Query Expansion with Semantic-Based Ellipsis Reduction for Conversational IR

Chia-Yuan Chang[1*], Hsien-Hao Chen[2*], Ning Chen[1*], Wei-Ting Chiang[1*], Chih-Hen Lee[1*],
Yu-Hsuan Tseng[1*]
Ming-Feng Tsai[2], Chuan-Ju Wang[1]
[1] Research Center for Information Technology Innovation, Academia Sinica
[2] Department of Computer Science, National Chengchi University

ABSTRACT

Word choice mismatch between query and documents is a common problem in QA/dialogue subjects such as the TREC Conversational Assistance Track (CAsT) 2020. We account for this kind of mismatch by expanding queries using semantic-based ellipsis reduction (SER), which involves gathering supplemental information from historical queries and potentially relevant documents.

We formulate information retrieval as (1) retrieving potential information and (2) reranking its priority. To explain the importance of query expansion and verify our method's effectiveness, we conduct experiments with diverse settings in the retrieval part, followed by a Transformer model for reranking. We also resolve coreferences by replacing pronouns with their coreferential antecedents using a Transformer-based model.

This work shows the importance of accounting for differences in wording and the potential of semantic-based approaches.

ACM Reference Format:
Chia-Yuan Chang[1*], Hsien-Hao Chen[2*], Ning Chen[1*], Wei-Ting Chiang[1*], Chih-Hen Lee[1*], Yu-Hsuan Tseng[1*] and Ming-Feng Tsai[2], Chuan-Ju Wang[1]. 2020. Query Expansion with Semantic-Based Ellipsis Reduction for Conversational IR. In *TREC '20: Text REtrieval Conference, 2020, Online.* ACM, New York, NY, USA, 4 pages. https://doi.org/00.0000/0000000.0000000

1 INTRODUCTION

Vocabulary mismatch is a common and inevitable phenomenon in conversational IR, given that many concepts can be expressed using more than one kind of word or description. Also, conversations contain many pronouns; although such coreferences make for more succinct language, they complicate comprehension with their resultant ambiguities, as demonstrated in Table 1. Moreover, omitting repeated subjects—that is, ellipsis—can result in profoundly wrong misinterpretations. Note that the TREC CAsT 2020 query turns also include elliptical queries. For instance, in the 83-3 query (see Table 1), the subject *bees* is omitted in the utterance, as is the corresponding pronoun.

We attempt to account for such mismatch by using techniques for relevance feedback; specifically, we reconcile word differences between queries and passages by expanding each query with keywords from retrieved passages. However, since Transformer-based models provide a better way to consider each word's co-occurrences, and since they use attention [1] to eliminate the window size limit inherent to n-gram-based models, we use the T5, a Transformer-based text-to-text model [2] to determine the correct pronoun reference given the current context and historical conversations. We also borrow from another Transformer-based model to add auxiliary information from historical queries and topic sentences in passages to resolve ellipses.

Number	Raw utterance
83-1	What are some interesting facts about bees?
83-2	Why doesn't it spoil?
83-3	Why are so many dying?
83-4	What can be done to stop it?
83-5	What has happened to their habitat?
83-6	What can I do to help with the problem?
83-7	What is the cause of CCD?
83-8	What would happen if they died out?

Table 1: TREC CAsT 2020 conversation utterances

Figure 1 shows the proposed pipeline: we first use a T5-based coreference query reformation (T5-CQR) model to resolve coreferences in queries. Using the resultant queries in which the pronouns have been replaced, we expand the query via RM3, the proposed semantic-based ellipsis reduction (SER) method, and a T5-based QA method. Last, in the reranking part, we train a T5 reranker based on the BERT reranker [3] to rerank the retrieved passages by relevance.

In summary, our contributions are as follows.

- We use a semantic-aware Transformer-based model to expand queries to account for word mismatches between queries and documents.
- To efficiently resolve coreferences, we propose the T5-CQR model, a Transformer-based sequence-to-sequence model fine-tuned on the CANARD dataset [4] for question-in-context rewriting.
- We propose semantic-based ellipsis reduction, a novel way to resolve ellipses by considering historical queries and related topic sentences at the semantic level.

2 METHODOLOGY

2.1 Coreference Query Reformation (CQR)

A simple way to approach the coreference problem is to rewrite sentences with their coreferential antecedent, as a type of sequence-to-sequence problem. Therefore, we adopt a Transformer-based

∗ These authors contributed equally to this work; author order was determined alphabetically.

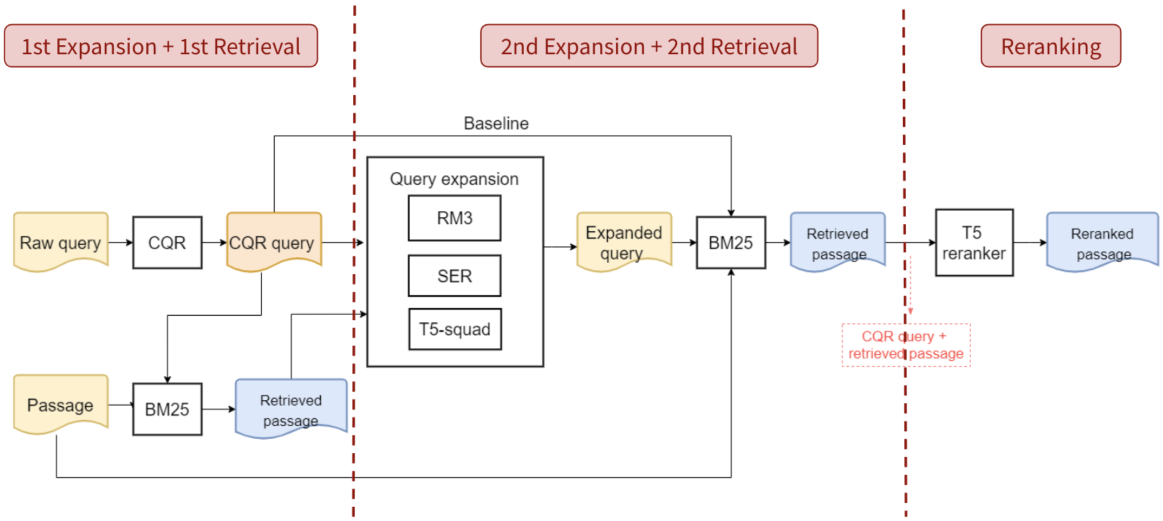

Figure 1: Proposed pipeline

model fine-tuned on a coreference-resolution dataset to produce queries with explicit antecedents.

2.2 Semantic-Based Ellipsis Reduction (SER)

To tackle the ellipsis problem at the semantic level, we propose semantic-based ellipsis reduction (SER), which supplements the original query with the historical queries and the related topic sentences in documents that are semantically associated with the original query.

In the first part, as shown in Figure 2, we calculate the similarity between the original query and the other queries in the same conversation, and extend the original query using the historical query with the highest similarity; this is referred to as the expanded query.

Second, to acquire more query-related semantic information from the passages, we extract a topic sentence from potentially related passages. We separate each retrieved passage into sentences and apply the doc2query model [5] to each sentence to obtain its underlying "latent query". We then calculate the similarity between each latent query and the original query to determine the topic sentence most relevant to the original query, which we append to the expanded query. Both similarities are calculated by a Transformer model fine-tuned on a natural-language inference dataset.

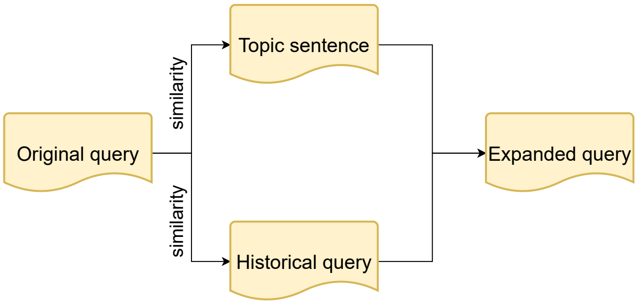

Figure 2: SER workflow

2.3 Retrieval

Okapi BM25, a modified version of TF-IDF, has been used for decades. Given its efficiency and competitiveness with most language models, we use it as the primary retrieval method.

2.4 Reranking

We follow the BERT reranker [3] settings by concatenating the query and each document with a [SEP] tag as the model input, ranking all the pairs with the corresponding score. However, to make full use of the encoder-decoder structure, we adopt T5 instead of the BERT model.

3 EXPERIMENTS

3.1 Evaluation and Settings

To compare our models with different query expansion methods and hyperparameters, we evaluated model performance on the TREC CAsT 2019 evaluation topics in terms of recall (R@1000 and R@2000) and mean average precision (mAP). We used BM25 as the retrieval method, and the T5-3B model [6] as the reranking model, which we fine-tuned only on the MS MARCO dataset but not TREC CAR, since the MS MARCO queries are more similar to the queries in TREC CAsT than TREC CAR.

The proposed baseline model includes coreference resolution via T5-CQR, retrieval via BM25, and reranking via the T5-reranker model. Also, as shown in Table 2, to improve recall in the retrieval stage, we used three different methods to expand queries of TREC CAsT raw utterance topics and validated these methods on the 2019 TREC CAsT evaluation topics.

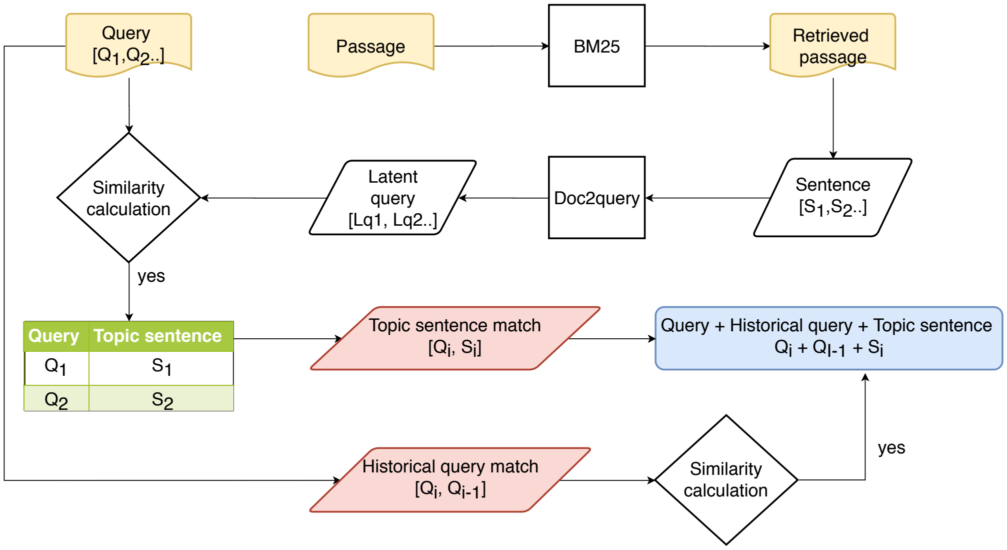

Figure 3: SER structure

3.2 CQR Model Settings

We fine-tuned our CQR model on the CANARD dataset [4], a conversational reading comprehension dataset containing dialogic question–answer pairs. This dataset can be used to evaluate query rewriting models that account for linguistic phenomena such as coreference and ellipsis resolution. When fine-tuning the T5-CQR model, for the input we used the previous dialogs followed by the current query, and for the target answer we used the rewritten, coreference-reformated query. For the ground truth, we used the reference rewrites. Note that questions and answers in the input dialog utterances were separated by the "|||" symbol, as recommended by the authors of CANARD. After fine-tuning, we assembled each query in each turn under each TREC CAsT topic with its previous queries. Using the assembled sentences as the input to the proposed CQR model, the outputs were the coreference-free queries.

3.3 SER Settings

As Figure 3 shows, to exploit semantic relations, we first encoded each coreference-solved query by a pre-trained sentence Transformer model, a Roberta-large based model [7] trained on the NLI and STSb datasets and optimized for semantic textual similarity. For each query, we calculated the cosine similarity between the current query and the previous query, and appended the previous query as historical information if the cosine similarity was between 0.8 and 0.9, as determined empirically. Also, we used BM25 with the Anserini toolkit [8] to filter the corpus into a smaller corpus for efficiency. We parsed the passages of this smaller corpus into sentences and extracted one latent query per sentence with the doc2query model trained on the MS MARCO and TREC CAR datasets. The Roberta-large-based model was adopted to encode the original query and the latent queries to calculate the similarity. The most similar latent query with a cosine similarity greater than 0.9 was extracted as the most related topic sentence. Thus we concatenated the original query, the historical information, and the topic sentence to form an expanded query.

We also leveraged the manual responses provided in the 2020 TREC CAsT topics using the following T5-based method to expand each query.

3.3.1 *T5-SQuAD Query Expansion.*

We used the manual results as the SQuAD-format content, which was one of the pre-trained tasks of the T5 model, and the raw utterance of each turn as the SQuAD-format question. We inputted this into the T5-3B model to yield the answers. In this query expansion method, we leveraged the answers as the words for query expansion.

	TREC 2019					**TREC 2020**				
Stage	Retrieval			Reranking		Retrieval + Reranking				
Raw utterances only	mAP@1000	R@1000	R@2000	mAP@1000	R@1000	mAP@1000	R@1000	NDCG@3	NDCG@5	NDCG@1000
Raw queries	0.1077	0.4182	0.4681	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Baseline	0.2497	0.7628	0.8260	**0.3724**	0.8060	**0.3096**	0.6106	**0.4579**	**0.4472**	0.4943
QE (RM3)	**0.2845**	**0.8024**	**0.8563**	0.3695	**0.8252**	0.3092	**0.6405**	0.4511	0.4362	**0.5003**
QE (SER)	0.2434	0.7674	0.8288	0.3713	0.8089	0.3090	0.6131	0.4576	0.4456	0.4934
Manually rewritten utterances	mAP@1000	R@1000	R@2000	mAP@1000	R@1000	mAP@1000	R@1000	NDCG@3	NDCG@5	NDCG@1000
QE (T5-SQuAD)	N/A	N/A	N/A	N/A	N/A	0.3102	0.6498	0.4663	0.4514	0.5131

Table 2: Results

3.4 Quantitative Analysis

3.4.1 Results on 2019 TREC CAsT evaluation set. The results shown in Table 2 are from using the raw utterances of the 2019 TREC CAsT evaluation set with various query expansion methods. In the retrieval stage, after rewriting raw queries with T5-CQR, the baseline model yields improvements in mAP and recall. T5-CQR improves the mAP, R@1000, and R@2000 at the retrieval stage by 132%, 82%, and 76%. With query expansion, SER slightly improves on the baseline model by 0.6% and 0.3% in R@1000 and R@2000 but decreases the mAP by 2.5%. RM3 query expansion, in turn, improves on the baseline model's R@1000 and R@2000 by 5.2% and 3.7%, and even increases the mAP by 14%. As the primary purpose of the retrieval stage is to increase recall, we believe that the above query expansion methods both yield marginal gains over the baseline.

Since higher recall likely benefits reranking, we retrieved the top 2000 passages at the retrieval stage. After re-ranking, we kept only the top 1000 passages as the final results. The baseline model without query expansion shows the best mAP after reranking, although the R@1000 is not as high as that of QE (RM3) or QE (SER). This may be due to the difference in distributions between the training and test datasets. When fine-tuning the T5-reranker, we used only MS MARCO, a question-answering dataset featuring real human-generated questions and answers similar to the TREC CAsT queries. However, since the answer passage set of TREC CAsT contains both MS MARCO and TREC CAR, the proposed T5-reranker may assign higher ranks to MS MARCO passages than to TREC CAR passages.

3.4.2 2020 TREC CAsT submission. We submitted four runs to TREC CAsT this year with the query expansion methods mentioned in the previous section, three of which use only the raw questions, and the last of which uses the manual questions. The results of the evaluation are shown in Table 2.

For the three runs using only raw utterances, the results are similar to the 2019 TREC CAsT evaluation set. That is, the baseline model without query expansion shows the best mAP although its R@1000 is lower than that of the other two results. For NDCG, the baseline model still outperforms at $NDCG$@3 and $NDCG$@5, which suggests that the top-3 and top-5 passages of the baseline model are more related than the other query-expanded passages. However, query expansion with RM3 boosts $NDCG$@1000 by 1.21%, perhaps because even though the pool with RM3 query expansion contains more related passages, the related passages rank lower at the retrieval stage.

For the run using manually rewritten utterances, we submitted only one result with T5-based QA-style query expansion. We still used the raw utterances and used the same model pipeline as that used for the raw-utterances-only run; the only difference is that we used the manually rewritten utterances for T5-squad query expansion. Compared to the baseline model, this query expansion slightly improves $NDCG$@3, $NDCG$@5, $NDCG$@1000, and mAP. These meager improvements suggest that such query expansion yields little new information.

4 CONCLUSION

This work demonstrates the importance of accounting for query-document mismatch in conversational systems. We propose CQR and SER, resolving three critical challenges—vocabulary mismatch, coreference, and subject ellipsis—and we conduct empirical studies to evaluate the performance of the proposed methods. CQR and SER improve retrieval results, and the T5 reranker further enhances the final performance.

ACKNOWLEDGMENTS

We thank Jheng-Hong Yang for valuable comments on the experience pipeline and experiment sharing.

REFERENCES

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[2] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2020.

[3] Rodrigo Nogueira and Kyunghyun Cho. Passage re-ranking with BERT. *arXiv preprint arXiv:1901.04085*, 2019.

[4] Ahmed Elgohary, Denis Peskov, and Jordan Boyd-Graber. Can you unpack that? Learning to rewrite questions-in-context. In *Empirical Methods in Natural Language Processing*, 2019.

[5] Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. Document expansion by query prediction. *ArXiv*, abs/1904.08375, 2019.

[6] Google Research. T5: Text-To-Text Transfer Transformer. https://github.com/google-research/text-to-text-transfer-transformer, 2019.

[7] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

[8] Peilin Yang, Hui Fang, and Jimmy Lin. 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*, SIGIR '17, pages 1253–1256, New York, NY, USA, 2017. Association for Computing Machinery.