Towards Robust Dense Retrieval via Local Ranking Alignment

Xuanang Chen[1,2], **Jian Luo**[1,2], **Ben He**[1,2*], **Le Sun**[2] and **Yingfei Sun**[1*]
[1]University of Chinese Academy of Sciences, Beijing, China
[2]Institute of Software, Chinese Academy of Sciences, Beijing, China
chenxuanang19@mails.ucas.ac.cn, luojian2021@iscas.ac.cn
benhe@ucas.ac.cn, sunle@iscas.ac.cn, yfsun@ucas.ac.cn

Abstract

Dense retrieval (DR) has extended the employment of pre-trained language models, like BERT, for text ranking. However, recent studies have raised the robustness issue of DR model against query variations, like query with typos, along with non-trivial performance losses. Herein, we argue that it would be beneficial to allow the DR model to learn to align the relative positions of query-passage pairs in the representation space, as query variations cause the query vector to drift away from its original position, affecting the subsequent DR effectiveness. To this end, we propose RoDR, a novel robust DR model that learns to calibrate the in-batch local ranking of query variation to that of original query for the DR space alignment. Extensive experiments on MS MARCO and ANTIQUE datasets show that RoDR significantly improves the retrieval results on both the original queries and different types of query variations. Meanwhile, RoDR provides a general query noise-tolerate learning framework that boosts the robustness and effectiveness of various existing DR models. Our code and models are openly available at https://github.com/cxa-unique/RoDR.

1 Introduction

Dense retrieval (DR) technique has been successfully applied to quite a few language systems, such as web search [Zhan *et al.*, 2020; Luan *et al.*, 2021; Xiong *et al.*, 2021; Qu *et al.*, 2021] and open-domain question answering [Karpukhin *et al.*, 2020; Xiong *et al.*, 2021; Qu *et al.*, 2021]. DR models employ pre-trained language models (PLMs), like BERT [Devlin *et al.*, 2019], to separately encode queries and passages into low-dimensional dense vectors, and adopt a lightweight similarity mechanism (e.g., dot product) for efficient online retrieval. However, recent work has raised the concern that DR models may not perform as well as expected when confronting queries with typos [Zhuang and Zuccon, 2021]. Aside from typos, for the same information need, queries created by users can also vary greatly, referred to as query variations [Bailey *et al.*, 2015]. Meanwhile, Penha *et al.* [2022] have demonstrated that query variations can significantly reduce the performance of a variety of neural re-ranking models. Nevertheless, it remains to be investigated whether and how various query variations (in addition to typos) weaken the DR model, and how to deal with these many types of query variations for more robust dense retrieval.

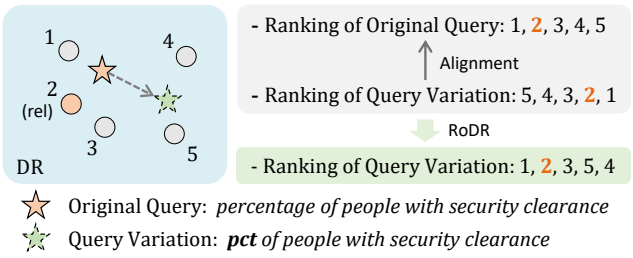

Figure 1: Example of local ranking alignment. In the representation space of DR, when replacing '*percentage*' with its abbreviation '*pct*', the query vector drifts away from its original position, resulting in effectiveness loss in that the rank of the relevant passage (id = 2) drops from the 2nd to the 4th. Our RoDR learns to maintain the relative positions of query-passage pairs in the DR space, by aligning the local ranking of query variation to that of original query as close as possible, for the improved DR robustness.

Different synthetic query variations have been used to assess the robustness of neural ranking models, including inserted typos [Wu *et al.*, 2021; Zhuang and Zuccon, 2021; Penha *et al.*, 2022], additional punctuation [Ma *et al.*, 2021], synonym substitution [Penha *et al.*, 2022], stopword removal [Penha *et al.*, 2022], back-translation [Ma *et al.*, 2021; Penha *et al.*, 2022] and random word swapping [Penha *et al.*, 2022]. In this paper, we summarize eight types of query variations and investigate their effects on the DR model. Given one original query set, the same type of query variation generator is used to modify all queries in this query set, yielding eight synthetic query variation sets for robustness testing. By comparing the performance of DR model on the original query set and synthetic query variation sets, we can examine the individual impacts of these query variation types. Unfortunately, relative to the original queries, significant effectiveness decreases can be observed on the query variations.

To address this issue, we propose RoDR, a novel variation-enhanced robust DR model by taking the relative positions of

*Corresponding author.

query-passage pairs in the DR space into account. Intuitively, dense retrieval is based on the similarities between the query-passage pairs in the representation space, and a query variation causes the query vector to shift away from its original position, resulting in skewed query-passage similarities, and consequently, degraded retrieval effectiveness.

As illustrated by the example in Figure 1, replacing '*percentage*' with its abbreviation '*pct*' causes the query vector to drift closer to the negative passages, diverting from the original well-learned representation. Thereby, our RoDR aims to maintain the spatial layout of the DR model by a local ranking alignment mechanism on the query variations. Specifically, the original queries are first used to train a standard DR model (denoted as DR_O) under the NLL objective, and then another DR model (denoted as DR_N) initialized from DR_O is further enhanced using the query variations, by minimizing the KL divergence on the similarity distributions of query-passage pairs between the DR_N and DR_O space. Additional to the NLL objective that maximizes the distances between positives and negatives to the query variations, our local ranking alignment mechanism aims to provide extra performance boost by functioning as a regularization loss that maintains the in-batch similarity ranking relationship for query-passage pairs with variations or not. Extensive experiments on MS MARCO [Nguyen *et al.*, 2016] and ANTIQUE [Hashemi *et al.*, 2020] datasets demonstrate that RoDR is able to achieve performance gains on both the original and variant queries. Moreover, RoDR provides a general framework for training a query noise-tolerant DR model. As shown by the evaluation results, RoDR leads to enhanced ranking effectiveness when it is applied to a variety of existing DR models, including ANCE [Xiong *et al.*, 2021], TAS-Balanced [Hofstätter *et al.*, 2021], and ADORE+STAR [Zhan *et al.*, 2021]. Our contributions are three-fold:

- We systematically investigate the impacts of eight types of query variations on dense retrieval (DR) models, and observe non-negligible performance decreases.
- A novel robust DR model RoDR is proposed, which learns to maintain the relative positions of query-passage pairs in the DR space by a local ranking alignment mechanism.
- Our RoDR provides a general query noise-tolerate learning framework that is applicable to various existing DR models, improving their robustness and effectiveness.

2 Preliminaries

As mentioned above, recent studies have highlighted the non-negligible negative impact of query variations on text ranking effectiveness [Zhuang and Zuccon, 2021; Penha *et al.*, 2022]. Herein, we first introduce eight types of operations that generate different query variations in Section 2.1, followed by a brief introduction to the dense retrieval in Section 2.2.

2.1 Query Variation Generation

MISSPELL: injected misspellings. Apart from the widely-studied Typos [Wu *et al.*, 2021; Zhuang and Zuccon, 2021], we also study OCR, Keyboard and SpellingError as defined in TextFlint [Gui *et al.*, 2021]. Specifically, out of the above four types of spelling errors, one type is randomly selected and injected into one random query word.

Variation Type	Query Text
ORIGINAL	what happens when stop drinking alcohol
MISSPELL	what happens when stpp drinking alcohol
EXTRAPUNC	what happens when stop drinking alcohol?
BACKTRANS	What happens if you stop drinking alcohol?
SWAPSYN-GLOVE	what happens when stop consuming alcohol
SWAPSYN-WNET	what happens when stop imbibe alcohol
TRANSTENSE	what happening when stop drinking alcohol
NOSTOPWORD	~~what~~ happens ~~when~~ stop drinking alcohol
SWAPWORDS	what drinking when stop happens alcohol

Table 1: Examples of the synthetic query variations generated upon the query (id = 667373) from MS MARCO passage Dev set.

EXTRAPUNC: extra punctuation. Akin to [Ma *et al.*, 2021], one to three identical and successive punctuation marks (including comma, period, question mark, and exclamation mark) are added at the end of the query.

BACKTRANS: back-translation. Akin to [Ma *et al.*, 2021; Penha *et al.*, 2022], the query in English is first translated to German and then is translated back to English, by the default machine translation model in TextFlint [Gui *et al.*, 2021].

SWAPSYN: swapped synonyms. Akin to [Penha *et al.*, 2022], one random word in the query is replaced by its synonym according to the Glove embeddings or WordNet, coined as **SWAPSYN-GLOVE** and **SWAPSYN-WNET**, respectively.

TRANSTENSE: transformed verb tense. The tense of one random verb in the query is transformed by TextFlint [Gui *et al.*, 2021]. No prior work has studied this type of query variation as far as we are aware of.

NOSTOPWORD: stopword removal. Akin to [Penha *et al.*, 2022], all stopwords in the query are deleted.

SWAPWORDS: swapped words. Akin to [Penha *et al.*, 2022], two randomly sampled words in the query are swapped by TextAttack [Morris *et al.*, 2020].

The examples for each type of query variations are listed in Table 1. Note that not all queries can obtain all eight types of variations. For example, if one query does not contain any stopword, the NOSTOPWORD variation is not applicable. Similarly, the BACKTRANS, SWAPSYN, TRANSTENSE and NOSTOPWORD variations could also be inapplicable to some certain queries. Additional to the synthetic query variations, the human validated query variations from [Penha *et al.*, 2022] are also included in our experiments.

2.2 Dense Retrieval

Dense retrieval (DR) usually employs the dual-encoder architecture, and optimizes the negative log likelihood (NLL) of the positive (relevant) passage among a set of negative (irrelevant) passages during its standard training procedure [Karpukhin *et al.*, 2020]. After that, all passages in corpus are encoded and indexed offline, and only a single query encoding inference is needed during online retrieval. Besides, as mentioned, previous work [Zhuang and Zuccon, 2021] has studied the impact of typos in query on the DR model, and

found that exposing query with typos (50% chances) directly to the DR model during training using NLL loss could reduce the negative impact of typos. In our work, additional to NLL, a local ranking alignment mechanism is also used to maintain the relative positions of query-passage pairs in the DR space, which is crucial to the subsequent DR retrieval.

3 Method

Let $\mathcal{D}_O = \{(q_i, p_i^+, p_{i,1}^-, \cdots, p_{i,n}^-)\}_{i=1}^m$ be the original training data, wherein each original query q_i is coupled with one positive passage p_i^+ and n negative passages $p_{i,j}^-$. The negative passages are usually collected from the corpus or top-ranked BM25 candidates [Karpukhin *et al.*, 2020]. As mentioned, we first use $\mathcal{D}_O$ through standard DR training procedure to obtain a DR model, denoted as DR_O. Then, another DR model (denoted as DR_N) initialized from DR_O is further optimized to improve the robustness against query variations under the guidance of DR_O, whose parameters are frozen during training. Let $\mathcal{D}_N = \{(q_i, \overline{q}_i, p_i^+, p_{i,1}^-, \cdots, p_{i,n}^-)\}_{i=1}^m$ be the training data used for local ranking alignment, wherein each original query q_i is coupled with one uniformly sampled query variation $\overline{q}_i$ from Section 2.1, and the negatives $p_{i,j}^-$ are all re-collected from the top-ranked candidates returned by DR_O, which act as the nearer neighbors as described later.

3.1 Local Ranking Alignment

As mentioned, in the representation space of DR_O, different query variations cause the query vector to drift away from its original position, resulting in the change of its relative positions among the neighboring passages and further a performance decrease. Thereby, it is necessary to maintain the relative positions of query-passage pairs in the DR space, which could be scrambled by the query variations. Herein, we propose to align the ordering of neighboring passages (queries) from DR_N to DR_O according to the in-batch local ranking of queries (passages). Specifically, in a batch of training samples $\mathcal{B} = \{(q_i, \overline{q}_i, p_i^+, p_{i,1}^-, \cdots, p_{i,n}^-)\}_{i=1}^{|\mathcal{B}|}$, an original query set $\mathcal{Q} = \{q_i\}_{i=1}^{|\mathcal{B}|}$, a query variation set $\overline{\mathcal{Q}} = \{\overline{q}_i\}_{i=1}^{|\mathcal{B}|}$, and a passage set $\mathcal{P} = \{p_i^+, p_{i,1}^-, \cdots, p_{i,n}^-\}_{i=1}^{|\mathcal{B}|} = \{p_j\}_{j=1}^{|\mathcal{P}|}$, wherein $|\mathcal{P}| = |\mathcal{B}| \cdot (n+1)$, can be all collected.

Query centering alignment. For each original query $q_i \in \mathcal{Q}$ and its variation $\overline{q}_i \in \overline{\mathcal{Q}}$, all in-batch positive and negative passages are collected as their shared neighbors (namely, $\mathcal{P}$), which also constitute the local ranking of both q_i and $\overline{q}_i$. In this local ranking, the rank of each passage for query variation $\overline{q}_i$ is calibrated to that for original query q_i. In order to be in line with the scoring mechanism of DR model, we also use dot product as the similarity measure for the local ranking. Besides, we use the similarity distribution to align the local ranking between q_i and $\overline{q}_i$, as the relative order of passages in the local ranking matters more than the exact similarity values when training a robust DR model. Consequently, for each $q_i \in \mathcal{Q}$ and $p_j \in \mathcal{P}$, we compute the conditional probability of passage p_j being close to the original query q_i in the representation space of DR_O, as defined in Eq. 1.

$$p_o(p_j|q_i) = \frac{e^{\text{sim}(q_i, p_j)}}{\sum_{k=1}^{|\mathcal{P}|} e^{\text{sim}(q_i, p_k)}} \tag{1}$$

wherein $\text{sim}(q,\ p)$ is the dot product similarity between the representations of query q and passage p.

Similarly, in the representation space of DR_N, the conditional probability of passage $p_j \in \mathcal{P}$ being close to the query variation $\overline{q}_i \in \overline{\mathcal{Q}}$ can be defined as Eq. 2.

$$p_n(p_j|\overline{q}_i) = \frac{e^{\text{sim}(\overline{q}_i, p_j)}}{\sum_{k=1}^{|\mathcal{P}|} e^{\text{sim}(\overline{q}_i, p_k)}} \tag{2}$$

To mitigate the impacts of the different query variations to the local ranking, conditional probabilities $p_o(p_j|q_i)$ and $p_n(p_j|\overline{q}_i)$ should be as close as possible. This alignment target is estimated by the Kullback-Leibler (KL) divergence between these two conditional probabilities:

$$\mathcal{L}_{p|q} = \frac{1}{|\mathcal{B}|} \sum_{i=1}^{|\mathcal{B}|} \sum_{j=1}^{|\mathcal{P}|} p_o(p_j|q_i) \log \frac{p_o(p_j|q_i)}{p_n(p_j|\overline{q}_i)} \tag{3}$$

Passage centering alignment. In the query centering alignment, we calibrate the relative positions of queries using the in-batch passages. Similarly, we can also calibrate the relative positions of passages using the in-batch queries. Specifically, for each passage $p_j \in \mathcal{P}$, all original queries in $\mathcal{Q}$ are served as the local ranking candidates in DR_O, and all query variations in $\overline{\mathcal{Q}}$ are served as the local ranking candidates in DR_N. Similar to Eq. 1-3, we align conditional probability $p_n(\overline{q}_i|p_j)$ to $p_o(q_i|p_j)$ by KL divergence as in Eq. 6.

$$p_o(q_i|p_j) = \frac{e^{\text{sim}(q_i, p_j)}}{\sum_{k=1}^{|\mathcal{B}|} e^{\text{sim}(q_k, p_j)}} \tag{4}$$

$$p_n(\overline{q}_i|p_j) = \frac{e^{\text{sim}(\overline{q}_i, p_j)}}{\sum_{k=1}^{|\mathcal{B}|} e^{\text{sim}(\overline{q}_k, p_j)}} \tag{5}$$

$$\mathcal{L}_{q|p} = \frac{1}{|\mathcal{P}|} \sum_{j=1}^{|\mathcal{P}|} \sum_{i=1}^{|\mathcal{B}|} p_o(q_i|p_j) \log \frac{p_o(q_i|p_j)}{p_n(\overline{q}_i|p_j)} \tag{6}$$

Nearer neighbors. To achieve better alignment, we use the top-1k passages returned by DR_O excluding those judged relevant to update the initial negatives in $\mathcal{D}_O$ to obtain the final $\mathcal{D}_N$. As the utilized local ranking involves all in-batch passages (especially negatives), these new negatives from DR_O act as the nearer neighbors for queries, which is beneficial to correctly calibrate the relative positions of query-passage pairs in the DR_N space. Besides, these new negatives also act as hard negatives in NLL objective, which have been shown useful to train an effective DR model [Xiong *et al.*, 2021].

3.2 Model Training

During training, along with the query centering and passage centering alignment described in Section 3.1, the NLL which teaches the DR model to pick out the positive passage from a set of negative passages for query variations is also used:

$$\mathcal{L}_{nll} = -\frac{1}{|\mathcal{B}|} \sum_{i=1}^{|\mathcal{B}|} \log \frac{e^{\text{sim}(\overline{q}_i, p_i^+)}}{e^{\text{sim}(\overline{q}_i, p_i^+)} + \sum_{j=1}^{\overline{n}} e^{\text{sim}(\overline{q}_i, p_{i,j}^-)}} \tag{7}$$

wherein $|\mathcal{B}|$ is the effective batch size, $\overline{n}$ is the number of effective negative passages, which is usually larger than n (in

$\mathcal{D}_N$) due to the use of in-batch negatives [Karpukhin *et al.*, 2020] or cross-batch negatives [Qu *et al.*, 2021].

Different to the standard DR training, the original query q_i in NLL loss is replaced by its variation $\overline{q}_i$. Besides, different to [Zhuang and Zuccon, 2021], a local ranking alignment mechanism (LRA) is used as a regularization item to maintain the relative positions of query-passage pairs in the DR space (namely, DR_N). As summarized in Eq. 8, we train DR_N using the weighted sum of NLL and LRA losses, wherein w_1, w_2 and w_3 are hyper-parameters to control the loss weights.

$$\mathcal{L} = w_1\mathcal{L}_{nll} + w_2\mathcal{L}_{p|q} + w_3\mathcal{L}_{q|p} \tag{8}$$

After the above query noise-tolerate training, the resulting DR_N model is coined as **RoDR**, and used for robust retrieval.

4 Experiments

4.1 Experimental Setup

Datasets and metrics. We employ both MS MARCO passage and document datasets for our experiments. The MS MARCO passage (document) corpus contains 8.8 (3.2) million passages (documents), from which we construct about 0.40 (0.37) million training samples. We use 6,980 (5,193) Dev queries in MS MARCO passage (document) dataset for evaluation, along with the official metric, namely, MRR@10 (MRR@100). Additionally, Recall@1000 (Recall@100) is also used to evaluate the recall quality for passage (document) retrieval. Statistical significance with paired two-tailed t-test is reported. Akin to [Wu *et al.*, 2021; Zhuang and Zuccon, 2021], the drop rate (loss ratio in performance) on one query variation set relative to the original query set (denoted as ORIGINAL) is also reported.

Query variations. We use eight types of query transformation as defined in Section 2.1 for query variation generation on MS MARCO Dev sets. Specifically, for each type of query transformation, we apply it to all queries in the Dev set to obtain a variation set, which is named by the acronym of the query variation (e.g., MISSPELL). As mentioned in Section 2.1, if one variation is not applicable to the given query, this query is kept original in this variation set. The statistics of query variations on MS MARCO passage and document Dev sets are summarized in Tables 2 and 3, respectively. Besides, we report the average results on these eight query variation sets, denoted as VARIATIONS (AVG.).

Manually validated query variations. Additional to our synthetic query variations, which could be occasionally inapplicable or incomprehensible, we also experiment with the manually validated query variations released in [Penha *et al.*, 2022]. This human validated query variation data is based on the 200 original queries from ANTIQUE dataset [Hashemi *et al.*, 2020] (namely, the 'antique/train/split200-valid' set available in `ir_datasets` [MacAvaney *et al.*, 2021]), denoted as ANTIQUE-ORIGINAL. There are five released query variation sets (e.g., `T5DescToTitle`, `BackTransl` as in [Penha *et al.*, 2022]), which are used to test the DR models trained on MS MARCO passage dataset in a zero-shot setting. The average results on these five validated query variation sets are reported, denoted as ANTIQUE-VARIATIONS (AVG.).

Training details. Our model training is based on the Tevatron toolkit [Gao *et al.*, 2022], with parameter-shared BERT-Base model as the query and passage (document) encoders. The max query length is 32, and the max passage (document) length is 128 (512). For MS MARCO passage (document) retrieval, the DR model is trained with the learning rate of 5e-6 and per-device batch size of 16 (2) for 4 epochs on one (four) GeForce RTX 3090 GPU(s). The negative passages (documents) in D_O are sampled from the provided official triples (the top-1k BM25 candidates), and there are 7 negatives in each training sample, namely, $n = 7$ in D_O and D_N. The loss weights w_1, w_2, w_3 in Eq. 8 are set as 1, 1, 0.2 (1, 0.1, 1) for MS MARCO passage (document) retrieval.

Models in comparison. As for baselines, we compare the following models: $\textbf{DR}_{\boldsymbol{OQ}}$ is a standard DR model trained using original queries (namely, D_O), without using any query variation. $\textbf{DR}_{\boldsymbol{QV}}$ is trained using query variations rather than original queries (namely, all original queries in D_O are replaced with their variations, instead of 50% used in [Zhuang and Zuccon, 2021], since our setting yields better results). $\textbf{DR}_{\boldsymbol{OQ \to QV}}$ is first trained using original queries before using query variations, in two training stages and for the same training steps with RoDR for fair comparison.

Plugging RoDR into existing DR models. As described, DR_{OQ} is the basic model of our vanilla RoDR, which is denoted as RoDR w/ DR_{OQ}. In addition to the standard DR training used in DR_{OQ}, quite a few recent works have proposed more effective DR models with more advanced training methods. Thereby, we also apply RoDR to a few existing DR models, including ANCE [Xiong *et al.*, 2021], TAS-Balanced [Hofstätter *et al.*, 2021] and ADORE+STAR [Zhan *et al.*, 2021]. Specifically, we use the released checkpoints of these DR models trained on MS MARCO passage dataset to initialize DR_O and DR_N, and then conduct query noise-tolerate training in the same way as described in Section 3. Herein, equal loss weights in Eq. 8 are used, and the learning rate is 3e-5, 5e-6 and 5e-7 for RoDR w/ ANCE, TAS-Balanced and ADORE+STAR, respectively. Other training configurations are the same as in RoDR w/ DR_{OQ}.

4.2 Results

The performance degradation on query variations. As seen in Tables 2 and 3, non-negligible degradation of DR_{OQ} to varying degrees is observed on all query variation sets relative to ORIGINAL set. On average, comparing ORIGINAL with VARIATIONS (AVG.), the performance of DR_{OQ} drops 19.4% and 7.26% in terms of MRR@10 and R@1000 on MS MARCO passage Dev set, 19.8% and 14.9% in terms of MRR@100 and R@100 on MS MARCO document Dev set. Among eight types of query variations, DR_{OQ} suffers from misspellings (MISSPELL) the most (over 50% MRR drops on both Dev sets), and suffers from extra punctuation (EXTRAPUNC) and transformed verb tense (TRANSTENSE) the least. Overall, the standard DR_{OQ} indeed suffers from effectiveness losses on various query variations.

Exposure of query variations to the DR model helps. As seen in Tables 2 and 3, when using various query variations during training, DR models including DR_{QV} and $\text{DR}_{OQ \to QV}$

Variation Set	#Ori.	#Vari.	DR_{OQ}		DR_{QV}		$DR_{OQ \to QV}$		RoDR w/ DR_{OQ}	
			MRR@10	R@1000	MRR@10	R@1000	MRR@10	R@1000	MRR@10	R@1000
ORIGINAL	6,980	0	32.4	95.4	32.5	95.1	33.0	95.4	**34.9**†	**96.4**†
VARIATIONS (AVG.)	-	-	26.1/-19.4%	88.5/-7.26%	27.8/-14.4%	90.7/-4.54%	28.2/-14.4%	91.0/-4.62%	**30.4**†/-13.0%	**93.0**†/-3.57%
MISSPELL	0	6,980	15.5/-52.0%	73.5/-23.0%	22.0/-32.4%	84.5/-11.1%	22.5/-31.7%	84.7/-11.2%	**25.0**†/-28.5%	**88.0**†/-8.71%
EXTRAPUNC	0	6,980	32.0/-1.17%	95.2/-0.19%	32.4/-0.24%	95.2/+0.12%	32.8/-0.56%	95.5/+0.10%	**34.7**†/-0.69%	**96.5**†/+0.11%
BACKTRANS	486	6,494	25.5/-21.2%	87.1/-8.72%	26.1/-19.8%	87.9/-7.50%	26.3/-20.5%	88.2/-7.49%	**28.2**†/-19.2%	**90.3**†/-6.33%
SWAPSYN-GLOVE	665	6,315	22.0/-32.0%	84.2/-11.7%	24.0/-26.1%	86.7/-8.78%	24.2/-26.6%	86.9/-8.85%	**26.5**†/-24.1%	**89.5**†/-7.11%
SWAPSYN-WNET	434	6,456	22.1/-31.7%	84.4/-11.5%	24.3/-25.1%	87.6/-7.81%	24.3/-26.4%	87.8/-7.95%	**27.5**†/-21.4%	**90.8**†/-5.83%
TRANSTENSE	4,199	2,781	31.6/-2.32%	95.0/-0.41%	32.0/-1.54%	94.9/-0.21%	32.6/-1.23%	95.1/-0.28%	**34.6**†/-0.96%	**96.3**†/-0.14%
NOSTOPWORD	859	6,121	29.0/-10.4%	93.7/-1.77%	30.6/-5.76%	94.3/-0.80%	31.2/-5.39%	94.6/-0.77%	**32.9**†/-5.75%	**96.1**†/-0.34%
SWAPWORDS	0	6,980	30.9/-4.43%	94.7/-0.76%	31.2/-3.97%	94.8/-0.22%	32.0/-3.16%	94.9/-0.53%	**33.9**†/-3.05%	**96.2**†/-0.19%

Table 2: Retrieval results on MS MARCO passage Dev set. Statistically significant improvements at p-value < 0.01 over $DR_{OQ \to QV}$ are marked with †. For each DR model, the drop rate on different variation sets relative to the ORIGINAL set are reported.

Variation Set	#Ori.	#Vari.	DR_{OQ}		DR_{QV}		$DR_{OQ \to QV}$		RoDR w/ DR_{OQ}	
			MRR@100	R@100	MRR@100	R@100	MRR@100	R@100	MRR@100	R@100
ORIGINAL	5,193	0	34.2	81.1	35.2	82.7	34.9	81.9	**38.6**†	**89.1**†
VARIATIONS (AVG.)	-	-	27.5/-19.8%	69.0/-14.9%	30.0/-14.7%	74.3/-10.1%	29.7/-14.8%	73.7/-9.91%	**34.0**†/-11.8%	**83.6**†/-6.19%
MISSPELL	0	5,193	16.4/-52.1%	48.0/-40.8%	24.1/-31.5%	65.3/-21.0%	23.7/-32.1%	65.0/-20.5%	**28.3**†/-26.6%	**76.2**†/-14.4%
EXTRAPUNC	0	5,193	33.0/-3.60%	78.7/-2.97%	34.0/-3.43%	80.2/-3.00%	33.8/-2.98%	79.5/-2.87%	**38.5**†/-0.33%	**89.1**†/-0.00%
BACKTRANS	350	4,843	27.3/-20.2%	68.7/-15.3%	28.6/-18.7%	71.9/-13.0%	28.2/-19.2%	71.1/-13.2%	**32.0**†/-17.2%	**80.1**†/-10.1%
SWAPSYN-GLOVE	498	4,695	22.3/-35.0%	59.7/-26.4%	25.0/-29.0%	66.3/-19.8%	24.5/-29.8%	65.3/-20.2%	**29.4**†/-23.7%	**77.4**†/-13.1%
SWAPSYN-WNET	318	4,875	23.4/-31.7%	61.4/-24.2%	26.3/-25.3%	68.1/-17.6%	26.0/-25.4%	67.6/-17.4%	**30.8**†/-20.3%	**79.4**†/-10.9%
TRANSTENSE	3,145	2,048	33.7/-1.53%	80.1/-1.16%	34.9/-0.87%	81.9/-0.89%	34.6/-0.68%	81.2/-0.80%	**38.3**†/-0.86%	**88.9**†/-0.22%
NOSTOPWORD	651	4,542	30.5/-10.8%	75.2/-7.27%	33.2/-5.69%	79.8/-3.45%	33.1/-5.05%	79.4/-3.01%	**37.3**†/-3.37%	**88.8**†/-0.28%
SWAPWORDS	0	5,193	33.1/-3.43%	80.1/-1.26%	34.0/-3.31%	81.1/-1.89%	33.6/-3.48%	80.8/-1.29%	**37.8**†/-1.98%	**88.6**†/-0.50%

Table 3: Retrieval results on MS MARCO document Dev set. The base model in significance test is DR_{QV}, others are the same as in Table 2.

Model	ANTIQUE-ORIGINAL		ANTIQUE-VARIATIONS (AVG.)	
	nDCG@10	R@1000	nDCG@10	R@1000
DR_{OQ}	27.1	59.8	25.5/-6.05%	57.4/-4.02%
DR_{QV}	27.3	60.2	25.6/-6.44%	58.0/-3.68%
$DR_{OQ \to QV}$	27.4	59.6	25.3/-7.70%	57.3/-3.98%
RoDR w/ DR_{OQ}	**29.0**†	**62.6**†	**27.3**†/-5.67%	**60.3**†/-3.72%

Table 4: Zero-shot retrieval results on ANTIQUE dataset. DR models are only trained on MS MARCO passage dataset. Statistically significant improvements at p-value < 0.01 over DR_{QV} are marked with †. The average drop rates on variation sets are reported.

are much more tolerant to all types of query variations. On average, when facing with query variation sets, DR_{QV} gains 1.7 (2.5) points over DR_{OQ} in terms of MRR@10 (MRR@100) on passage (document) Dev set. Besides, compared to DR_{OQ}, the average drop rates of DR_{QV} and $DR_{OQ \to QV}$ are reduced on both Dev sets (e.g., from 19.4% to 14.4% for DR_{QV} on passage Dev set in terms of MRR@10). On the whole, simply adding query variations into the training data is helpful.

RoDR produces significantly better retrieval results and smaller drop rates. The retrieval results of vanilla RoDR w/ DR_{OQ} are summarized in the last column of Tables 2 and 3. It can be seen that RoDR w/ DR_{OQ} not only performs significantly better on all query variation sets, but also on the original query set, which indicates the super effectiveness of RoDR framework. Besides, compared to all three baseline models, RoDR w/ DR_{OQ} can also reach smaller drop rate (which reflects the robustness of DR model) from original query set to various query variation sets. Thereby, we demonstrate that RoDR boosts not only the retrieval effectiveness, but also the retrieval robustness on query variations.

RoDR still performs better on manually validated query variations. The zero-shot retrieval results on ANTIQUE are listed in Table 4. We can see that DR_{QV} and $DR_{OQ \to QV}$ perform on par with DR_{OQ} on both ANTIQUE-ORIGINAL and ANTIQUE-VARIATIONS (AVG.). It indicates that simple exposure of query variations provides no benefit to DR model on ANTIQUE. In contrast, RoDR w/ DR_{OQ} produces significantly better retrieval quality on both ANTIQUE-ORIGINAL and ANTIQUE-VARIATIONS (AVG.). Meanwhile, RoDR w/ DR_{OQ} still reaches smaller average drop rate on validated query variations (e.g., 5.67% in terms of nDCG@10).

RoDR provides extra performance boosts over existing DR models. As seen in Table 5, although ANCE, TAS-Balanced and ADORE+STAR with more advanced training strategies can outperform DR_{OQ}, they still suffer from the query variations (e.g., 15%-17% average drop rate in MRR@10 on MS MARCO, and 6%-12% average drop rate in nDCG@10 on ANTIQUE). Meanwhile, when RoDR is applied to these three DR models, their effectiveness and robustness can be further improved (e.g., about one point gains in MRR@10 on MS MARCO ORIGINAL set, and smaller average drop rates on both MS MARCO and ANTIQUE).

4.3 Analysis

Ablation study on loss function of RoDR. As seen in Table 6, on ORIGINAL set, the contribution of passage centering alignment (namely, $\mathcal{L}_{q|p}$) is not strikingly high, but it is still beneficial to the query variation sets. Meanwhile, without both query and passage centering alignments (namely, $\mathcal{L}_{p|q}$

Model	MS MARCO Passage Dev				ANTIQUE			
	ORIGINAL		VARIATIONS (AVG.)		ANTIQUE-ORIGINAL		ANTIQUE-VARIATIONS (AVG.)	
	MRR@10	R@1000	MRR@10	R@1000	nDCG@10	R@1000	nDCG@10	R@1000
ANCE	33.0	95.9	27.5/-16.7%	90.1/-6.08%	31.4	59.1	29.3/-6.70%	56.6/-4.20%
RoDR w/ ANCE	**34.1**†	**96.3**†	**30.6**†/-10.0%	**93.2**†/-3.26%	**31.7**	**64.2**†	**29.5**/-7.09%	**61.6**†/-4.03%
TAS-Balanced	34.4	**97.7**	29.1/-15.4%	93.2/-4.62%	22.1	**59.7**	19.5/-11.4%	**57.4**/-3.86%
RoDR w/ TAS-Balanced	**35.5**†	97.5	**30.9**†/-12.8%	**94.2**/-3.41%	**24.2**†	59.4	**22.2**†/-8.19%	57.3/-3.53%
ADORE+STAR	34.7	**96.9**	29.3/-15.6%	92.1/-4.88%	31.0	63.6	28.3/-8.68%	60.7/-4.44%
RoDR w/ ADORE+STAR	**35.4**†	96.8	**31.0**†/-12.4%	**93.3**†/-3.57%	**31.5**	**64.4**	**29.2**†/-7.52%	**61.8**†/-4.14%

Table 5: Retrieval results of RoDR based on existing DR models on MS MARCO passage Dev set and zero-shot retrieval results on ANTIQUE dataset. Statistically significant gains by RoDR at p-value < 0.05 are marked with †. The average drop rates on variation sets are reported.

Model	ORIGINAL		VARIATIONS (AVG.)	
	MRR@10	R@1000	MRR@10	R@1000
RoDR w/ DR_{OQ}	34.9	96.4	30.4	93.0
w/o $\mathcal{L}_{p\|q}$	34.5$^\downarrow$	96.3	29.9$^\downarrow$	92.8$^\downarrow$
w/o $\mathcal{L}_{q\|p}$	34.8	96.4	30.3$^\downarrow$	92.9
w/o $\mathcal{L}_{p\|q}$ and $\mathcal{L}_{q\|p}$	34.3$^\downarrow$	96.3	29.7$^\downarrow$	92.7$^\downarrow$
w/o $\mathcal{L}_{nll}$	33.8$^\downarrow$	95.9$^\downarrow$	29.6$^\downarrow$	92.2$^\downarrow$

Table 6: Ablation study of RoDR on MS MARCO passage Dev set. Statistically significant drops at p-value < 0.05 are marked with ↓.

and $\mathcal{L}_{q|p}$), the performance of RoDR on both ORIGINAL and VARIATIONS (AVG.) sets is greatly affected. Moreover, NLL on query variations (namely, $\mathcal{L}_{nll}$) is also important in RoDR since its removal leads to marked effectiveness loss.

The learning curve of RoDR. Furthermore, we present the learning curve of RoDR w/ DR_{OQ} in Figure 2 to show the effect of the local ranking alignment mechanism (abbreviated as LRA). When used with the NLL loss, LRA acts as a positional regularizer (in the DR space) over $\mathcal{L}_{nll}$ to prevent overfitting on the query variations. Besides, the performance on MS MARCO passage Dev set (both ORIGINAL and VARIATIONS (AVG.)) is always better when LRA is enabled. Thus, our LRA is complementary to NLL, and using both for training can achieve more robust DR retrieval.

The training efficiency. Although using augmented query variations for training can improve the robustness performance, it indeed increases the training time of DR models. In our experiments, for MS MARCO passage retrieval, it takes about 10hrs for both DR_{OQ} and DR_{QV}, about 20hrs for $DR_{OQ \to QV}$, and about 22.7hrs for RoDR w/ DR_{OQ}. In other words, the training time from scratch of RoDR is normally double when compared to the standard DR model. However, when based on the existing DR models, it is generally cost-efficient to apply RoDR for robustness enhancement.

5 Related Work

Dense retrieval systems have been widely studied and developed. Reimers and Gurevych [2019] present Sentence-BERT in a siamese architecture to derive semantically meaningful sentence embeddings for textual similarity matching. Karpukhin *et al.* [2020] use BM25 negatives to better fine-tune the DR model and successfully apply it to the open question answering task. Luan *et al.* [2021] find that dense retrieval is not effective when the text sequences are too long.

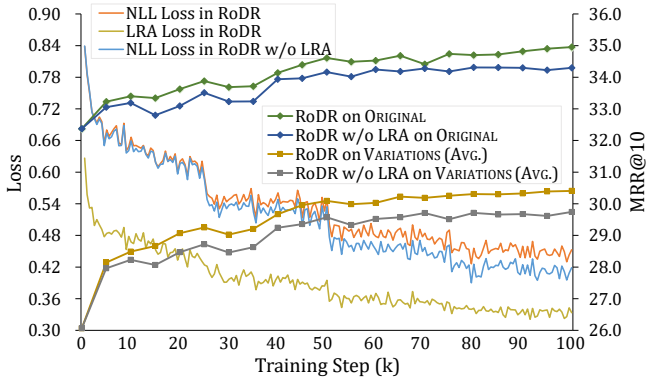

Figure 2: The learning curve of RoDR, when LRA is enabled or not.

Xiong *et al.* [2021] use global negatives from the DR model to enhance itself. Then, more and more studies [Qu *et al.*, 2021; Hofstätter *et al.*, 2021; Ren *et al.*, 2021a; Zhan *et al.*, 2021; Ren *et al.*, 2021b; Li *et al.*, 2021] work toward more powerful DR models by a series of sophisticated training procedures.

The robustness issue of DR model has also been raised. Zhuang and Zuccon *et al.* [2021] have shown that DR model degrades dramatically when facing with the query with typos. In fact, the typos in query not only affect the performance of DR, but also have serious impacts on different neural ranking models [Wu *et al.*, 2021]. Meanwhile, apart from typos, other various query variations are also found to have negative influence on the performance of neural ranker [Ma *et al.*, 2021; Penha *et al.*, 2022]. In our work, additional to typos, a variety of query variations are included to study the robustness of DR model. Moreover, this work proposes a novel RoDR model, which provides a general learning framework that boosts the robustness of various existing DR models.

6 Conclusion

In this work, we first summarize eight types of query variations, which are used for the systematic investigation on the robustness of DR model. Evaluation results show that DR model degrades significantly under various query variations. Then, to address this, we propose a novel robust DR model, named RoDR, to boost the performance by maintaining the relative positions of query-passage pairs in the DR space. In future work, we plan to extend our study to other related tasks, like open question answering [Karpukhin *et al.*, 2020].

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant no. U1936207 and 62076233, and the University of Chinese Academy of Sciences.

References

[Bailey *et al.*, 2015] Peter Bailey, Alistair Moffat, Falk Scholer, and Paul Thomas. User variability and IR system evaluation. In *SIGIR*, pages 625–634, 2015.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186, 2019.

[Gao *et al.*, 2022] Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. Tevatron: An efficient and flexible toolkit for dense retrieval. *CoRR*, abs/2203.05765, 2022.

[Gui *et al.*, 2021] Tao Gui, Xiao Wang, Qi Zhang, Qin Liu, Yicheng Zou, Xin Zhou, Rui Zheng, Chong Zhang, Qinzhuo Wu, Jiacheng Ye, Zexiong Pang, Yongxin Zhang, Zhengyan Li, Ruotian Ma, Zichu Fei, Ruijian Cai, Jun Zhao, Xinwu Hu, Zhiheng Yan, Yiding Tan, Yuan Hu, Qiyuan Bian, Zhihua Liu, Bolin Zhu, Shan Qin, Xiaoyu Xing, Jinlan Fu, Yue Zhang, Minlong Peng, Xiaoqing Zheng, Yaqian Zhou, Zhongyu Wei, Xipeng Qiu, and Xuanjing Huang. TextFlint: Unified multilingual robustness evaluation toolkit for natural language processing. *CoRR*, abs/2103.11441, 2021.

[Hashemi *et al.*, 2020] Helia Hashemi, Mohammad Aliannejadi, Hamed Zamani, and W. Bruce Croft. ANTIQUE: A non-factoid question answering benchmark. In *ECIR*, pages 166–173, 2020.

[Hofstätter *et al.*, 2021] Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. Efficiently teaching an effective dense retriever with balanced topic aware sampling. In *SIGIR*, pages 113–122, 2021.

[Karpukhin *et al.*, 2020] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *EMNLP*, pages 6769–6781, 2020.

[Li *et al.*, 2021] Yizhi Li, Zhenghao Liu, Chenyan Xiong, and Zhiyuan Liu. More robust dense retrieval with contrastive dual learning. In *ICTIR*, pages 287–296, 2021.

[Luan *et al.*, 2021] Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. Sparse, dense, and attentional representations for text retrieval. *TACL*, pages 329–345, 2021.

[Ma *et al.*, 2021] Xiaofei Ma, Cícero Nogueira dos Santos, and Andrew O. Arnold. Contrastive fine-tuning improves robustness for neural rankers. In *ACL/IJCNLP (Findings)*, pages 570–582, 2021.

[MacAvaney *et al.*, 2021] Sean MacAvaney, Andrew Yates, Sergey Feldman, Doug Downey, Arman Cohan, and Nazli Goharian. Simplified data wrangling with ir_datasets. In *SIGIR*, pages 2429–2436, 2021.

[Morris *et al.*, 2020] John X. Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. TextAttack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP. In *EMNLP (Demos)*, pages 119–126, 2020.

[Nguyen *et al.*, 2016] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. MS MARCO: A human generated machine reading comprehension dataset. In *CoCo@NIPS*, CEUR Workshop Proceedings, 2016.

[Penha *et al.*, 2022] Gustavo Penha, Arthur Câmara, and Claudia Hauff. Evaluating the robustness of retrieval pipelines with query variation generators. In *ECIR*, pages 397–412, 2022.

[Qu *et al.*, 2021] Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering. In *NAACL-HLT*, pages 5835–5847, 2021.

[Reimers and Gurevych, 2019] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *EMNLP/IJCNLP*, pages 3980–3990, 2019.

[Ren *et al.*, 2021a] Ruiyang Ren, Shangwen Lv, Yingqi Qu, Jing Liu, Wayne Xin Zhao, Qiaoqiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. PAIR: leveraging passage-centric similarity relation for improving dense passage retrieval. In *ACL/IJCNLP (Findings)*, pages 2173–2183, 2021.

[Ren *et al.*, 2021b] Ruiyang Ren, Yingqi Qu, Jing Liu, Wayne Xin Zhao, Qiaoqiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. RocketQAv2: A joint training method for dense passage retrieval and passage re-ranking. In *EMNLP*, pages 2825–2835, 2021.

[Wu *et al.*, 2021] Chen Wu, Ruqing Zhang, Jiafeng Guo, Yixing Fan, and Xueqi Cheng. Are neural ranking models robust? *CoRR*, abs/2108.05018, 2021.

[Xiong *et al.*, 2021] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *ICLR*, 2021.

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

[Zhan *et al.*, 2021] Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, and Shaoping Ma. Optimizing dense retrieval model training with hard negatives. In *SIGIR*, pages 1503–1512, 2021.

[Zhuang and Zuccon, 2021] Shengyao Zhuang and Guido Zuccon. Dealing with typos for bert-based passage retrieval and ranking. In *EMNLP*, pages 2836–2842, 2021.