwapNeighbor, and SwapAdjacent. Please refer to Appendix A.3 for examples of the misspelled queries.

2.2 Similarity Score Calculation

The goal of this step is to compute similarity score distributions between queries and passages for passage retrieval and query retrieval tasks.

Let $S(\cdot, \cdot)$ denote the score distribution function:

$$S(a, \mathbf{B}) = \left\{ b_i \in B \;\middle|\; \frac{exp(a \cdot b_i)}{\sum_{b_j \in \mathbf{B}} exp(a \cdot b_j)} \right\} \quad (1)$$

where $\mathbf{P} = \{p_1, p_2, ..., p_M\}$ is a set of M passages and $\mathbf{Q}'_k = \{q'_{1,k}, q'_{2,k}, ..., q'_{N,k}\}$ is the k^{th} set of misspelled queries in $\boldsymbol{\mathcal{Q}}'$. We compute two groups of score distributions as follow:

- Passage retrieval: we calculate score distributions in a query-to-passages direction for each original query $s_p = S(q_n, \mathbf{P})$ and misspelled query $s'^{k}_{p} = S(q'_{n,k}, \mathbf{P})$.
- Query retrieval: we calculate score distributions in a passage-to-queries direction for original queries $s_q = S(p_m, \mathbf{Q})$ and each set of misspelled queries $s'^{k}_{q} = S(p_m, \mathbf{Q}'_k)$.

In this way, we produce four different score distributions (s_p, s'^{k}_{p}, s_q, s'^{k}_{q}) for our training objective.

2.3 Dual Self-Teaching Loss Calculation

We design the *Dual Self-Teaching loss* ($\mathcal{L}_{\text{DST}}$) to capture the three desired properties: *Alignment*, *Robustness*, and *Contrast*.

$$\mathcal{L}_{\text{DST}} = \underbrace{(1-\beta)\mathcal{L}_{\text{DCE}}}_{\text{Dual Cross-Entropy}} + \underbrace{\beta\mathcal{L}_{\text{DKL}}}_{\text{Dual KL-Divergence}} \quad (2)$$

Dual Cross-Entropy loss ($\mathcal{L}_{\text{DCE}}$) satisfies the *Alignment* and *Contrast* properties by utilizing cross-entropy losses to learn score distributions of the original queries for passage retrieval (s_p) and query retrieval (s_q) given labels y_p and y_q.

$$\mathcal{L}_{\text{DCE}} = \underbrace{(1-\gamma)\mathcal{L}^{(P)}_{\text{CE}}(s_p, y_p)}_{\text{Passage Retrieval}} + \underbrace{\gamma\mathcal{L}^{(Q)}_{\text{CE}}(s_q, y_q)}_{\text{Query Retrieval}} \quad (3)$$

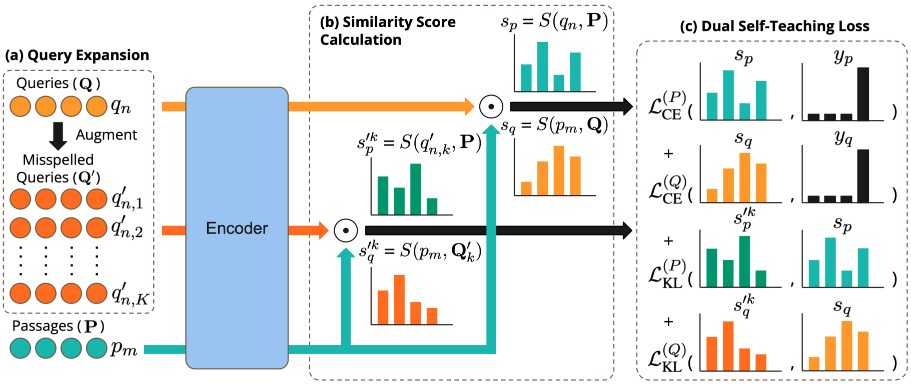

(a) Query Expansion
Queries ($\mathbf{Q}$)
q_n
Augment
Misspelled Queries ($\mathbf{Q}'$)
$q'_{n,1}$
$q'_{n,2}$
$q'_{n,K}$
Passages ($\mathbf{P}$)
p_m
Encoder
(b) Similarity Score Calculation
$s_p = S(q_n, \mathbf{P})$
$s_q = S(p_m, \mathbf{Q})$
$s'^k_p = S(q'_{n,k}, \mathbf{P})$
$s'^k_q = S(p_m, \mathbf{Q}'_k)$
(c) Dual Self-Teaching Loss
$\mathcal{L}^{(P)}_{\text{CE}}($ s_p , y_p $)$
$+$ $\mathcal{L}^{(Q)}_{\text{CE}}($ s_q , y_q $)$
$+$ $\mathcal{L}^{(P)}_{\text{KL}}($ s'^k_p , s_p $)$
$+$ $\mathcal{L}^{(Q)}_{\text{KL}}($ s'^k_q , s_q $)$

Figure 1: The proposed training pipeline consists of three steps: (a) Query Expansion, (b) Similarity Score Calculation, and (c) Dual Self-Teaching Loss Calculation.

Minimizing the $\mathcal{L}^{(P)}_{\text{CE}}$ term will increase the similarity scores between queries and their relevant passages to be higher than other irrelevant passages by separating the relevant and irrelevant passages from one another. Minimizing the $\mathcal{L}^{(Q)}_{\text{CE}}$ term will increase the similarity scores between passages and their relevant queries to be higher than other irrelevant queries by separating the relevant and irrelevant queries from one another. In this manner, minimizing one of the two terms will align queries with their corresponding passages, satisfying the *Alignment* property. Moreover, minimizing both terms will separate queries that refer to different passages and passages that belong to different queries, satisfying the *Contrast* property.

Dual KL-Divergence loss ($\mathcal{L}_{\text{DKL}}$) aims to fulfill the *Robustness* property by using KL losses to match score distributions of misspelled queries $\{s'^1_p, s'^2_p, ..., s'^K_p\}$ and $\{s'^1_q, s'^2_q, ..., s'^K_q\}$ to the score distributions of the original query s_p and s_q.

$$\mathcal{L}_{\text{DKL}} = \frac{1}{K}\sum_{k=1}^{K} \underbrace{(1-\sigma)\mathcal{L}^{(P)}_{\text{KL}}(s'^k_p, s_p)}_{\text{Passage Retrieval Consistency}} + \underbrace{\sigma\mathcal{L}^{(Q)}_{\text{KL}}(s'^k_q, s_q)}_{\text{Query Retrieval Consistency}} \tag{4}$$

Minimizing $\mathcal{L}^{(P)}_{\text{KL}}$ and $\mathcal{L}^{(Q)}_{\text{KL}}$ will reduce the discrepancy between misspelled and non-misspelled queries for both query-to-passages and passage-to-queries score distributions. This way, we implicitly align representations of the misspelled queries to the original queries, satisfying the *Robustness* property. To stabilize training, we apply stop-gradient to the score distributions of the original queries (s_p and s_q) in the $\mathcal{L}_{\text{DKL}}$. The β, γ, and σ are the balancing coefficients selected by hyper-parameter tuning on a development set. With this loss combination, we achieve all three desired properties.

3 Experimental Settings

3.1 Training Details

We experiment on two pre-trained language models, BERT (Devlin et al., 2019) and Character-BERT (El Boukkouri et al., 2020). We train both models only on the training set of MS MARCO dataset (Nguyen et al., 2016). Moreover, the training data, provided by the Tevatron toolkit (Gao et al., 2022), also contains hard negative passages. We include the training set details and hyper-parameter settings in Appendix A.1.

3.2 Competitive Methods

To show the effectiveness of our method, we compare our work with the following baseline and competitive training methods.

- *DPR* (Karpukhin et al., 2020) is a baseline training method which trains dense retrieval merely on non-misspelled queries using $\mathcal{L}^{(P)}_{\text{CE}}$ loss.
- *DPR+Aug* (Zhuang and Zuccon, 2021) is the Typos-aware Training method which trains dense retrieval on both misspelled and non-misspelled queries using $\mathcal{L}^{(P)}_{\text{CE}}$ loss.
- *DPR+Aug+CL* (Sidiropoulos and Kanoulas, 2022) employs additional contrastive loss to train the misspelled queries.
- *DPR+ST* (Zhuang and Zuccon, 2022) is the Self-Teaching method which trains dense retrieval on both misspelled and non-misspelled queries using $\mathcal{L}^{(P)}_{\text{CE}}$ and $\mathcal{L}^{(P)}_{\text{KL}}$ losses.

Note that, their query augmentation method is identical to the Query Expansion with $K = 1$. We retrain all models using the same setting described in the previous section.

	BERT-based					CharacterBERT-based				
	MS MARCO		DL-typo			MS MARCO		DL-typo		
Methods	MRR@10	R@1000	nDCG@10	MRR	MAP	MRR@10	R@1000	nDCG@10	MRR	MAP
DPR	.143 (.331)	.696 (**.954**)	.276 (**.682**)	.431 (**.873**)	.175 (.563)	.162 (.321)	.726 (.945)	.268 (.643)	.376 (.832)	.212 (.503)
+ Aug	.227 (.334)	.857 (.950)	.398 (**.682**)	.530 (.806)	.286 (.565)	.258 (.326)	.883 (.946)	.414 (.631)	.578 (.783)	.318 (.512)
+ Aug + CL	.234 (.335)	.867 (.951)	.387 (.668)	.536 (.864)	.267 (.544)	.263 (.330)	.894 (.947)	.466 (**.677**)	.635 (.819)	.360 (**.544**)
+ ST	.237 (.333)	.874 (.950)	.392 (.677)	.525 (.852)	.283 (.557)	.274 (**.332**)	.900 (.947)	.469 (.650)	.619 (.810)	.359 (.517)
+ DST (our)	**.260**†(**.336**)	**.894**†(**.954**)	**.432** (.673)	**.558** (.833)	**.343**†(**.568**)	**.288**†(**.332**)	**.918**†(**.949**)	**.529**†(.673)	**.742**†(**.854**)	**.403** (.537)

Table 1: Results of different training methods on misspelled and non-misspelled queries. We report the results in the format of "`misspelled query performance (non-misspelled query performance)`". We emphasize the best score with bold text and the second-best score with underlined text. We use † to denote DST results that significantly outperform the second-best result ($p < 0.05$).

3.3 Dataset and Evaluation

Datasets. We evaluate the effectiveness of DST on two passage retrieval datasets, MS MARCO and DL-typo (Zhuang and Zuccon, 2022), each with misspelled and non-misspelled queries. There are 8.8 million candidate passages for both datasets. The development set of MS MARCO contains 6,980 non-misspelled queries. To obtain misspelled queries, we use the typos generator method proposed by Zhuang and Zuccon (2021) to generate 10 misspelled variations for each original query. The DL-typo provides 60 real misspelled queries and 60 corresponding non-misspelled queries that are corrected manually.

Evaluation. We use the standard metrics originally used by each dataset's creators. For MS MARCO, each misspelled query performance is the average of 10 measurements. We employ Ranx evaluation library (Bassani, 2022) to measure performance and statistical significance. Specifically, we use a two-tailed paired t-test with Bonferroni correction to measure the statistical significance ($p < 0.05$).

4 Experimental Results

4.1 Main Results

As shown in Table 1, the results indicate that DST outperforms competitive methods for misspelled queries in every cases without sacrificing performance for non-misspelled queries in eight out of ten cases. We observe some performance trade-offs for the BERT-based model in non-misspelling scores (nDCG@10 and MRR) of the DL-typo dataset. Aside from that, there is no performance trade-off for the CharacterBERT-based model. These outcomes conform with the observation in Figure 2 (Appendix A.4) that DST improves the *Robustness* and *Contrast* of misspelled queries.

4.2 Loss Ablation Study

In this experiment, we study the benefit of each term in DST by training dense retrieval models on variant loss combinations with $K = 40$.

$\mathcal{L}_{\text{CE}}^{(P)}$	$\mathcal{L}_{\text{CE}}^{(Q)}$	$\mathcal{L}_{\text{KL}}^{(P)}$	$\mathcal{L}_{\text{KL}}^{(Q)}$	MRR@10
✓	✓	✓	✓	**.260** (.336)
✓	✓	✓		.257 (.335)
✓	✓		✓	.228 (.326)
✓		✓	✓	.251 (**.337**)
	✓	✓	✓	.087 (.114)
✓		✓		.249 (.336)
	✓		✓	.120 (.158)

Table 2: Loss ablation study results on MS MARCO.

The results in Table 2 reveal that robustness terms ($\mathcal{L}_{\text{KL}}^{(P)}$ and $\mathcal{L}_{\text{KL}}^{(Q)}$) positively contribute to the performance of misspelled and non-misspelled queries, with the $\mathcal{L}_{\text{KL}}^{(P)}$ being more important. The passage retrieval loss ($\mathcal{L}_{\text{CE}}^{(P)}$) is very important for retrieval performance, whereas the query retrieval loss ($\mathcal{L}_{\text{CE}}^{(Q)}$) improves performance for misspelled queries. Disabling query retrieval terms ($\mathcal{L}_{\text{CE}}^{(Q)}$ and $\mathcal{L}_{\text{KL}}^{(Q)}$) greatly reduces performances for misspelled queries. The passage retrieval terms ($\mathcal{L}_{\text{CE}}^{(P)}$ and $\mathcal{L}_{\text{KL}}^{(P)}$) remain the most essential and irreplaceable by the query retrieval terms.

5 Conclusion

This paper aims to address the misspelling problem in dense retrieval. We formulate three desired properties for making dense retrieval robust to misspellings: *Alignment*, *Robustness*, and *Contrast*. Unlike previous methods, which only focus on the *Alignment* and *Robustness* properties, Our method considers all the desired properties. The empirical results show that our method performs best against misspelled queries, revealing the importance of the *Contrast* property.

References

Elias Bassani. 2022. ranx: A blazing-fast python library for ranking evaluation and comparison. In *ECIR (2)*, volume 13186 of *Lecture Notes in Computer Science*, pages 259–264. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Hicham El Boukkouri, Olivier Ferret, Thomas Lavergne, Hiroshi Noji, Pierre Zweigenbaum, and Jun'ichi Tsujii. 2020. CharacterBERT: Reconciling ELMo and BERT for word-level open-vocabulary representations from characters. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6903–6915, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Luyu Gao, Xueguang Ma, Jimmy J. Lin, and Jamie Callan. 2022. Tevatron: An efficient and flexible toolkit for dense retrieval. *ArXiv*, abs/2203.05765.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 39–48, New York, NY, USA. Association for Computing Machinery.

Yizhi Li, Zhenghao Liu, Chenyan Xiong, and Zhiyuan Liu. 2021. More robust dense retrieval with contrastive dual learning. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '21, page 287–296, New York, NY, USA. Association for Computing Machinery.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016*, volume 1773 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Gustavo Penha, Arthur Câmara, and Claudia Hauff. 2022. Evaluating the robustness of retrieval pipelines with query variation generators. In *Advances in Information Retrieval*, pages 397–412, Cham. Springer International Publishing.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2021. RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5835–5847, Online. Association for Computational Linguistics.

Ruiyang Ren, Shangwen Lv, Yingqi Qu, Jing Liu, Wayne Xin Zhao, QiaoQiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2021a. PAIR: Leveraging passage-centric similarity relation for improving dense passage retrieval. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2173–2183, Online. Association for Computational Linguistics.

Ruiyang Ren, Yingqi Qu, Jing Liu, Wayne Xin Zhao, QiaoQiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2021b. RocketQAv2: A joint training method for dense passage retrieval and passage re-ranking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2825–2835, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Georgios Sidiropoulos and Evangelos Kanoulas. 2022. Analysing the robustness of dual encoders for dense retrieval against misspellings. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yingce Xia, Tao Qin, Wei Chen, Jiang Bian, Nenghai Yu, and Tie-Yan Liu. 2017. Dual supervised learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3789–3798. PMLR.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and

Arnold Overwijk. 2021. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations*.

Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2020. Repbert: Contextualized text embeddings for first-stage retrieval. *CoRR*, abs/2006.15498.

Shengyao Zhuang and Guido Zuccon. 2021. Dealing with typos for BERT-based passage retrieval and ranking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2836–2842, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shengyao Zhuang and Guido Zuccon. 2022. Characterbert and self-teaching for improving the robustness of dense retrievers on queries with typos. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 1444–1454, New York, NY, USA. Association for Computing Machinery.

A Appendix

A.1 Training Setup and Hyperparameters

The MS MARCO is a large scale English language dataset for machine reading comprehension (MRC). The dataset consists of anonymized queries sampled from Bing's search query logs, each with human generated answers. The training set we used contains 400,782 training samples, each of which consists of a query, positive passage, and a set of hard negative passages, which we randomly select 7 hard negative passages for each training sample. We set a batch size to 16 and use in-batch negative sampling for each training sample, therefore, we obtain 7 + 8 * 15 = 127 negative passages for each training sample. We use the AdamW optimizer and learning rate of $1\mathrm{e}{-5}$ for 150,000 steps with a linear learning rate warm-up over the first 10,000 steps, and a linear learning rate decay over the rest of the training steps. For our training method, we set the hyper-parameters $\beta = 0.5$, $\gamma = 0.5$, $\sigma = 0.2$, and the query expansion size $K = 40$. Using a Tesla V100 32G GPU, the BERT-based model training time is around 31 hours, while the CharacterBERT-based model training time is roughly 56 hours.

A.2 Query Expansion Size Study

To study the benefit of query expansion and find the optimal expansion size, we measure performance of dense retrieval models trained with DST using the query expansion size K of 1, 10, 20, 40, 60. Note that, the query augmentation method used in previous works is a special case of the query expansion when $K = 1$. We report the results using MRR@10 for the development set of MS MARCO dataset. We also report training time to show trade-offs between performance and computation.

Queries	K				
	1	10	20	40	60
Original	.334	.334	.335	**.336**	.332
Misspelled	.251	.258	**.260**	**.260**	**.260**
Training time (hr)	18	20	23	31	39

Table 3: Results of query expansion size study. We train all models in this experiment on Tesla V100 32G GPU.

As shown in Table 3, the results indicate that increasing K improves the performance of both misspelled and non-misspelled queries, but only up to a certain point, after which the performance begins to decline. We observe that setting $K = 40$ produces the best results, and there is no further performance improvement after this point. In addition, the $K = 1$ result demonstrates the performance of our method when utilizing the same query augmentation method as the previous methods.

A.3 Query Expansion Examples

Table 4 provides examples of misspelled queries generated by the Query Expansion for each original query.

Original query:

what is the goddess of agriculture in greek mythology

Misspelled queries:

what is the goddoess of agriculture in greek mythology

what is the goddess of agriulture in greek mythology

what is the goddess of agriculture in greek mythologo

what is the goddses of agriculture in greek mythology

what is the goddess of agriculture in greek myhhology

what is the goddess of agriculture in greeck mythology

what is the goddess of agriculture in greek myhology

what is the goddess of agriculture in grvek mythology

what is the goddess of agricultrue in greek mythology

what is the goddess of ahriculture in greek mythology

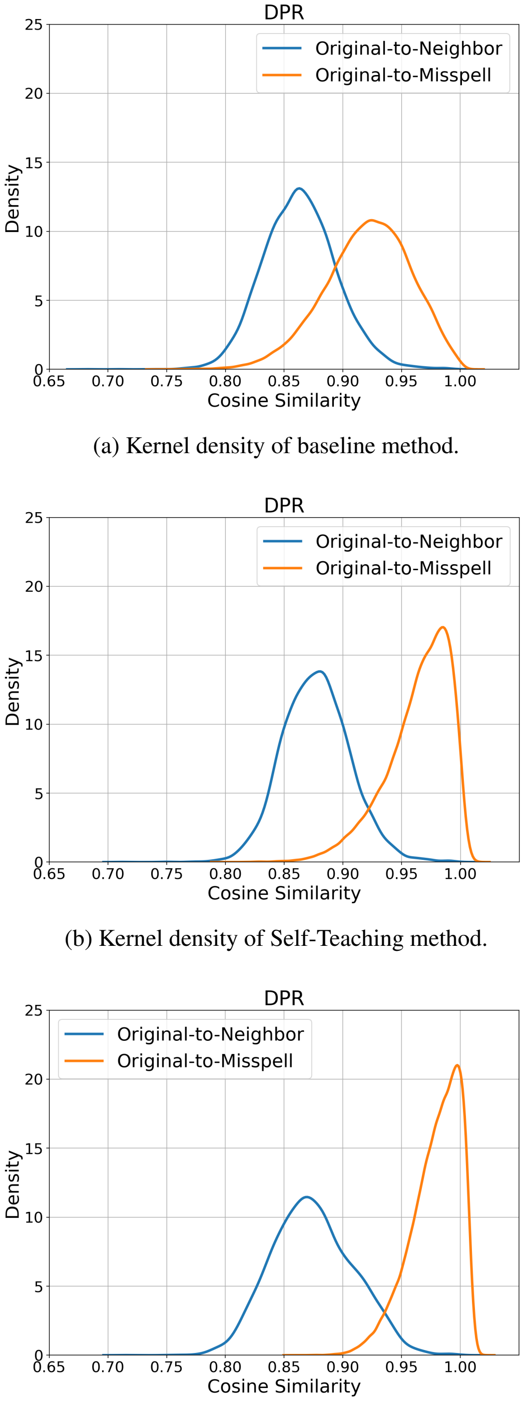
Table 4: The outputs of Query Expansion with $K = 10$. We use different colors to indicate different types of typo: RandInsert, RandDelete, RandSub, SwapNeighbor, and SwapAdjacent.

A.4 Query Distributions

The purpose of this section is to study the impact of our training method to the *Robustness* and *Contrast* of misspelled queries. We also compare our method against the baseline and competitive methods to show the effectiveness. The *Robustness* and *Contrast* of misspelled queries are illustrated using the following kernel density graphs:

- Original-to-Misspell: the cosine similarity distribution between original and misspelled queries.
- Original-to-Neighbor: the cosine similarity distribution between original and neighbor queries.

The *Robustness* property is emphasized by the Original-to-Misspell distribution having high cosine similarity. On the other hand, the *Contrast* property is emphasized by the Original-to-Misspell and Original-to-Neighbor distributions having small overlapping. The results in Figure 2 show that our method produces the best *Robustness* and *Contrast* properties for misspelled queries, in comparison to other methods.



(a) Kernel density of baseline method.

(b) Kernel density of Self-Teaching method.

(c) Kernel density of Dual Self-Teaching method (our).

Figure 2: Kernel density of Original-to-Neighbor (orange) and Original-to-Misspell (blue) of different training methods.

A.5 Limitations

In the following part, we list the limitations of the proposed method.

- The Query Expansion is designed for the English alphabet; therefore, other languages with different alphabets, such as Thai and Chinese, will require further work.
- The training strategy may not be suitable for languages with limited resources since it relies on fine-tuning a pre-trained language model using a large passage retrieval dataset.

A.6 Licenses

Datasets: The MS MARCO dataset is available under the MIT license, and the DL-typo dataset is available under the Apache license 2.0. These licenses allow users to use the datasets under non-restrictive agreements.
Softwares: We employ Hugging Face (Wolf et al., 2020) and Tevatron (Gao et al., 2022) libraries to train dense retrieval models. We utilize Ranx library (Bassani, 2022) to evaluate retrieval performance. These libraries are available under the Apache license 2.0 which allows both academic and commercial usages. For this reason, we release our code under the Apache license 2.0 to make our code fully accessible and compatible with the other codes we use.