Abstract

Recently, dense passage retrieval has become a mainstream approach to finding relevant information in various natural language processing tasks. A number of studies have been devoted to improving the widely adopted dual-encoder architecture. However, most of the previous studies only consider query-centric similarity relation when learning the dual-encoder retriever. In order to capture more comprehensive similarity relations, we propose a novel approach that leverages both query-centric and passage-centric similarity relations (called PAIR) for dense passage retrieval. To implement our approach, we make three major technical contributions by introducing formal formulations of the two kinds of similarity relations, generating high-quality pseudo labeled data via knowledge distillation, and designing an effective two-stage training procedure that incorporates passage-centric similarity relation constraint. Extensive experiments show that our approach significantly outperforms previous state-of-the-art models on both MSMARCO and Natural Questions datasets.

1 Introduction

With the recent advances of pre-trained language models, dense passage retrieval techniques (representing queries and passages in low-dimensional semantic space) have significantly outperformed traditional term-based techniques (Guu et al., 2020; Karpukhin et al., 2020). As the key step of finding the relevant information, it has been shown that dense passage retrieval can effectively improve the performance in a variety of tasks, including question answering (Lee et al., 2019; Xiong et al., 2020b), information retrieval (Luan et al., 2021; Khattab and Zaharia, 2020), dialogue (Ji et al., 2014; Henderson et al., 2017) and entity linking (Gillick et al., 2019; Wu et al., 2020).

Typically, the dual-encoder architecture is used to learn the dense representations of queries and passages, and the dot-product similarity between the representations of queries and passages becomes ranking measurement for retrieval. A number of studies have been devoted to improving this architecture (Guu et al., 2020; Karpukhin et al., 2020; Xiong et al., 2020a) for dense passage retrieval. Previous studies mainly consider learning query-centric similarity relation, where it tries to increase the similarity \( s(q, p^+) \) between a query and a positive (i.e., relevant) passage meanwhile decrease the similarity \( s(q, p^-) \) between the query and a negative (i.e., irrelevant) passage. We argue that query-centric similarity relation ignores the relation between passages, and it brings difficulty to discriminate between positive and negative passages. To illustrate this, we present an example in Figure 1, where a query \( q \) and two passages \( p^+ \) and \( p^- \) are related to the same passage.

![Figure 1](image-url)
and $p^-$ are given. As we can see in Figure 1(a), although query-centric similarity relation can enforce $s(q, p^+) > s(q, p^-)$ and identify the positive passages in this case, the distance (i.e., dissimilarity) between positive and negative passages is small. When a new query is issued, it is difficult to discriminate between positive passage $p^+$ and negative passage $p^-$. Considering this problem, we propose to further learn passage-centric similarity relation for enhancing the dual-encoder architecture. The basic idea is shown in Figure 1(b), where we set an additional similarity relation constraint $s(p^+, q) > s(p^+, p^-)$: the similarity between query $q$ and positive passage $p^+$ should be larger than that between positive passage $p^+$ and negative passage $p^-$. In this way, it is able to better learn the similarity relations among query, positive passages and negative passages. Although the idea is appealing, it is not easy to implement due to three major issues. First, it is unclear how to formalize and learn both query-centric and passage-centric similarity relations. Second, it requires large-scale and high-quality training data to incorporate passage-centric similarity relation. However, it is expensive to manually label data. Additionally, there might be a large number of unlabeled positives even in the existing manually labeled datasets (Qu et al., 2020), and it is likely to bring false negatives when sampling hard negatives. Finally, learning passage-centric similarity relation (an auxiliary task) is not directly related to the query-centric similarity relation (a target task). In terms of multi-task viewpoint, multi-task models often perform worse than their single-task counterparts (Alonso and Plank, 2017; McCann et al., 2018; Clark et al., 2019). Hence, it needs a more elaborate design for the training procedure.

To this end, in this paper, we propose a novel approach that leverages both query-centric and Passage-centric Similarity Relations (called PAIR) for dense passage retrieval. In order to address the aforementioned issues, we have made three important technical contributions. First, we design formal loss functions to characterize both query-centric and passage-centric similarity relations. Second, we propose to generate pseudo-labeled data via knowledge distillation. Third, we devise a two-stage training procedure that utilizes passage-centric similarity relation during pre-training and then fine-tunes the dual-encoder according to the task goal. The improvements in the three aspects make it possible to effectively leverage both kinds of similarity relations for improving dense passage retrieval.

The contributions of this paper can be summarized as follows:

- We propose an approach that simultaneously learns query-centric and passage-centric similarity relations for dense passage retrieval. It is the first time that passage-centric similarity relation has been considered for this task.
- We make three major technical contributions by introducing formal formulations, generating high-quality pseudo-labeled data and designing an effective training procedure.
- Extensive experiments show that our approach significantly outperforms previous state-of-the-art models on both MSMARCO and Natural Questions datasets.

2 Related Work

Recently, dense passage retrieval has demonstrated better performance than traditional sparse retrieval methods (e.g., TF-IDF and BM25). Different from sparse retrieval, dense passage retrieval represents queries and passages into low-dimensional vectors (Zhao et al., 2022; Guu et al., 2020; Karpukhin et al., 2020), typically in a dual-encoder architecture, and uses dot product as the similarity measurement for retrieval. The existing approaches for dense passage retrieval can be divided into two categories: (1) unsupervised pre-training for retrieval (2) fine-tuning only on labeled data.

In the first category, different pre-training tasks for retrieval were proposed. Lee et al. (2019) proposed a specific approach to pre-training the retriever with an unsupervised task, namely Inverse Cloze Task (ICT), and then jointly fine-tuned the retriever and a reader on labeled data. REALM (Guu et al., 2020) proposed a new pre-training approach, which jointly trained a masked language model and a neural retriever. Different from them, our proposed approach utilizes the pseudo-labeled data via knowledge distillation in the pre-training stage, and the quality of the generated data is high (see Section 4.6).

In the second category, the existing approaches fine-tuned pre-trained language models on labeled
data (Karpukhin et al., 2020; Luan et al., 2021). Both DPR (Karpukhin et al., 2020) and MEBERT (Luan et al., 2021) used in-batch random sampling and hard negative sampling by BM25, while ANCE (Xiong et al., 2020a), NPRINC (Lu et al., 2020) and RocketQA (Qu et al., 2020) explored more sophisticated hard negative sampling approach. Izacard and Grave (2020) and Yang et al. (2020) leveraged a reader and a cross-encoder for knowledge distillation on labeled data, respectively. RocketQA found large batch size can significantly improve the retrieval performance of dual-encoders. ColBERT (Khattab and Zaharia, 2020) incorporated light-weight attention-based re-ranking while increasing the space complexity.

The existing studies mainly focus on learning the similarity relation between the queries and the passages, while ignoring the relation among passages. It makes the model difficult to discriminate the positive passages and negative passages. In this paper, we propose an approach simultaneously learn query-centric and passage-centric similarity relations.

3 Methodology

In this section, we present an approach that leverages both query-centric and PAssage-centric sImilarity Relations (called PAIR) for dense passage retrieval.

3.1 Overview

The task of dense passage retrieval (Karpukhin et al., 2020) is described as follows. Given a query \( q \), we aim to retrieve \( k \) most relevant passages \( \{p_j\}_{j=1}^k \) from a large collection of \( M \) passages.

For this task, the dual-encoder architecture is widely adopted (Karpukhin et al., 2020; Qu et al., 2020), where two separate encoders \( E_Q(\cdot) \) and \( E_P(\cdot) \) are used to represent the query \( q \) and the passage \( p \) into \( d \)-dimensional vectors in different representation spaces. Then a dot product is performed to measure the similarity between \( q \) and \( p \) based on their embeddings:

\[
s(q,p) = E_Q(q)^\top \cdot E_P(p).
\]

Previous studies mainly capture the query-centric similarity relation. As shown in Figure 1, passage-centric similarity relation reflects important evidence for improving the retrieval performance. Therefore, we extend the original query-centric learning framework by leveraging the passage-centric similarity relation.

To develop our approach, we need to address the issues described in Section 1, and we consider three aspects to extend. First, we design a new loss function that considers both query-centric and passage-centric similarity relations. Second, we utilize knowledge distillation to obtain large-scale and high-quality pseudo-labeled data to capture more comprehensive similarity relations. Third, we design a two-stage training procedure to effectively learn the passage-centric similarity relation and improve the final retrieval performance.

3.2 Defining the Loss Functions

Our approach considers two kinds of losses, namely query-centric loss and passage-centric loss, as shown in Figure 2. The two kinds of losses are characterized by the two different similarity relations, query-centric similarity relation and passage-centric similarity relation.

**Query-centric Loss** The query-centric similarity relation regards the query \( q \) as the center and pushes the negative passages \( p^- \) farther than the positive passages \( p^+ \). That is:

\[
s^{(Q)}(q,p^+) \succ s^{(Q)}(q,p^-),
\]

where \( s^{(Q)}(q,p^+) \) and \( s^{(Q)}(q,p^-) \) represent the similarities for the relevant and irrelevant passages to query \( q \), and they are defined the same as \( s(q,p) \) in Eq. (1). Following (Karpukhin et al., 2020; Qu et al., 2020), we learn the query-centric similarity relation by optimizing query-centric loss that is the negative log likelihood of the positive passage:

\[
L_Q = -\frac{1}{N} \sum_{(q,p^+)} \log \frac{e^{s^{(Q)}(q,p^+)}}{e^{s^{(Q)}(q,p^+)} + \sum_{p^-} e^{s^{(Q)}(q,p^-)}}.
\]
The query-centric loss function, which is the negative log likelihood of discourse similarity relations by optimizing the passage-centric similarity relation, is given by:

\[ \mathcal{L}_Q = -\alpha \log \mathcal{E}_Q(p^+, q) + \beta \log \mathcal{E}_Q(p^+, p^-) \]

where \( \mathcal{E}_Q \) is the well-trained cross-encoder, i.e., \( \mathcal{E}_Q(p^+, q) = \mathcal{E}_Q(p^+, p^-) \) when \( p^- \) is a highly similar yet irrelevant passage.

The passage-centric similarity relation, our approach will be more capable of discriminating between a positive passage and negatives even in the existing manually labeled datasets (Qu et al., 2020), and it is likely to bring false negatives when sampling hard negatives. Hence, we propose to generate pseudo-labeled training data via knowledge distillation.

### 3.3 Generating the Pseudo-labeled Training Data via Knowledge Distillation

By optimizing both query-centric loss and passage-centric loss, we can capture more comprehensive similarity relations. However, more similarity relation constraints require large-scale and high-quality training data for optimization. Additionally, there might be a large number of unlabeled positives even in the existing manually labeled datasets (Qu et al., 2020), and it is likely to bring false negatives when sampling hard negatives. Hence, we propose to generate pseudo-labeled training data via knowledge distillation.

**Cross-encoder Teacher Model** The teacher model is used to generate large-scale pseudo-labeled data. Following RocketQA (Qu et al., 2020), we adopt the cross-encoder architecture to implement the teacher, which takes as input the concatenation of query and passage and models the semantic interaction between query and passage representations. Such an architecture has been demonstrated to be more effective than the dual-encoder architecture in characterizing query-passage relevance (Yang et al., 2020). We follow Qu et al. (2020) to train the cross-encoder teacher with the labeled data.

**Generating Pseudo Labels** In this paper, we follow Qu et al. (2020) to obtain positives and hard negatives\(^2\) for unlabeled queries\(^3\). First, we retrieve the top-\(k\) candidate passages of unlabeled queries from the corpus by an efficient retriever DPR (Karpukhin et al., 2020), and score them by the well-trained cross-encoder (i.e., teacher model). We set two values \( s_{pos} \) and \( s_{neg} \) (\( s_{pos} > s_{neg} \)) as the positive and hard negative thresholds, for queries and passages, respectively. In this case, different encoders may project queries and passages into two different spaces. However, to simultaneously model the query-centric similarity relation and the passage-centric similarity relation, the representations of queries and passages should be in the same space. Otherwise, the similarity between passages and the similarity between queries and passages are not comparable. Therefore, we propose using the encoders that share the same parameters and structures for both queries and passages, i.e., \( E_Q(\cdot) = E_P(\cdot) \).

---

\(^2\)Xiong et al. (2020a) and Karpukhin et al. (2020) demonstrate the importance of hard negatives.

\(^3\)We obtain easy negatives from in-batch sampling.
respectively. Then, given each query, a candidate passage with a score above $s_{pos}$ or below $s_{neg}$ will be considered as positive or negative. Note that we also apply this on labeled corpus to obtain more positives and reliable hard negatives. Because there might be a large number of unlabeled positives even in the existing manually labeled datasets (Qu et al., 2020) and it is likely to bring false negatives in hard negative sampling.

### 3.4 Two-stage Training Procedure

Although passage-centric similarity relation (Eq. (5)) is able to incorporate additional relevance evidence, it is not directly related to the final task goal (i.e., query-centric similarity relation). Therefore, we design a two-stage training procedure that incorporates the passage-centric loss in the pre-training stage, and then only optimize the tasks-specific loss (i.e., query-centric loss) in the fine-tuning stage. We present an illustration for the two-stage training procedure in Figure 3. Next, we present the detailed training procedure.

**Pre-training**  In the pre-training stage, we train the dual-encoder by optimizing the loss function $L$ in Eq. (6) (i.e., a combination of query-centric loss and passage-centric loss). The pseudo-labeled data from unlabeled corpus is adopted as the pre-training data (Section 3.3).

**Fine-tuning**  In the fine-tuning stage, we only fine-tune the dual-encoder (pre-trained in the first stage) according to the query-centric loss $L_Q$ in Eq. (3). In this way, our approach focuses on learning the task-specific loss, yielding better retrieval performance. In this stage, we use both ground-truth labels and pseudo labels derived from the labeled corpus for training.

### 4 Experiments

In this section, we first describe the experimental settings, then report the main experimental results, ablation study and detailed analysis.
Table 2: Experimental results on MSMARCO and Natural Questions datasets. Note that we copy the results from original papers and we leave it blank if the original paper does not report the result.

| Methods                          | PLM       | MSMARCO Dev | Natural Questions Test |
|----------------------------------|-----------|-------------|------------------------|
|                                  |           | MRR@10  | R@50  | R@1000 | R@5 | R@20 | R@100 |
| BM25 (anserini) (Yang et al., 2017) | -         | 18.7   | 59.2  | 85.7   | -   | 59.1 | 73.7  |
| doc2query (Nogueira et al., 2019b) | -         | 21.5   | 64.4  | 89.1   | -   | -    | -     |
| DeepCT (Dai and Callan, 2019)    | -         | 24.3   | 69.0  | 91.0   | -   | -    | -     |
| docTTTTTquery (Nogueira et al., 2019a) | -         | 27.7   | 75.6  | 94.7   | -   | -    | -     |
| GAR (Mao et al., 2020)           | -         | -      | -     | -      | -   | 74.4 | 85.3  |
| DPR (single) (Karpukhin et al., 2020) | BERTbase | -      | -     | -      | -   | 78.4 | 85.4  |
| DPR-E                            | ERNIEbase | 32.5   | 82.2  | 97.3   | 68.4 | 80.7 | 87.3  |
| ANCE (single) (Xiong et al., 2020a) | RoBERTa_base | 33.0  | -     | 95.9   | -   | 81.9 | 87.5  |
| ME-BERT (Luan et al., 2021)      | BERTlarge | 34.3   | -     | -      | -   | -    | -     |
| NPRINC (Lu et al., 2020)         | BERTbase  | 31.1   | -     | 97.7   | 73.3 | 82.8 | 88.4  |
| ColBERT (Khattab and Zaharia, 2020) | BERTbase | 36.0   | 82.9  | 96.8   | -   | -    | -     |
| RocketQA (Qu et al., 2020)       | ERNIEbase | 37.0   | 85.5  | 97.9   | 74.0 | 82.7 | 88.5  |
| PAIR (Ours)                      | ERNIEbase | 37.9   | 86.4  | 98.2   | 74.9 | 83.5 | 89.1  |

4http://answers.yahoo.com/
Table 3: The ablation study and controlled experiments of different variants of PAIR on Natural Questions.

| Methods   | R@5  | R@20 | R@100 |
|-----------|------|------|-------|
| Complete (PAIR) | 74.9 | 83.5 | 89.1  |
| w/o PSR  | 73.6 | 83.3 | 88.8  |
| w/o KD  | 70.9 | 82.7 | 88.1  |
| w/ PSR FT | 74.6 | 83.4 | 89.0  |
| w/o SP  | 74.0 | 83.4 | 88.9  |
| w/o PT  | 73.0 | 82.8 | 88.5  |

Instead of BERT, to examine the effects of pre-trained LMs.

Table 2 presents the main experimental results.

(1) We can see that PAIR significantly outperforms all the baselines on both MSMARCO and NQ datasets. The major difference between our approach and baselines lies in that we incorporate both query-centric and passage-centric similarity relations, which can capture more comprehensive semantic relations. Meanwhile, we incorporate the augmented data via knowledge distillation.

(2) We notice that baseline methods use different pre-trained LMs, as shown in the second column of Table 2. In PAIR, we use the ERNIE-base. To examine the effects of ERNIE-base, we implement DPR-E by replacing BERT-base used in DPR as ERNIE-base. From Table 2, we can observe that PAIR significantly outperforms DPR-E, although they employ the same pre-trained LM.

(3) Another observation is that the dense retrievers are overall better than the sparse retrievers. Such a finding has also been reported in prior studies (Karpukhin et al., 2020; Xiong et al., 2020a; Luan et al., 2021), which indicates the effectiveness of the dense retrieval approach.

4.4 Ablation Study

In this section, we conduct ablation study to examine the effectiveness of each strategy in our proposed approach. We only report the results on the NQ, while the results on the MSMARCO are similar and omitted here due to limited space.

Here, we consider five variants based on our approach for comparison:

(a) w/o PSR removes the loss for passage-centric similarity relation in the pre-training stage;

(b) w/o KD removes the knowledge distillation for obtaining pseudo-labeled data and only uses the labeled data (MSMARCO and NQ) for both pre-training stage and fine-tuning stage;

(c) w/ PSR FT adds the loss for passage-centric similarity relation in the fine-tuning stage;

(d) w/o SP uses separate encoders for queries and passages instead of encoders with shared parameters;

(e) w/o PT removes the pre-training stage.

Table 3 presents the results for the ablation study. We can observe the following findings:

- The performance drops in w/o PSR, demonstrating the effectiveness of learning passage-centric similarity relation;

- The performance drops in w/o KD, demonstrating the necessity and effectiveness of the knowledge distillation for obtaining large-scale and high-quality pseudo-labeled data, since the passage-centric loss tries to distinguish highly similar but semantically different passages;

- The performance slightly drops in w/ PSR FT, because passage-centric loss is not directly related to the target task (i.e., query-based retrieval), which suggests that passage-centric loss should be only used in the pre-training stage;

- The performance drops in w/o SP, demonstrating the effectiveness of dual-encoders with shared parameters;

- The performance significantly drops in w/o PT, demonstrating the importance of our pre-training procedure.

4.5 Analysis on Passage-centric Similarity Relation

The previous results demonstrate the effectiveness of our proposed approach PAIR. Here, we further analyze the effect of passage-centric loss (Eq. (5)) in a more intuitive way. To examine this, we prepare two variants of our approach,
Table 4: The comparison of the top-1 passages retrieved by PAIR and PAIR¬PSR, respectively. The bold words represent the main topics in queries and passages. The italic words with wavy underline are the right answers. The words with straight underline among passages have many words in common and may mislead the model PAIR¬PSR to select the wrong passage.

Table 5: The data quality and retrieval performance in different thresholds on NQ. Acc_{pos} denotes accuracy of positives and Acc_{neg} denotes accuracy of negatives.

4.6 Analysis on Knowledge Distillation

In this section, we examine the influence of the thresholds on pseudo-labeled data via knowledge distillation, including the data quality and the retrieval performance. We conduct the analyses by using different positive thresholds $s_{pos}$ and negative thresholds $s_{neg}$ (See Section 3.3).

We first manually evaluate the quality of the pseudo-labeled data via knowledge distillation w.r.t. different threshold settings (i.e., the combinations of $s_{neg}$ and $s_{pos}$). For each threshold setting, we randomly select 100 queries, each of which corresponding to a positive passage and a hard-negative passage. In total, we have 4 threshold settings (as shown in Table 5) and 800 query-passage pairs. We ask two experts to manually annotate the query-passage pairs and evaluate the quality of pseudo-labeled data, the Cohen’s Kappa of experts is 0.9. As shown in the first two columns
of Table 5, we can observe that when $s_{pos} = 0.9$ and $s_{neg} = 0.1$, the data quality is relatively good. Additionally, when setting a low value of $s_{pos}$ and a high value of $s_{neg}$, the data quality becomes worse.

The last three columns of Table 5 also present the retrieval performance w.r.t. different threshold settings. When choosing a low value of $s_{pos}$ and a high value of $s_{neg}$, the retrieval performance drops. Hence, our approach is configured with a strict threshold setting ($s_{pos} = 0.9$, $s_{neg} = 0.1$) in experiments to achieve good performance.

5 Conclusion and Future Work

This paper presented a novel dense passage retrieval approach that leverages both query-centric and passage-centric similarity relations for capturing more comprehensive semantic relations. To implement our approach, we made three important technical contributions in the loss formulation, training data augmentation and effective training procedure. Extensive results demonstrated the effectiveness of our approach. To our knowledge, it is the first time that passage-centric similarity relation has been considered for dense passage retrieval. We believe such an idea itself is worth exploring in designing new ranking mechanism. In future work, we will design more principle ranking functions and apply current retrieval approach to downstream tasks such as question answering and passage re-ranking.

6 Ethical Impact

The technique of dense passage retrieval is effective for question answering, where the majority of questions are informational queries. Semantic crowdedness problem of passages, and term mismatch between questions and passages are typical problems, which bring barriers for the machine to accurately find the information. Our technique contributes toward the goal of asking machines to find the answer passages to natural language questions from a large collection of documents. With these advantages also come potential downsides: Wikipedia or any potential external knowledge source will probably never fully cover the breadth of user questions. The goal is still far from being achieved, and more efforts from the community is needed for us to get there.

Acknowledgement

This work is partially supported by the National Key Research and Development Project of China (No.2018AA0101900), National Natural Science Foundation of China under Grant No. 61872369 and Beijing Outstanding Young Scientist Program under Grant No. BJJWZJH012019100020098.

References

Héctor Martínez Alonso and Barbara Plank. 2017. When is multitask learning effective? semantic sequence prediction under varying data conditions. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 1: Long Papers, pages 44–53.

Kevin Clark, Minh-Thang Luong, Urvashi Khandelwal, Christopher D. Manning, and Quoc V. Le. 2019. Bam! born-again multi-task networks for natural language understanding. In Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28-August 2, 2019, Volume 1: Long Papers, pages 5931–5937.

Nick Craswell, Daniel Campos, Bhaskar Mitra, Eminine Yilmaz, and Bodo Billerbeck. 2020. ORCAS: 20 million clicked query-document pairs for analyzing search. In CIKM ’20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020, pages 2983–2989.

Zhuyun Dai and Jamie Callan. 2019. Deeper text understanding for IR with contextual neural language modeling. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019, pages 985–988.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pages 4171–4186.

Daniel Gillick, Sayali Kulkarni, Larry Lansing, Alessandro Presta, Jason Baldrige, Eugene Ie, and Diego Garcia-Olano. 2019. Learning dense representations for entity retrieval. In Proceedings of the 23rd Conference on Computational Natural Language Learning, CoNLL 2019, Hong Kong, China, November 3-4, 2019, pages 528–537.
Matthew L. Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruigi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *CoRR*, abs/1705.00652.

Gautier Izacard and Edouard Grave. 2020. Distilling knowledge from reader to retriever for question answering. *CoRR*, abs/2012.04584.

Zongcheng Ji, Zhengdong Lu, and Hang Li. 2014. An information retrieval approach to short text conversation. *CoRR*, abs/1408.6988.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1601–1611.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. 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 2020, Online, November 16-20, 2020*, pages 6769–6781.

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 2020, Virtual Event, China, July 25-30, 2020*, pages 39–48.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Lilian Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Trans. Assoc. Comput. Linguistics*, 7:452–466.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 6086–6096.

Jing Lu, Gustavo Hernández Ábrego, Ji Ma, Jianmo Ni, and Yinfei Yang. 2020. Neural passage retrieval with improved negative contrast. *CoRR*, abs/2010.12523.

Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. 2021. Sparse, dense, and attentional representations for text retrieval. *Transactions of the Association for Computational Linguistics*, 9:329–345.

Yanjun Ma, Dianhai Yu, Tian Wu, and Haifeng Wang. 2019. Paddlepaddle: An open-source deep learning platform from industrial practice. *Frontiers of Data Domputing*, 1(1):105–115.

Yuning Mao, Pengcheng He, Xiaodong Liu, Ye-long Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. 2020. Generation-augmented retrieval for open-domain question answering. *CoRR*, abs/2009.08553.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering. *CoRR*, abs/1806.08730.

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.

Rodrigo Nogueira, Jimmy Lin, and AI Epistemic. 2019a. From doc2query to doctttquery. *Online preprint*.

Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019b. Document expansion by query prediction. *CoRR*, abs/1904.08375.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyian Ren, Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2020. Rocketqa: An optimized training approach to dense passage retrieval for open-domain question answering. *CoRR*, abs/2010.08191.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100, 000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2383–2392.

Yu Sun, Shuohuan Wang, Yu-Kun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. ERNIE 2.0: A continual pre-training framework for language understanding. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020*, New York, NY, USA, February 7-12, 2020, pages 8968–8975.
Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable zero-shot entity linking with dense entity retrieval. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020, pages 6397–6407.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020a. Approximate nearest neighbor negative contrastive learning for dense text retrieval. CoRR, abs/2007.00808.

Wenhan Xiong, Xiang Lorraine Li, Srinivasan Iyer, Jingfei Du, Patrick S. H. Lewis, William Yang Wang, Yashar Mehdad, Wen-tau Yih, Sebastian Riedel, Douwe Kiela, and Barlas Oguz. 2020b. Answering complex open-domain questions with multi-hop dense retrieval. CoRR, abs/2009.12756.

Peilin Yang, Hui Fang, and Jimmy Lin. 2017. Anserini: Enabling the use of lucene for information retrieval research. In Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017, pages 1253–1256.

Yinfei Yang, Ning Jin, Kuo Lin, Mandy Guo, and Daniel Cer. 2020. Neural retrieval for question answering with cross-attention supervised data augmentation. CoRR, abs/2009.13815.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018, pages 2369–2380.

Yang You, Jing Li, Sashank J. Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. 2020. Large batch optimization for deep learning: Training BERT in 76 minutes. In 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020.

Wayne Xin Zhao, Jing Liu, Ruiyang Ren, and Ji-Rong Wen. 2022. Dense text retrieval based on pre-trained language models: A survey. arXiv preprint arXiv:2211.14876.