Context Propagation in Conversational Search Utterances

Participation of the CNR Team in CAsT 2022

Ida Mele[1], Cristina Ioana Muntean[2],
Franco Maria Nardini[2], Raffaele Perego[2], and Nicola Tonellotto[3]

[1] IASI-CNR, Rome, Italy
[2] ISTI-CNR, Pisa, Italy
[3] University of Pisa, Italy
ida.mele@iasi.cnr.it cristina.muntean@isti.cnr.it
francomaria.nardini@isti.cnr.it raffaele.perego@isti.cnr.it
nicola.tonellotto@unipi.it

Abstract. Every year, NIST organizes the **T**ext **RE**trieval **C**onference (TREC) which gathers competitions for forecasting research on text retrieval. Since 2019, it has included a contest on conversational assistant systems, called **C**onversational **As**sistant **T**rack (CAsT) with the purpose of helping research on conversational information systems. CAsT provides test collections for open-domain conversational seeking where the users can ask multiple questions to the system and get answers like in a multi-turn conversation. For our participation in CAsT 2022, we implemented an architecture consisting of two steps: utterance rewriting and passage retrieval. Each run is based on a different utterance rewriting technique for enriching the raw utterance with context extracted from the previous utterances and/or from the replies in the conversation. Three of our approaches are completely automatic, while another one uses the manually rewritten utterances provided by the organizers of TREC CAsT 2022.

1 Introduction

Conversational Information Seeking (CIS) is an emerging area of research that poses new challenges in information retrieval both in terms of effectiveness [7] and efficiency [2]. The increasing popularity of CIS is due to the advances in automatic speech recognition and understanding tools that are largely employed in smart home assistants, smartphones, and wearable devices.

Thanks to TREC CAsT [1], the researchers can experiment with their methodologies that aim to improve the automatic understanding of the users' requests and to find relevant responses using contextual information. The typical scenario is a conversational system that helps the user to fulfill her information need by answering vocal questions. The search goes on as a multi-turn dialogue between the user and the system, where the requests are often general and vague due to the ambiguity of natural language as well as the lacking of context. The missing context is often a subject mentioned before in the conversation (e.g., in the previous requests or replies), and the user refers to it indirectly with pronouns. The operation of adding context to ambiguous and incomplete requests is challenging due to the complexity of understanding the semantic meaning of previous questions and their answers. Furthermore, another challenge is represented by the fact that the system response is not just a list of relevant documents, but, rather it is constrained to a short text passage, which can be a summary of sentences extracted from the documents relevant to the user request. The text passage must be brief as it is returned to the user through a voice interface or a small screen of a mobile.

This year, CAsT proposes two tasks. The primary task is response retrieval focusing on providing fluent and relevant responses that may summarize more passages coming from different documents. Plus, a novel mixed-initiative subtask where, for each conversation turn, the system may reply to a user request or may ask a question for clarification. This question is chosen from a pool of questions for each turn, resulting in a dialogue tree representing all the possible conversations. Compared to last year, CAsT 2022 provides a series of text responses for each turn. Each response can be a passage or a summary generated from one or more passages and has at least one passage called *provenance* for evaluating the provenance ranking. Instead of one conversation per topic, each topic has multiple conversations and information needs on a shared topic (i.e., a dialogue tree). Lastly, the mixed-initiative sub-task evaluates the ability of systems to use mixed-initiative. As a consequence, CAsT allowed three submission classes: (1) *automatic*, where raw utterances are reformulated with automatic rewriting or expansion methods, (2) *automatic_MI*, where the response ranking is from the mixed-initiative sub-task after using feedback, clarification, etc., and (3) *manual*, where human assessors manually rewrite raw utterances.

We only participate in the primary task for information-seeking conversations and submitted one manual run and 3 automatic runs explained in Sec. 3.

2 Dataset

TREC CAsT 2022 has provided a dataset including evaluation topics (i.e., search conversations), a mixed-initiative question pool for the mixed-initiative subtask, and three document collections.

For each evaluation topic, CAsT 2022 provides a dialogue tree representing all possible conversations between the user and the system.

The three document collections are: (1) *KILT Wikipedia* dump from 2019/08/01 consisting of 5M articles, (2) *MS MARCO V2 document corpus* used in the 2021 TREC Deep Learning Track and consisting of 11.9M documents from Bing search, and (3) *TREC Washington Post collection* (V4 2020) consisting of 728,626 news articles from 2012 to 2020 (this data requires a signed license agreement with NIST).

The core task of the system is to return a response after every turn using context extracted from the previous-turn utterances or replies. For each turn, the system returns a response that is fluent and suitable for the users. It should not contain irrelevant information or repetitions, and it should be short so that it can be shown on a small screen or read vocally to the user (e.g., response text is limited to a maximum of 250 words as measured by SpaCy v3.3). Each response can have one or more source passages as provenance and this information is used for evaluating the retrieval performance of the system.

The documents are split into passages (up to 250 words), and the passage segmentation is performed using SpaCy's SentenceRecognizer pipeline component for sentence detection with a fixed non-overlapping passage size. CAsT organizers gave to the participants the option of processing the collection themselves to generate passage splits using the provided tools or requesting the processed corpus from the organizers. We requested the processed corpus and run our experiments on it.

3 Methodologies

Our framework consists of two steps: *utterance rewriting* and *passage retrieval.* All our methods rely on a Python NLP toolkit for extracting various linguistic features from the utterances [3]. We perform utterance rewriting to enrich the raw utterance with the missing context, then we use the rewritten utterances to retrieve the passages. For indexing and querying the collections, we used PyTerrier 0.7.1, based on Terrier 5.6, employing traditional unsupervised sparse retrieval (e.g., DPH hypergeometric weighting model).

We assume that a user has an information need that intends to fulfill by asking questions (a.k.a. utterances) to a conversational search system. A raw utterance, u_i, represents the natural language question issued by the user to the system. This is the input of our automatic utterance rewriting module whose output is an enriched utterance, $\hat{u}_i$, used to retrieve passages from the document collections. The purpose of the utterance rewriting module is to add missing context to the raw utterance so that the user can get a good answer to her request. Our runs are inspired by our previous works on topic propagation in multi-turn conversational searches [4, 5].

- `CNR-run1`. This run automatically rewrites the current utterance by adding the topics extracted from the previous automatically rewritten utterance provided by CAsT 2022.
- `CNR-run2`. This run adds to the current utterance the topics extracted from the previous manually rewritten utterance provided by CAsT 2022.
- `CNR-run3`. This run enriches the current utterance with the first sentence of the response to the previous utterance.
- `CNR-run4`. This run enriches the current utterance with the top-5 frequent terms extracted from the response to the previous utterance.

In all our runs, the topics are extracted from utterances using SpaCy noun chunks (objects or subjects).

4 Experimental Results

CAsT 2022 provided for the runs two different evaluations: a *lenient* evaluation where passages at least "slightly meet" the need of the request at that turn (relevance level 1), and a *strict* evaluation where passages must "moderately meet" the need of the request at that turn (relevance level 2).

In Table 1, we report the values of the *Mean Average Precision* (MAP) and the *normalized Discounted Cumulative Gain* (nDCG@20) for our four runs. Plus, we report the worst, median, and best performances provided by CAsT 2022 for each query and averaged over all the queries.

As expected, the worst results are achieved by CNR-run1 as it enriches the current utterance with context extracted from the previous automatically rewritten utterance, and it does not take into account the response. On the other hand, CNR-run2 performs pretty well as it uses context from the previous manually rewritten utterance provided by CAsT 2022. We can also notice that the values of nDCG@20 achieved by our runs are close to the median values. Overall, this performance could benefit from a further step of re-ranking.

Table 1. Performance of CAsT 2022 runs: averaged over all queries

	Lenient		Strict	
	MAP	nDCG@20	MAP	nDCG@20
CNR-run1	0.0796	0.1524	0.0593	0.1524
CNR-run2	0.0951	0.1832	0.0758	0.1832
CNR-run3	0.0867	0.1671	0.0679	0.1671
CNR-run4	0.0843	0.1710	0.0717	0.1710
Avg-Min	0.018	0.035	0.012	0.035
Avg-Median	0.176	0.320	0.147	0.320
Avg-Max	0.439	0.667	0.426	0.667

Re-ranking. We used the model by Nogueira and Cho [6] to re-rank the results from the previous stage. The model fine-tunes the BERT base pre-trained model for re-ranking on the MSMARCO passage retrieval dataset. For each query, we used as input for the re-ranking step the top 200 results. The performance of our runs after re-ranking is shown in Table 2.

Table 2. Performance of our runs after re-ranking

	MAP	nDCG@20	recip-rank	P@1
CNR-run1	0.0844	0.1847	0.5069	0.4198
CNR-run2	0.1037	0.2200	0.5730	0.4568
CNR-run3	0.0875	0.1917	0.5222	0.4259
CNR-run4	0.0802	0.1851	0.5051	0.4074

5 Conclusions and Future Work

We have presented the methodologies implemented for our participation in CAsT 2022. Our approaches aim to enrich the raw utterances using topical keywords extracted from the previous utterances or from their responses.

As future work we would like to improve the utterance-dependency understanding in order to better capture the dependencies between the current utterance and those of the previous turns as well as their responses with the purpose of improving the automatic rewriting and enrichment of raw utterances.

References

1. TREC CAsT. `http://www.treccast.ai/`.
2. O. Frieder, I. Mele, C.I. Muntean, FM Nardini, R. Perego, and N. Tonellotto. Caching historical embeddings in conversational search. *ACM Trans. Web*, dec 2022. Just Accepted.
3. SpaCy library. `https://spacy.io/usage/linguistic-features`.
4. I. Mele, C. I. Muntean, F. M. Nardini, R. Perego, N. Tonellotto, and O. Frieder. Topic Propagation in Conversational Search. In *SIGIR 2020*, pages 2057–2060. ACM, 2020.
5. I. Mele, C. I. Muntean, F. M. Nardini, R. Perego, N. Tonellotto, and O. Frieder. Adaptive utterance rewriting for conversational search. In *IPM 2021*. Elsevier, 2021.
6. R. Nogueira and K. Cho. Passage Re-ranking with BERT. *arXiv preprint arXiv:1901.04085*, 2019.
7. H. Zamani, J. R Trippas, J. Dalton, and F. Radlinski. Conversational information seeking. *arXiv preprint arXiv:2201.08808*, 2022.