Finding Context through Utterance Dependencies in Search Conversations

Participation of the CNR Team in CAsT 2021

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

To help research on Conversational Information Seeking, TREC has organized a competition on conversational assistant systems, called **C**onversational **As**sistant **T**rack (CAsT). It provides test collections for open-domain conversational search systems. For our participation in CAsT 2021, we implemented a three-step architecture consisting of: (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 and/or replies in the conversation. Two of our approaches use only raw utterances and other two use utterances plus the canonical responses of the automatically rewritten utterances provided by CAsT 2021. Our approaches also rely on utterances manually classified by human assessors using a taxonomy defined ad hoc for this task.

1 Introduction

The increasing popularity of conversational assistant systems as well as the advances of automatic-speech recognition and understanding tools have brought novel attention to Conversational Information Seeking (CIS).

A conversational assistant system helps the user in different activities such as checking the weather forecast, searching for information, or performing e-commerce transactions. Such systems are used in wearable devices, smartphones (e.g., Apple Siri, Google Assistant, Microsoft Cortana), and smart home devices (e.g., Google Home, Amazon Alexa).

The ability of conversational assistant systems to support conversational information seeking is still limited due to the complexity of the search task. Indeed, information seeking often evolves as a multi-turn dialogue between the user and the system, so the search goes on as natural-language questions (i.e., utterances) and answers. The retrieval of documents relevant to an utterance is difficult due

to ambiguity of natural language as well as the lacking of context (a subject may be mentioned before in the conversation). The operation of adding context to ambiguous/incomplete utterances is challenging due to the complexity of understanding the semantic meaning of previous utterances and their answers.

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

2 Dataset

The TREC Conversational Assistant Track[4] (CAsT) 2021 provided a dataset including search conversations and document collections. Compared to previous years, CAsT 2021 is based on three collections: (1) English Wikipedia (KILT Wikipedia dump from 2019/08/01) consisting of 5M articles, (2) MS MARCO Web documents (first version) consisting of 3.2M documents from Bing search, and (3) TREC Washington Post collection (V4 2020) consisting of 728,626 news articles from 2012 to 2020.[5] Documents are split into passages, and the passage segmentation is performed using tools (available in the TREC CAsT tools) using SpaCy sentence detection with a fixed non-overlapping passage size.

CAsT 2021 dataset is made of 26 conversations, each having from 6 to 13 utterances for a total of 239 utterances. The dataset also provides canonical system responses for the utterances.

An example of conversation is as follows: (1) *"I'd like to learn more about frogs. What's the biggest one?"*, (2) *"What's been done to protect them?"*, (3) *"Has that been effective?"*, (4) *"How can I help?"*, and (5) *"Okay, that's the biggest. What is the smallest?"* . While the first utterance is relatively easy to process by an Information Retrieval (IR) system, the follow-up utterances have references to previous subjects mentioned in the conversation or in the answers. For example, *"What's been done to protect them?"*, "them" refers to the biggest frogs. Hence, the second utterance lacks context and adding the missing keywords is mandatory for a better retrieval of the relevant documents. Third utterance also lacks context and it depends on the answer to the previous utterance since "that" refers to the strategy to protect the frogs. Also, we can observe a topic shift in the fifth utterance as the user is interested in knowing more about smallest frogs. Even in this case, the utterance needs to be rewritten and enriched with the keyword "frogs" to be successfully answered by an automatic IR system.

2.1 Search Conversations

By carefully inspecting the utterances in the CAsT 2021 dataset, we noticed some common patterns in the conversations:

[4] http://www.trecCAsT.ai/
[5] This data requires a signed license agreement with NIST.

(a) Some utterances do not lack context, and we define them as *self-explanatory* utterances. As an example, the first utterance of each conversation is always self-explanatory, but there could be other self-explanatory utterances in the middle of the conversation. Very often these utterances introduce a subtopic exploration or even a topic shift.
(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 must be enriched with context extracted from the previous utterances and also from their responses. We will refer to these utterances as depending on *previous responses*.

2.2 Utterance Labeling

Given our previous observations, we asked human assessors to manually check the raw utterances along with the manually rewritten utterances with the purpose of labeling raw utterances based on their dependencies. In particular, assessors familiar with the challenges in conversational search evaluated the 239 utterances from 26 conversations using the following 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 canonical response. Where * defines whether the utterance depends on the first or the previous utterance, e.g., FT or PT.

An example of manually labeled conversation is reported in Table 1. Notice that in some cases, the human assessors use the labels FT-PR (i.e., 125_2) or PT-PR (e.g., 125_4, 125_5) to specify that the current utterance depends on the topics from the first/previous utterance and from the response, too.

3 Methodologies

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

Table 1. Example of manually labeled conversation

ID	Manual	Raw	Label
125_1	I'd like to learn more about frogs. What's the biggest one?	I'd like to learn more about frogs. What's the biggest one?	SE
125_2	What has been done to protect Goliath frogs?	What's been done to protect them?	FT-PR
125_3	Has the Equatorial Guinean government's conservation measures to protect Goliath frogs been effective?	Has that been effective?	PT-PR
125_4	How can I help with protecting Goliath frogs?	How can I help?	PT-PR
125_5	What is the smallest frog?	Okay, that's the biggest. What is the smallest?	FT
125_6	Why would leaf litter affect the size of a frog?	Why would leaf litter affect its size?	FT

All our methods employ a Python NLP toolkit for extracting various linguistic features from the utterances[6] and perform utterance rewriting to enrich 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 for passage re-ranking [3].

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 collections. The purpose of the utterance rewriting module is adding missing context to the raw utterance so that the user can get a good answer to her request.

Runs with Unsupervised Utterance Rewriting. These runs are inspired by our work on topic propagation in multi-turn conversational searches [1]. They use the raw utterances only.

- `CNR-run1.` The approach automatically rewrites the utterance by adding the topics extracted from the first utterance and the previous utterance. The idea behind this approach is that the first utterance has the general topic of the conversation, while the previous one represents the most recent context. This approach has a drawback since it always propagates the context of the very first-turn utterance. This can lead to noisy results, especially for those cases where the focus of interest may change during the conversation (e.g., topic shift, subtopic exploration).
- `CNR-run3.` This run tries to address the weakness of the previous run, avoiding the dependency with the first utterance. The run adds the topics extracted from the previous automatically rewritten utterances provided by CAsT 2021.

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

Runs with Utterance Rewriting based on Classification. These runs are inspired by our work on adaptive topic propagation in conversational utterances [2]. These runs perform the automatic rewriting of raw utterances using the utterance classification explained in Section 2.2. The labels represent the dependencies between the current utterance and the previous utterances as well as their canonical responses. 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 FT-PR, the utterance is rewritten using the topic extracted from the first 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.
- When the label is PT-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.

The runs perform utterance rewriting as follows:

– `CNR-run2.` This run uses raw utterances plus canonical responses (when needed). For each utterance, the approach checks the label and enriches the current utterance with the topics extracted from the utterance and the response of the previous turns for which there is a dependency.
– `CNR-run4.` As for run2, this run uses raw utterances plus canonical responses (when needed) by checking the utterance label. Differently from the previous run, the topics are extracted from previous automatically rewritten utterances provided by CAsT 2021 for which there is a dependency.

In all our runs, the topics are extracted from utterances using Spacy noun chunks (objects or subjects). In those cases where the utterance also depends on a previous response, the approach adds the named entities extracted from the candidate response by TagMe[7] with threshold set to 0.1. 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 `CNR-run1&3`, `CNR-run2&4` are both based on manual labels, they use context from canonical responses (when needed), and they extract topics from the utterances of the previous turns (enriched or not).

[7] `https://pypi.org/project/tagme/0.1.2/`

4 Experimental settings

Metrics. The effectiveness of the rewriting techniques is evaluated with traditional TREC metrics. In particular, the Average Precision for cutoff at 500 (`AP@500`) and the normalized Discounted Cumulative Gain (`nDCG`) for cutoffs at 3, 5, and 500. 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. In all our runs, we used Anserini BM25 with RM3 query expansion. In particular, for the first-stage retrieval, we used BM25 with parameters $b = 0.9$ and $k1 = 2.0$, chosen after a fine-tuning on MSMARCO-docs collection for the retrieval task with $5,192$ queries from the `DEV` set. The query expansion is done with 10 keywords taken from the top-10 results with the original query weight set to 0.5.

Neural re-ranking. We used the model by Nogueira and Cho [3] to re-rank the results from the previous stage. The model fine-tunes the BERT base pretrained model for re-ranking on the MSMARCO passage retrieval dataset. For each query, Anserini retrieves $1K$ results which are the input for the re-ranking step.

5 Experimental Results

In Table 2, we report the values of the following metrics `nDCG@k` (with $k =$ 3, 5, and 500) and `AP@500` for our four runs. As we can see, the worst results are achieved by `CNR-run1` as it does not use any utterance classification and any context from the canonical responses of the previous utterances. On the other hand, `CNR-run3` performs pretty well as it uses context from the previous automatic rewritten utterance provided by CAsT 2021.

Better performances are achieved by `CNR-run2` and `CNR-run4` as they enrich the raw utterances leveraging the utterance classification and adding the context extracted from the previous canonical responses when needed. Still, they cannot outperform `CNR-run3`.

CAsT 2021 also provided for each query/utterance the worst, median, and best performance for 10 raw runs, 27 canonical runs, and 13 manual runs. We computed the average over all the queries, and the results are shown in Table 3.

As expected, the performances of the two unsupervised runs (`CNR-run1` and `CNR-run3`) using raw utterances are close to the raw median values reported in Table 3. While the performances of `CNR-run2` and `CNR-run4` are close to the canonical median values.

6 Conclusions and Future Work

In this report, we have presented the methodologies implemented for our participation in CAsT 2021. Our approaches aim to enrich the raw utterances using topical keywords extracted from the previous utterances and their responses.

Table 2. Performance of our runs at CAsT 2021

Run	nDCG@3	nDCG@5	nDCG@500	AP@500
Run1 (raw)	0.2983	0.2897	0.1956	0.1122
Run2 (canonical)	0.3035	0.2936	0.2018	0.1218
Run3 (raw)	0.3490	0.3395	0.2218	0.1256
Run4 (canonical)	0.3327	0.3281	0.2201	0.1279

Table 3. Performance of CAsT 2021 runs: averaged over all queries

Run		nDCG@3	nDCG@5	nDCG@500	AP@500
Raw	worst	0.057	0.066	0.078	0.030
	median	0.338	0.336	0.334	0.176
	best	0.675	0.635	0.656	0.433
Canonical	worst	0.017	0.024	0.063	0.017
	median	0.380	0.384	0.454	0.244
	best	0.809	0.770	0.766	0.545
Manual	worst	0.088	0.111	0.118	0.041
	median	0.555	0.550	0.612	0.371
	best	0.780	0.761	0.765	0.535

As future work, we plan to improve the utterance classification in order to better capture the dependencies between the current utterance and the utterances of the previous turns as well as their canonical responses. Also, we plan to use neural expansion methods with the purpose of improving our automatic rewriting techniques.

References

1. 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.
2. 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.
3. R. Nogueira and K. Cho. Passage Re-ranking with BERT. *arXiv preprint arXiv:1901.04085*, 2019.