Predicting Relevant Conversation Turns for Improved Retrieval in Multi-Turn Conversational Search

Esteban A. Ríssola[1]*, Mohammad Aliannejadi[2]**,
Manajit Chakraborty[1], Fabio Crestani[1]

[1] Università della Svizzera italiana (USI), Lugano, Switzerland
rissoe@usi.ch, chakrm@usi.ch, fabio.crestani@usi.ch
[2] University of Amsterdam, Amsterdam, The Netherlands
m.aliannejadi@uva.nl

Abstract. This technical report presents the work of Università della Svizzera italiana in TREC CAsT 2019. TREC CAsT was set up to advance research on conversational search systems. The goal of the track is to create a reusable benchmark for open-domain information-centric conversational dialogues and to establish a concrete and standard collection of data with information needs to make systems directly comparable. Given the complexity of natural language and the evolution of user's information need in a conversation with multiple turns, finding relevant context is not always straightforward. We developed a neural model for identifying relevant turn(s) corresponding to the given turn. Our model reformulates the information need of the user to take into account the conversational context to enhance the ad-hoc passage retrieval performance. Two of our runs also employ neural re-ranking of the passages post-retrieval. One of our runs was able to achieve above-median performance.

Keywords: Conversational Search, Multi-turn Conversations, Information-Seeking

1 Introduction

Conversational Information Seeking (CIS) has been highlighted as an important emerging research area in the SWIRL 2018 workshop report[3] on future trends in Information Retrieval. CIS is timely and important with increased adoption of a new generation of conversational "assistant" systems, including Amazon Alexa, Cortana, Bixby, Google Assistant, and many others. The goal of the TREC CAsT track [3] is to pursue CIS research and create a large-scale reusable test

* Corresponding Author

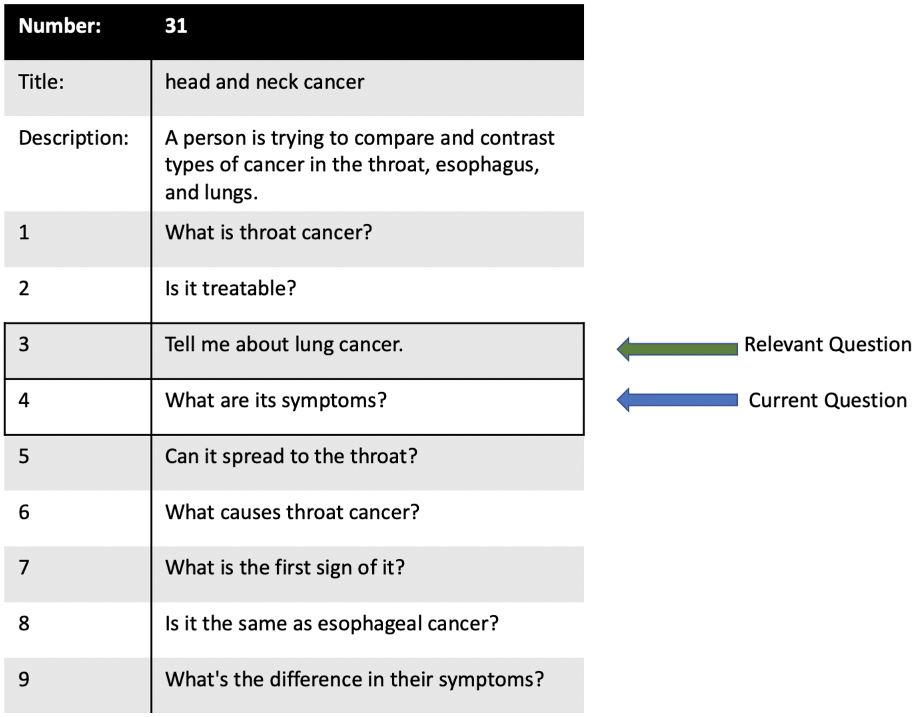
** Work done while Mohammad Aliannejadi was affiliated with Università della Svizzera italiana (USI).

[3] http://sigir.org/wp-content/uploads/2018/07/p034.pdf

collection for open-domain conversational search systems. The primary initial focus is targeted towards system understanding of information needs in a conversational format and finding relevant responses using contextual information. In particular, the track is motivated by long-running and complex tasks requiring multiple turns (possibly multiple sessions).

Although much work has been done on studying single-turn conversations, several challenges emerge for a conversational system in a multi-turn dialogue. Multiple turns in a conversation can be used to understand the user information need more effectively [9]. To understand the dependency of conversation turns, we have annotated each turn of the conversation with related turns in the conversation's context. The turn-relevance annotation enables us to understand and visualize the conversation evolution and dependencies. An utterance is *relevant* to the *current* one when:

- it is needed to clarify the *current* utterance or
- it augments the information need for the *current* utterance or
- it entails the *current* utterance.



Number:	31
Title:	head and neck cancer
Description:	A person is trying to compare and contrast types of cancer in the throat, esophagus, and lungs.
1	What is throat cancer?
2	Is it treatable?
3	Tell me about lung cancer.
4	What are its symptoms?
5	Can it spread to the throat?
6	What causes throat cancer?
7	What is the first sign of it?
8	Is it the same as esophageal cancer?
9	What's the difference in their symptoms?

Fig. 1. Identifying Relevant Utterances (Questions)

As an example, consider the following conversation extracted from the TREC CAsT 2019 evaluation dataset in Figure 1. As can be observed, utterance u_4 cannot be answered unless additional information from previous utterances, *i.e.,* the context, is included in the question. In this case, utterance u_3 provides the necessary context and thus, is considered *relevant* to u_4 [1]. In this work, we propose a simple yet effective neural model for predicting relevant and complementary conversation turns. The relevant utterances are then used to expand the current utterance to build a self-sufficient query, which is then passed to an ad-hoc Information Retrieval (IR) system to provide a ranked list of passages.

2 Modeling

We submitted four runs to the track. Two of those runs employed neural reranking after the ad-hoc retrieval stage. The retrieval step is preceded by a reformulation of utterances using the proposed relevance utterance prediction.

2.1 Dataset

The task was organized to focus on candidate information ranking in context with two main goals:

- *Read the dialogue context*: Track the evolution of the information need in the conversation, identifying salient information needed for the current turn in the conversation.
- *Retrieve Candidate Response Information*: Perform retrieval over a large collection of paragraphs (or knowledge base content) to identify relevant information.

As part of the track dataset, organizers released three open-sourced collection namely MS MARCO (MAchine Reading COmprehension) Ranking passages [7], TREC CAR 2018 [5] paragraph collection and News article from Washington Post (WAPO)[4]. MS MARCO dataset was released by Microsoft in 2018 and comprises of 8,841,823 passages – extracted from 3,563,535 web documents retrieved by Bing. TREC-CAR (Complex Answer Retrieval) collection is a corpus of 20 million paragraphs harvested from a snapshot of Wikipedia. The TREC CAsT 2019 track came with its own set of 30 training conversations and 50 evaluation conversations[5].

2.2 Relevant Utterance Prediction

For training the relevant utterance classifier, we needed a pre-labeled set of questions that were marked *relevant* to the current question. This labeling was done by three human annotators who were provided with the current utterance

[4] https://ir.nist.gov/wapo/

[5] https://github.com/daltonj/treccastweb/tree/master/2019/data

and all the previous utterances to that point. The annotators were supposed to select one or more utterance(s) that seemed relevant to or would help clarify the current question. After independent labeling, we computed the percentage agreement among the annotators, and if at least two of the three annotators agreed on the same set of relevant question(s) (*i.e.,* the agreement score was greater than or equal to 66.67%) in the first round, we recorded them. In the second round, for questions that had an agreement score below 66.67%, the three annotators deliberated and arrived at an agreement on the relevance of questions. Of the total 748 questions (from 80 conversations), there was an agreement on 625 questions after the first round. The rest 123 utterance annotations were resolved in the second round by rigorous discussion. This exercise was carried out for both the training and evaluation conversations.

We employed a high-dimensional language and position representation (pre-trained BERT-Base) to predict the relevant utterances corresponding to the current utterance. We initialize the BERT parameters with the model that is pre-trained for the language modeling task on Wikipedia. BERT has recently outperformed state-of-the-art models in a number of language understanding and retrieval tasks [4]. Rectified linear unit (ReLU) is employed as the activation function in the hidden layers, and a softmax function is applied on the output layer to compute the probability of each label (*i.e.,* relevant or non-relevant). To train our classifier, we used a cross-entropy loss function and followed the point-wise learning approach.

2.3 Retrieval and Re-ranking

Before feeding the utterances as queries to the document retrieval model, we performed two steps of query processing and reformulation, as follows:

1. We observed from the data that a large share of utterances was dependent on previous turns in the conversation. Hence, it is imperative to perform coreference resolutions on such utterances to make them a complete question in itself. To this end, we use AllenNLP [6] coreference resolution tool. For self-contained utterances and first utterances in the conversations, the output of the tool is the same as the original utterance.
2. On this modified query, we perform stopword removal, removal of special characters, and tokenization.
3. To this modified query, we added the relevant utterances selected by the relevance prediction model to make the query complete. Two of the runs also include adding the first utterance to the modified query since it is supposed to introduce the premise for the complete conversation.

After reformulating the utterances, we pass the resulting query set to various document retrieval models. For document (passage) retrieval, we employed Galago[6] and indexed the MS MARCO collection. We used Okapi-BM25 [10] term

[6] https://www.lemurproject.org/galago.php

matching retrieval model to retrieve the passages. Two of the runs that we submitted had an additional step of re-ranking the retrieved passages. For this, we employed the neural re-ranking model based on BERT, as proposed by Nogueira and Cho [8]. The job of the re-ranker is to estimate a score s_i of how relevant a candidate passage d_i is to a query q. The model is fed the query as sentence A and the passage text as sentence B. The query is truncated to have at most 64 tokens. The passage text is also truncated, such that the concatenation of query, passage, and separator tokens have a maximum length of 512 tokens. The implementation uses a BERT LARGE model as a binary classification model *i.e.,* it uses the [CLS] vector as input to a single layer neural network to obtain the probability of the passage being relevant. We compute this probability for each passage independently and obtain the final list of passages by ranking them with respect to these probabilities.
We start training from a pre-trained BERT model and fine-tune it to our re-ranking task using the cross-entropy loss:

$$L = -\sum_{j \in J_{pos}} log(s_j) - \sum_{j \in J_{neg}} log(1 - s_j), \tag{1}$$

where J_{pos} is the set of indexes of the relevant passages and J_{neg} is the set of indexes of non-relevant passages in top-1,000 documents retrieved with BM25.

The four runs that were submitted to the track are listed below:

1. *galago_rel_q*: In this run, we only pass the reformulated query by expanding the current utterance with relevant utterance(s) to the retrieval system.
2. *bertrr_rel_q*: For this run, we pass the reformulated query by expanding the current utterance with relevant utterance(s) to the retrieval system. After this, the retrieval results are re-ranked using BERT.
3. *galago_rel_1st*: In this run, the reformulation involves adding the relevant utterance(s) along with the first utterance in the conversation to the current one. The idea behind is that usually, the first turn in the conversation sets the context of the conversation and hence may be useful in clarifying the current utterance.
4. *bertrr_rel_1st*: This is similar to *galago_rel_1st*, with the only difference being that the retrieved results are re-ranked using BERT.

3 Results and Conclusion

Table 1 shows the overall average performance of our submitted runs against three metrics proposed by the track committee. We compare the results against the average median performance of all submitted runs to TREC CAsT. From the table, we can observe that the only run bertrr_rel_1st supersedes the TREC median performance on two metrics, MAP@5 and NDCG@5. This highlights the importance of the first utterance in a conversation. On further analysis, we found that over 50% of the relevant utterances were found at the first position in

Table 1. Passage retrieval performance of the submitted runs before and after the bug fix.

	MAP@5	**NDCG@5**	**NDCG@10**
galago_rel_q	0.0251	0.1823	0.1900
bertrr_rel_q	0.0406	0.2977	0.2977
galago_rel_1st	0.0271	0.1952	0.1981
bertrr_rel_1st	**0.0432**	**0.3027**	0.2994
TREC CAsT median	0.0337	0.2656	**0.3622**

a conversation, which corroborates our initial hypothesis that the first question sets the premise for the rest of the conversation.

Another observation from the table is that the neural re-ranking approaches perform better than the simple ad-hoc retrieval systems reinstating our belief that neural models such as BERT are better at capturing the relevance of passages to queries than simple heuristic or IR models.

In this technical report, we presented the outcome of the participation of Università della Svizzera italiana in TREC CAsT 2019. The proposed model reformulates the information need of the user to take into account the conversational context to enhance the ad-hoc passage retrieval performance. To this end, conversations released by the TREC CAsT 2019 have been annotated to identify the relevant utterance from a conversation's context. Based on the labels obtained, we employed a high-dimensional language and position representation to predict the relevant utterances corresponding to the current utterance. Results showed that adding the predicted relevant utterance(s) along with the first utterance in the conversation to the current one in combination with neural passage re-ranking provided the best performance.

4 Conclusions and Future Work

In this paper, we presented a BERT fine-tuning model to predict the relevance of the utterances. Our proposed model outperformed competitive classification baselines. Also, we demonstrated its effectiveness in improving the performance of document retrieval models. As future work, we plan to perform a similar analysis on information-seeking conversational systems with the ability of asking clarifying questions [2]. The proposed task in the TREC CAsT track is aiming to evaluate systems that retrieve responses to the user's utterances. However, as indicatd by the organizers, one of the future plans is to include a subtask aiming to evaluate clarifying questions in a conversation.

References

1. Aliannejadi, M., Chakraborty, M., Ríssola, E.A., Crestani, F.: Harnessing evolution of multi-turn conversations for effective answer retrieval. In: Proceedings of

the 2020 Conference on Conference Human Information Interaction and Retrieval, CHIIR 2020, Vancouver, British Columbia, Canada, March 14-18, 2020 (2020)

2. Aliannejadi, M., Zamani, H., Crestani, F., Croft, W.B.: Asking Clarifying Questions in Open-Domain Information-Seeking Conversations. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019. pp. 475–484 (2019)
3. Dalton, J., Xiong, C., Callan, J.: Cast 2019: The conversational assistance track overview. In: Proceedings of the Twenty-Eighth Text REtrieval Conference, TREC 2019, Gaithersburg, Maryland, USA, November 13-15, 2019 (2019)
4. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers). pp. 4171–4186 (2019)
5. Dietz, L., Gamari, B., Dalton, J., Craswell, N.: TREC Complex Answer Retrieval Overview. In: Proceedings of the Twenty-Seventh Text REtrieval Conference, TREC 2018, Gaithersburg, Maryland, USA, November 14-16, 2018 (2018)
6. Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N.F., Peters, M., Schmitz, M., Zettlemoyer, L.S.: AllenNLP: A Deep Semantic Natural Language Processing Platform (2017)
7. Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., Deng, L.: MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. In: Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016. (2016)
8. Nogueira, R., Cho, K.: Passage re-ranking with BERT. CoRR abs/1901.04085 (2019), `http://arxiv.org/abs/1901.04085`
9. Radlinski, F., Craswell, N.: A Theoretical Framework for Conversational Search. In: Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval, CHIIR 2017, Oslo, Norway, March 7-11, 2017. pp. 117–126 (2017)
10. Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M., Gatford, M.: Okapi at TREC-3. In: Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994. pp. 109–126 (1994)