ConvAI3: Generating Clarifying Questions for Open-Domain Dialogue Systems (ClariQ)

Mohammad Aliannejadi[1], Julia Kiseleva[2], Aleksandr Chuklin[3], Jeff Dalton[4], and Mikhail Burtsev[5]

[1]University of Amsterdam, Amsterdam, The Netherlands,
m.aliannejadi@uva.nl
[2]Microsoft Research AI, Seattle, USA,
julia.kiseleva@microsoft.com
[3]Google Research, Zürich, Switzerland, chuklin@google.com
[4]University of Glasgow, Glasgow, UK,
jeff.dalton@glasgow.ac.uk
[5]MIPT, Moscow, Russia, burtcev.ms@mipt.ru

July 8, 2020

1 Introduction

This document presents a detailed description of the challenge on clarifying questions for dialogue systems (ClariQ) *[pronounce as Claire-ee-que]*. The challenge is organized as part of the Conversational AI challenge series (ConvAI3)[1] at Search-oriented Conversational AI (SCAI) EMNLP workshop in 2020.[2]. The main aim of the conversational systems is to return an appropriate answer in response to the user requests. However, some user requests might be ambiguous. In IR settings such a situation is handled mainly thought the diversification of search result page [4]. It is however much more challenging in dialogue settings. Hence, we aim to study the following situation for dialogue settings:

- a user is asking an ambiguous question (where ambiguous question is a question to which one can return > 1 possible answers);
- the system must identify that the question is ambiguous, and, instead of trying to answer it directly, ask a good clarifying question.

The main research questions we aim to answer as part of the challenge are the following:

[1]http://convai.io/
[2]https://scai.info

- RQ1: When to ask clarifying questions during dialogues?
- RQ2: How to generate the clarifying questions?

The rest of the document is organized as follows: Section 2 describes previous efforts and datasets related to the clarifying questions. We outline the design of our ClariQ challenge in Section 3.

2 Related work

There were a number of attempts to study clarifying questions recently.
[2] and [5] studied how users of Stack Exchange. [2] focused on characteristics, forms, and general patterns of clarifying questions. [5] designed a model to rank a candidate set of clarification questions by their usefulness to the given post at Stack Exchange. As result, the dataset extracted from Stack Exchange was released, but it covers specific narrow topics. Therefore, we are limited to understand clarifying questions for open-domain conversations.

Information retrieval (IR) community recently has payed close attention to the problem of generating clarifying questions in open-domain settings
[6], [1], and [8].
Where the general settings are the following: (1) a user is issuing a keyword query, which is ambiguous, and (2) a search engine's goal is to suggest a conversational clarifying question to help to find the required information. Aliannejadi et al. [1] shared a *Qulac* dataset with the community which allows follow-up studies. As a follow-up Zamani et al. [8] released a *MIMICS* dataset [9], which consists of queries, issued by real users, and behavioral signals such as clicks.

ClariQ challenges is closely related to QA domain [3]. For example, [7] made an attempt to understand an unclear question. Hence, we can assume that some solution for ClariQ might benefit from utilizing some techniques from Q&A and adapted to conversational settings.

Therefore, we can conclude that understanding and generating clarification questions has also been recognized as a major component in conversational information-seeking systems. There were a number of efforts of sharing datasets with the community to facilitate further research in this direction. Hence, our efforts to set-up challenge ClariQ on generating clarifying questions in dialogue settings are timely.

3 Challenge Design

The ClariQ challenge is run in two stages. At Stage 1 (described in Section 3.1) participants are provided static datasets consisting mainly of an initial user request, clarifying question and user answer, which is suitable for initial training, validating and testing. At Stage 2 (described in Section 3.2), we bring a human in the loop. Namely, the TOP-N systems, resulted from Stage 1, are exposed to the real users.

Table 1: Statistics of SCAI Challenge Data - Train.

# topics	237
# faceted topics	141
# ambiguous topics	57
# single topics	39
# facets	891
# informational facets	577
# navigational facets	185
# questions	3,304
# question-answer pairs	11,489
Average terms per question	9.70 ± 2.64
Average terms per answer	7.68 ± 4.75

3.1 Stage 1: initial dataset

Taking inspiration from the Qulac dataset [1] [3], we have crowdsourced a new dataset to study clarifying questions that is suitable for conversational settings. Namely, the collected dataset consists of:

- **User Request:** an initial user request in the conversational form, e.g. *What is Fickle Creek Farm*, with a label reflects if clarification is needed ranged from 1 to 4;
- **Clarification questions:** a set of possible clarifying questions, e.g. *do you want to know the location of fickle creek farm*;
- **User Answers:** each questions is supplied with a user answer, e.g. *no i want to find out where can i purchase fickle creek farm products*.

For training, the collected dataset is split into training (70%) and validation (30%) sets. For testing, the participants are supplied with: (1) a set of user requests in conversational form and (2) a set a set of questions (i.e., question bank) which contains all the questions that we have collected for the collection. Therefore to answer our research questions in Section 1 we suggest the following two tasks:

- To answer **RQ1**: Given a user request, return a score from 1 to 4 indicating the necessity of asking clarifying questions.
- To answer **RQ2**: Given a user request which needs clarification, return the most suitable clarifying question.

Table 3.1 provides statistics for the train set.

As system automatic evaluation metrics we use MRR, P@[1,3,5,10,20], nDCG@[1,3,5,20]. These metrics are computed as follows: a selected clarifying question, together

[3] Qulac is based on the TREC Web Track 2009-2012

with its corresponding answer are added to the original request. The updated query is then used to retrieve (or re-rank) documents from the collection. The quality of a question is then evaluated by taking into account how much the question and its answer affect the performance of document retrieval. Models are also evaluated in how well they are able to rank relevant questions higher than other questions in the question bank. For this task, that we call 'question relevance', the models are evaluated in terms of Recall@[10,20,30]. Since the precision of models is evaluated in the document relevance task, here we focus only on recall.

The datasets and the scripts for automatic evaluation can be found at the following repository – `https://github.com/aliannejadi/ClariQ`.

3.2 Stage 2: human-in-the-loop

At Stage 2 the participating systems are put in front of human users. The systems are rated on their overall performance. At each dialog step, a system should give either a factual answer to the user's request or ask for a clarification question. Therefore, the participants would need to:

- ensure their system can answer simple user questions
- make their own decisions on when clarification might be appropriate
- provide clarification question whenever appropriate
- interpret user's answer to the clarifying question

The participants would need to strike a balance between asking too many questions and providing irrelevant answers.

Note that the setup of this stage is quite different from the Stage 1. Participating systems would likely need to operate as a *generative* model, rather than a *retrieval* model. One option would be to cast the problem as generative from the beginning, and solve the retrieval part of Stage 1, e.g., by ranking the offered candidates by their likelihood.

Alternatively, one may solve Stage 2 by retrieving a list of candidate answers (e.g., by invoking Wikipedia API or the Chat Noir[4] API that we describe above) and ranking them as in Stage 1.

3.3 User-based evaluation

We use the real humans to interact with the systems. As a result, we get the following tuples `(Conversation history up until now, System's response, Ratings for the response)` as well as overall rating for the interaction. The users will rate each answer of the system on relevance and naturalness.

[4] `https://www.chatnoir.eu`

- **Relevance.** Is the particular answer or clarifying question relevant to the user's information need (e.g., does it help ranking?). This can be used on the utterance level, independent of the dialog.
- **Naturalness.** Is the clarifying question natural *in the context of the dialog*?

The separation between the two is to a large extent motivated by the setup, where we assume the user has an information need, but we don't want the system to approach it as a "yes/no" puzzle. An example of relevant but unnatural conversation:

```
User> Zurich zoo
Syst> Do you want to know the opening hours?
User> No.
Syst> Do you want to know how many elephants does it have?
User> You are reading my mind!

%% The user does want to learn about the number of elephants there,
%% but doesn't say so explicitly.
```

Acknowledgements

The challenge is organized as a joint effort by the University of Amsterdam, Microsoft, Google, University of Glasgow, and MIPT. We would like to thank Microsoft for their generous support of data annotation costs. We would also like to thank the Webis Group for giving us access to ChatNoir search API. Thanks to the crowd workers for their invaluable help in annotating ClariQ. We would also like to stress that all content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

References

[1] Aliannejadi, M., Zamani, H., Crestani, F., and Croft, W. B. (2019). Asking clarifying questions in open-domain information-seeking conversations. In *International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, SIGIR '19.

[2] Braslavski, P., Savenkov, D., Agichtein, E., and Dubatovka, A. (2017). What do you mean exactly? analyzing clarification questions in cqa. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, pages 345–348.

[3] Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., et al. (2019). Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

[4] Radlinski, F. and Dumais, S. (2006). Improving personalized web search using result diversification. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 691–692.

[5] Rao, S. and Daumé III, H. (2018). Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information. *arXiv preprint arXiv:1805.04655*.

[6] Rosset, C., Xiong, C., Song, X., Campos, D., Craswell, N., Tiwary, S., and Bennett, P. (2020). Leading conversational search by suggesting useful questions. In *The Web Conference 2020 (formerly WWW conference)*.

[7] Trienes, J. and Balog, K. (2019). Identifying unclear questions in community question answering websites. In *European Conference on Information Retrieval*, pages 276–289. Springer.

[8] Zamani, H., Dumais, S., Craswell, N., Bennett, P., and Lueck, G. (2020a). Generating clarifying questions for information retrieval. In *Proceedings of The Web Conference 2020*, pages 418–428.

[9] Zamani, H., Lueck, G., Chen, E., Quispe, R., Luu, F., and Craswell, N. (2020b). Mimics: A large-scale data collection for search clarification. *arXiv preprint arXiv:2006.10174*.