Representation Decoupling for Open-Domain Passage Retrieval

Bohong Wu, Zhuosheng Zhang, Jinyuan Wang, Hai Zhao

Department of Computer Science and Engineering, Shanghai Jiao Tong University
{chengzhipanpan, zhangzs, steve_wang}@sjtu.edu.cn, zhaohai@cs.sjtu.edu.cn

Abstract

Training dense passage representations via contrastive learning (CL) has been shown effective for Open-Domain Passage Retrieval (ODPR). Recent studies mainly focus on optimizing this CL framework by improving the sampling strategy or extra pretraining. Different from previous studies, this work devotes itself to investigating the influence of conflicts in the widely used CL strategy in ODPR, motivated by our observation that a passage can be organized by multiple semantically different sentences, thus modeling such a passage as a unified dense vector is not optimal. We call such conflicts Contrastive Conflicts. In this work, we propose to solve it with a representation decoupling method, by decoupling the passage representations into contextual sentence-level ones, and design specific CL strategies to mediate these conflicts. Experiments on widely used datasets including Natural Questions, Trivia QA, and SQuAD verify the effectiveness of our method, especially on the dataset where the conflicting problem is severe. Our method also presents good transferability across the datasets, which further supports our idea of mediating Contrastive Conflicts.

1 Introduction

Compared to Close-Domain Question Answering tasks which ask models to give prediction when the corresponding contexts are given, Open-Domain Question Answering (ODQA) is more challenging, which is computationally the largest natural language understanding (NLU) task. Open-Domain Question Answering (ODQA) is more challenging, which is computationally the largest natural language understanding (NLU) task. Therefore, as the core component of a preliminary step of ODQA, Open-Domain Passage Retrieval (ODPR) is attracting the attention of researchers for its widely usage both academically and industrially. ODPR aims to retrieve the most relevant passages as evidence pieces of a given question, with the requirements of both retrieval recall for QA quality, and low latency for faster response.

Traditional retrieval approaches mainly focus on the lexical matching, including TF-IDF and BM25. These methods neglect non-lexical semantic similarity, resulting in suboptimal retrieval performance. Researchers recently are trying to build dense representations for both questions and Wikipedia passages using Bi-Encoders. Bi-Encoders encode the passages and questions separately, and retrieve evidence passages using similarity function like the inner product or cosine similarity. Given that the representations of Wikipedia passages could be precomputed, the retrieval speed of the Bi-Encoder architecture could be on par with lexical approaches, meeting the industrial requirement.

For Bi-Encoder based methods, to train a fine-grained dense representation for both passages and questions, previous approaches mainly focus on pretraining the Bi-Encoders...
with a specially designed pretraining objective, Inverse Cloze Task (ICT), firstly proposed by ORQA [Lee, Chang, and Toutanova 2019]. More recently, DPR [Karpukhin et al. 2020] adopts a simple but effective contrastive learning (CL) framework, achieving impressive performance without any pretraining. Concretely, for each question \( q \), several positive passages \( p^+ \) and hard negative passages \( p^- \) produced by BM25 are pre-extracted. By feeding the Bi-Encoder with \( (q, p^+, p^-) \) triples, DPR simultaneously maximizes the similarity between the representation of a given question \( q \) and the corresponding gold passages \( p^+ \), and minimizes the similarity between the representations of this question \( q \) and all other hard negative passages \( p^- \). Following this training paradigm, many researchers are seeking further improvements for DPR from the perspective of sampling strategy [Xiong et al. 2020; Lu et al. 2020; Tang et al. 2021; Qu et al. 2021] or extra pretraining (Sachan et al. 2021).

However, these studies fail to realize that there exist severe drawbacks in current CL framework adopted by DPR. Essentially, as illustrated in Figure 1 each passage \( p \) is composed of multiple sentences, upon which multiple semantically faraway questions can be derived, which forms a question set \( Q = \{q_1, q_2, ..., q_k\} \). Under our investigation, such a one-to-many problem is causing severe conflicting problems in current CL framework, which we refer to as Contrastive Conflicts. To the best of our knowledge, this is the first work that formally studies the conflicting problems in the CL framework of ODPR. Here, we distinguish two Contrastive Conflicts problems.

- **Transitivity of Similarity** The goal of CL framework in DPR is to maximize the similarity between the question representation and gold passage representation. As illustrated in Figure 2 under Contrastive Conflicts, the current CL framework will unintentionally maximize the similarity between different question representations derived from the same passage, even if they might be semantically faraway, confusing the training of the question encoder, which is exactly the cause of low performance on SQuAD [Rajpurkar et al. 2016] for DPR (SQuAD has an average of 2.66 questions derived from each passage).

- **Multiple References in Large Batch Size** According to [Karpukhin et al. 2020], the performance of DPR highly benefits from large batch size in the CL framework. However, under Contrastive Conflicts, one passage could be the positive passage \( p^+ \) of multiple questions (i.e. the question set \( Q \)). Therefore, a large batch size will increase the probability that some questions of \( Q \) might occur in the same batch. With the widely adopted in-batch negative technique [Karpukhin et al. 2020; Lee et al. 2021], such \( p^+ \) will be simultaneously referred to as both the positive sample and the negative sample, making the training procedure unstable.

Therefore, this paper adopts a decoupling strategy that effectively breaks down dense passage representations into sentence-level ones, which we refer to as Dense Contextual Sentence Representation (DCSR). Since it’s hard to derive semantically faraway questions from one single sentence, we decouple one passage representation into several contextual sentence representations. We don’t simply break down passages into sentence pieces, as this will lose the contextual information. Instead, we insert several special indicator tokens “\(<sent>>” at the sentence boundaries, and use random sampling to create positive and negative samples based on these indicator tokens. We note that as we are modeling the retrieval approach in a totally different granularity from DPR, simply using random sentences in BM25 negative passages is not optimal for DCSR, resulting into easy negatives. Therefore, we further introduce in-passage negative sampling strategy, which samples neighbouring sentences of the positive one in the same passage, to create hard negative samples. Finally, concrete experiments have verified the effectiveness of our DCSR from both retrieval recall and transferability, especially on the datasets where the Contrastive Conflicts is the most severe (i.e. SQuAD).

Our main contributions in this paper can be summarized as follows.

- We investigate the defects of current CL framework in training dense passage representation in Open-Domain Passage Retrieval.
- To mediate the Contrastive Conflicts problem, we propose to index the Wikipedia corpus using contextual sentences instead of passages by inserting indicator tokens “\(<sent>>” at the boundaries of sentences. We also design a specific CL strategy in training the contextual sentence representations.
- Experiments show that our DCSR significantly outperforms the original DPR design, especially on datasets where the Contrastive Conflicts is the most severe. Also, our DCSR enjoys better transferability than the original DPR, indicating that our DCSR better captures the universality of the concerned task datasets.

## 2 Related Work

### 2.1 Open-Domain Passage Retrieval

As the core component of Open-Domain Question Answering, Open-Domain Passage Retrieval has been a hot research topic. It requires a system to extract evidence passages for a specific question from a large passage corpus like Wikipedia, which is closer to the real scenario in daily life. It is challenging as it requires both high retrieval accuracy and specifically low latency for practical usage. Traditional approaches like TF-IDF [Ramos et al. 2003], BM25...
Contrastive learning (CL) recently is attracting researchers’ attention in all area. After witnessing its superiority in Computer Vision tasks (Chen et al. 2020; He et al. 2020), researchers in NLP are also applying this techniques (Wu et al. 2020; Karpukhin et al. 2020; Yan et al. 2021; Giorgi et al. 2021; Gao, Yao, and Chen 2021). For the concern of ODPR, the research lines of CL can be divided into two types: (i) Improving the sampling strategies for positive samples and hard negative samples. According to (Manmatha et al. 2017), the quality of positive samples and negative samples are of vital importance in the CL framework. Therefore, many researchers seek better sampling strategies to improve the retrieval performance (Xiong et al. 2020), (ii) Improving the CL framework. DensePhrase (Lee et al. 2021) uses memory bank like MOCO (He et al. 2020) to increase the number of in-batch negative samples without increasing the GPU memory usage, as the experimental analysis in DPR has already revealed the importance of training batch size. Larger batch size indicates larger number of in-batch negative samples, which is helpful for the Bi-Encoder to learn a fine-grained passage representation. Also, as DensePhrase aims to produce phrase-based answers in an end-to-end way, its retrieval processing is done on the basis of phrase level but not passage level, which directly alters the granularity of the CL framework.

Our DCSR follows the second research line. We investigate a special phenomenon, Contrastive Conflicts in the CL framework, and experimentally verify the effectiveness of mediating such conflicts by a decoupling method.

3 Methods

3.1 Contrastive Conflicts in DPR

Existing CL framework aims to maximize the similarity between the representations of each question and its corresponding gold passages.

Suppose there is a batch of $n$ questions, $n$ corresponding gold passages and in total $k$ hard negative passages. Denote the questions in batch as $q_1, q_2, ... q_n$, their corresponding gold passages as $gp_1, gp_2, ... , gp_n$, and hard negative passages as $np_1, np_2, ... , np_k$. Two separate PLMs are first used separately to acquire representations for questions and passages $\{h_{q_1}, h_{gp_1}, h_{gp_2}, ... , h_{np_1}, h_{np_2}, ... \}$. The training objective for each question sample $q_i$ of original DPR is shown in Eq (1):

$$L(q_i, gp_1, \cdots, gp_n, np_1, \cdots, np_k) = -\log \frac{e^{\text{sim}(h_{q_i}, h_{gp_j})}}{\sum_{j=1}^{n} e^{\text{sim}(h_{q_i}, h_{gp_j})} + \sum_{j=1}^{k} e^{\text{sim}(h_{q_i}, h_{np_j})}} \tag{1}$$

The sim($\cdot$) could be any similarity operator that calculates the similarity between the question representation $h_{q_i}$ and the passage representation $h_{np_j}$.

Minimizing the objective in Eq (1) is the same as (i) maximizing the similarity between each $h_{q_i}$ and $h_{gp_j}$, pair, and (ii) minimizing the similarity between $h_{q_i}$ and all other $h_{gp_j}$, $i \neq j$ and $h_{np_j}$. As discussed above, this training paradigm will cause conflicts under current contrastive learning framework. In the real scenario, multiple questions may be derived from the same gold passage. This is causing two kinds of conflicts in training. (i) The similarity between two semantically faraway questions may be unintendedly maximized via same gold passage, as both questions are expected to have high similarity with the same passage (transitivity of similarity function). (ii) If the questions $q_1, q_2, ... , q_i$ derived from the same passage $p_j$ appears in the same training batch (this is highly possible as large batch size is desired in the contrastive learning framework), such a passage will be referred as both positive sample and negative sample for each question, making the training procedure of the Bi-Encoder unstable.

3.2 Passage Representation Decoupling

The cause of the Contrastive Conflicts lies in the fact that each passage is always composed of several sentences, while these sentences may not always stick to the same topic, as depicted in Figure 1. Therefore, we propose to decouple the passage representation into contextual sentence-level ones for better retrieval performance.

Since contextual information is also important in passage retrieval, simply breaking down passages into sentences and indexing them independently is infeasible. Instead, following (Beltagy, Peters, and Cohan 2020; Lee et al. 2020; Wu, ...
In-Passage Negatives

To handle the circumstance where batch negatives as additional easy negatives. (Karpukhin et al. 2020; Lee et al. 2021), we introduce in-passage negatives to maximize the difference between contextual sentences representations within the same passage. Concretely, we randomly sample one sentence that is at least one sentence away from the gold sentence that contains answer, to ensure that the correlation between the negative and positive sentence is minimized (i.e. a random sentence from $P / \{P_{s^+_1}, P_{s^+_2}, P_{s^+_{n+1}}\}$). Note that a positive passage might not contain such sentence. If it doesn’t exist, this in-passage negative sentence is substituted by another easy negative sentence from the corresponding BM25 negative passage (a random sentence from $N$). These in-passage negatives function as hard negative samples in our CL framework.

3.3 Retrieval

For retrieval, we first use FAISS (Johnson, Douze, and Jégou 2019) to calculate the matching scores between the question and all the contextual sentence indexes. As one passage have multiple keys in the indexes, we retrieve top $100 \times 4$ (4 is the average number of sentences per passage) contextual sentences for inference. To change these sentence-level scores into passage-level ones, we adopt a probabilistic design for ranking passages, which we refer to as Score Normalization.

**Score Normalization**

After getting the scores for each contextual sentences to each question by FAISS, we first use a Softmax operation to normalize all these similarity scores into passage-level probabilities. Suppose one passage $P$ with several contextual sentences $s_1, s_2, \ldots, s_n$, and probability for each sentence $p_{s_1}, p_{s_2}, \ldots, p_{s_n}$, therefore, we can calculate the probability that the answer is in passage $P$ by Equation (2)

$$HasAns(P) = 1 - \prod_{i=1}^{n}(1 - p_{s_i})$$

We then rank all the retrieved passages by $HasAns(P)$, and select the top 100 passages for evaluation and comparison in our following experiments.
Different from other frontier researches which mainly devote themselves either in investigating better negative sampling strategy, like ANCE [Xiong et al. 2020], NPRINC [Lu et al. 2020], etc., or in extra pretraining [Sachan et al. 2021], our DCSR directly optimizes the Bi-Encoder architecture adopted in DPR. Therefore, our DCSR could be naturally incorporated with these researches and achieve better results further. Due to computational resource limitation, we do not intend to replicate these frontier researches.

4.3 Ablation Study

To illustrate the efficacy of the previously proposed negative sampling strategy, we performed an ablation study on a subset of OpenQA Wikipedia corpus\footnote{Because evaluating on the whole Wikipedia corpus takes too much resource and time (over 1 day per experiment per dataset).}. We sample 1/20 of the whole corpus, which results in a collection of 1.05 million passages in total. As reference, we rerun the original DPR code and also list their results in the table. We compare the following negative sampling strategies of our DCSR.

\begin{itemize}
  \item \textbf{+ 1 BM25 random} In this setting, we randomly sample (i) one gold sentence from the positive passage as the positive sample, and (ii) one negative sentence from the negative passage as the negative sample per question.
  \item \textbf{+ 2 BM25 random} In this setting, we randomly sample (i) one gold sentence from the positive passage as the positive sample, and (ii) two negative sentences from two different negative passages as two negative samples per question.
  \item \textbf{+ 1 in-passage & + 1 BM25 random} To encourage one passage to generate diverse sentence representations, we randomly sample (i) one gold sentence from the positive passage as the positive sample, (ii) one negative sentence from the positive passage as the first negative sample, and (iii) one negative sentence from the negative passage as the second negative sample per question.
\end{itemize}

The results are shown in Table \ref{tab:ablation} (i) Under the circumstance where only 1.05 million passages are indexed, variants of our DCSR generally perform significantly better than the original DPR implementation, especially on NQ dataset (over 1% improvement on both Top-20 and Top-100) and SQuAD dataset (8.0% improvement on Top-20 and 4.9% improvement on Top-100), which verifies the effectiveness of solving the \textit{Contrastive Conflicts} problem. (ii) Further, we found that increasing the number of negative samples helps little, but even introduces slight performance degradation on several metrics. (iii) The in-passage negative sampling strategy consistently helps in boosting the performance of nearly all datasets on all metrics.

We also implement “+nq-adv-train” on NQ dataset as an example to verify that our DCSR is compatible with sampling methods. Here, “+ nq-adv-train” is a simple negative sampling strategy provided in the DPR repo\footnote{https://github.com/facebookresearch/DPR} which uses the previous best retriever checkpoint to retrieve a set of new semantically hard negative passages that BM25 cannot provide. We directly use this augmented dataset provided by DPR and train our DCSR. We remark that this augmented dataset is still sub-optimal for DCSR, as these

\begin{table}[h]
\centering
\begin{tabular}{l|cccc}
\hline
& 1 & 2 & 3 & \geq 4 \\
\hline
SQuAD & 8,482 & 6,065 & 5,013 & 6,754 \quad 2.66 \\
Trivia & 43,401 & 5,308 & 1,206 & 587 \quad 1.20 \\
NQ & 32,158 & 4,971 & 1,670 & 1,871 \quad 1.45 \\
\hline
\end{tabular}
\caption{Occurrence of \textit{Contrastive Conflicts} in training sets.}
\end{table}

4 Experiments

4.1 Datasets

\textbf{OpenQA Dataset} OpenQA \cite{lee2019openqa} collects over 21 million 100-token passages from Wikipedia to simulate the open-domain passage corpus. OpenQA also collects question-answer pairs from existing datasets, including SQuAD \cite{rajpurkar2016squad}, TriviaQA \cite{joshi2017triviaqa}, Natural Questions \cite{kwiatkowski2019natural}, WebQuestions \cite{berant2013webquestions} and TRECQA \cite{baudis2015trec}. We test our DCSR on SQuAD, TriviaQA and NQ. Details of these three datasets are shown in Table \ref{tab:dataset}.

For previously concerned \textit{Contrastive Conflicts} problem, we also analyze the existence frequency of the conflicting phenomenon for each dataset. We count the number of questions for each passage, i.e., the times that this passage is referred to as the positive sample. The corresponding results are shown in Table \ref{tab:occurrence}. From this table, we can see that of all three datasets we choose, SQuAD is most severely affected where only 1.05 million passages are indexed, variants of our DCSR generally perform significantly better than the original DPR implementation, especially on NQ dataset (over 1% improvement on both Top-20 and Top-100) and SQuAD dataset (8.0% improvement on Top-20 and 4.9% improvement on Top-100), which verifies the effectiveness of solving the \textit{Contrastive Conflicts} problem. (ii) Further, we found that increasing the number of negative samples helps little, but even introduces slight performance degradation on several metrics. (iii) The in-passage negative sampling strategy consistently helps in boosting the performance of nearly all datasets on all metrics.

\begin{table}[h]
\centering
\begin{tabular}{l|cccc}
\hline
& 1 & 2 & 3 & \geq 4 \\
\hline
SQuAD & 8,482 & 6,065 & 5,013 & 6,754 \\
Trivia & 43,401 & 5,308 & 1,206 & 587 \\
NQ & 32,158 & 4,971 & 1,670 & 1,871 \\
\hline
\end{tabular}
\caption{Number of question-answer pairs in OpenQA.}
\end{table}

4.2 Main Results on Passage Retrieval

\textbf{Training Details} In our main experiments, we follow the hyperparameter setting in DPR \cite{karpukhin2020dense} to acquire comparable performance, i.e., an initial learning rate of 2e-5 for 40 epochs on each dataset. We use 8 Tesla V100 SXM2 to train the Bi-Encoder with a batch size of 16.

\textbf{Analysis} Table \ref{tab:main-results} shows our main results on OpenQA. Consistent with the core aim of this paper that DCSR solves \textit{Contrastive Conflicts}, our method achieves significantly better results on all the metrics when compared to original DPR baseline, especially on the dataset that is severely affected by \textit{Contrastive Conflicts}, i.e., SQuAD, with over 10% performance gain on the Top-20 metric, and over 7% on the Top-100 metric.

\begin{table}[h]
\centering
\begin{tabular}{l|ccc}
\hline
Datasets & Train & Dev & Test \\
\hline
SQuAD & 70,096 & 8,886 & 10,570 \\
Trivia & 60,413 & 8,837 & 11,313 \\
NQ & 58,880 & 8,757 & 3,610 \\
\hline
\end{tabular}
\caption{Table 2: Occurrence of \textit{Contrastive Conflicts} in training sets.}
\end{table}

\begin{table}[h]
\centering
\begin{tabular}{l|cccc}
\hline
Datasets & Train & Dev & Test \\
\hline
SQuAD & 70,096 & 8,886 & 10,570 \\
Trivia & 60,413 & 8,837 & 11,313 \\
NQ & 58,880 & 8,757 & 3,610 \\
\hline
\end{tabular}
\caption{Table 1: Number of question-answer pairs in OpenQA.}
\end{table}

\begin{table}[h]
\centering
\begin{tabular}{l|ccccc}
\hline
& 1 & 2 & 3 & \geq 4 & Avg \\
\hline
SQuAD & 8,482 & 6,065 & 5,013 & 6,754 & 2.66 \\
Trivia & 43,401 & 5,308 & 1,206 & 587 & 1.20 \\
NQ & 32,158 & 4,971 & 1,670 & 1,871 & 1.45 \\
\hline
\end{tabular}
\caption{Table 2: Occurrence of \textit{Contrastive Conflicts} in training sets.}
\end{table}
Table 3: Retriever Performance Comparison on the test sets. “†”: For SQuAD dataset on DPR, we are not able to reproduce the original results from the authors’ official code, as they do not provide their hyperparameter setting on SQuAD. We rerun DPR on SQuAD and report its performance based on our reproduction. The parameter settings are shared between our DPR reproduction and DCSR to ensure fairness.

Table 4: Ablation study of the retriever performance on a subset of OpenQA Wikipedia corpus, with over 1.05 million passages.

Figure 4: Average cosine similarity scores for passages with a certain number of questions on SQuAD dataset.

4.4 Similarity Analysis
As shown in Figure 4, the original DPR training paradigm unintentionally pushed the questions derived from the same passage, making the training of the question encoder defective. To verify the effectiveness of our DCSR, we cluster all the passages in SQuAD by the number of questions derived from them, and calculate the average of minimum cosine similarity per passage in each cluster. The results are shown in Figure 4 (i) The average cosine similarity keeps going down as the question number increases. This is because the more questions that are derived from the same passage, the more possible that two of these questions are semantically faraway. (ii) Our DCSR gets consistently lower average cosine similarity than the original DPR, indicating that our design has better capability in distinguishing the questions derived from the same passage.

4.5 Transferability
To further verify that our DCSR is more suitable in training dense representations, especially under the Contrastive Conflicts circumstance, we conduct experiments to test the transferability between original DPR and our DCSR. Similarly, instead of running such experiments on the entire Wikipedia corpus, we sample 1/20 of the corpus, which results in a collection of 1.05 million passages in total. We test the transferability result from SQuAD to Trivia and from NQ to Trivia, as compared to Trivia, both SQuAD and NQ suf-
error more from the Contrastive Conflicts problem. The results are shown in Table 5.

From Table 5 when compared to DPR, our model enjoys significantly better transferability. In both scenarios, DPR shows over 2% performance gap in all metrics of the transferability tests, indicating that our DCSR performs much better in generalization across the datasets. By decoupling passages into contextual sentence-level representations, our DCSR models well the universality across the datasets, and shows much better transfer capability than DPR.

5 Pros and Cons

To analyze the retrieval performance difference between DPR and DCSR, we especially focus on the different Top 1 predictions on SQuAD. We count the number of winning times for each baseline, where DCSR significantly outperforms DPR (893 vs. 161), shown in Figure 5.

**DCSR winning cases** On the question *Who was the NFL Commissioner in early 2012?*, the strengths of our DCSR are listed as follows.

- **Capability of utilizing contextual information.** The key phrase 2012 and NFL is faraway from Commissioner Roger Goodell, while our DCSR is still capable of capturing such distant contextual information.

- **Locating the exact sentence of the answer.** This is an obvious feature of DCSR, as we are modelling on the granularity of contextual sentences.

On the contrary, due to Contrastive Conflicts, the question encoder of DPR is severely affected that it cannot generate fine-grained question representation. Therefore, on this question, DPR can only find out one key phrase commissioner, falling into a totally wrong prediction.

**DCSR losing cases** On the question *Super Bowl 50 decided the NFL champion for what season?*, our DCSR has already found a contextual sentence that is very close to the given question, with several key phrases detected. However, this contextual sentence is actually a low quality index, as it suddenly reaches the end of passage. This is caused by the brute force segmentation strategy of OpenQA, which focuses on the passage level and restricts the length of each passage to 100. In this paper, we perform sentence split directly on these broken passages, which as a result breaks down many sentences into low quality indexes, affecting the final retrieval performance. We leave it for future work.

6 Conclusion

In this paper, we analyze the Contrastive Conflicts problem in current contrastive learning framework adopted in the Open-Domain Passage Retrieval area. To solve such problem, we propose Dense Contextual Sentence Representation (DCSR). Concretely, we first decouple the original passage representations into contextual sentence-level ones, and then refine the original contrastive learning framework by creating sentence-aware positive and negative samples. Our DCSR achieves significant performance gain compared to original DPR baseline, especially on datasets with severe conflicting problem. Extensive experiments shows that our DCSR also enjoys better transferability, indicating that DCSR well captures the universality in different datasets.
Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2383–2392. Austin, Texas: Association for Computational Linguistics.

Ramos, J.; et al. 2003. Using tf-idf to determine word relevance in document queries. In Proceedings of the first instructional conference on machine learning, volume 242, 29–48. Citeseer.

Robertson, S.; and Zaragoza, H. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. Foundations and Trends in Information Retrieval, 3(4): 333–389.

Sachan, D. S.; Patwary, M.; Shoeybi, M.; Kant, N.; Ping, W.; Hamilton, W. L.; and Catanzaro, B. 2021. End-to-End Training of Neural Retrievers for Open-Domain Question Answering. arXiv:2101.00408.

Tang, H.; Sun, X.; Jin, B.; Wang, J.; Zhang, F.; and Wu, W. 2021. Improving Document Representations by Generating Pseudo Query Embeddings for Dense Retrieval. arXiv:2105.03599.

Vig, J.; and Ramea, K. 2019. Comparison of transfer-learning approaches for response selection in multi-turn conversations.

Wang, B.; Yao, T.; Zhang, Q.; Xu, J.; Tian, Z.; Liu, K.; and Zhao, J. 2019. Document Gated Reader for Open-Domain Question Answering. In Piwowarski, B.; Chevalier, M.; Gaussier, É.; Maarek, Y.; Nie, J.; and Scholer, F., eds., Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019, 85–94. ACM.

Wang, S.; Yu, M.; Guo, X.; Wang, Z.; Klinger, T.; Zhang, W.; Chang, S.; Tesauro, G.; Zhou, B.; and Jiang, J. 2018. R 3: Reinforced Ranker-Reader for Open-Domain Question Answering. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 32.

Wolf, T.; Sanh, V.; Chaumond, J.; and Delangue, C. 2019. TransferTransfo: A Transfer Learning Approach for Neural Network Based Conversational Agents. arXiv:1901.08149.

Wu, B.; Zhang, Z.; and Zhao, H. 2021. Graph-free Multi-hop Reading Comprehension: A Select-to-Guide Strategy. arXiv:2107.11823.

Wu, Z.; Wang, S.; Gu, J.; Khabsa, M.; Sun, F.; and Ma, H. 2020. CLEAR: Contrastive Learning for Sentence Representation. arXiv:2012.15466.

Xiong, L.; Xiong, C.; Li, Y.; Tang, K.-F.; Liu, J.; Bennett, P. N.; Ahmed, J.; and Overwijk, A. 2020. Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. In International Conference on Learning Representations.

Yan, Y.; Li, R.; Wang, S.; Zhang, F.; Wu, W.; and Xu, W. 2021. ConSERT: A Contrastive Framework for Self-Supervised Sentence Representation Transfer. arXiv:2105.11741.

Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W.; Salakhutdinov, R.; and Manning, C. D. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2369–2380. Brussels, Belgium: Association for Computational Linguistics.