Topical Enrichment of Conversational Search Utterances

Participation of the `HPCLab-CNR` Team in `CAsT 2020`

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

The TREC Conversational Assistant Track (CAsT) provides test collections for open-domain conversational search systems with the purpose of pursuing research on Conversational Information Seeking (CIS). For our participation in CAsT 2020, we implemented a modular architecture consisting of three steps: (i) automatic utterance rewriting, (ii) first-stage retrieval of candidate passages, and (iii) neural re-ranking of candidate passages. Each run is based on a different utterance rewriting technique for enriching the raw utterance with context extracted from the previous utterances in the conversation. Two of our approaches are completely unsupervised, while the other two rely on utterances manually classified by human assessors. These approaches also employ the canonical responses for the automatically rewritten utterances provided by CAsT 2020.

1 Introduction

Conversational Information Seeking (CIS) has recently gained interest due to the popularity of conversational assistant systems and to the recent advances in automatic speech recognition and understanding. Conversational assistant systems help users in a wide range of activities such as checking the weather forecast, managing music streaming services, or performing e-commerce transactions. They are used in chatbots and smart home devices (e.g., Google Home, Amazon Alexa) as well as in wearable devices and smartphones (e.g., Apple Siri, Google Assistant, Microsoft Cortana).

Although conversational assistants are good at performing simple well-defined actions, their ability to support conversational information seeking is still very limited. Indeed, the seeking of information evolves as a dialogue between the user and the system, and the search goes on as turns of user natural-language questions, i.e., utterances. The retrieval of documents relevant to an utterance is difficult due to the informal use of natural language in the speech and the lacking of context. Moreover, adding context to ambiguous utterances is tricky due to the complexity of understanding the semantic meaning of previous utterances and their answers.

Thanks to CAsT data, researchers can experiment with their methodologies that aim to improve the automatic understanding of the users' information needs and to find the relevant responses using contextual information.

2 Dataset

The TREC Conversational Assistant Track[4] (CAsT) 2020 provided a dataset including search conversations and document collections. Compared to last year (CAsT 2019), CAsT 2020 is based on two collections: (1) TREC CAR (Complex Answer Retrieval) containing $\sim 29M$ of passages extracted from approximately $5M$ Wikipedia articles, and (2) MS-MARCO (MAchine Reading COmprensation) made of $\sim 8M$ passages from answer candidates of the Bing search engine.

CAsT 2020 dataset provides 25 conversations, each having from 6 to 13 utterances for a total of 216 utterances. Compared to CAsT 2019, the CAsT 2020 conversations do not include topic titles or descriptions, so topics unfold throughout the dialogue. Also, CAsT 2020 dataset provides a canonical system response of the previous turn utterance.

An example of conversation is as follows: (1) *"How do you make Japanese Yakiniku?"*, (2) *"Can the sauce be used in other dishes?"*, (3) *"What are the best Yakiniku restaurants in Tokyo?"*, and (4) *"Tell me about three star Michelin sushi restaurants there"*. While the first and third utterances are relatively easy to process by an Information Retrieval (IR) system, in the second utterance there is a reference to the subject of the conversation, i.e., Japanese Yakiniku, and the fourth utterance refers to a specific location, i.e., Tokyo, just introduced in the conversation. Even in this short piece of a conversation, it is possible to identify different kinds of utterances. The first and third utterances do not require any rewriting to be successfully answered. On the other hand, the second utterance lacks context and adding the keywords "Japanese Yakiniku" is mandatory for retrieving relevant documents. We also observe a topic shift in the third utterance, i.e., Japanese Yakiniku → Tokyo, that makes the fourth utterance referring to a newly introduced topic (Tokyo) and not to the previous one (Japanese Yakiniku). Even in this case, the utterance needs to be rewritten and enriched with the keyword "Tokyo" to be successfully answered by an automatic IR system.

2.1 Search Conversations

By carefully inspecting the utterances in the CAsT 2020 dataset, we noticed some common patterns in the conversations:

(a) Some utterances are self-explanatory as they do not lack context. In particular, the first utterance of each conversation is self-explanatory, but sometimes we can observe them in the middle of the conversation. Very often self-explanatory utterances in the middle of the conversation introduce a subtopic exploration or even a topic shift.

[4] http://www.trecCAsT.ai/

(b) In some cases, the topic of the first utterance dominates the conversation. Follow-up utterances are not self-explanatory and refer to the topic introduced at the beginning of the conversation. These utterances depend on the *first topic* of the conversation.

(c) In other cases, utterances are not self-explanatory and refer to some topics mentioned in a previous utterance (different from the first utterance). Hence, these utterances need to be enriched with some context extracted from the *previous utterances* in the conversation.

(d) Similarly to (c), the utterances are not self-explanatory and refer to some topics mentioned in previous utterances and/or in their answers. These utterances are even more tricky as they cannot be enriched with context extracted from only the previous utterances but rather from their results. We will refer to these utterances as depending on *previous responses*.

2.2 Utterance Labeling

Given our previous observations, we asked human assessors to manual label raw utterances. In particular, assessors familiar with the challenges in conversational search evaluated 216 utterances from 25 conversations using four labels:

- Self-Explanatory (SE): the utterance is self-explanatory, so the context is fully provided;
- First Topic (FT): the utterance misses context which depends on the first utterance;
- Previous Topic (PT): the utterance misses context which depends on the previous utterance;
- Previous Response (PR): the utterance misses context which depends on the previous utterance as well as its canonical response.

For most of the utterances (i.e., 162 out of 216) we could see a full agreement among the assessors. To measure the inter-annotator agreement, we also computed the Fleiss' Kappa [2] that gives a measure of how consistent are the assessors' ratings. It is computed as $\kappa = \frac{\bar{P}-\bar{P}_e}{1-\bar{P}_e}$, where $1 - \bar{P}_e$ gives the degree of agreement that is attainable above chance, and $\bar{P} - \bar{P}_e$ gives the degree of agreement actually achieved above chance. If $\kappa = 1$ the agreement is complete, while $\kappa \leq 0$ means no agreement. We registered a value of 0.82 which, according to the table for interpreting κ values provided in [3], corresponds to an "almost perfect" agreement.

3 Methodologies

Our framework consists of three steps: (1) *utterance rewriting*, (2) *candidate passage retrieval*, and (3) *neural re-ranking*.

All our methods employ a Python NLP toolkit for extracting various linguistic features from the utterances[5] and perform utterance rewriting to enrich

[5] SPACY library available at `https://spacy.io/usage/linguistic-features`.

the raw utterance with the missing context. After utterance rewriting, in the first-stage retrieval, we use the rewritten utterances to retrieve the candidate passages and narrow down the search space. Then, neural re-ranking exploits a contextualized language model based on BERT to re-rank the passages [5].

3.1 Automatic Utterance Rewriting

We assume that a user has an information need that intends to fulfill by issuing utterances to a conversational IR 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 candidate passages from the document collection. The purpose of the utterance rewriting module is to add the missing context to the raw utterance so that the user can get a good response to her request.

Runs with Unsupervised Utterance Rewriting. These runs are inspired by our work on investigating topic propagation in multi-turn conversations [4]. They use the raw utterances from the previous turns of the conversation.

- `HPCLab-CNR-run2` rewrites the raw utterance by using the topics extracted from all the raw utterances of the previous turns. These topics are noun chunks, different from pronouns, which are either subjects or objects, while verbs are not used. The drawback of this approach is that propagating all the context seen during the conversation can lead to noisy results, especially for those conversations where the focus of interest may change.
- `HPCLab-CNR-run4` tries to address the main weakness of `run2` as it rewrites the utterance by using the topics extracted from two main sources: (i) the raw utterance of the previous turn (*previous topic*); (ii) the raw utterance that has introduced a topic shift (*focus topic*). To detect the possible topic shifts, our approach employs some clue phrases (e.g., "tell me about", "what about"). Namely, at the beginning of the conversation, the focus topic is represented by the topic extracted from the first utterance, and when a topic shift occurs the focus topic is updated with the topic extracted from the utterance where the topic shift is detected. The topics in `run4` are noun chunks (objects or subjects) plus the full verbs.

Runs with Utterance Rewriting based on Classification. These runs perform the automatic rewriting of raw utterances using the utterance classification explained in Section 2.2. The classification is used to determine the best enrichment for the current utterance. In particular:

- If the raw utterance is labeled as SE, no rewriting is applied.
- If the raw utterance is labeled as FT, it is enriched with the topic extracted from the first utterance of the conversation.
- When the utterance label is PT, the rewriting is performed using the topic extracted from the previous enriched utterance.

- When the label is PR, the utterance is rewritten using the topic extracted from the previous enriched utterance. Plus, the context (e.g., topics or keywords) from the canonical response of the previous automatically rewritten utterance is added at the end of the enriched utterance.

These runs differ from each other because they extract the topics from utterances and the context from canonical responses in different ways:

- `HPCLab-CNR-run1` extracts the topics from the enriched utterance using the noun chunks (objects or subjects). The context from the canonical response for the automatically rewritten utterance of the previous turn (provided by CAsT 2020) is represented by the keywords of the canonical response (after stopword removal).
- `HPCLab-CNR-run3`, topics are the noun chunks (objects or subjects) that are recognized as named entities by TagMe[6] (with threshold = 0.1). They are extracted from both the enriched utterances and their canonical responses. Using only named entities has the advantage to clean a noisy context, although, in some cases, the set of recognized named entities can be empty which may lead to poor context enrichment.

Compared to `run2` and `run4`, these approaches are based on manual labels, they use context from canonical responses (when needed), and they extract topics from the enriched utterances of the previous turn. As we will see in Section 5, the approaches implemented for `run1` and `run3` perform better compared to the completely unsupervised approaches.

4 Experimental setting

Metrics. The effectiveness of the rewriting techniques is evaluated with traditional TREC metrics. In particular, Mean Average Precision (`map`) and normalized Discounted Cumulative Gain (`nDCG`) for cutoffs at 3, 5, and 1K. The use of small cutoffs, such as 3 and 5, is common for the conversational search task since the user expects to receive one crisp answer rather than a long list of potentially relevant results.

First-stage retrieval. For indexing and querying the CAsT dataset, we used Indri[7]. We indexed the two datasets by removing stopwords and we applied the *Krovetz* stemmer for stemming. As Indri querying method we used the Indri language model based on Dirichlet smoothing with parameter $\mu = 2500$. We also applied pseudo-relevance feedback (PRF) based on the RM3 algorithm [1] using 20 keywords taken from the top 20 results and $\gamma = 0.5$.

Neural re-ranking. We used the model by Nogueira and Cho [5] to re-rank the results from the previous stage. The model fine-tunes the BERT base pre-trained model for re-ranking on the MS-MARCO passage retrieval dataset. For each query, Indri retrieves $1K$ results which are the input for the re-ranking step.

[6] https://pypi.org/project/tagme/0.1.2/

[7] https://www.lemurproject.org/indri.php

5 Experimental Results

In Table 1, we report the values of the following metrics map@1K and nDCG@k (with $k = 3$, 5, and 1K) for our four runs. As we can see, the worst results are achieved by run2 and run4 as they do not use any utterance classification and any context from the canonical response for the previous utterance.

Better performances are achieved by run1 and run3 as they enrich the raw utterances leveraging the utterance classification and add the context extracted from the previous canonical response.

CAsT 2020 also provided for each query/utterance the worst, median, and best performance for 35 raw runs, 8 canonical runs, and 12 manual runs. We computed the average over all the queries, and the results are shown in Table 2.

As expected, the performances of the two unsupervised runs (run2 and run4) are close to the raw median values reported in Table 2. While the performances of run1 and run3 are close to the canonical median values. Moreover, the nDCG@3 values achieved with run1 and run3 are 0.313 and 0.331, respectively. They are slightly better than the CAsT 2020 BERT baseline for automatic-canonical results (nDCG@3 = 0.3).

Table 1. Performance of our runs

Run	nDCG@3	nDCG@5	nDCG@1K	map@1K
Run1 (canonical)	0.313	0.304	0.403	0.220
Run2 (raw)	0.275	0.270	0.360	0.185
Run3 (canonical)	0.331	0.319	0.422	0.236
Run4 (raw)	0.292	0.275	0.358	0.194

Table 2. Performance of CAsT 2020 runs: averaged over all queries

Run		nDCG@3	nDCG@5	nDCG@1K	map@1K
Raw	worst	0.006	0.007	0.046	0.006
	median	0.279	0.273	0.375	0.180
	best	0.733	0.687	0.710	0.492
Canonical	worst	0.058	0.066	0.142	0.057
	median	0.337	0.328	0.426	0.233
	best	0.634	0.606	0.657	0.440
Manual	worst	0.016	0.023	0.174	0.029
	median	0.414	0.398	0.489	0.263
	best	0.724	0.683	0.698	0.478

5.1 Automatic Utterance Rewriting Limits

By inspecting the rewritten utterances, we could notice some limitations of our automatic approaches. First of all, we need to replace acronyms to have a better understanding of the utterance, e.g., in "Why did it replace PB?" → *PB* refers to a variety of oranges *Parson Brown.* Also, some topic shifts are hard to find and simple clues do not help much. For example, consider this part of a conversation (1) *"What is the climate like in Utah?"*, (2) *"How does Salt Lake City differ?"*, (3) *"What is its main economic activity?"* → *its* refers to *Salt Lake City* rather than *Utah*, but this topic shift may be undetected.

6 Conclusions

In this report, we have presented the methodologies implemented for our participation in CAsT 2020.

As future work, we plan to develop an improved rewriting approach that automatically replaces acronyms (e.g., *PB* with *Parson Brown*). Also, we will improve our function for detecting topic shifts and extracting context from canonical responses. Lastly, we would like to experiment with automatic classifications of utterances where classifiers can be trained on a set of manually labeled utterances to predict the labels of future ones.

References

1. C. L. A. Clark, S. Büttcher, and G. V. Cormack. *Information Retrieval: Implementing and Evaluating Search Engines.* MIT Press, 2010.
2. J. L. Fleiss. Measuring Nominal Scale Agreement among Many Raters. *Psychological Bulletin*, 76(5):378–382, 1971.
3. J. R. Landis and G. G. Koch. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174, 1977.
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. R. Nogueira and K. Cho. Passage Re-ranking with BERT. *arXiv preprint arXiv:1901.04085*, 2019.