On Interfacing Tip-of-the-tongue References In Movie Cases

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

A critical challenge in a language-based interface arises when human users try to refer to an item with imprecision or incorrect specifications, while the task requires to retrieve its information to fulfill the task. This is known as the concept of "Tip-of-the-Tongue" (ToT) known-item retrieval, arising when users seek information from their vague memories but cannot quite recall specific, reliable identifying information (RII). For example, consider interfacing a request to find information about a movie that the user watched. If users utter the exact movie title, director's name, and key characteristics of the movie, this would serve as an RII, sufficient for a retriever to find the correct movie. However, in reality, individuals often only hold partial or somewhat fuzzy memories of the plots and characters instead. Therefore, we aim to bridge the gap between RII and unreliable identifying information (UII), addressing both insufficiency and irrelevancy in UII.

Our first contribution involves formulating the problem as a retrieval, of finding a relevant document with a much shorter query, where a query can be interpreted as insufficient UII. Inspired by a self-supervised learning approach, we extract UII surrogates from the corpus and pair them with the document. These pairs are used for training and enriching document representations to handle insufficiency. Second, to overcome irrelevant UII, we simulate such queries by augmentation with cropping and adversarial perturbation with a learning curriculum. 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

A significant challenge in a language-based interface arises when users try to refer to an item with imprecise or incorrect details, while the task requires retrieving its information to complete it. This challenge can be specified as Tip-of-the-Tongue (ToT) known-item retrieval, where users attempt to identify items from their previous experiences but struggle to recall reliable identifying information (RII) (Arguello et al., 2021).

We focus on the ToT references in movie cases. For instance, imagine a user mentions a previously watched movie in the interface, and the task requires obtaining related information to complete it. If the person can recall both the title and additional details like the director's name, release year, and key characteristics of the movie, a retriever can locate the exact movie. In essence, this combined information constitutes RII.

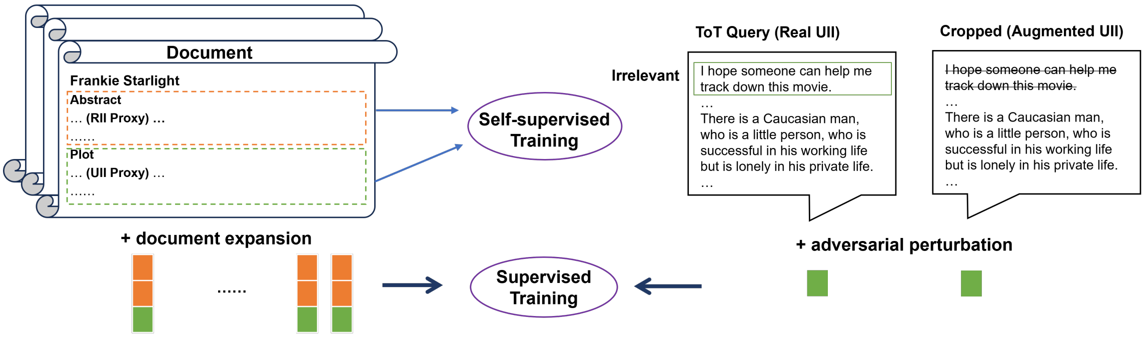
In practice, individuals are more prone to struggle with recalling RII from memory. If they do manage to recall it, they might not provide sufficient detail, and instead, recollect only fragments of the film's plot, standout scenes, or a rough estimate of when they watched it. This scenario is in contrast to RII, as it falls under unreliable identifying information (UII). In such cases, a retriever would encounter difficulty locating the movie.

We categorize UII into two types: **(1) Insufficient UII**. Users ask queries with incomplete information therefore relying solely on partial information may not lead to intended referencing. **(2) Irrelevant UII**. Users may also introduce errors or unverifiable information in their queries, further complicating the retrieval process.

Our initial contribution tackles the challenge of **Insufficient UII**. Drawing inspiration from Information Retrieval (IR) techniques, where a user query, typically much shorter than the document, can be interpreted as a UII, we have devised 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 UII surrogates

[*] Corresponding author.

Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 2: Short Papers), pages 31–39
November 1–4, 2023. ©2023 Association for Computational Linguistics



I hope someone can help me track down this movie.
...
There is a Caucasian man, who is a little person, who is successful in his working life but is lonely in his private life.
...

~~I hope someone can help me track down this movie.~~
...
There is a Caucasian man, who is a little person, who is successful in his working life but is lonely in his private life.
...

Figure 1: An overview of our model. The orange color represents the RII and green represents the UII. To bridge the gap between RII and UII, we 1) use a self-supervised method to extract potential queries for both training and enriching documents, and 2) simulate UIIs in diverse ways.

from documents, and pair them for training for both retrieval and document representation. Figure 1 provides an overview of our approach, wherein we extract RII and UII surrogates from a corpus for self-supervised learning.

However, it's important to note that while this simulates insufficient UII, real-life TREC-ToT queries often contain irrelevant details. We should deal with such irrelevant UII to better solve TREC-ToT queries.

Our second contribution is focused on improving training for **Irrelevant UII**. TREC-ToT queries inherently contain irrelevant UII, so they can be used to instruct the model about irrelevancy. However, this straightforward approach may not effectively encourage the model to understand irrelevant UII due to limitations in the quantity of such data. To tackle this challenge, we introduce a diverse query simulation strategy that generates UII from TREC-ToT queries, aiming to simulate real-life UII. 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 handle both insufficient and irrelevant UII, ultimately improving the retrieval process.

We validate the effectiveness of our approach on benchmark dataset TREC-ToT (Arguello et al., 2021). Our empirical results demonstrate that our model can improve retrieval performances significantly. Our model surpasses state-of-the-art methods, including GPT-4. The results of the TREC-ToT competition also showcase the competitive performance of our model. In addition, we design a comprehensive analysis to underscore our method's strength in the domain of ToT.

2 Related Work

Our study is related to tip-of-tongue retrieval and long-term memory of humans, which we describe below.

Tip of Tongue Retrieval Tip of tongue retrieval task is defined as 'an item identification task where the searcher has previously experienced or consumed the item but cannot recall a reliable identifier' (Arguello et al., 2021). Recently, many public datasets based on community questions and answers have been released in the field of movie (Arguello et al., 2021; Bhargav et al., 2022), book (Lin et al., 2023; Bhargav et al., 2022), and music (Bhargav et al., 2023). In this study, we focus on the ToT retrieval focusing on movies as it has been the main focus of recent research (Arguello et al., 2021; Bhargav et al., 2022; Meier et al., 2021; Fröbe et al., 2023), and for the psychological in-laboratory experiments (Furman et al., 2007)

Human Long-Term Memory Human memory is prone to degradation as time passes (Rubin and Wenzel, 1996). Psychological investigations have studied the extent of information retention. They have illuminated a common phenomenon: individuals tend to lose specific details while retaining more general attributes (Rubin, 1977). Moreover, studies showed the susceptibility of human memory to false recollections (Neisser and Harsch, 1992; Patihis et al., 2013). Remarkably, this tendency for false memories is not confined to individuals with exceptional memory skills but affects all of us (Patihis et al., 2013). Hence, our approach centers on addressing two key factors that render information retrieved from memory unreliable: insufficiency and irrelevancy.

3 Methods

The original problem is to find RII from UII. UII represents the users' (unsuccessful) recall attempts, which may contain insufficient or irrelevant details. This problem can be reformulated as text information retrieval, where RII serves as a document d, and UII serves as a query q.

3.1 Insufficiency

We first target insufficiency in UII by a self-supervised extracting method to extract insufficient queries from a corpus. These potential queries serve both to train a retriever and enhance document representations to address insufficiency effectively.

3.1.1 Self-supervised Q-D Extraction

Given our formulation as a retrieval, we can consider self-supervising the training of a dense retriever for retrieving 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).

Inspired, we suggest extracting the surrogate of an insufficient query from the corpus itself. 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. 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 insufficient for precise identification. For example in Wikipedia, the abstract of each article provides sufficient information. The other sections such as plots or settings become unreliable when it comes to sufficiency. 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 (see the left side of Figure 1 for example).

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.

Let $E(d)$ and $E(q)$ be a document and a query representation attained from an encoder. In order to enhance the alignment between $E(d)$ and $E(q)$, we apply contrastive loss function:

$$\mathcal{L}_{co} = -\sum_{i \in I} \log \frac{f(d_i, q_i)}{\sum_{a \in I \setminus \{i\}} f(d_i, q_a) + \sum_{a \in I} f(d_i, q_a)}$$
$$\text{where } f(x, y) = \exp\left(E(x) \cdot E(y)\right).$$

The symbol I is an index set of minibatch. In addition to the contrastive loss, we incorporate the MLM loss $\mathcal{L}_{mlm}$. Consequently, the total loss of the first stage training is given by:

$$\mathcal{L}_{pre} = \mathcal{L}_{co} + \mathcal{L}_{mlm}$$

3.1.2 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 queries. Specifically, our technique involves appending an extracted query to a document. In cases where the combined document surpasses the capacity of our encoder, we partition the query into separate segments, each of which is then added to the document. This yields multiple expanded documents and consequently generates multiple representations of a single document. Following this, we calculate the relevance score for each document utilizing the MaxSim operator. For additional information, please refer to Appendix A.

[*] With the application of our method, the nDCG score of masked-autoencoding is 0.2227 (Table 5) while the score of contrastive learning is 0.2687 (Table 3)

3.2 Irrelevancy

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 insufficient information.

To enhance the generalizability of retrieval models to accommodate UIIs that may contain irrelevancies, we simulate to inject irrelevancy and train the model with them to robustify the model to deal with such irrelevancies. The simulation process is done in two phases as depicted Figure 1(right): cropping irrelevant sentences and injecting adversarial noises.

3.2.1 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 provided by Arguello et al. (2021). 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 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.

3.2.2 Phase II: Virtual Adversarial Training with Curriculum Learning

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 minimax 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 $P^{dq}_{\delta_d\delta_q}$:

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

The second solution is to introduce a mechanism for controlling the noise level within the context

[†] 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.

of curriculum learning (Jiang et al., 2017). This is because perturbing queries that are already noisy can potentially result in entirely incorrect semantics. In our approach, we introduce a coefficient α_q that controls the amount of adversarial loss, and this coefficient can vary for each query. Initially, we train the model without VAT and compute the model's performance for each training query. We then set α_q to 0 for a certain percentage ($\beta\%$) of the training queries where the model's performance is low, and 1 for the rest. This curriculum enables us to generate adversarial examples only for queries where a substantial portion of relevant information is assured. The resulting loss function is:

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

4 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.

4.1 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 For the data sampling, we heuristically filter the sections of Wikipedia with the dictionary in Appendix C. From those sections, we use 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. We apply the learning scheme of Condenser (Gao and Callan, 2021). It has basically the same structure as the BERT-base, with an additional loss function and head that efficiently aggregates text information into the dense representation. 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. We utilize eight RTX 3090 GPUs. To accommodate this large batch size on the available GPU VRAM size, we employ GradCache (Gao et al., 2021) with a chunk size of 16. Upon the completion of self-supervised training, we remove the additional Condenser head and treat the model in the same manner as a standard BERT model.

Supervised Training In the supervised training stage, we exclusively utilize 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. For the supervised training process, we set the maximum number of segments of each document to 2. 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 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). For curriculum learning, We trained a model without perturbation and used its nDCG score as a performance metric on training queries. β is selected from $\{0.1\ 0.3\ 0.5\}$ and we report the results of each value in Appendix D.

4.2 Results

Table 1 presents the results on the dev set of the TREC-ToT retrieval task. We use official baselines released in TREC 2023 Tip-of-the-Tongue (ToT) Track[¶]. It includes **BM25**, dense passage retrieval (**DPR**) based on DistillBERT-base (Sanh et al., 2019), and another DPR model that re-runs the task using negatives generated from the model's output. We mark it as **DPR (hard-negative)**. They also provide **GPT-4** (OpenAI, 2023) as a baseline, which only retrieves the titles of top-20 document candidates by prompting.

Our model demonstrates a significant performance improvement, claiming the top rank across all metrics except for GPT-4. Regarding GPT-4, our model achieves superior performance in R@100, R@1000, and nDCG because of the inherent limitations of GPT-4, which struggles with ranking a large number of documents. It's noteworthy that our model even outperforms GPT-4 in Recall@10, despite being smaller in size. This success can be attributed to the fact that GPT-4, in its retrieval process, focuses solely on movie titles. Such an approach provides insufficient information for precisely identifying a specific movie, giving our model a notable edge.

[‡] Excerpted from TREC-ToT benchmark.

[§] Only a maximum of 20 candidates are generated.

[¶] https://github.com/TREC-ToT/bench/

	TREC-ToT dev set					
	nDCG	Recall@1	Recall@10	Recall@100	Recall@1000	MRR
BM25[‡]	0.1314	0.0800	0.0933	0.1800	0.4067	0.0881
DPR (DistilBERT$_{\text{base}}$)[‡]	0.1313	0.0267	0.1333	0.2733	0.5333	0.0606
DPR (hard-negative)[‡]	0.1627	0.0400	0.1467	0.3600	0.6600	0.0743
DPR (BERT$_{\text{base}}$)	0.1433	0.0533	0.1067	0.2733	0.5733	0.0713
GPT-4[‡] [§]	0.2407	**0.1800**	0.2867	-	-	**0.2180**
Ours	**0.3052**	0.1400	**0.3000**	**0.5600**	**0.8200**	0.2054

Table 1: Results on the dev set of TREC-ToT. The best-performing model is marked in bold. Our methods outperform both dense and sparse retrieval baselines and are comparable to GPT-4.

	TREC-ToT test set		
	nDCG	nDCG@100	Recall@5
Avg. of medians	0.146	0.0828	0.1133
Ours (ablated)	**0.3301**	**0.2417**	**0.2867**

Table 2: Results on the test set of TREC-ToT. We compare our ablated version with the medians of participating runs regarding nDCG@1000, nDCG@100, and Recall@5 since those are the only available values for the test set.

Table 2 presents the results of our approach on the TREC-ToT test set, highlighting its competitiveness in the competition. As only the median values for each question from participating runs are available, and given that we submitted a version without curriculum learning in the adversarial stage, we calculated the average of these median values and compared it with our ablated version. The results show that our model significantly outperforms the median results across all metrics.

4.3 Ablation Study

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

Table 3: Ablation study for our approach. 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 3 illustrates how four approaches for addressing the UII 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.

5 Analysis

5.1 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 4: 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, a common method used to address data insufficiency by generating relevant queries using generative language model (Nogueira et al., 2019; Gospodinov et al., 2023). Specifically, we generated one query per document using a finetuned **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 the use of other components except for self-supervised training to isolate its own effect.

Table 4 demonstrates that our extracting method outperforms the other extracting techniques. We hypothesize that this difference arises from the fact that other methods do not differentiate between RII and UII when generating self-supervised pairs.

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

Table 5: 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 5 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.

5.2 Analysis of Adversarial Perturbation

	nDCG	MRR
(a)	0.2739	0.1673
(b)	0.2912	0.187
Ours	**0.3052**	**0.2054**
(d)	0.2753	0.1743

Table 6: Comparison of performances with different learning curricula. (a) is a model without a curriculum. (b) is the model that removes adversarial loss for relevant queries. Our model removes adversarial loss for already noisy queries. (d) is a model that increases batch size instead of adversarial training.

In this section, we verify our argument that adversarial training with a curriculum enhances its performance by transforming relevant queries into irrelevant ones. We evaluate the performance of models employing distinct learning curricula. The first model (a) is a baseline that doesn't employ curricula; it uniformly sets α_q to 1 for all queries. The second model (b) takes the reverse approach. It excludes the adversarial loss for training instances when the nDCG score of the query is high, signifying 'relevant' queries.

The results are in Table 6. Our findings indicate that our proposed curriculum consistently outperforms both baseline models (a) and (b). It is important to note that penalizing the relevant queries (b) exhibits worse performance even compared to model (a). This indicates that the indiscriminate application of adversarial perturbation might be harmful due to the presence of originally noisy queries. These results support our hypothesis that adversarial training with curriculum can enhance the performance of retrieval models by simulating irrelevant queries from relevant ones.

Finally, 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 6-(d). 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.

6 Conclusion

This paper tackled the challenge of using imprecise or incorrectly specified items in language-based interfaces. It happens when users reference items from vague memories. We aimed to bridge the gap between RII and UII in language-based interfaces. We introduced self-supervised learning, extracting potential insufficient queries from a corpus. These queries were used to train a retriever and enhance document representations. To combat irrelevancy, we proposed a query simulation strategy. This involved augmentation by cropping and adversarial perturbation with a learning curriculum. Our results showcased the effectiveness of this two-fold approach, consistently outperforming state-of-the-art methods, including GPT-4. Our model also performed competitively in the TREC-ToT competition.

While our current experimental focus has been on the domain of movies, an exciting avenue for future exploration involves extending our method to diverse domains, such as the realm of book referencing. The applicability of our approach is not limited to specific domains, as the concepts of RII and UII transcend various subject matters.

7 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 $P'_j = \{p_{j1}, \ldots p_{jK}\}$ using following equation:

$$s(q_i, P'_j) = \max_{1 \leq k \leq K} E(q_i) \cdot E(p_{jk})$$

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']

	nDCG	MRR
β: 0.1	**0.3052**	**0.2054**
β: 0.3	0.2911	0.1899
β: 0.5	0.3048	0.204

Table 7: Model performances with different β values.

D Model Performances with Different β Values

Table 7 shows the model performances with different β values that are used for hyperparameter selection.

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*.

Samarth Bhargav, Anne Schuth, and Claudia Hauff. 2023. When the music stops: Tip-of-the-tongue retrieval for music. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM.

Samarth Bhargav, Georgios Sidiropoulos, and Evangelos Kanoulas. 2022. 'it's on the tip of my tongue' a new dataset for known-item retrieval. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 48–56.

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*.

Maik Fröbe, Eric Oliver Schmidt, and Matthias Hagen. 2023. A large-scale dataset for known-item question performance prediction.

Orit Furman, Nimrod Dorfman, Uri Hasson, Lila Davachi, and Yadin Dudai. 2007. They saw a movie: long-term memory for an extended audiovisual narrative. *Learning & memory*, 14(6):457–467.

Luyu Gao and Jamie Callan. 2021. Condenser: a pre-training architecture for dense retrieval. In *Conference on Empirical Methods in Natural Language Processing*.

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.

Luyu Gao, Yunyi Zhang, Jiawei Han, and Jamie Callan. 2021. Scaling deep contrastive learning batch size under memory limited setup. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 316–321, Online. Association for Computational Linguistics.

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 pretrained natural language models through principled regularized optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2177–2190.

Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. 2017. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International Conference on Machine Learning.*

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.

Kevin Lin, Kyle Lo, Joseph E Gonzalez, and Dan Klein. 2023. Decomposing complex queries for tip-of-the-tongue retrieval. *arXiv preprint arXiv:2305.15053.*

Florian Meier, Toine Bogers, Maria Gäde, and Line Ebdrup Thomsen. 2021. Towards understanding complex known-item requests on reddit. *Proceedings of the 32nd ACM Conference on Hypertext and Social Media.*

Ulric Neisser and Nicole Harsch. 1992. Phantom flashbulbs: False recollections of hearing the news about challenger. *Affect and accuracy in recall: Studies of "flashbulb" memories*, 4:9–31.

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.*

OpenAI. 2023. Gpt-4 technical report.

Lawrence Patihis, Steven J Frenda, Aurora KR LePort, Nicole Petersen, Rebecca M Nichols, Craig EL Stark, James L McGaugh, and Elizabeth F Loftus. 2013. False memories in highly superior autobiographical memory individuals. *Proceedings of the National Academy of Sciences*, 110(52):20947–20952.

David C. Rubin. 1977. Very long-term memory for prose and verse. *Journal of Verbal Learning and Verbal Behavior*, 16:611–621.

David C. Rubin and Amy Wenzel. 1996. One hundred years of forgetting : A quantitative description of retention.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108.*

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.*