Sentence-aware Contrastive Learning for Open-Domain Passage Retrieval

Anonymous ACL submission

Abstract

Training dense passage representations via contrastive learning has been shown effective for Open-Domain Passage Retrieval (ODPR). Existing studies focus on further optimizing by improving negative sampling strategy or extra pretraining. However, these studies keep unknown in capturing passage with internal representation conflicts from improper modeling granularity. This work thus presents a refined model on the basis of a smaller granularity, contextual sentences, to alleviate the concerned conflicts. In detail, we introduce an in-passage negative sampling strategy to encourage a diverse generation of sentence representations within the same passage. Experiments on three benchmark datasets verify the efficacy of our method, especially on datasets where conflicts are severe. Extensive experiments further present good transferability of our method across datasets.

1 Introduction

Open-Domain Passage Retrieval (ODPR) has recently attracted the attention of researchers for its wide usage both academically and industrially (Lee et al., 2019; Yang et al., 2017). Provided with an extremely large text corpus that composed of millions of passages, ODPR aims to retrieve a collection of the most relevant passages as the evidences of a given question.

With recent success in pretrained language models (PrLMs) like BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), dense retrieval techniques have achieved significant better results than traditional lexical based methods, including TF-IDF (Ramos et al., 2003) and BM25 (Robertson and Zaragoza, 2009), which totally neglect semantic similarity. Thanks to the Bi-Encoder structure, dense methods (Lee et al., 2019; Guu et al., 2020; Karpukhin et al., 2020) encode the Wikipedia passages and questions separately, and retrieve evidence passages using similarity functions like the inner product or cosine similarity. Given that the representations of Wikipedia passages could be pre-computed, the retrieval speed of dense approaches could be on par with lexical ones.

Previous approaches often pretrain the Bi-Encoders with a specially designed pretraining objective, Inverse Cloze Task (ICT) (Lee et al., 2019). More recently, DPR (Karpukhin et al., 2020) adopts a simple but effective contrastive learning 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 q and corresponding p^+, and minimizes the similarity between the representations of q and all p^-. Following such contrastive learning framework, 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), or even using knowledge distillation (Izacard and Grave, 2021; Yang et al., 2021).

However, these studies fail to realize that there exist severe drawbacks in the current contrastive learning 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 $\mathcal{Q} = \{q_1, q_2, ..., q_k\}$. Under our investigation, such a *one-to-many problem* is causing severe conflicting problems in the current contrastive learning 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 contrastive learning framework of dense passage retrieval. Here, we distinguish two kinds of *Contrastive Conflicts*.

- **Transitivity of Similarity** The goal of the con-

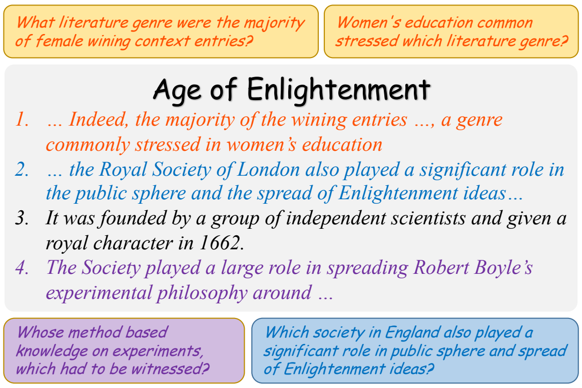

Figure 1: A sample from SQuAD. Different colors indicate the questions/sentences focus on different topics.

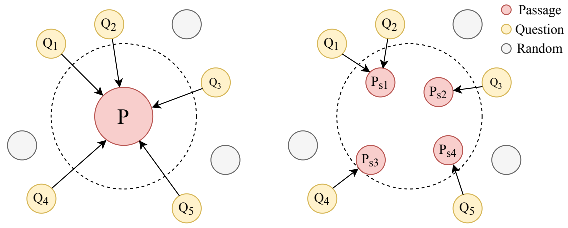

Figure 2: Visualization of contrastive conflicts in DPR (left) and solution provided by our method (right).

trastive learning framework in DPR is to maximize the similarity between the representation of the question and its corresponding gold passage. As illustrated in Figure 2, under *Contrastive Conflicts*, the current contrastive learning framework will unintendedly maximize the similarity between different question representations derived from the same passage, even if they might be semantically different, which is exactly the cause of low performance on SQuAD (Rajpurkar et al., 2016) for DPR (SQuAD has an average of 2.66 questions per 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 contrastive learning framework. However, under *Contrastive Conflicts*, one passage could be the positive passage p^+ of multiple questions (i.e. the question set $\mathcal{Q}$). Therefore, a large batch size will increase the probability that some questions of $\mathcal{Q}$ 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 for every q in $\mathcal{Q}$, which is logically unreasonable.

Since *one-to-many problem* is the direct cause of both conflicts, this paper presents a simple but effective strategy that breaks down dense passage representations into contextual sentence level ones, which we refer to as **D**ense **C**ontextual **S**entence **R**epresentation (DCSR). Unlike long passages, it is hard to derive semantically faraway questions from one short sentence. Therefore, by modeling ODPR in smaller units like contextual sentences, we fundamentally alleviate *Contrastive Conflicts* by solving the *one-to-many problem*. Note that we do not simply encode each sentence separately. Instead, we encode the passage as a whole and use sentence indicator tokens to acquire the sentence representations within the passage, to preserve the contextual information. We further introduce the in-passage negative sampling strategy, which samples neighboring sentences of the positive one in the same passage to create hard negative samples. Finally, concrete experiments have verified the effectiveness of our proposed method from both retrieval accuracy and transferability, especially on datasets where *Contrastive Conflicts* are severe.

Contributions (i) We investigate the defects of the current contrastive learning framework in training dense passage representation in Open-Domain Passage Retrieval. (ii) To handle *Contrastive Conflicts*, we propose to index the Wikipedia corpus using contextual sentences instead of passages. We also propose the in-passage negative sampling strategy in training the contextual sentence representations. (iii) Experiments show that our proposed method significantly outperforms original baseline, especially on datasets where *Contrastive Conflicts* are severe. Extensive experiments also present better transferability of our DCSR, indicating that our method captures the universality of the concerned task datasets.

2 Related Work

Open-Domain Passage Retrieval Open-Domain Passage Retrieval has been a hot research topic in recent years. It requires a system to extract evidence passages for a specific question from a large passage corpus like Wikipedia, and 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 (Robertson and Zaragoza, 2009) retrieve the evidence passages based on the lexical match between questions and passages. Although these lexical approaches meet the requirement of low latency, they fail to capture non-lexical semantic similarity, thus performing unsatisfying on retrieval accuracy.

With recent advances of pretrained language models (PrLMs) like BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), a series of neural approaches based on cross-encoders are proposed (Vig and Ramea, 2019; Wolf et al., 2019). Although enjoying satisfying retrieval accuracy, the retrieval latency is often hard to tolerate in practical use. More recently, the Bi-Encoder structure has captured the researchers' attention. With Bi-Encoder, the representations of the corpus at scale can be precomputed, enabling it to meet the requirement of low latency in passage retrieval. Lee et al. (2019) first proposes to pretrain the Bi-Encoder with *Inverse Cloze Task* (ICT). Later, DPR (Karpukhin et al., 2020) introduces a contrastive learning framework to train dense passage representation, and has achieved impressive performance on both retrieval accuracy and latency. Based on DPR, many works make further improvements either by introducing better 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), or even distilling knowledge from cross-encoders (Izacard and Grave, 2021; Yang et al., 2021).

Our method follows the contrastive learning research line of ODPR. Different from previous works that focus on either improving the quality of negative sampling or using extra pretraining, we make improvements by directly optimizing the modeling granularity with an elaborately designed contrastive learning training strategy.

Contrastive Learning Contrastive learning 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 technique (Wu et al., 2020; Karpukhin et al., 2020; Yan et al., 2021; Giorgi et al., 2021; Gao et al., 2021). For the concern of ODPR, the research lines of contrastive learning 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 contrastive learning framework. Therefore, many researchers seek better sampling strategies to improve the retrieval performance (Xiong et al., 2020). (ii) Improving the contrastive learning 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, and models retrieval process on the phrase level but not passage level, achieving impressive performance.

Our proposed method follows the second research line. We investigate a special phenomenon, *Contrastive Conflicts* in the contrastive learning framework, and experimentally verify the effectiveness of mediating such conflicts by modeling ODPR in a smaller granularity. More similar to our work, Akkalyoncu Yilmaz et al. (2019) also proposes to improve dense passage retrieval based on sentence-level evidences, but their work is not in the research line of contrastive learning, and focuses more on passage re-ranking after retrieval but not retrieval itself.

3 Methods

3.1 Contrastive Learning Framework

Existing contrastive learning 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 PrLMs are first used separately to acquire representations for questions and passages $\{h_{q_1}, h_{q_2}, ...; h_{gp_1}, h_{gp_2}, ...; h_{np_1}, h_{np_2}, ...\}$. The training objective for each question sample q_i of original DPR is shown in Eq (1):

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

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

Minimizing the objective in Eq (1) is the same as (i) maximizing the similarity between each h_{q_i} and h_{gp_i} pair, and (ii) minimizing the similarity between h_{q_i} and all other h_{gp_j} $(i \neq j)$ and h_{np_k}. As discussed previously, this training paradigm will cause conflicts under current contrastive learning

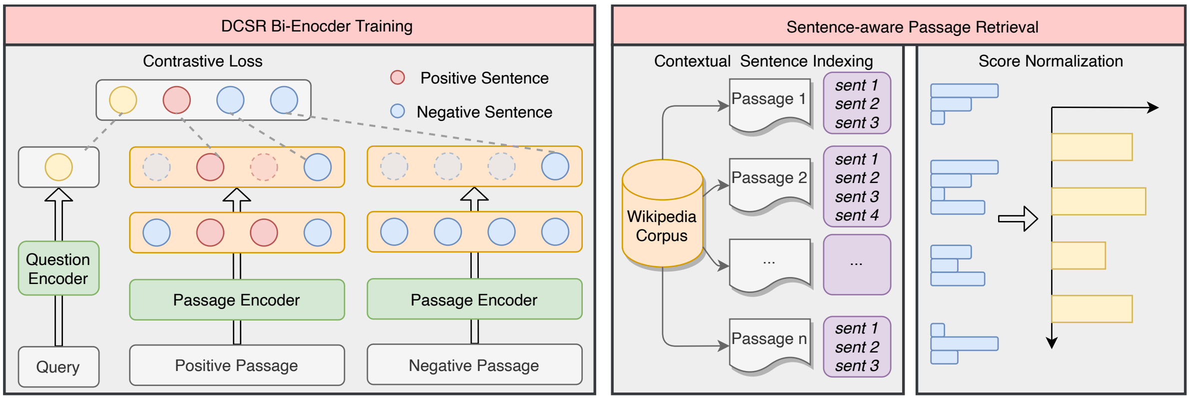

Figure 3: An illustration of our DCSR processing pipeline. The left part shows the contrastive training paradigm of our method, and the right part presents the inference pipeline.

framework due to (i) Transitivity of Similarity, and (ii) Multiple References in Large Batch Size.

3.2 Dense Contextual Sentence Representation

The cause of the *Contrastive Conflicts* lies in *one-to-many problem*, that most of the passages are often organized by multiple sentences, while these sentences may not always stick to the same topic, as depicted in Figure 1. Therefore, we propose to model passage retrieval in a smaller granularity, i.e. contextual sentences, to alleviate the occurrence of *one-to-many problem*.

Since contextual information is also important in passage retrieval, simply breaking down passages into sentences and encoding them independently is infeasible. Instead, following (Beltagy et al., 2020; Lee et al., 2020; Wu et al., 2021), we insert a special <sent> token at the sentence boundaries in each passage, and encode the passage as a whole to preserve the contextual information, which results in the following format of input for each passage:

[CLS] <sent> sent_1 <sent> sent_2 ... [SEP]

We then use BERT (Devlin et al., 2019) as encoder to get the contextual sentence representations by these indicator <sent> tokens. For convenience of illustration, taking a give query q into consideration, we denote the corresponding positive passage in the training batch as p^+, which consists of several sentences:

$$\mathcal{P} = \{p_{s_1^-}, p_{s_2^-}, ...p_{s_i^+}, ...p_{s_{k-1}^-}, p_{s_k^-}\}$$

Similarly, we denote the corresponding BM25 negative passage as:

$$\mathcal{N} = \{n_{s_1^-}, n_{s_2^-}, ...n_{s_i^-}, ...n_{s_{k-1}^-}, n_{s_k^-}\}$$

Here $(*)^{-/+}$ means whether the sentence or passage contains the gold answer. We refine the original contrastive learning framework by creating sentence-aware positive and negative samples. The whole training pipeline is shown in the left part of Figure 3.

3.2.1 Positives and Easy Negatives

Following Karpukhin et al. (2020), we use BM25 to retrieve hard negative passages for each question. To build a contrastive learning framework based on contextual sentences, we consider the sentence that contains the gold answer as the positive sentence (i.e. $p_{s_i^+}$), and randomly sample several negative sentences (random sentences from $\mathcal{N}$) from a BM25 random negative passage. Also, following (Karpukhin et al., 2020; Lee et al., 2021), we introduce in-batch negatives as additional easy negatives.

3.2.2 In-Passage Negatives

To handle the circumstance where multiple semantically faraway questions may be derived from one single passage, we hope to encourage the passage encoder to generate contextual sentence representations as diverse as possible for sentences in the same passage. Noticing that not all the sentences in the passage contain the gold answer and stick to the topic related to the given query, we further introduce in-passage negatives to maximize the difference between contextual sentences representations within the same passage. Concretely, we randomly sample one sentence that does not contain the gold answer (i.e. a random sentence from $\mathcal{P}/\{P_{s_i^+}\}$). Note that a positive passage might not contain such sentence. If it does not exist, this in-passage negative sentence is substituted by another easy negative sentence from the corresponding BM25 negative passage (a random sentence from $\mathcal{N}$). These in-passage negatives function as hard negative samples in our contrastive learning framework.

3.3 Retrieval

For retrieval, we first use FAISS (Johnson et al., 2019) to calculate the matching scores between the question and all the contextual sentence indexes. As one passage has multiple keys in the indexes, we retrieve top $100 \times k$ (k 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 probabilities. Suppose one passage $\mathcal{P}$ with several sentences $s_1, s_2, ..., s_n$, and denote the probability for each sentence that contains the answer as $p_{s_1}, p_{s_2}, ..., p_{s_n}$, we can calculate the probability that *the answer is in passage* $\mathcal{P}$ by Equation 2.

$$HasAns(\mathcal{P}) = 1 - \prod_{i=1}^{n}(1 - p_{s_i}) \quad (2)$$

We then re-rank all the retrieved passages by $HasAns(\mathcal{P})$, and select the top 100 passages for evaluation in our following experiments.

4 Experiments

4.1 Datasets

OpenQA Dataset OpenQA (Lee et al., 2019) 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 (Rajpurkar et al., 2016), TriviaQA (Joshi et al., 2017), Natural Questions (Kwiatkowski et al., 2019), WebQuestions (Berant et al., 2013) and TREC (Baudiš and Šedivỳ, 2015).

We experiment our proposed method on SQuAD, TriviaQA and NQ. For the previously concerned *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 1. From this table, we can see that of all three datasets we choose, SQuAD is most severely affected by the *Contrastive Conflicts* problem, that many passages occur multiple times as the positive passages for different questions. These statistics are consistent with the fact that DPR performs the worst on SQuAD, while acceptable on Trivia and NQ.

	1	2	3	≥ 4	Avg
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

Table 1: Occurrence of *one-to-many problem* in training sets.

4.2 Training and Implementation Details

Hyperparameters In our main experiments, we follow the hyperparameter setting in DPR (Karpukhin et al., 2020) to acquire comparable performance, i.e. an initial learning rate of 2e-5 for 40 epochs on each dataset. We use 8 Tesla V100 GPUs to train the Bi-Encoder with a batch size of 16 on each GPU.

Extra Cost Although we are modeling passage retrieval in a totally different granularity, our method adds little extra computation overhead compared to DPR. For model complexity, our proposed method adopts exactly the same model structure as DPR does, meaning that there are no additional parameters introduced. For training time, the negative sentences in our method are randomly sampled from the negative passage in DPR. Therefore, the extra time burden brought by our method is only caused by the sampling procedure, which is negligible.

Training Settings To have a comprehensive comparison with DPR, we train DCSR under three different settings. (i) *Single*, where each dataset is both trained and evaluated under their own domain. (ii) *Multi*, where we use a combination of the NQ, Trivia and SQuAD datasets to train a universal Bi-Encoder, and evaluate its performance on the test sets of all three datasets. (iii) *Adversarial Training*, which is a simple negative sampling strategy. We first use the original dataset to train a DPR or DCSR checkpoint, and use such checkpoint to acquire semantically hard negative passages from the

Model	Top-20			Top-100		
	NQ	Trivia	SQuAD	NQ	Trivia	SQuAD
Base Architecture Comparison – Single						
DPR (Karpukhin et al., 2020)	78.4	79.4	52.8†	85.4	85.0	71.0†
DCSR (Ours)	**78.9**(+0.5)	**79.7**(+0.3)	**63.7**(+10.9)	**86.5**(+1.1)	**85.2**(+0.2)	**78.1**(+7.1)
Base Architecture Comparison – Multi						
DPR (Karpukhin et al., 2020)	**79.4**	78.8	51.6	86.0	84.7	67.6
DCSR (Ours)	79.1(-0.3)	**79.6**(+0.8)	**63.8**(+12.2)	**86.6**(+0.6)	**85.2**(+0.5)	**77.6**(+10.0)

Table 2: Retriever Performance Comparison on the test sets. "†": For SQuAD dataset on DPR in the *Single* setting, we are not able to reproduce the original results from the official DPR code[1][2]. Instead, we rerun DPR on SQuAD in the *Single* setting and report its performance based on our reproduction. The parameter settings are shared between our DPR reproduction and DCSR to ensure fairness. Other statistics are taken from Karpukhin et al. (2020).

Model		Top-20		Top-100	
		NQ	Trivia	NQ	Trivia
DPR	+ adv-train	81.3	-	87.3	-
	+ ANCE (Xiong et al., 2020)	**81.9**	**80.3**	**87.5**	85.3
DCSR	+ adv-train	81.4	80.0	**87.5**	**85.7**

Table 3: Performance Comparison when incorporated with negative sampling strategy.

whole Wikipedia corpus.

4.3 Main Results on Passage Retrieval

Table 2 shows our main results on OpenQA.

For the *Single* setting, (i) Consistent with the core aim of this paper that our proposed sentence-aware contrastive learning solves *Contrastive Conflicts*, DCSR achieves significantly better results than DPR especially on the dataset that is severely affected by *Contrastive Conflicts*. For example, on the SQuAD dataset, our method achieves 10.9% performance gain on the Top-20 metric, and 7.1% performance gain on the Top-100 metric. (ii) For datasets that are less affected by *Contrastive Conflicts*, like NQ and Trivia, we still achieve slight performance gain on all metrics.

For the *Multi* setting, DPR on Trivia and SQuAD suffers from a significant performance drop compared to *Single* setting, while our model is only slightly affected. It indicates that our proposed sentence-aware contrastive learning not only solves the *Contrastive Conflicts*, but also captures the universality of datasets from different domains.

[1] Code in https://github.com/facebookresearch/DPR.

[2] It is an issue that is shared by researchers on github. More discussion about this result will be discussed in Appendix B.

4.4 Incorporated with Negative Sampling

Different from other frontier researches which mainly devote themselves either to investigating better negative sampling strategies, like ANCE (Xiong et al., 2020), NPRINC (Lu et al., 2020), etc., or to extra pretraining (Sachan et al., 2021), or to distilling knowledge from cross-encoders (Izacard and Grave, 2021; Yang et al., 2021), our proposed method directly optimizes the modeling granularity in DPR. Therefore, our method could be naturally incorporated with these researches and achieve better results further. Due to computational resource limitation, we do not intend to replicate all these methods, but use *adversarial training* as an example. Following ANCE (Xiong et al., 2020), we conduct experiments on NQ and Trivia to show the compatibility of our method, listed in Table 3. With such a simple negative sampling strategy, our DCSR achieves comparable results with its DPR counterpart.

4.5 Ablation Study

To illustrate the efficacy of the previously proposed negative sampling strategy, we conduct an ablation study on a subset of OpenQA Wikipedia corpus[1]. We sample 1/20 of the whole corpus, which results in a collection of 1.05 million passages in total. As reference, we reproduce DPR and also list their results in Table 4. We compare the following negative sampling strategies of our proposed method.

+ 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

[1] Because evaluating on the whole Wikipedia corpus takes too much resource and time (over 1 day per experiment per dataset).

Model		Top-20 NQ	Top-20 Trivia	Top-20 SQuAD	Top-100 NQ	Top-100 Trivia	Top-100 SQuAD
DPR	(Karpukhin et al., 2020)	43.7	62.1	46.5	54.0	72.4	63.6
DCSR	+ 1 BM25 random	44.5	63.1	51.1	54.5	72.9	66.6
	+ 2 BM25 random	44.0	**63.5**	50.3	54.7	72.9	65.1
	+ 1 in-passage & +1 BM25 random	**45.2**	63.4	**54.5**	**55.3**	**73.2**	**68.5**

Table 4: Ablations of Negative Sampling Strategy on Wikipedia subset (1/20 of the whole corpus) in the *Single* Setting.

sentence from the negative passage as the negative sample per question.

+ 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.

+ 1 in-passage & + 1 BM25 random In this setting, 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.

Ablations of Negative Sampling Strategy The results are shown in Table 4. (i) Under the circumstance where only 1.05 million passages are indexed, variants of our DCSR generally perform significantly better than DPR baseline, 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 *Contrastive Conflicts*. (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, especially on the SQuAD dataset, which is consistent with our motivation for in-passage negatives, which is to encourage a diverse generation of contextual sentence representations within the same passage in solving the *one-to-many problem*.

Ablations of Training Data The results are shown in Table 5. (i) We first directly use the augmented adversarial training dataset provided by DPR (marked as *DPR-hard*) and train our DCSR, having achieved even better results on the NQ dataset. This augmented dataset is sub-optimal

Model	Top-20 NQ	Top-20 Trivia	Top-100 NQ	Top-100 Trivia
$\text{DPR}_{raw-data}$	43.7	62.1	54.0	72.4
$\text{DPR}_{\textit{DPR-hard}}$	47.6	-	56.5	-
$\text{DCSR}_{\textit{DPR-hard}}$	47.6	-	57.0	-
$\text{DCSR}_{\textit{DCSR-hard}}$	**48.8**	**66.2**	**57.1**	**75.0**

Table 5: Ablations of Training Data. For Trivia, *DPR-hard* is not provided in the original paper.

for our model, as these hard negative samples are passage-specific, while our model prefers sentence-specific ones. (ii) We then use our previous best DCSR checkpoint to retrieve a set of sentence-specific hard negatives (marked as *DCSR-hard*) and train a new DCSR, which achieves further performance gain on both metrics on NQ dataset.

5 Discussion

In this section, we discuss the transferability difference and the influence of Wikipedia corpus size on both DPR and our DCSR. More discussions from different aspects are presented in the Appendices, including (i) Validation accuracy on dev sets in Appendix A, which is also a strong evidence of alleviating *Contrastive Conflicts*. (ii) Error analysis for SQuAD in Appendix B, which further shows the generalization ability of our method. (iii) Case study in Appendix C, which discusses the future improvement of DCSR.

5.1 Transferability

To further verify that our learned DCSR is more suitable in Open-Domain Passage Retrieval, especially under the *Contrastive Conflicts* circumstance, we conduct experiments to test the transferability between 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

Model	SQuAD-to-Trivia Top 20	diff	Top 100	diff	NQ-to-Trivia Top 20	diff	Top 100	diff
DPR	48.7/62.1	↓13.4	64.5/72.4	↓7.9	48.8/62.1	↓13.3	62.7/72.4	↓9.7
DCSR	54.0/63.4	↓**9.4**	67.8/73.2	↓**5.4**	52.7/63.4	↓**10.7**	65.9/73.2	↓**7.3**

Table 6: Transferability comparing our methods with DPR. We train the retriever model on the SQuAD dataset or the NQ dataset, and evaluate it on Trivia QA (statistics on the left). For reference, we also list the performance where the retriever model is both trained and evaluated on the Trivia QA (statistics on the right).

test the transferability result from SQuAD to Trivia and from NQ to Trivia, as compared to Trivia, both SQuAD and NQ suffer more from *Contrastive Conflicts*. The results are shown in Table 6.

From Table 6, 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 method performs much better in generalization across the datasets. This phenomenon once again confirms our theorem, that by modeling passage retrieval in the granularity of contextual sentences, our DCSR well models the universality across the datasets, and shows much better transferability than DPR.

Model	Top-20 NQ	Top-20 Trivia	Top-100 NQ	Top-100 Trivia	Wiki
DPR	25.5	39.4	36.7	51.9	
DCSR	**27.8**	**41.0**	**39.0**	**53.6**	0.10M
Δ	+2.3	+1.6	+2.3	+1.7	
DPR	43.7	62.1	54.0	72.4	
DCSR	**45.2**	**63.4**	**55.3**	**73.2**	1.05M
Δ	+1.5	+1.3	+1.3	+0.8	
DPR	78.4	79.4	85.4	85.0	
DCSR	**78.9**	**79.7**	**86.5**	**85.2**	21.0M
Δ	+0.5	+0.3	+1.1	+0.2	

Table 7: Retrieval performance when the size of Wikipedia Corpus is varied.

5.2 Corpus Size

In our extensive experiments, we further found out that our method can achieve overwhelming better performance than DPR on smaller corpus. In this experiment, we take *the first 0.1 million*, *the first 1.05 million* and *all passages* from the original Wikipedia corpus, and conduct dense retrieval on these three corpora varied in size. The statistic results are shown in Table 7.

From Table 7, first of all, our model achieves better performance than DPR in all settings, where such improvement is more significant in smaller corpus. On the setting where only 0.1 million passages are indexed in the corpus, our model achieves over 2.0% exact improvement on all metrics on both NQ and Trivia. We speculate this is because of the following two strengths of our method.

- The alleviation of *Contrastive Conflicts*, which we have analyzed previously.
- Modeling passage retrieval using contextual sentences enables a diverse generation of indexes. Some sentences may not be the core aim of their corresponding passages, but can still be the clue for some questions.

Secondly, we can discover that the performance gap between DPR and DCSR is decreasing when the size of Wikipedia corpus increases. This is because with the expansion of indexing corpus, many questions that cannot be solved in the small corpus setting may find much more closely related passages in the large corpus setting, which gradually neutralizes the positive effect brought by the second strength of our proposed method discussed above. Still, our model achieves better performance under the full Wikipedia setting on all datasets and all metrics.

6 Conclusion

In this paper, we make a thorough analysis on the *Contrastive Conflicts* issue in the current open-domain passage retrieval. To well address the issue, we propose an enhanced sentence-aware conflict learning method by carefully generating sentence-aware positive and negative samples. We show that the dense contextual sentence representation learned from our proposed method achieves significant performance gain compared to the original baseline, especially on datasets with severe conflicts. Extensive experiments show that our proposed method also enjoys better transferability, and well captures the universality in different datasets.

References

Zeynep Akkalyoncu Yilmaz, Wei Yang, Haotian Zhang, and Jimmy Lin. 2019. Cross-domain modeling of sentence-level evidence for document retrieval. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3490–3496, Hong Kong, China. Association for Computational Linguistics.

Petr Baudiš and Jan Šedivỳ. 2015. Modeling of the question answering task in the yodaqa system. In *International Conference of the cross-language evaluation Forum for European languages*, pages 222–228. Springer.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *ArXiv preprint*, abs/2004.05150.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on Freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544, Seattle, Washington, USA. Association for Computational Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Empirical Methods in Natural Language Processing (EMNLP)*.

John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. 2021. DeCLUTR: Deep contrastive learning for unsupervised textual representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 879–895, Online. Association for Computational Linguistics.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 3929–3938. PMLR.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9726–9735. IEEE.

Gautier Izacard and Edouard Grave. 2021. Distilling knowledge from reader to retriever for question answering. In *ICLR 2021: The Ninth International Conference on Learning Representations*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*.

Mandar Joshi, Eunsol Choi, Daniel 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 (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

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

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion 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. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Haejun Lee, Drew A. Hudson, Kangwook Lee, and Christopher D. Manning. 2020. SLM: Learning a discourse language representation with sentence unshuffling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1551–1562, Online. Association for Computational Linguistics.

Jinhyuk Lee, Mujeen Sung, Jaewoo Kang, and Danqi Chen. 2021. Learning dense representations of phrases at scale. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6634–6647, Online. Association for Computational Linguistics.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv preprint*, abs/1907.11692.

Jing Lu, Gustavo Hernandez Abrego, Ji Ma, Jianmo Ni, and Yinfei Yang. 2020. Neural passage retrieval with improved negative contrast. *ArXiv preprint*, abs/2010.12523.

R. Manmatha, Chao-Yuan Wu, Alexander J. Smola, and Philipp Krähenbühl. 2017. Sampling matters in deep embedding learning. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2859–2867. IEEE Computer Society.

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

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Juan Ramos 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, pages 29–48. Citeseer.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Devendra Sachan, Mostofa Patwary, Mohammad Shoeybi, Neel Kant, Wei Ping, William L. Hamilton, and Bryan Catanzaro. 2021. End-to-end training of neural retrievers for open-domain question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6648–6662, Online. Association for Computational Linguistics.

Hongyin Tang, Xingwu Sun, Beihong Jin, Jingang Wang, Fuzheng Zhang, and Wei Wu. 2021. Improving document representations by generating pseudo query embeddings for dense retrieval. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5054–5064, Online. Association for Computational Linguistics.

Jesse Vig and Kalai Ramea. 2019. Comparison of transfer-learning approaches for response selection in multi-turn conversations. In *Workshop on DSTC7*.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *ArXiv preprint*, abs/1901.08149.

Bohong Wu, Zhuosheng Zhang, and Hai Zhao. 2021. Graph-free multi-hop reading comprehension: A select-to-guide strategy. *ArXiv preprint*, abs/2107.11823.

Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. 2020. Clear: Contrastive learning for sentence representation.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations*.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. ConSERT: A contrastive framework for self-supervised sentence representation transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5065–5075, Online. Association for Computational Linguistics.

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*, pages 1253–1256.

Yinfei Yang, Ning Jin, Kuo Lin, Mandy Guo, and Daniel Cer. 2021. Neural retrieval for question answering with cross-attention supervised data augmentation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 263–268, Online. Association for Computational Linguistics.

A Validation Accuracy

One may argue that the improvement of DCSR might be due to the expansion of indexing corpus (which we have discussed in previous sections), but not the alleviation of *Contrastive Conflicts*. In this section, we present the validation accuracy comparison during the training process between DPR and our DCSR, which is a strong evidence that DCSR well handles the problem of *Contrastive Conflicts*.

Under 8 V100 GPUs with a batch size of 16 on each GPU, the validation process could be viewed as a tiny retrieval process for both DPR and DCSR. To maintain a similar validation environment for fair comparison, we use the *+1 BM25 random* version of DCSR, which results in 8*16=128 questions and 2*8*16=256 contextual sentences in one batch. Therefore, the validation process could be interpreted as *retrieving the most relevant contextual sentence for each question in a corpus of 256 sentences*. Under such a validation task, the size of the indexing corpus is restricted to the same for both DPR and DCSR.

The result is shown in Figure 4. For both Trivia and NQ, DCSR performs consistently better than DPR with a small accuracy margin. On SQuAD, especially, our DCSR can achieve higher validation accuracy than DPR with only one single epoch, and achieves nearly 20% final validation accuracy improvement. This phenomenon further verifies that improvement of DCSR is also achieved by improving the training strategy which alleviates *Contrastive Conflicts*, but not only the expansion of the indexing corpus.

B Error Analysis for SQuAD

Although achieving overwhelmingly better performance on SQuAD than DPR, our DCSR on SQuAD still lags far behind its counterparts on NQ or Trivia. Interestingly, we found that the results on SQuAD dev sets are pretty good and comparable to the results on NQ or Trivia. The results of both DPR and DCSR on dev set and test set performance are shown in Table 8.

By analyzing the training instances, we observe that there exists a severe distribution bias problem in SQuAD: SQuAD-dev and SQuAD-train share a great number of positive passages. In fact, almost all positive passages in the SQuAD-dev could also be found in SQuAD-train. Of all 7921 questions that have at least one positive passage containing the answer in SQuAD-dev, 7624 (96.25%) of these passages' titles could be found in the positive passages of SQuAD-train. More surprisingly, 6973 (88.03%) of these passages are shared between SQuAD-train and SQuAD-dev. However, this feature is exactly what SQuAD-test does not have, resulting in relatively poor performance. But again, this phenomenon reveals another strength of our DCSR, that it enjoys better generalization ability than DPR, thus is more robust in practical use.

Model	SQuAD-dev Top-1	Top-5	Top-20	Top-100
DPR	15.8	34.5	52.8	71.0
DCSR	**26.9**	**47.4**	**63.7**	**78.1**
Model	SQuAD-test Top-1	Top-5	Top-20	Top-100
DPR	42.5	66.8	76.2	85.0
DCSR	**49.5**	**69.6**	**79.6**	**86.4**

Table 8: Performance comparison on both SQuAD-test and SQuAD-dev.

C Case Study

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.

C.1 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 *Commisioner 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 modeling 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.

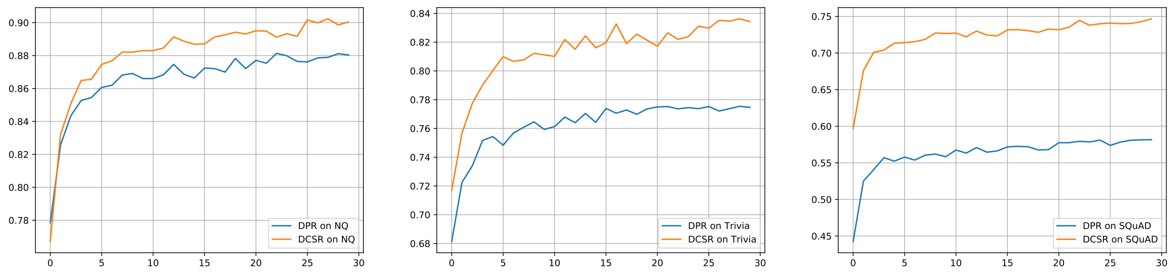

Figure 4: Dev accuracy in training the encoder of DCSR, including NQ (left), Trivia (middle) and SQuAD (right).

Error Type	Example Case
DCSR wins (893 times)	**Question:** Who was the ***NFL Commissioner*** in early ***2012***? **Answer:** ["Roger Goodell", "Goodell"]
	DPR prediction: ***Arena Football League (wiki:692135)*** … which also owns the NHL's Washington Capitals, NBA's Washington Wizards, and WNBA's Washington Mystics, were "close to a deal" in bring a new expansion franchise to the Verizon Center. On March 10, 2016, ***AFL commissioner*** Scott Butera announced that the deal was finalized and that the new Washington, D.C., team would begin play in 2017. On July 14, 2016, … **DCSR prediction:** ***Pro Bowl (wiki:1946017, sentence 4)*** 1. … include one pre-game story on the event in 2012), the players on the field appear to be taking … 2. In the ***2012*** game, the lack of defensive effort was apparent, not only to anyone watching… 3. ***One NFL player*** watching the game said, "They probably should have just put flags on them," … 4. ***Commissioner Roger Goodell*** stated that the game needed to improve, otherwise it would be eliminated. 5. It is worth noting that entire teams have…
DCSR loses (161 times)	**Question:** ***Super Bowl 50*** decided the ***NFL champion*** for what season? **Answer:** ["2015", "the 2015 season"]
	DPR prediction: ***Super Bowl 50 (wiki:18084463)*** ***Super Bowl 50*** was an American football game to ***determine the champion of the National Football League (NFL)*** for ***the 2015 season***. The American Football Conference (AFC) champions Denver Broncos defeated the National Football Conference (NFC) champions Carolina Panthers, … **DCSR prediction:** ***Vince Lombardi Trophy (wiki:1938593, sentence 5)*** 1. …, began appearing on the trophy, still with a frosted appearance. 2. Other than the logo, the trophy has had no significant changes made since the first Super Bowl. 3. While no franchise possesses all four versions, the Green Bay Packers, New England Patriots, … 4. The Super Bowl is currently played in early February (the game originally took place in … ***5.*** ***Super Bowl 50***, which was played on February 7, 2016, ***determined the league champion*** (end of passage)

Figure 5: Error Case Study of Our DCSR on SQuAD. Green color represents the correct clues and correct answers, while red color represents wrong ones.

C.2 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 the 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 do not intend to refine the splition strategy to have a fair comparison with DPR, and leave it for future investigation.