TriSampler: A Better Negative Sampling Principle for Dense Retrieval

Anonymous ACL submission

Abstract

Negative Sampling is an essential technique for dense retrieval that can be utilized to effectively train retrieval models, which significantly effects the retrieval performance. While existing negative sampling methods have already achieved promising results by leveraging hard negatives, there still lacks a general principle to guide negative sampling, including negative candidate construction and negative sampling distribution design. To address it, we conduct a theoretical analysis of negative sampling in dense retrieval and propose *the quasi-triangular principle* to illustrate the triangular-like relationship among query, positive document, and negative document. Relying on this principle, we develop a simple yet effective negative sampling method, TriSampler , which aims to sample more informative negatives within a constrained region. Experimental results indicate that our TriSampler can achieve superior retrieval performance across various representative retrieval models.

1 Introduction

Recently, dense retrieval has gained tremendous attention due to its excellent performance in real-world downstream applications, such as open-domain question answer (Karpukhin et al., 2020), web search (Xiong et al., 2020), and conversational search (Yu et al., 2021). In dense retrieval, the retrieval models must distinguish relevant documents for a specific query from all other negative documents in the entire corpus. Due to a large number of negative documents, it is not feasible to take advantage of all of them. Consequently, negative sampling is keypoint to address the above issue.

Previous efforts have investigated massive negative sampling methods to sample negatives for dense retrieval, such as in-batch negatives, random negatives, hard negatives, and debiased negatives. Inspired by contrastive learning (Oord et al., 2018; He et al., 2020; Chen et al., 2020), dense retrieval models adopt in-batch negatives, a special case of random negatives, to enhance training efficiency, which reuses samples within the current batch and not requires additional sampling operations. However, results in several works (Faghri et al., 2017; Kalantidis et al., 2020; Robinson et al., 2020; Karpukhin et al., 2020; Gao et al., 2021) show that such easy random negatives may not provide sufficient information for model training and result in sub-optimal retrieval performance. To address this issue, hard negative sampling methods (Karpukhin et al., 2020; Xiong et al., 2020; Zhan et al., 2021; Qu et al., 2020; Sun et al., 2022) have been effectively exploited to improve performance, which aim to sample top-k hard negatives based on the current model or an auxiliary retrieval model. A critical challenge associated with hard negatives is the potential presence of false negatives, which may degrade performance (Schroff et al., 2015; Chuang et al., 2020; Qu et al., 2020; Zhou et al., 2022).

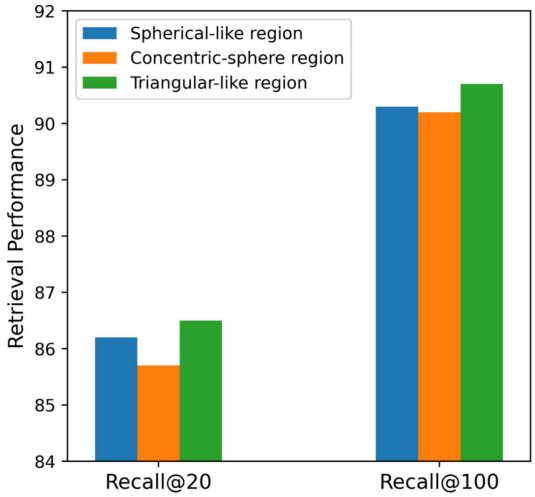


Figure 1: Insight experiments to illustrate the significance of the quasi-triangular principle.

Although prior works have employed various negative sampling methods to achieve promising retrieval results, a general principle guiding nega-

tive sampling remains unclear that should clearly quantify the relationship among query, positive document, and negative document. It is necessary to propose an explicit negative sampling method that can sample more informative negatives within a constraint region (i.e. relationship) based on this principle.

To gain a better understanding of the negative sampling principle, we design two extended experiments: (1) sampling negatives from a spherical-like region where the query acts as the center and the positive similarity serves as the radius; (2) sampling negatives on the concentric-sphere region that is centered on the positive document. Such a spherical-like region allows for a more controllable negative sampling method since it restricts the sampling space aroud the query. This concentric-sphere region aims to sample negatives that are related but not too similar to the positive document. As shown in Figure 1, constraining the sampling of negatives within a triangular-like region can bring about improvements in retrieval performance. By doing so, the retrieval model can effectively distinguish relevant (positive) and irrelevant (negative) documents since the sampled negatives are drawn from a more confined and meaningful region. Therefore, the insights gained from the above experiments suggest that the principle of constraining negatives within a triangular-like region is beneficial for retrieval performance.

In this paper, we propose a general negative sampling principle called *quasi-triangular principle* to constrain the sampled negatives within a triangular-like region. To implement this principle, we develop a straightforward and effective negative sampling method TriSampler, comprising of negative candidate construction and negative sampling distribution implementation. Experimental results in four retrieval benchmarks demonstrate that TriSampler can achieve better retrieval performance compared to other negative sampling methods. Moreover, TriSampler exhibits its adaptiveness and compatibility with a range of classical retrieval models, including AR2, ANCE, and RocketQA.

2 Related Work

Dense retrieval. Dense retrieval (Lee et al., 2019; Karpukhin et al., 2020; Xiong et al., 2020; Khattab and Zaharia, 2020) shows tremendous success in many downstream tasks (e.g. open-domain QA and web search) compared with the traditional sparse retrieval models (e.g. TF-IDF and BM25). The primary paradigm is to model semantic interaction between queries and passages based on the learned representations. Most dense retrieval models leverage the pretrained language models to learn latent semantic representations for both queries and passages. Lee et al. (2019) first proposed the dual-encoder retrieval architecture based on BERT, paving the way for a new retrieval approach. In order to model fine-grained semantic interaction between queries and passages, Poly-encoder (Humeau et al., 2019), ColBERT (Khattab and Zaharia, 2020), and ME-BERT (Luan et al., 2021) explored multi-representation dual-encoder to enhance retrieval performance. Besides, knowledge distillation has become a vital technique to enhance the capacity of the dual-encoder by distilling knowledge from a more powerful reader to a classical retriever (Qu et al., 2020; Ren et al., 2021b; Lin et al., 2020; Hofstätter et al., 2021).

Recently, massive works have investigated task-related pre-training methods for dense retrieval models (Gao and Callan, 2021a,b; Wang et al., 2021; Ren et al., 2021a; Oğuz et al., 2021; Meng et al., 2021). Condenser (Gao and Callan, 2021a) proposed the Condenser architecture to enforce the late backbone layers to aggregate the whole information. coCondenser (Gao and Callan, 2021b) leveraged contrastive learning to incorporate a query-agnostic contrastive loss. PAIR (Ren et al., 2021a) and DPR-PAQ (Oğuz et al., 2021) also designed special tasks in pre-training to enhance retrieval models. Additionally, jointly training retrieval models with the rerank model can bring about better performance. Sachan et al. (2021) proposed an end-to-end training method to jointly or individually model the retrieved documents. Zhang et al. (2021) adopted adversarial training to model the retriever and the reranker.

Negative sampling in dense retrieval. Several recent works (Karpukhin et al., 2020; Xiong et al., 2020; Qu et al., 2020; Zhan et al., 2021) demonstrate that hard negative sampling plays a crucial role in enhancing dense retrieval. Previous studies on negative sampling can be roughly categorized into three categories: (1) random sampling is the simplest way to obtain negatives. As an efficient random sampling method, in-batch negatives are widely used in dense retrieval models (Karpukhin et al., 2020; Zhan et al., 2021). Such an approach is sub-optimal because random negatives have been

proven to be too easy for learning effective models. RocketQA (Qu et al., 2020) adopted cross-batch negatives to increase the number of random negatives, resulting in better performance. (2) hard negative sampling can improve model generalization and accelerate convergence. DPR (Karpukhin et al., 2020) additionally integrated hard negative passages from BM25 into in-batch negatives for dense passage retrieval. ANCE (Xiong et al., 2020) verified that global hard negatives obtained from the current retrieval model can significantly enhance the retrieval performance. ADORE (Zhan et al., 2021) proposed a dynamic negative sampling method to train retrieval models. ANCE-Tele (Sun et al., 2022) combined past iterations by a momentum queue and future iterations by a look-head operation to select hard negatives for stable training. (3) debiased hard negative sampling can efficiently alleviate false negatives. RocketQA (Qu et al., 2020) utilized a well-trained cross-encoder to select hard negatives for the dual-encoder training. SimANS (Zhou et al., 2022) proposed ambiguous negatives to reweight the relevant score with the positives. Different from the abovementioned methods, our TriSampler aims to sample negatives within a triangular-like region based on a general *quasi-triangular principle*, which constraints the range of negative candidates and provides more informative negatives for model training.

3 Understanding Negative Sampling

In this section, we first review the preliminary for dense retrieval and then analyze the vital role of negative sampling in dense retrieval from the perspective of objective.

3.1 Preliminary for Dense Retrieval

Previous dense retrieval works (Karpukhin et al., 2020; Xiong et al., 2020) aim to distinguish the most relevant documents $\mathcal{D}^+$ from a large document corpus $\mathcal{D}$ for a given query q. Typically, these retrieval models leverage negative sampling method to sample several negatives to substitute the entire corpus for model training, thus significantly reducing training costs. The objective function for dense retrieval can be simplified as:

$$\mathcal{L} = \sum_{q} \sum_{d^+ \in \mathcal{D}^+} \sum_{d^- \in \mathcal{D}^-} l(s(\mathbf{h}_q, \mathbf{h}_{d^+}), s(\mathbf{h}_q, \mathbf{h}_{d^-})) \tag{1}$$

where $l(\cdot)$ represents a loss function, such as cross entropy or hinge loss, $s(\cdot)$ denotes the dot product used to measure the similarity metric, $\mathbf{h}_q$ and $\mathbf{h}_d$ represent query embedding and document embedding that are encoded by a query encoder and a document encoder respectively. The pre-trained language models (PLMs) (Devlin et al., 2018; Liu et al., 2019; Zhang et al., 2019) serve as dual-encoder and the representations of the [CLS] token are leveraged as embeddings.

The construction of negative candidates $\mathcal{D}^-$ depends on either the current retrieval model or sparse retrieval model (BM25). The final negatives are then sampled based on different negative sampling distributions.

3.2 Analysis for Negative Sampling

A representative dense retrieval model is trained on training triples $\{(q, d^+, \{d^-\}_{i=1}^n)\}$ where (q, d^+) is a positive query-document pair and $\{d^-\}_{i=1}^n$ are the sampled negative irrelevant documents. A conventional contrastive loss for dense retrieval can be formulated as:

$$\mathcal{L} = -\log \frac{exp(s^+)}{exp(s^+) + \sum_{i=1}^n exp(s_i^-)} \tag{2}$$

where s^+ denotes positive similarity score between the query and the corresponding positive document $s(\mathbf{h}_q, \mathbf{h}_{d^+})$, s^- represents negative similarity score between the query and negative document $s(\mathbf{h}_q, \mathbf{h}_{d^-})$.

The gradient of the above contrastive loss can be split into two parts in terms of s^+ and s_j^-:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial s^+} &= -\frac{\sum_{i=1}^n exp(s_i^-)}{exp(s^+) + \sum_{i=1}^n exp(s_i^-)}, \\ \frac{\partial \mathcal{L}}{\partial s_j^-} &= \frac{exp(s_j^-)}{exp(s^+) + \sum_{i=1}^n exp(s_i^-)} \end{aligned} \tag{3}$$

According to Equation 3, the gradient with respect to the negative document is proportional to the negative similarity score $exp(s_j^-)$. The negatives obtained through random sampling possess very low similarity scores, leading to gradients close to zero and providing minimal contribution to model training. The negatives that are sampled from the top K nearest irrelevant documents can provide larger similarity scores, facilitating the retrieval model to achieve faster convergence. However, the gradients with respect to the positive document will be bounded into a fixed value when negative similarity scores are much larger than the positive ones. As a result, the negatives should

satisfy a specific constraint relationship. A simple relationship is $s^+ \approx s^-$ where negatives are sampled from the spherical-like region. Such sampled negatives provide more information for accelerating model convergence, effectively alleviating zero or fixed-value gradients.

The above analysis clearly demonstrates that negatives within the constraint region $s^+ \approx s^-$ can eliminate excessively hard or easy negatives. Besides, we incorporate the similarity score between the positive document and negatives into negative sampling. The specific relationship among query, positive document, and negative document will be discussed in Section 4.1.

4 Method

As analyzed in Section 3.2, a promising negative sampling method should be satisfied the constraint region $s^+ \approx s^-$, suggesting that sampling negatives from the spherical-like region. However, the entire spherical region is vast for negative sampling, for example, negatives far away from the positive document may not provide more valuable information since the retrieval model should be able to distinguish positive and negative documents. Therefore, we propose the *quasi-triangular principle* where the sampled negatives are constrained within a triangular-like region. Based on this principle, we develop a simple and effective negative sampling method TriSampler.

4.1 The Principle of Negative Sampling

Here, we propose *the quasi-triangular principle* to simulate the pairwise relationship among a training triple (q, d^+, d^-) for improving negative sampling in dense retrieval. The principle constrains the region of sampled negatives within a triangular-like region rather than the entire spherical-like region. Figure 2 demonstrates the planar projection of a sphere where the angular θ in the triangular-like region can be defined as:

$$\theta = |\arccos(\frac{s(\mathbf{h}_q,\mathbf{h}_{d^+})}{||\mathbf{h}_q||\cdot||\mathbf{h}_{d^+}||}) - \arccos(\frac{s(\mathbf{h}_q,\mathbf{h}_{d^-})}{||\mathbf{h}_q||\cdot||\mathbf{h}_{d^-}||}) \quad (4)$$

In the triangular-like region, the boundary for negatives is $\theta = 60°$. Compared with the whole spherical-like region, this constraint further pushes the negatives closer to the positive documents. Such a region simultaneously ensures that negatives possess high similarity with both the query and positive document, which helps to alleviate issues related to false negatives that are too close to the query and uninformative negatives that are far from the query and the positive.

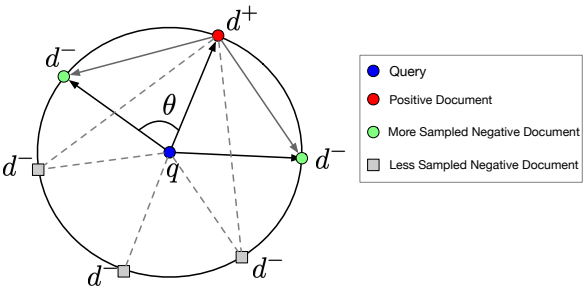


Figure 2: The proposed quasi-triangular principle for negative sampling in dense retrieval.

4.2 Negative Candidates

To provide more informative negative candidates, we follow the *quasi-triangular principle* to construct negative candidates $\mathcal{D}_q^-$ within a triangular-like region for any certain query q. Specifically, we first sample the top-ranked irrelevant documents in terms of the query based on the current retrieval model, which is widely used in previous hard negative selection methods (Xiong et al., 2020; Zhan et al., 2021; Zhang et al., 2021; Zhou et al., 2022). Then, we can obtain the relevant scores between the positive document and the above top-ranked documents. After that, based on the abovementioned relevant scores $s(\mathbf{h}_q, \mathbf{h}_{d^-})$ and $s(\mathbf{h}_{d^+}, \mathbf{h}_{d^-})$ where $d^- \in \mathbf{TopK}_{s(q,\mathcal{D}^-)}$, we derive the following criteria for constructing a more informative negative candidate set that satisfies the *quasi-triangular principle*:

- Negative candidates should conform to the first range constraint $s(\mathbf{h}_q, \mathbf{h}_{d^+}) \approx s(\mathbf{h}_q, \mathbf{h}_{d^-})$, which can effectively eliminate too hard or too easy negatives;
- Negative candidates should be in line with the second range constraint $s(\mathbf{h}_{d^+}, \mathbf{h}_{d^-}) \geq s(\mathbf{h}_q, \mathbf{h}_{d^-})$, which can provide more informative negative candidates.

4.3 Negative Sampling Distribution

The primary goal of negative sampling method is to design an effective distribution for sampling high-quality negatives from the negative candidates. Based on *the quasi-triangular principle*, we formulate the first distribution for the range constraint $s(\mathbf{h}_q, \mathbf{h}_{d^+}) \approx s(\mathbf{h}_q, \mathbf{h}_{d^-})$ as:

$$p_{d^-}^{(q)} \propto exp(-\frac{1}{4} * (s^- - s^+)^2) \quad (5)$$

where s^- and s^+ represent $s(\mathbf{h}_q, \mathbf{h}_{d^-})$ and $s(\mathbf{h}_q, \mathbf{h}_{d^+})$ respectively. Such a distribution eliminates too hard negatives (i.e. false negatives) and further consolidates the first range constraint $s^+ \approx s^-$. The resulting distribution is performed on the top-ranked negative candidates $\mathbf{TopK}_{s(q,\mathcal{D}^-)}$ to obtain transitional negatives $\tilde{\mathcal{D}}_q^-$.

In the second range constraint, we devise a new distribution to obtain the final negatives for retrieval model training. In specific, negatives that are close to positive should be assigned with higher sampling probabilities among the triangular-like region. Thus, the distribution can be represented as:

$$p_{d^-} \propto \text{ReLU}(s(\mathbf{h}_{d^+}, \mathbf{h}_{d^-}) - s(\mathbf{h}_q, \mathbf{h}_{d^-})) \quad (6)$$

where this distribution is conducted on transitional negatives $\tilde{\mathcal{D}}_q^-$.

The key insight of using the RuLU function is that it can exclude negatives that are not in the triangular-like region and further guarantee that negatives that are closer to positive possess higher sampling probabilities. In this way, the sampled negatives can satisfy *the quasi-triangular principle*, providing more informative negatives to enhance retrieval performance for dense retrieval models.

4.4 Discussion

In this work, we propose *the quasi-triangular principle* to guide negative sampling in dense retrieval and design a negative sampling method TriSampler to sample more informative and valuable negatives. TriSampler is a general method that can be directly applied to existing dense retrieval models by substituting the default negative sampling method. Algorithm 1 represents the overall training process of TriSampler. Here, we discuss the connection and discrimination between TriSampler and previous negative sampling methods.

- **TriSampler vs RandNS.** RandNS (Huang et al., 2020) is a basic method that randomly samples negatives from a huge set of negative candidates. TriSampler relies on the *quasi-triangular principle* to sample more informative negatives within the triangular-like region. Different from RandNS that assigns equal weights for each negative, TriSampler leverages a well-designed distribution to sample negatives.

- **TriSampler vs TopNS.** TopNS aims to sample top-k ones from all ranked negatives based on a dynamic-trained dense retrieval model (Xiong et al., 2020; Zhan et al., 2021) or a sparse retrieval model (Karpukhin et al., 2020) (BM25). Unlike TopNS which has a higher risk of false negatives, TriSampler eliminates too hard negatives via a constraint triangular-like region.

Algorithm 1: Algorithm of TriSampler

Input: Positive query-documents $\{(q, \mathcal{D}^+)\}$, document corpus $\mathcal{D}$.

1 Build ANN index on $\mathcal{D}$.
2 Generate the top-ranked negative candidates $\mathbf{TopK}_{s(q,\mathcal{D}^-)}$ from $\mathcal{D}$.
3 Sample transitional negatives $\tilde{\mathcal{D}}_q^-$ from $\mathbf{TopK}_{s(q,\mathcal{D}^-)}$ with distribution $p_{d^-}^{(q)}$.
4 Sample final negatives $\hat{\mathcal{D}}_q^- = \{d^-\}_{i=1}^n$ based on distribution p_{d^-} from $\tilde{\mathcal{D}}_q^-$.
5 Construct training data $\{(q, \mathcal{D}^+, \hat{\mathcal{D}}_q^-)\}$

- **TriSampler vs SimANS.** SimANS (Zhou et al., 2022) designs a negative sampling distribution to sample ambiguous negatives, which avoids sampling negatives that are either too hard or too easy. Similar to SimANS, TriSampler also devises two distributions for the constraint region. The main difference between these is that TriSampler limits its negatives within a triangular-like region while SimANS leverages top-ranked negatives as the sampling region.

- **TriSampler vs ANCE-Tele.** ANCE-Tele (Sun et al., 2022) combines three types of negatives (standard ANCE negatives, momentum negatives, and lookahead negatives) to form negative candidates and then randomly sample negatives from the above candidates. Different from ANCE-Tele, TriSampler constraints the sampling region within a triangular-like region and employs two specifically-designed distributions for sampling.

5 Experiments

5.1 Experimental Setup

Datasets. We conduct experiments on the first retrieval stage of four benchmarks: three passage retrieval datasets: MS MARCO passage (MS Pas) (Nguyen et al., 2016), Natural Questions (NQ) (Kwiatkowski et al., 2019), and TriviaQA (TQA) (Joshi et al., 2017), and a document retrieval dataset: MS MARCO document (MS

Doc) (Nguyen et al., 2016). The statistics of each dataset is illustrated in Table 1.

Datasets	Training	Dev	Test	Documents
NQ	58,880	8,757	3,610	21,015,324
TQA	60,413	8,837	11,313	21,015,324
MS Pas	502,939	6,980	-	8,841,823
MS Doc	367,013	5,193	-	3,213,835

Table 1: The statistics of four retrieval datasets.

Evaluation metrics. We evaluate retrieval performance using official evaluation methodologies, such as MMR@10 and R@k. For the NQ and TQA datasets, R@20 and R@100 serve as metrics to measure whether the top-k retrieved passages contain the answer span. For the MS MARCO datasets, we evaluate the results on their dev datasets in terms of MRR@10 and R@50 for MS Pas dataset, MRR@10 and R@100 for MS Doc dataset.

Baselines. We compare our proposed TriSampler with previously established baselines for retrieval benchmarks. Baselines can be generally divided into the following categories.

- **Sparse Retrieval.** The compared sparse retrieval models contains BM25 (Yang et al., 2017) and improved variants of BM25 models that incorporate pretrained language models, such as doc2query (Nogueira et al., 2019a), DeepCT (Dai and Callan, 2019), docTTTTTquery (Nogueira et al., 2019b), and GAR (Mao et al., 2020).

- **Dense Retrieval.** Massive dense retrieval baselines have investigated a variety of training methods to improve the retrieval performance, such as hard negative sampling (Karpukhin et al., 2020; Xiong et al., 2020; Zhan et al., 2021; Zhou et al., 2022), distillation (Qu et al., 2020; Lu et al., 2022; Ren et al., 2021b), integrating rerankers into retrievers (Zhang et al., 2021), pre-training (Ren et al., 2021a; Gao and Callan, 2021b,a), etc. Among these, hard negative sampling is a particularly important strategy. DPR (Karpukhin et al., 2020), RocketQA (Qu et al., 2020), ANCE (Xiong et al., 2020), ADORE (Zhan et al., 2021), and SimANS (Zhou et al., 2022) attempt to design various negative sampling methods to obtain top-k hard negatives.

Implementation Details. We implement TriSampler based on SOTA dense retrieval model AR2 (Zhang et al., 2021) and run all experiments on 8 NVIDIA Tesla A100 GPUs. Following AR2, ERNIE-2.0-base (Sun et al., 2020) serves as a backbone model to encode queries and passages. Similar to SimANS (Zhou et al., 2022), we directly utilize checkpoints in the AR2 model to continue training with our proposed TriSampler. For MS Doc dataset, the model parameters are initialized with STAR (Zhan et al., 2021). In our experiments, the ratio of positive to negative pairs is set to $1:15$, the inner product is leveraged to estimate the relevance score and Faiss (Johnson et al., 2019) is adopted for efficient similarity search. We utilize the top-200 passages for NQ and TQA datasets and the top-400 documents for MS Pas and MS Doc datasets as negative candidates.

5.2 Overall Results

Our TriSampler achieves a better retrieval performance than all baselines on all metrics (See Table 2 and Table 3). The improvements primarily stem from the superiority of the *quasi-triangular principle* over previous hard negative sampling methods. Since the measurement principle between query-negatives and pos_passage-negatives may share a *quasi-triangular principle* (See Section 3.2), previous methods are unable to capture this principle or even overlook the impact of pos_passage-negatives. Our TriSampler aims to construct negative candidates based on the abovementioned principle. Moreover, the newly designed negative sampling distribution focuses on sampling informative negatives that are simultaneously close to both the query and the positive passage, effectively providing high-quality negatives to accelerate model convergence. To verify the adaptiveness of TriSampler to different genres of dense retrieval models, we also integrate it into two representative retrieval models: ANCE (Xiong et al., 2020) and RocketQA (Qu et al., 2020). For fairness, we only substitute the default negative sampling method in these two retrieval models with TriSampler. Experimental results in Table 2 show that our TriSampler is a general method that can be naturally applied to various dense retrieval models. Such a method can provide more informative negatives to consistently improve downstream performance in dense retrieval.

5.3 Why TriSampler performs better?

Perspective of candidates. To deepen the understanding of TriSampler, we vary the selection methods of negative candidates and conduct

Method	NQ		TQA		MS Pas	
	R@20	R@100	R@20	R@100	MRR@10	R@50
BM25 (Yang et al., 2017)	59.1	73.7	66.9	76.7	18.7	59.2
doc2query (Nogueira et al., 2019b)	-	-	-	-	21.5	64.4
DeepCT (Dai and Callan, 2019)	-	-	-	-	24.3	69.0
docTTTTTquery (Nogueira et al., 2019a)					27.7	75.6
GAR (Mao et al., 2020)	74.4	85.3	80.4	85.7	-	-
DPR (Karpukhin et al., 2020)	78.4	85.4	79.3	84.9	-	-
ME-BERT (Luan et al., 2021)	-	-	-	-	33.8	-
Joint top-k (Sachan et al., 2021)	81.8	87.8	81.3	86.3	-	-
Individual top-k (Sachan et al., 2021)	84.0	89.2	83.1	87.0	-	-
RocketQAv2 (Ren et al., 2021b)	83.7	89.0	-	-	38.8	86.2
PAIR (Ren et al., 2021a)	83.5	89.1	-	-	37.9	86.4
DPR-PAQ (Oğuz et al., 2021)	84.0	89.2	-	-	31.1	-
Condenser (Gao and Callan, 2021a)	83.2	88.4	81.9	86.2	36.6	-
coCondenser (Gao and Callan, 2021b)	84.3	89.0	83.2	87.3	38.2	-
ANCE-Tele (Sun et al., 2022)	84.9	89.7	83.4	87.3	39.1	-
ERNIE-Search (Lu et al., 2022)	85.3	89.7	-	-	40.1	-
AR2+SimANS (Zhou et al., 2022)	86.2	90.3	84.6	88.1	40.9	88.7
ANCE (Xiong et al., 2020)	81.9	87.5	80.3	85.3	33.0	81.1
ANCE + TriSampler	**83.8**	**89.1**	**83.4**	**87.2**	**35.8**	**83.4**
RocketQA (Qu et al., 2020)	82.7	88.5	-	-	37.0	85.5
RocketQA + TriSampler	**85.3**	**89.6**	-	-	**38.3**	**86.0**
AR2 (Zhang et al., 2021)	86.0	90.1	84.4	87.9	39.5	87.8
AR2 + TriSampler	**86.5**	**90.7**	**85.0**	**88.5**	**41.4**	**89.1**

Table 2: Results on three retrieval benchmarks, including NQ test set, TQA test set, and MS Pas dev set. The results of baselines are directly obtained from the original papers and results not provided are marked as "-".

Method	MRR@100	R@100
BM25 (Yang et al., 2017)	27.9	80.7
DPR (Karpukhin et al., 2020)	32.0	86.4
ANCE (Xiong et al., 2020)	37.7	89.4
STAR (Zhan et al., 2021)	39.0	91.3
ADORE (Zhan et al., 2021)	40.5	91.9
AR2 (Zhang et al., 2021)	41.8	91.4
AR2+SimANS (Zhou et al., 2022)	43.1	92.3
AR2+TriSampler	**43.8**	**93.1**

Table 3: Experimental performance on MS Doc dev set.

two extended experiments on the NQ dataset and the MS Pas dataset using the AR2 retrieval model: (1) top-k query-document ranked negative candidates $\mathcal{D}_q^- = \mathbf{TopK}_{s(q,\mathcal{D}^-)}$; (2) top-k document-document ranked negative candidates $\mathcal{D}_q^- = \mathbf{TopK}_{s(d^+,\mathcal{D}^-)}$.

As shown in Table 4, TriSampler surpasses all other variants of negative candidate selection methods, indicating the effectiveness of our TriSampler. For $\mathbf{TopK}_{s(q,\mathcal{D}^-)}$ and $\mathbf{TopK}_{s(d^+,\mathcal{D}^-)}$, they seem to only account for the impact of the query or positive document on negatives, ignoring the triangular-like relationship outlined in Section 4.1. TriSampler combines these two methods based on *the quasi-triangular principle*, which alleviates the excessive reliance on the query and constrains the region of negative candidates. Consequently, TriSampler can achieve enhanced performance, suggesting that the triangular-like relationship is a valuable constraint for selecting negative candidates.

Method	NQ		MS Pas	
	R@20	R@100	MRR@10	R@50
$\mathbf{TopK}_{s(q,\mathcal{D}^-)}$	86.2	90.3	40.9	88.7
$\mathbf{TopK}_{s(d^+,\mathcal{D}^-)}$	85.5	90.4	40.3	88.5
TriSampler	86.5	90.7	41.4	89.1

Table 4: Various negative candidate selection methods on the NQ dataset and the MS Pas dataset.

Perspective of distributions. To demonstrate the effectiveness of the negative sampling distribution proposed in TriSampler, we evaluate the retrieval performances on three variations of TriSampler on

the MS Pas dataset: (1) Uniform sampling that assigns negative candidates with equal weights; (2) TopK Sampling that leverages the relevant score as sampling weights; (3) Debiased Sampling that computes sampling weights by reducing the impact of the positive relevant score.

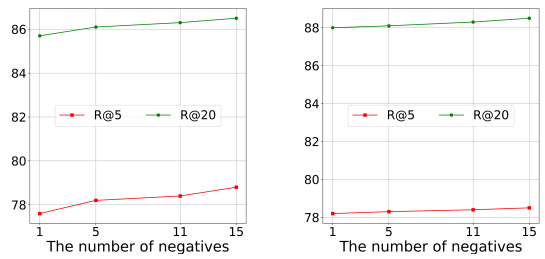
Table 5 reveals that TriSampler outperforms the other variant negative sampling distributions. According to Equation (5), the negative sampling distribution suggested by TriSampler adheres to *the quasi-triangular principle*. This principle allocates higher sampling probabilities to negatives that are closer to the positive document within a restricted region. This observation confirms that a well-designed sampling distribution can indeed contribute to enhanced performance.

Method	MRR@10	R@50	R@1k
Uniform Sampling	39.7	87.9	98.6
TopK Sampling	40.6	88.6	98.7
Debiased Sampling	41.1	88.9	98.8
TriSampler	41.4	89.1	98.9

Table 5: Various negative sampling distributions on the MS Pas dataset.

5.4 Further Analysis

Impact of negative sample size. We further investigate the impact of negative sample size k on retrieval performance using the AR2 model. We vary k in the range of $\{1, 5, 11, 15\}$ and conduct experiments on the NQ and the TQA datasets. As depicted in Figure 3, retrieval performance consistently enhances with the increasing number of negatives, verifying the significance of negative sample size in improving performance. These experimental results align with findings from RocketQA, which also suggest that increasing the number of negatives contributes to better retrieval performance.



(a) NQ (b) TQA

Figure 3: The impact of negative sample size on the NQ dataset.

Training efficiency comparison. To explore the training efficiency of TriSampler, we test the wall-clock time cost, including the cost of training per batch Cost_D and the cost of training instances construction Cost_C. As shown in Table 6, it is obviously observed that the training cost of TriSampler is slightly higher compared with SimANS. Although TriSampler requires more time to construct training instances, the cost is distributed across $t = 2000$ training steps, resulting in a per-batch cost of $\text{Cost}_{C/t} = 0.055s$. Thus, the overall cost for training each batch has increased only slightly. However, the total training time to reach optimal performance is reduced because our TriSampler achieves faster convergency (See Figure 4). To sum up, TriSampler demonstrates improved efficiency gains in comparison to SimANS.

Method	Cost_D	Cost_C	$\text{Cost}_{C/t}$	Cost_{all}
AR2+SimANS	2.9s	85s	0.043s	2.943s
AR2+TriSampler	3.0s	110s	0.055s	3.055s

Table 6: Training efficiency comparison on the NQ dataset.

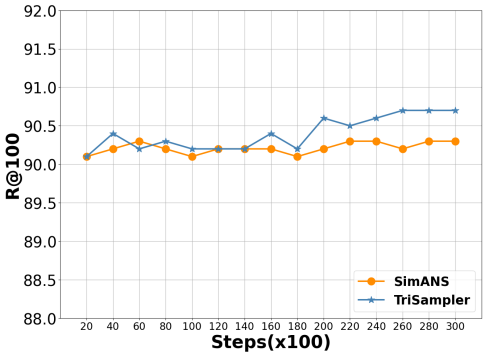


Figure 4: Training convergency curves comparison between SimANS and TriSampler on the NQ dataset.

6 Conclusion

In this paper, we investigate the fundamental principle that negative sampling should satisfy in dense retrieval. First, we analyze negative sampling from the perspective of objective. Next, we propose a general principle to guide negative sampling, termed *the quasi-triangular principle*. This principle suggests that the sampled negatives should be constrained within a triangular-like region. Finally, building upon this principle, we propose a negative sampling method TriSampler to sample more informative negatives within the constrained region. Experiments on four benchmark datasets show that TriSampler can achieve better retrieval performance compared with other methods.

Limitations

Although our study presents an effective negative sampling principle for guiding the selection of negatives in dense retrieval, we recognize two limitations in our work. First, we have only evaluated our TriSampler on benchmark datasets. In future work, we hope to apply TriSampler to real-world industrial dataset and investigate the applicability of *the quasi-triangular principle* across different domains. Second, the proposed TriSampler method may not be the optimal solution based on *the quasi-triangular principle*. Further research is needed to devise a better approach for achieving negative sampling within the constrained region.

Despite these limitations, we offer a general principle for negative sampling in dense retrieval that can serves as a foundation for future research.

References

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. 2020. Debiased contrastive learning. *Advances in neural information processing systems*, 33:8765–8775.

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*, pages 985–988.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2017. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*.

Luyu Gao and Jamie Callan. 2021a. Is your language model ready for dense representation fine-tuning. *arXiv preprint arXiv:2104.08253*.

Luyu Gao and Jamie Callan. 2021b. Unsupervised corpus aware language model pre-training for dense passage retrieval. *arXiv preprint arXiv:2108.05540*.

Luyu Gao, Zhuyun Dai, Tongfei Chen, Zhen Fan, Benjamin Van Durme, and Jamie Callan. 2021. Complement lexical retrieval model with semantic residual embeddings. In *European Conference on Information Retrieval*, pages 146–160. Springer.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021. Efficiently teaching an effective dense retriever with balanced topic aware sampling. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 113–122.

Jui-Ting Huang, Ashish Sharma, Shuying Sun, Li Xia, David Zhang, Philip Pronin, Janani Padmanabhan, Giuseppe Ottaviano, and Linjun Yang. 2020. Embedding-based retrieval in facebook search. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2553–2561.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2019. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring. *arXiv preprint arXiv:1905.01969*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. 2020. Hard negative mixing for contrastive learning. *Advances in Neural Information Processing Systems*, 33:21798–21809.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

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*, pages 39–48.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. *arXiv preprint arXiv:1906.00300*.

Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. 2020. Distilling dense representations for ranking using tightly-coupled teachers. *arXiv preprint arXiv:2010.11386*.

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 arXiv:1907.11692*.

Yuxiang Lu, Yiding Liu, Jiaxiang Liu, Yunsheng Shi, Zhengjie Huang, Shikun Feng Yu Sun, Hao Tian, Hua Wu, Shuaiqiang Wang, Dawei Yin, et al. 2022. Ernie-search: Bridging cross-encoder with dual-encoder via self on-the-fly distillation for dense passage retrieval. *arXiv preprint arXiv:2205.09153*.

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.

Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. 2020. Generation-augmented retrieval for open-domain question answering. *arXiv preprint arXiv:2009.08553*.

Yu Meng, Chenyan Xiong, Payal Bajaj, Paul Bennett, Jiawei Han, Xia Song, et al. 2021. Coco-lm: Correcting and contrasting text sequences for language model pretraining. *Advances in Neural Information Processing Systems*, 34:23102–23114.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset. In *CoCo@ NIPs*.

Rodrigo Nogueira, Jimmy Lin, and AI Epistemic. 2019a. From doc2query to doctttttquery. *Online preprint*, 6.

Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019b. Document expansion by query prediction. *arXiv preprint arXiv:1904.08375*.

Barlas Oğuz, Kushal Lakhotia, Anchit Gupta, Patrick Lewis, Vladimir Karpukhin, Aleksandra Piktus, Xilun Chen, Sebastian Riedel, Wen-tau Yih, Sonal Gupta, et al. 2021. Domain-matched pretraining tasks for dense retrieval. *arXiv preprint arXiv:2107.13602*.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

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

Ruiyang Ren, Shangwen Lv, Yingqi Qu, Jing Liu, Wayne Xin Zhao, QiaoQiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2021a. Pair: Leveraging passage-centric similarity relation for improving dense passage retrieval. *arXiv preprint arXiv:2108.06027*.

Ruiyang Ren, Yingqi Qu, Jing Liu, Wayne Xin Zhao, Qiaoqiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2021b. Rocketqav2: A joint training method for dense passage retrieval and passage re-ranking. *arXiv preprint arXiv:2110.07367*.

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2020. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*.

Devendra Singh 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. *arXiv preprint arXiv:2101.00408*.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

Si Sun, Chenyan Xiong, Yue Yu, Arnold Overwijk, Zhiyuan Liu, and Jie Bao. 2022. Reduce catastrophic forgetting of dense retrieval training with teleportation negatives. *arXiv preprint arXiv:2210.17167*.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8968–8975.

Kexin Wang, Nils Reimers, and Iryna Gurevych. 2021. Tsdae: Using transformer-based sequential denoising auto-encoder for unsupervised sentence embedding learning. *arXiv preprint arXiv:2104.06979*.

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

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.

Shi Yu, Zhenghao Liu, Chenyan Xiong, Tao Feng, and Zhiyuan Liu. 2021. Few-shot conversational dense retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 829–838.

Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, and Shaoping Ma. 2021. Optimizing dense retrieval model training with hard negatives. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1503–1512.

Hang Zhang, Yeyun Gong, Yelong Shen, Jiancheng Lv, Nan Duan, and Weizhu Chen. 2021. Adversarial retriever-ranker for dense text retrieval. *arXiv preprint arXiv:2110.03611*.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. Ernie: Enhanced language representation with informative entities. *arXiv preprint arXiv:1905.07129*.

Kun Zhou, Yeyun Gong, Xiao Liu, Wayne Xin Zhao, Yelong Shen, Anlei Dong, Jingwen Lu, Rangan Majumder, Ji-Rong Wen, Nan Duan, et al. 2022. Simans: Simple ambiguous negatives sampling for dense text retrieval. *arXiv preprint arXiv:2210.11773*.