SNU LDILAB @ TREC Tip of the tongue 2023

Jongho Kim
Interdisciplinary Program in Artificial Intelligence, Seoul National University / Korea
jongh97@snu.ac.kr

Soona Hong
Seoul National University / Korea
hongsoona@snu.ac.kr

Seung-won Hwang[*]
Seoul National University / Korea
seungwonh@snu.ac.kr

Abstract

This paper describes our participation in the TREC 2023 Tip-of-the-Tongue (ToT) Track. Our first contribution involves formulating the problem as a retrieval, of finding a relevant document with a much shorter query. Inspired by a self-supervised learning approach, we extract ToT query surrogates from the corpus and pair them with the document. These pairs are used for self-supervised training and then enriching document representations to handle insufficiency. Second, we augment ToT queries with cropping and adversarial perturbation. Our results in the ToT benchmark show that our model outperforms state-of-the-art methods including GPT-4 and performs competitively in the TREC-ToT competition.

1 Introduction

This paper explains our TREC 2023 Tip-of-the-Tongue Track submission. Tip-of-the-Tongue (ToT) known-item retrieval is the task where users attempt to identify items from their previous experiences but struggle to recall reliable identifying information (Arguello et al., 2021).

To target the task, our initial contribution draws inspiration from Information Retrieval (IR) techniques. Given that a user query is usually much shorter than the document and shares the characteristic of being 'underspecified,' akin to the ToT scenario, we have introduced a self-supervised learning method. Specifically, IR systems are supervised by relevant query-document pairs annotated by humans, which can be complemented by extracting potential queries from the corpus. Similarly, we can extract query surrogates from documents, and pair them for training for both retrieval and document representation.

However, it's important to note that while this simulates underspecified queries, real-life TREC-ToT queries often contain irrelevant details. Our second contribution is focused on improving training for irrelevant queries. To tackle this challenge, we introduce a diverse query augmentation strategy that generates varying degrees of irrelevant queries from TREC-ToT queries. It enhances our model's robustness against irrelevant information within queries, improving its ability to handle situations where users introduce errors or unverifiable details. Such a two-fold approach ensures that our model can effectively improve the retrieval process.

Our empirical results demonstrate that our model can improve retrieval performances significantly. In addition, we design a comprehensive analysis to underscore our method's strength in the domain of ToT.

2 Methods

This problem can be reformulated as text information retrieval, which retrieves a document d when given a query q.

2.1 Underspecified Query

We first target the task by a self-supervised extracting method to extract ToT query surrogates from a corpus. These potential queries serve both to train a retriever and enhance document representations to address underspecified queries effectively.

2.1.1 Self-supervised Q-D Extraction

Given our formulation as a retrieval, we can consider self-supervising the training of a dense retriever for retrieving document d from query q: A common approach is to extract random spans from the corpus and pair them with d as queries. Pairs from the same d are considered positive pairs, while those from different d are considered negatives. This approach helps increase the dataset size without demanding excessive labeled data (Lee et al., 2019; Izacard et al., 2022; Gao and Callan, 2022; Wu et al., 2023).

[*] Corresponding author.

Inspired, we suggest extracting the surrogate of a ToT query from the corpus itself. This allows us to generate query-document pairs that are better aligned with ToT retrieval. In our target problem of overcoming the gap in movie referencing, the source of reliable information comes from the essential details about movies, such as title, director's name, release year, and key characteristics of the movie. In Wikipedia, the abstract of each article provides them. Such information allows us to precisely identify a single movie based on its unique characteristics. On the other side, while the description of plots or settings in the corpus provides factual information, some movies may share similar narrative elements, making these sections reliable to a certain extent but underspecified for precise identification. For example, the other sections such as plots or settings become unreliable. Some movies may share similar narrative elements, making these sections underspecified for precise identification. Therefore, we extract such reliable information (e.g. abstract) as a document d and other information as a query surrogate q to initially train the model.

In detail, we heuristically filter the sections of Wikipedia with the dictionary in Appendix C. From those sections, we use the abstract as a document and other sections as a query. A total number of 146,928 query-document pairs are generated in a self-supervised way.

For self-supervised learning, we adopt the coCondenser framework (Gao and Callan, 2022). The rationale behind selecting it is based on empirical evaluations comparing primary self-supervised learning approaches in IR. Two leading approaches are contrastive learning (Gao and Callan, 2022) and masked auto-encoding (Wu et al., 2023), with our findings suggesting that the latter is suboptimal for our target task [*]. Therefore, we select coCondenser which shows its effectiveness using contrastive learning.

2.1.2 Enriching Document Representations

Next, we use the extracted queries to expand a document representation. Recent solutions have used document expansion to deal with under-specified queries, enhancing the representation by adding potential queries to the documents. For example, they apply generation models to suggest queries relevant to the document which are then indexed along with the original document (Nogueira et al., 2019).

We found that document expansion can be seamlessly accomplished using the extracted query surrogates. Specifically, our technique involves appending the query surrogate to a document. If there is more than one query surrogate, which means there is more than one section in the Wikipedia document, we append each section separately. This yields multiple expanded documents and consequently generates multiple representations of a single document. In cases where the appended document surpasses the capacity of our encoder, we again partition the section into separate segments, each of which is then added to the document. Following this, we calculate the relevance score for each document utilizing the MaxSim operator. For additional information, please refer to Appendix A.

2.2 Irrelevant Query

Meanwhile, real-world TREC-ToT queries frequently feature irrelevant information that needs to be addressed for ToT query solution. Sentences may include irrelevant information, either due to redundancy (e.g., 'thanks'), or the information absent in the text (e.g., audio or visual information). It's crucial to extend our focus beyond addressing underspecified information.

To enhance the generalizability of retrieval models to such noises, we augment the dataset to create both more relevant and irrelevant queries. The simulation process is done in two phases cropping irrelevant sentences to generate queries without noise, and inject adversarial perturbation to generate queries with more noise.

2.2.1 Phase I: Cropping ToT Queries for Augmentation

To augment queries with more relevance, we employ a process of cropping irrelevant sentences based on annotations. To figure out irrelevant parts in queries, we utilize sentence-level annotations. These annotations indicate whether a sentence in a query contains information related to characters, genres, locations within the movie, and more. Moreover, they provide the ablation experiments by omitting sentences associated with each piece of information. Through the result of the ablation, shown in Appendix B, we identify a dictionary of sentences of irrelevance, which can be appended to the original ones, creating training data that covers

[*] With the application of our method, the nDCG score of masked-autoencoding is 0.2227 (Table 4) while the score of contrastive learning is 0.2687 (Table 2)

a wide range of irrelevancies.

To guarantee that augmented queries from the same original query are not used as in-batch negative during contrastive learning, we sample the data for each batch in a round-robin fashion. This augmented dataset fosters the model's ability to discern the relevance of individual pieces of information. Therefore, the model can produce consistent predictions even in the presence of irrelevant information.

2.2.2 Phase II: Virtual Adversarial Training

Query augmentation, building on original TREC-ToT queries, has its limitations, of containing less irrelevancy. To control the level of irrelevancy in queries, we propose incorporating adversarial perturbations into the input of the model. This helps simulate the presence of irrelevancy in queries, using unlabeled data.

The intuition behind this approach is grounded in prior research, which has demonstrated that introducing perturbations to the combination of transformer layers can provide the model with diverse semantics (Kanashiro Pereira et al., 2021). Moreover, they demonstrated that training with inputs perturbed adversarially is a promising approach for improving the model's generalizability (Zhu et al., 2019; Jiang et al., 2020).

Building on this idea, we introduce adversarial perturbations to each transformer layer and the input embeddings of the encoder. The introduced noise can cause queries to display minor variations in semantics while still referring to the same movie document.

The challenge is the degradation of labeling accuracy after queries are perturbed [†]. For the first solution, we use model prediction as a virtual label and adopt virtual adversarial training (VAT) (Jiang et al., 2020).

Let δ_q and δ_d be the perturbations for each q and d. We define probability distributions of model predictions P^{dq} and $P^{dq}_{\delta_d \delta_q}$ as follows:

P^{dq} is the probability distribution of relevance scores between document d and query q whose embeddings are $E(d)$ and $E(q)$. $P^{dq}_{\delta_d \delta_q}$ is the probability of relevance considering perturbed embeddings of the document and query.

Now, we can formulate the VAT loss as a mini-max problem. First, we seek the adversarial noises, δ_d and δ_q, which maximize the Kullback-Leibler (**KL**) divergence between the original probability distribution P^{dq} and the perturbed probability distribution $P^{dq}_{\delta_d \delta_q}$:

$$\delta_d, \delta_q = \underset{\|\delta_d\|_\infty < \epsilon, \|\delta_q\|_\infty < \epsilon}{\operatorname{argmax}} \mathbf{KL}\left[P^{dq} \parallel P^{dq}_{\delta_d \delta_q}\right]$$

Then the VAT loss, denoted as $\mathcal{L}adv(d,q)$, is computed as the KL divergence between P^{dq} and $P^{dq}_{\delta_d \delta_q}$:

$$\mathcal{L}_{adv}(d,q) = \mathbf{KL}\left[P^{dq} \parallel P^{dq}_{\delta_d \delta_q}\right]$$

When we denote the original retrieval loss as $\mathcal{L}_{ori}$, the resulting loss function is:

$$\mathcal{L} = \mathcal{L}_{ori} + \mathcal{L}_{adv}(d,q)$$

3 Experiments

In this section, we initially provide a detailed description of the experimental setup. Following that, we assess the effectiveness of our methods by evaluating our model on the TREC-ToT dataset.

3.1 Experimental Setups

Our training process consists of two main stages: self-supervised training and supervised training. We employ the BERT-base (Devlin et al., 2019) model as the backbone model for our approach.

Self-supervised Training We first train the model with the extracted queries and documents solely from the corpus. We employ the AdamW optimizer with a learning rate of 1e-4, a weight decay of 0.01, and a linear learning rate decay schedule. Our model is trained for 8 epochs, with a batch size of 4096 and a maximum token length of 512 tokens.

Fine-tuning In this stage, we fine-tune the model on the TREC-ToT dataset. Since we observed the model's performance degrades when training with BM25 hard negatives, we randomly selected negatives from the corpus. We set the maximum number of segments of each document to 2 for training and 4 for evaluation. The learning rate is set to 2e-5, utilizing the Adam optimizer. We incorporate a linear learning rate decay schedule with a warm-up factor of 0.1. The model is trained for a total of 20 epochs with a batch size of 16, and the best

[†] In IR, randomly sampled documents are labeled to be negative for the query. However, when the query is adversarially perturbed regarding labels, it may create a higher chance of potentially relevant documents mistakenly labeled as negatives.

	TREC-ToT test set			
	NDCG@10	NDCG@1000	MRR@1000	Recall@1000
pre_aug_vat_max4	0.2471	**0.3301**	**0.2263**	0.8467
pre_aug_vat_max4_origin	0.2352	0.3224	0.2138	**0.8533**
pre_aug_vat	**0.2498**	0.3206	0.2204	0.8200

Table 1: Results on the test set of TREC-ToT.

checkpoint is selected based on the Mean Reciprocal Rank (MRR). For VAT, we set the perturbation size δ to 1×10^{-5}, perturbation step 1, the step size 1×10^{-3}, and the variance 10^{-5} for initializing perturbation following Jiang et al. (2020).

3.2 Results

Table 1 presents the results of our approach on the TREC-ToT test set, highlighting its competitiveness in the competition. pre_aug_vat_max4 uses a maximum of 4 document representations for each document during training and 2 representations during inference. pre_aug_vat_max4_origin doesn't use sentence annotations during inference and uses the original queries. pre_aug_vat uses a maximum of 2 document representations for each document during training and 2 representations during inference.

3.3 Ablation Study

		nDCG	MRR
	DPR	0.1433	0.0713
Underspecified	+(a)	0.2687	0.1691
	+(b)	0.2698	0.185
Irrelevant	+(c)	0.2860	0.1860
	+(d)	**0.3052**	**0.2054**

Table 2: Ablation study for our approach on the dev set. From top to bottom, our components are added sequentially. Each alphabet corresponds to (a) self-supervised Q-D matching, (b) enriching document representation, (c) query augmentation, and (d) VAT with curriculum learning. With each addition, the performance consistently improves.

We conducted an ablation study on the components of our method. Starting from the DPR model based on the BERT-base model, we incrementally add components of our model. We mark each component as (a) self-supervised Q-D extraction, (b) enriching document representation, (c) query augmentation, and (d) VAT with curriculum learning. Table 2 illustrates how four approaches complement each other. Self-supervised training significantly enhances performance, resulting in an 87.5% improvement in nDCG score compared to baseline. Moreover, the addition of each component further increases the nDCG score, highlighting the effectiveness of our approach's elements.

4 Analysis

4.1 Comparison between Different Self-supervised Training

	nDCG	MRR
Random	0.2492	0.1611
ICT	0.2588	0.1655
GenQ	0.2498	0.1514
Ours	**0.2687**	**0.1691**

Table 3: Performances on TREC-ToT among span generation methods. Our method based on the section outperforms the others.

Other Extracting Strategies To evaluate the effectiveness of our extracting method during self-supervised training, we conducted a comparison with models trained on data created by **random** span selection and inverse cloze task (**ICT**) (Lee et al., 2019). We also compare our method with pseudo query generation, which generates relevant queries using generative language model (Nogueira et al., 2019; Gospodinov et al., 2023). Specifically, we generated one query per document using a fine-tuned **GenQ** (Thakur et al., 2021) model. We employ the model initially trained on the MS-MARCO dataset and fine-tuned to the TREC-ToT dataset. We report the performance excluding using other components except for self-supervised training to isolate its effect.

Table 3 demonstrates that our extracting method outperforms the other extracting techniques. We hypothesize that this difference arises from the fact that other methods do not explicitly consider the ToT scenario when generating self-supervised pairs.

	nDCG	MRR
Mix	0.1917	0.1067
Ours	**0.2227**	**0.1317**

Table 4: Evaluation of sampling strategy in the context of CoT-MAE. The results show that our data sampling method is generalizable to different training methods.

Generality of Extracting Strategy To verify the generalizability of our self-supervised data generation method, we conducted experiments with the CoT-MAE framework (Wu et al., 2023) which trains the retriever with a masked auto-encoding method. Original CoT-MAE makes pairs by **mix**ed generation method of Near, Olap, and Rand strategies. Table 4 indicates that our strategy is effective in the CoT-MAE framework, as evidenced by the performance gap between the original strategy of CoT-MAE and our strategy.

4.2 Analysis of Adversarial Perturbation

	nDCG	MRR
(a)	0.2753	0.1743
Ours	**0.2912**	**0.187**

Table 5: Comparison of performances with different learning curricula. (a) is a model that increases batch size instead of adversarial training.

We address the misconception that adversarial training's reliance on GPU memory is a drawback, especially when compared to dense retrieval's potential advantages from larger batch sizes. Our assertion is that, given a comparable amount of GPU resources, adversarial training benefits the model more than increasing the batch size.

We conducted an experiment involving a model with an increased batch size, as an alternative to adversarial training. The result is in Table 5-(a). Despite the larger batch size, it yields only marginal benefits when compared to our adversarial training method. It underscores the efficiency and effectiveness of our proposed adversarial training approach when operating within similar GPU resource constraints.

5 Conclusion

Participating in the TREC 2023 Tip-of-the-Tongue Track, we tackled the challenge of ToT known item retrieval. We introduced self-supervised learning, extracting potential underspecified queries from a corpus. These queries were used to train a retriever and enhance document representations. To combat irrelevancy, we proposed a diverse augmentation strategy. This involved cropping and adversarial perturbation. Our results showcased the effectiveness of this two-fold approach, performing competitively in the TREC-ToT competition.

6 Acknowledgement

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) [NO.2021-0-01343, Artificial Intelligence Graduate School Program (Seoul National University)].

A Details of Enriching Document Reprsentations

This section explains the details of our document expansion.

We truncate part of the input to fit within the encoder's max token length. In case the length of the original document d_i exceeds 1000, we truncate the document, resulting in d_i'. Additionally, we divide q_i into segments q_{ij}' using a stride of $\mathrm{maxlen} - \mathrm{len}(d_i')$. As a result, each divided passage p_{ij} consists of the concatenation of d_i' and q_{ij}'.

$$p_{ij} = \{d_i' | q_{ij}'\}$$

Here, the symbol | means the concatenating operator.

We calculate a score $s(q_i, P_j')$ for a query q_i and a divided document $P_j' = \{p_{j1}, \dots p_{jK}\}$ using following equation:

$$s(q_i, P_j') = \max_{1 \le k \le K} E(q_i) \cdot E(p_{jk})$$

B Dictionary of irrelevant information in queries

[‘music compare’, ‘production visual’, ‘music specific’, ‘production audio’, ‘production camera angle’, ‘quote’, ‘origin language’, ‘release date’]

C Dictionary for Filtering Wikipedia

[‘synopsis’, ‘plot’, ‘episode’, ‘premise’, ‘summary’, ‘storyline’, ‘content’, ‘setting’, ‘character’, ‘abstract’, ‘story’, ‘overview’, ‘segment’, ‘films’]

References

Jaime Arguello, Adam Ferguson, Emery Fine, Bhaskar Mitra, Hamed Zamani, and Fernando Diaz. 2021. Tip of the tongue known-item retrieval: A case study in movie identification. *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Luyu Gao and Jamie Callan. 2022. Unsupervised corpus aware language model pre-training for dense passage retrieval. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2843–2853.

Mitko Gospodinov, Sean MacAvaney, and Craig Macdonald. 2023. Doc2query–: When less is more. In *European Conference on Information Retrieval*, pages 414–422.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. 2020. Smart: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2177–2190.

Lis Kanashiro Pereira, Yuki Taya, and Ichiro Kobayashi. 2021. Multi-layer random perturbation training for improving model generalization efficiently. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 303–310, Punta Cana, Dominican Republic. 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.

Rodrigo Nogueira, Jimmy Lin, and AI Epistemic. From doc2query to docttttquery.

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

Nandan Thakur, Nils Reimers, Andreas Ruckl'e, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *ArXiv*, abs/2104.08663.

Xing Wu, Guangyuan Ma, Meng Lin, Zijia Lin, Zhongyuan Wang, and Songlin Hu. 2023. Contextual masked auto-encoder for dense passage retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 4738–4746.

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. 2019. Freelb: Enhanced adversarial training for natural language understanding. *arXiv preprint arXiv:1909.11764*.