Transform Retrieval for Textual Entailment in RAG

Xin Liang
Machine Intelligence Laboratory
College of Computer Science
Sichuan University
liangxin1@stu.scu.edu.cn

Quan Guo*
College of Artificial Intelligence
Guangxi Minzu University
Machine Intelligence Laboratory
College of Computer Science
Sichuan University
guoquan@gxmzu.edu.cn

Abstract

In this paper, we introduce Transform Retrieval, a novel approach aimed at improving Textual Entailment Retrieval within the framework of Retrieval-Augmented Generation (RAG). While RAG has shown promise in enhancing Large Language Models by retrieving relevant documents to extract specific knowledge or mitigate hallucination, current retrieval methods often prioritize relevance without ensuring the retrieved documents semantically support answering the queries. Transform Retrieval addresses this gap by transforming query embeddings to better align with semantic entailment without re-encoding the document corpus. We achieve this by using a transform model and employing a contrastive learning strategy to optimize the alignment between transformed query embeddings and document embeddings for better entailment. We evaluated the framework using BERT as frozen pre-trained encoder and compared it with a fully fine-tuned skyline model. Experimental results show that Transform Retrieval with simple MLP consistently approaches the skyline across multiple datasets, demonstrating the method's effectiveness. The high performance on HotpotQA highlights its strength in many-to-many retrieval scenarios.

1 Introduction

Large language models (LLMs) have shown significant potential across a spectrum of downstream tasks in NLP, especially in open-domain question-answering. However, they are prone to generating inaccurate responses due to a lack of knowledge and the hallucination problem. A commonly adopted solution to enhance answer generation is to use Retrieval-Augmented Generation (RAG), which integrates the strengths of information retrieval (IR) and LLMs and has emerged as a prominent technique in Artificial Intelligence Generated Content (AIGC). Specifically, RAG uses dense retrieval in IR to retrieve relevant documents, forms a prompt with the question, which is then fed into LLMs, and ultimately generates better and more accurate answers.

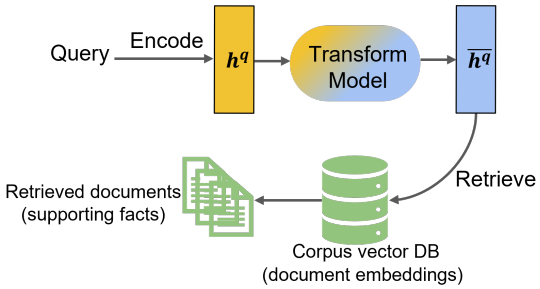

Figure 1: The proposed transform retrieval framework. The model first transforms query embedding to semantic entailment embedding and then retrieves the supported documents.

RAG usually retrieves documents by embedding vectors in a vector database with Approximate Nearest Neighbor (ANN) algorithms. Numerous efforts have been made to improve RAG for better supporting LLM in conversation (Rackauckas, 2024; Sarthi et al., 2024; Lyu et al., 2023; Asai et al., 2023; Chen et al., 2023).

An ideal retrieved document should provide supporting facts for the query, which can be identified by a semantic entailment relationship in Natural Language Inference (NLI) (Dagan et al., 2005). NLI determines whether the given hypothesis document logically follows (entailment), unfollows (contradiction), or is undetermined to (neutral) the premise document. Based on this intuition, we define a task called *Textual Entailment Retrieval (TER)*. A common solution is to train a discriminative model to classify the pair of documents into one of the above categories or fine-tune premise and hypothesis embedding for semantic entailment objective (Reimers, 2019). However, in the RAG scenario, due to the large number of documents in

*Correspondence: Quan Guo guoquan@gxmzu.edu.cn

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 595–599

the vector database (e.g., all 6M Wikipedia documents), such methods always struggle with efficiency. Discriminative models are intractable in inference time efficiency because the total inference time grows linearly as the number of vectors in the database. Fine-tuning the embedding leads to re-encoding all the documents, which is invasive and can incur significant computational and storage costs, exposing RAG to the risk of degeneration of other properties of existing embedding.

In this paper, we aim to mitigate the phenomenon of relevance without support in the relevancy search stage of RAG. Concretely, in the typical RAG process, only pre-trained language model embeddings and some similarity metric functions (e.g., cosine similarity) are used, which often leads to the retrieval of documents that are merely semantically related to the query rather than semantically entailed, meaning the retrieved documents do not necessarily provide the supporting facts required to answer the query. Motivated by SimSiam (Chen and He, 2021) architecture in visual encoding, we propose a Transform Retrieval framework to address this problem under an inference time efficiency concern. As shown in Figure 1, the core idea is to transform query embedding to a semantic entailment embedding relative to its entailed documents. Our method transforms the query embeddings, leaving the huge amount of document embeddings in the database unchanged. More importantly, transform retrieval can be built on top of any existing embedding, allowing RAG to enjoy the efficiency of ANN search.

We summarize the contributions as follows:

- We formulate the task of TER and investigate the limitations of commonly used embedding models and discriminative NLI models.
- We introduce a Transform Retrieval framework for TER task, which aims to mitigate the mismatch between query embeddings and document embeddings in terms of relevance and entailment in an efficient and non-invasive manner.
- We conducted experiments on different datasets, showing that our proposed method improves the performance of TER, validating its effectiveness in enhancing both relevance and entailment.

2 Preliminary

The goal of TER is to retrieve some supported document within the given query in the corpus vector database. Moreover, we treat the user's queries as hypotheses and the documents in the corpus as premises. Given a query q and documents D then TER is formulated as follows:

$$\begin{aligned} &\text{TER}(D|q) = \{d_1, d_2, \ldots, d_m\}, \\ &d_k \rightarrow q, \text{ for } k \in \{1, \ldots, m\}. \end{aligned} \tag{1}$$

We proposed a transform embedding framework with a transform model to manage TER as shown in Figure 1. Formally, we only transform query embedding without altering the existing document embeddings and use a common similarity metric in the retrieval stage, which is formulated as follows.

$$\begin{aligned} h^q &= Enc(q), \\ \overline{h^q} &= \Psi(h^q), \\ \text{TER}(D|q) &= \text{sim}(Enc(D), \overline{h^q}). \end{aligned} \tag{2}$$

where $Enc(\cdot)$ is any model can get sentence embedding, Ψ is the transform model and $sim(\cdot)$ is the similarity metrics such as cosine similarity.

Overall, the RAG process within our approach is similar to the Fact-checking method (Muharram and Purwarianti, 2024). However, the latter introduces additional steps after similarity retrieval, which reduces efficiency.

3 Transform Retrieval

The overall architecture is shown in Figure 2. We introduce a Transform Model to transform the query embedding for TER in the original embedding space. The Transform Model is parameterized and can be trained by contrastive learning.

3.1 Model

General purpose embedding models inadequately capture semantic similarity and perform poorly on the conveyance of semantic entailment. We take a similar approach as SimCSE (Gao et al., 2021), using a contrastive framework to get better sentence embedding. However, instead of optimizing the original BERT embedding space, our approach employs a transform model to transform the original embedding similarity matching into semantic entailment matching. As shown in Figure 2, only the transform model is trained, and the Encoder model (BERT, for instance) is frozen. For the transform model, we experiment with MLP and VAE in the Experiments section.

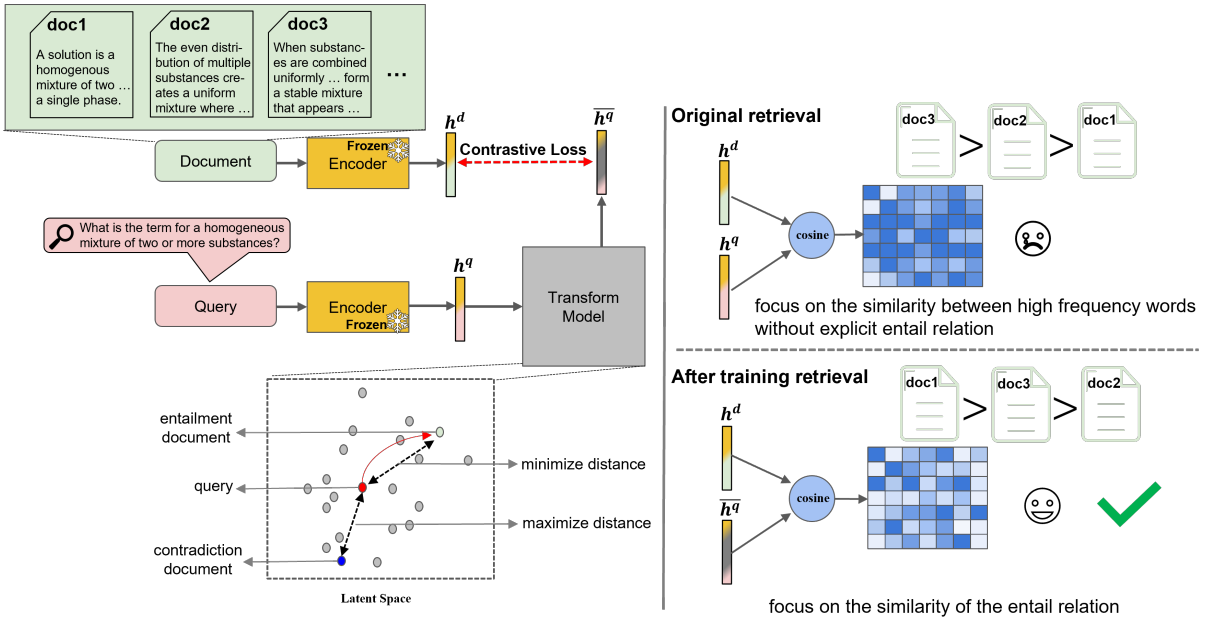

Figure 2: The overall architecture of Transform Retrieval. The query embedding is passed to the Transform model, and the contrastive loss between the transformed query embedding and the document embedding is used to optimize the transform model, which will transform the original query embedding to the desired query embedding for textual entailment retrieval.

3.2 Contrastive Learning

In contrastive learning, we utilize the supervised contrastive loss (Khosla et al., 2020) to push query embedding closer to its corresponding entail document embedding while keeping it away from contradicting document embedding. Given the query embeddings H^q and the document embeddings H^d, the contrastive loss $\mathcal{L}_{contra}$ is defined as:

$$\mathcal{L}_{\text{contra}} = \sum_{i \in \overline{H^q}} \frac{1}{|P(i)|} \sum_{p \in P(i)} \mathcal{L}^p_{\text{contra}}, \tag{3}$$

$$\mathcal{L}^p_{\text{contra}} = -\log \frac{\exp\left(\text{sim}\left(\overline{\boldsymbol{h}^q_i}, \boldsymbol{h}^d_p\right)/\tau\right)}{\sum_{a \in H^d} \exp\left(\text{sim}\left(\overline{\boldsymbol{h}^q_i}, \boldsymbol{h}^d_a\right)/\tau\right)}, \tag{4}$$

where $P(i) \equiv \left\{p \in H^d : h^d_p \to \overline{h^q_i}\right\}$ is the set of indices of all positives in the same batch distinct from i, and $|P(i)|$ is its cardinality. τ is a temperature hyperparameter and $sim\,(\cdot)$ is the cosine similarity. The transform model can be trained using conventional gradient descent with the above loss.

4 Experiments

We conduct experiments with transform retrieval. We use the selected encoder model with cosine similarity as a baseline and an offline deterministic semantic entailment model, namely SimCSE, as the skyline. Models are evaluated against three datasets. The main result is reported in Table 1, and we will analyze the results in the following subsections.

4.1 Datasets

Due to a lack of existing benchmarks, we conducted experiments on three synthetic TER datasets derived from NLI datasets. These datasets were constructed by filtering existing NLI datasets to identify instances where the hypothesis takes the form of a question, followed by selecting samples labeled with entailment.

Specifically, SciTail-TER was created from SciTail (Khot et al., 2018) that derived from approaches treating multiple-choice question-answering. HotpotQA-TER was created from the HotpotQA (Yang et al., 2018) dataset by utilizing the distractor version, and we only selected the first sentence of the supporting sentences. Since the original dataset does not include a test set, we allocated 40% of the validation set to serve as the test

Dataset	Model	R@1 ↑		R@3 ↑		R@5 ↑		MRR ↑	
SciTail-TER	BERT (Baseline)	3.4162	(72%)	7.2669	(77%)	10.1290	(80%)	31.9734	(86%)
	MLP	**3.4515**	(73%)	**8.2905**	(88%)	**10.7600**	(85%)	**33.2360**	(89%)
	VAE	3.2544	(69%)	6.8127	(72%)	9.3259	(73%)	2.9914	(8%)
	SimCSE (Skyline)	4.7145		9.4146		12.6382		36.9430	
HotpotQA-TER	BERT (Baseline)	28.3592	(55%)	39.8379	(63%)	44.9696	(66%)	36.4364	(61%)
	MLP	**42.2687**	(82%)	**59.7232**	(94%)	**66.6779**	(98%)	**53.5513**	(90%)
	VAE	16.6780	(32%)	28.4600	(45%)	35.4490	(52%)	25.9390	(43%)
	SimCSE (Skyline)	51.3167		63.1668		68.0284		59.2555	
SQuAD-ID-TER	BERT (Baseline)	1.3055	(24%)	1.3055	(24%)	1.4571	(26%)	1.6002	(26%)
	MLP	**4.1691**	(78%)	**4.1860**	(78%)	**4.2112**	(75%)	**4.9579**	(82%)
	VAE	0.2189	(4%)	0.2190	(4%)	0.2190	(3%)	0.3012	(4%)
	SimCSE (Skyline)	5.3314		5.3398		5.6009		6.0305	

Table 1: Evaluation of Textual Entailment Retrieval on three synthetic datasets, comparing baseline and proposed models to the skyline. The table shows top-k recalls and MRR, along with percentages relative to the skyline.

Name	#Training	#Validating	#Testing
SciTail-TER	8,600	657	842
HotpotQA-TER	90,447	4,443	2,962
SQuAD-ID-TER	118,445	11,874	11,873

Table 2: Statistics of the Synthetic Datasets

set. SQuAD-ID-TER is derived from the SQuAD-ID-NLI dataset, which is collected from the original SQuAD (Rajpurkar, 2016) dataset. The characteristics of the synthetic datasets are detailed in Table 2.

4.2 Implementation Details

We use Sentence-BERT (Reimers, 2019) checkpoint *bert-base-uncased* as the encoder and the baseline. The dimension of the sentence embedding h is set to 768. The architecture of the MLP comprises an input layer of size 768, followed by two hidden layers with sizes 2048 and 4096, respectively, and a final output layer of size 768. For VAE, we set the VAE encoder and decoder as each 6-layer TransformerEncoder with 8 heads. The latent dimension of VAE is 128.

Following the IR evaluation setting, we evaluate model performance with $Recall@k$, which identifies the correct answer found within the top-k retrieved passages, and with mean reciprocal rank (MRR) for the top 1 result.

4.3 Results and Analysis

Table 1 displays the experimental results on the three synthetic datasets, showing that our proposed method is effective in TER and outperforms the baseline. Specifically, for all three datasets, from small to large, our model (MLP) achieves better recall than the original model (BERT), which suggests that our approach can be adapted to a variety of scenarios with a wide range of data distributions.

Note that the SimCSE presented in Table 1 was fully fine-tuned on a large-scale NLI dataset utilizing the BERT model without specific adaptation to our datasets. Consequently, it serves as a skyline (performance upper bound) for comparative analysis. It is crucial to emphasize that our datasets exclusively comprise entailment pairs. The results reveal a marginal performance disparity between our proposed method and SimCSE, which further demonstrates the effectiveness of the transform retrieval.

The HotpotQA-TER, compared to the remaining two datasets, contains a large amount of one-to-many premise-hypotheses pairs, so its recall metric is higher. The Transform Retrieval method achieves the best improvement on HotpotQA-TER, which we speculated is because our method is more suitable for datasets with non-specific relationships, i.e., each query has multiple supported documents, and the document corpus is rich in information. At the same time, this setting exists abundantly in real RAG applications, which indicates that our method is more practical.

However, in the results presented in Table 1, VAE does not yield better TER improvement results, even worse than the baseline results. We believe that this is because there is a large gap between the BERT embedding space and the Gaussian distribution, and it is difficult to establish the transformation path in the two representation spaces using ordinary generative models such as VAE. Therefore, VAE, when used as a transform model, fails to build up the transition field between expected embeddings well. Perhaps other genera-

tive models would yield good results, but we leave this as an open problem.

5 Conclusions

In this work, we propose a novel approach for textual retrieval named Transform Retrieval, which enhances performance in semantic entailment retrieval in RAG by merely transforming query embedding with transform models trained by contrastive learning. The framework maintains efficient retrieval capabilities and low resource consumption. Our experiments demonstrate that our approach is effective and efficient in TER and has a promising use case in real-world RAG scenarios.

Limitations

Our proposed method has only experimented on our synthesized datasets without measuring the effectiveness in real RAG scenarios. For the transform model, we only explored two types of models, MLP and VAE, and there are other types of models to be explored in the future. We look forward to discussing results on a broader range of transform models.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*.

Tong Chen, Hongwei Wang, Sihao Chen, Wenhao Yu, Kaixin Ma, Xinran Zhao, Dong Yu, and Hongming Zhang. 2023. Dense x retrieval: What retrieval granularity should we use? *arXiv preprint arXiv:2312.06648*.

Xinlei Chen and Kaiming He. 2021. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine learning challenges workshop*, pages 177–190. Springer.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. Scitail: A textual entailment dataset from science question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Xiaozhong Lyu, Stefan Grafberger, Samantha Biegel, Shaopeng Wei, Meng Cao, Sebastian Schelter, and Ce Zhang. 2023. Improving retrieval-augmented large language models via data importance learning. *arXiv preprint arXiv:2307.03027*.

Arief Purnama Muharram and Ayu Purwarianti. 2024. Enhancing natural language inference performance with knowledge graph for covid-19 automated fact-checking in indonesian language. *arXiv preprint arXiv:2409.00061*.

Zackary Rackauckas. 2024. Rag-fusion: a new take on retrieval-augmented generation. *arXiv preprint arXiv:2402.03367*.

P Rajpurkar. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

N Reimers. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D Manning. 2024. Raptor: Recursive abstractive processing for tree-organized retrieval. *arXiv preprint arXiv:2401.18059*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.