What Do You Mean Exactly? Analyzing Clarification Questions in CQA

Pavel Braslavski*
Ural Federal University
pbras@yandex.ru

Denis Savenkov
Emory University
dsavenk@emory.edu

Eugene Agichtein
Emory University
eugene@mathcs.emory.edu

Alina Dubatovka
Saint Petersburg University
alina.dubatovka@gmail.com

ABSTRACT

Search as a dialogue is an emerging paradigm that is fueled by the proliferation of mobile devices and technological advances, e.g., in speech recognition and natural language processing. Such an interface allows search systems to engage in a dialogue with users aimed at fulfilling their information needs. One key capability required to make such search dialogues effective is asking *clarification questions* (CLARQ) proactively, when a user's intent is not clear, which could help the system provide more useful responses. With this in mind, we explore the dialogues between the users on a community question answering (CQA) website as a rich repository of information-seeking interactions. In particular, we study the clarification questions asked by CQA users in two different domains, analyze their behavior, and the types of clarification questions asked. Our results suggest that the types of CLARQ are very diverse, while the questions themselves tend to be specific and require both domain- and general knowledge. However, focusing on popular CLARQ types and domains can be fruitful. As a first step towards automatic generation of clarification questions, we explore the problem of predicting the specific subject of a clarification question. Our findings can be useful for future improvements of intelligent dialog search and question answering systems.

1. INTRODUCTION

Proliferation of mobile devices and more "natural" interfaces [4] are changing the way people search for information on the web. Many experts envision that search in the near future will be a dialog between a user and an intelligent assistant, rather than just "ten blue links" in response to a one-shot keyword query.[1] Participants of the SWIRL'2012 workshop foresaw a fusion of traditional IR and QA [1]: "Dialogue would be initiated by the searcher and proactively by the system. The dialogue would be about questions and answers, with the aim of refining the understanding of questions and improving the quality of answers." Today we can witness this trend embodied in such products as Apple Siri, Microsoft Cortana, Google's Allo, and others.

To build a system that can be engaged in a dialog with the user and ask probing questions proactively, the most natural option is to learn from the human interaction data. Community question answering (CQA) sites, such as Yahoo!Answers[2], Quora[3], and Stack Exchange[4], allow users to post questions on various topics to other community members, vote for questions and answers, as well as gain scores for their activities. CQA platforms received a considerable deal of popularity and collected a vast amount of user-generated content.

In this paper, we make the first attempt to examine the clarification questions (CLARQ) that users ask on the Stack Exchange community question answering (CQA) platform. We analyze Stack Exchange data in two domains corresponding to about 300K questions and comments. The contributions of this study are threefold:

- To learn about user behavior associated with CLARQ and about their role in CQA communications. We find that CLARQ are quite common on Stack Exchange, and therefore represent a good source of data for analysis.
- To study the types of CLARQ users asks in different situations. We classify clarification questions into several categories according to their target as well as syntactic patterns, which help define the space of CLARQ for future research;
- To make the first step towards automatic generation of CLARQ: we build a model to predict the subject of a particular popular type of clarification questions, which shows the potential of such approach for future research.

2. RELATED WORK

Several studies paved the way towards conversational answer retrieval and explored various types of clarifications in response to user questions. Kotov and Zhai [6] introduced

*Work was performed while the author was a Fulbright visiting scholar at Emory IRLab.

[1] http://time.com/google-now/

[2] http://answers.yahoo.com

[3] https://www.quora.com/

[4] http://stackexchange.com/

CHIIR '17, March 07 - 11, 2017, Oslo, Norway

ISBN 978-1-4503-4677-1/17/03…$15.00

DOI: http://dx.doi.org/10.1145/3020165.3022149

a concept of *question-guided search*, which can be seen as a variant of query suggestion scenario: in response to initial query the user is presented with a list of natural language questions that reflect possible aspects of the information need behind the query. Tang et al. [12] proposed a method for refinement question generation based on similar questions retrieved from a question archive, a thesaurus and question templates. Sajjad et al. [10] described a framework for search over a collection of items with textual descriptions exemplified with xbox avatar assets. Initially, attribute–value pairs were extracted from crowdsourced descriptions. In online phase intermediate search results are analyzed and yes/no questions about attributes and values are generated sequentially in order to bisect the result set and finally come to the sought item. Gangadharaiah and Narayanaswamy [3] elaborated a similar approach to search results refinement through clarification questions. The authors considered customer support scenario using forum data. Noun phrases, attribute–value pairs, and action tuples are extracted from forum collection in offline phase. In online phase answers to automatically generated questions help reduce the answer candidates set. These works demonstrate successful cases of CLARQ for search tasks. Our study looks into uses of CLARQ in answer seeking dialogs between real users, which could be useful for future improvements of the systems.

Kato et al. [5] investigated clarification questions in context of an enterprise Q&A instant messaging in software domain. Analysis has shown that about one third of all dialogues have clarification requests; 8.2% of all utterances in the log are related to clarifications. The authors developed a question classifier that prompted the asker to provide clarifications in case the request was identified as underspecified. This work is closest to our study: we also analyze human Q&A behavior. However, we deal with a different type of data – CQA archives – and investigate if the resource can be potentially useful for a different application – QA.

The ability to ask clarification questions is one of the key desired components of conversational search systems [9], and can be used for multiple tasks, e.g., to resolve anaphora and coreferences [8]. In spoken dialog systems, clarification questions can be used to resolve speech recognition uncertainty, either of individual words, or of whole utterances [11].

A large body of work deals with community question answering data, data from Stack Exchange sites in particular. For example, Anderson et al. [2] showed that long-term value of a question and its answers on Stack Overflow was positively correlated with the number of comments on the answers and the time for highest-score answer to arrive. Our work complements these studies and focus on clarification comments to gather insights, useful for automatic conversational search system development.

3. DATA

We took two Stack Exchange sites – Home improvements (DIY)[5] and Arqade (GAMES)[6]. These two domains are quite different – the former is devoted to purely practical real-world problems, the latter – to the virtual world of video games. Stack Exchange users can comment on the questions and answers; sometimes it leads to multi-turn forum-like discussions (see Fig. 1). The data dumps provided by Stack Exchange[7] cover a period of 5.5 years – from July 2010 to January 2016.

We define CLARQ in a straightforward manner: sentences in comments to the initial questions ending with question mark, provided by the users different from the asker of the initial question, four words and longer. This heuristic is not perfect, as clarification requests can be formulated as declarative sentence, e.g., *Please provide details...* or question mark can be just missed. At the same time, these interrogative comments may be rhetorical questions, or not on the initial question's subject. Nevertheless, manual inspection showed that this definition of CLARQ is operational and allows extraction of CLARQ with precision acceptable for an exploratory study.

Basic statistics of the two datasets are reported in Table 1.

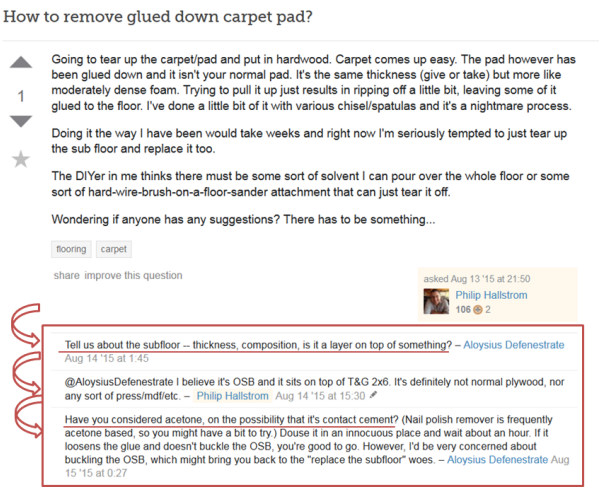

Figure 1: Screenshot of a DIY question page with threaded conversation in comments

	DIY	*GAMES*
# of questions	20,702	62,511
# of answers	36,580	105,167
# of accepted answers	8,381	40,049
# of comments	87,238	228,074
average question length in words	130.8	86.5
average comment length in words	33.8	25.8
# of comments on questions	37,296	96,247
# of non-asker comments on questions	27,873	72,495
# of comments on questions with '?'	11,040	21,448
CLARQ followed by asker's comments	3,679	8,021
CLARQ followed by post editing	4,270	9,038
CLARQ followed by post editing by asker	1,631	3,772

Table 1: Stack Exchange datasets statistics

4. RESULTS

4.1 User behavior

As Table 1 shows, the presence of CLARQ in CQA is substantial. Many characteristics such as questions/answers, accepted answers/all answers ratios are similar for both domains. Questions and comments in DIY are longer then in GAMES, which is expected: DIY implies richer and more diverse contexts. Askers are engaged in communication even

[5]http://diy.stackexchange.com/
[6]http://gaming.stackexchange.com/
[7]https://archive.org/details/stackexchange

after the initial question is posted: they comment on questions and edit them (however, questions are edited by community members more often). Although there are more comments on questions in DIY, GAMES seem to be somewhat more conversational: askers respond to questions on questions more often. Interestingly, thousands of initial questions are followed by clarifications, and in many cases these are followed by the original question being edited, presumably in response to the clarification request.

Unfortunately, we see that questions followed by CLARQ do not differ much from questions without any comments in length – a simple assumption that CLARQ are targeted at short underspecified questions does not hold. The hypothesis that questions asked by novice, less experienced community members (based on users' ratings) receive more CLARQ is not supported either. We also did not find any topical specificity of questions with CLARQ based on tags.

Figure 2 shows rating distribution of the GAMES users, who ask CLARQ and those who provide accepted answers (i.e., the best answerers in the community). We can observe that distribution for the former group is shifted towards higher scores (DIY exhibits a very similar distribution). However, the users who ask for clarifications provide answers for the initial questions very rarely. This observation suggests that CLARQ in CQA are a form of 'quality control measures' undertaken by most experienced users.

4.2 Question types and patterns

In order to investigate CLARQ breakdown by type, we sampled 294 comments on questions from both domains, and two authors performed manual annotation analogously to [5], see results in Table 2. Comments that are not aimed at clarifying the main question contain rhetorical or humorous questions, questions to previous comments, citations of other questions on the platform, etc. Interestingly, the breakdown of CLARQ by type is roughly the same as in [5].

Further, upon examining the sample we identified a list of common three-word question starting patterns, and calculated their frequency in the whole dataset, see Table 3. As can be seen from the Table 3, some patterns are quite indicative for the clarification type (e.g., *what kind of* corresponds to *More info* category, whereas *have you tried* – to *Experience*). This observation suggests that recognition of clarification question type is a feasible task.

Category	*%*	*Example*
More Info	28.6	What OS are you using?
Check	29.3	Are you on a 64-bit system?
Reason	8.5	What is the reason you want a drip pan?
General	10.2	Can you add more details to this question?
Selection	9.9	Are you using latex or oil based Kilz?
Experience	10.2	Have you tried to update video card drivers?
Not a CLARQ	13.9	Out of curiosity what elo are you?

Table 2: Questions in comments by type. Some comments contain several clarQ of different types, so the sum is more the 100%.

4.3 Clarification subject prediction

As we have shown, there are many different kinds of questions that users ask in comments. Many of them address a certain ambiguity present in questions, e.g., *what kind of* questions inquire about a subtype of a mentioned object. These questions are quite common (Table 3) and have a simple structure, which make them a quite appealing target for automatic question generation. The first step for such question generation is to predict the object to ask about. We collected questions, which received at least one *what (kind|type) of* CLARQ in DIY. From these comments and questions we extracted noun phrases using Stanford CoreNLP parser [7], and kept only those questions that have a common pair of noun phrases in the question and comment. We formulated the task as the noun phrase ranking problem, where the noun phrase from the comment should be placed higher in the list than other noun phrases from the question. Each candidate phrase was represented with the following set of features:

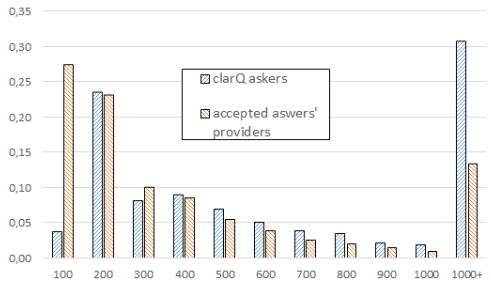

Figure 2: Distribution of users' reputation scores in the groups of accepted answers' providers and commentators on questions (GAMES)

- **prior**: number of times the noun phrase was used in comments (separate from the training and test sets)
- **topicality**: number of occurrences of the phrase in the current question (in title and body together)
- **position**: position of the first occurrence of the noun phrase relative to the beginning of the question
- **entropy**: collection-based statistic, computed using all noun phrases that contain the given noun phrase, which estimates the number of different modifications of the current noun phrase object
- **length**: number of words in the noun phrase

To train the ranking model we used a random forest algorithm implemented in the RankLib library[8]. We optimized DCG@10 and Table 4 summarizes the performance metrics on 10-fold cross validation. As we can see, even with a limited set of features our model was able to place the true subject of a clarification question above other candidates in 35% of the cases. To study the contributions of different feature groups we conducted a series of experiments to train the model with each group of features individually. The results in Table 4 suggest, that the number of occurrences of a phrase and the position of the first occurrence are strong features, and confirms our intuition that CLARQ are usually asked about the main topic of the question. However, some noun phrases are more ambiguous in general, therefore the prior feature also contributed significantly to the quality of the model.

Overall, our experiment showed promising results for predicting the subject for certain type of CLARQ. As a next

[8]https://sourceforge.net/p/lemur/wiki/RankLib

Pattern	DIY	GAMES	Example
have you (tried|considered)	256	1,123	Have you tried reinstalling the game yet?
do you have	592	692	Do you have enough disk space left?
do you mean	248	552	Do you mean a separate tub and shower?
are you sure	206	366	Are you sure you have timber frame construction?
what (is|are) the	558	361	What is the slope of the floor?
what do you	103	284	What do you mean by squeaking?
(are|is) there any	154	147	Are there any airflow ducts in the room already?
can you (post|provide)	204	125	can you post some pictures?
how X (is|are)	290	117	how old is the water heater?
is it a	186	112	is it a constant 18-22 fps?
what (kind|type) of	344	106	What kind of texture is on the wallpaper?
why do you	73	101	Why do you need to run it from the Flash drive?
have you checked	66	98	Have you checked the frequency of the outlets?
is it possible	78	84	Is it possible the tank is just over filling?
do you know	120	64	Do you know the manufacturer of the fixture?

Table 3: Question patterns in comments (sorted by frequency in GAMES)

Model	*P@1*	*MAP*	*RR@10*	*ERR@10*
random	0.077	0.215	0.231	0.015
+ entropy	0.143	0.334	0.350	0.024
+ length	0.148	0.337	0.345	0.024
+ position	0.165	0.335	0.357	0.024
+ prior	0.214	0.402	0.427	0.030
+ topicality	0.319	0.426	0.473	0.032
all features	0.350	0.508	0.549	0.038

Table 4: Performance metrics (P@1 – precision at 1, MAP – mean average precision, RR@10 – reciprocal rank at 10, ERR@10 – expected reciprocal rank) of the ranking model for "ambiguous" noun phrase selection problem

step, our model can be combined with an ambiguity detection classifier, which would trigger clarification as a response from a conversational search agent.

5. DISCUSSION AND FUTURE WORK

As a step towards general-purpose interactive QA system, we analyzed clarification questions asked by CQA users. In particular, we examined user interactions related to CLARQ, as well as the role and place of these questions in CQA. We analyzed a large sample of CLARQ according to their type, and identified most common syntactic patterns in a large collection of CLARQ. Finally, we conducted an experiment aimed at automatically detecting the subject of clarification question of a particular type.

Based on our analyses, we can conclude that CLARQ are common in CQA archives, and introduce a valuable resource for user behavior studies and QA research. Clarification questions asked by community members are an important component in maintaining the quality of user-generated content in CQA. Furthermore, we see that clarification questions are quite diverse in topic and style, are are highly dependent on context and individual characteristics of the users. However, there are several types of questions and syntactic patterns that are common within each domain. As a first step towards automatically generating clarification questions, we show promising results on identifying the subject of CLARQ based on a small set of shallow features. Our findings suggest that CQA data may be useful for research in the field of interactive QA.

There is still a long way to go towards automatic generation of clarification questions. This capability would require identification of user questions which are indeed ambiguous, choosing an aspect and type of clarification, and generating the text. These steps imply naturally an analysis of answer candidates, which was not addressed in the current work. A significant portion of CLARQ deals with properties, attributes, relations and types of the objects mentioned in the initial question. This suggests that domain-specific knowledge-based approach to CLARQ generation can be promising.

These two tasks – the use of candidate answers and knowledge bases – define the directions for future research in the area of clarification questions generation for interactive QA.

6. REFERENCES

[1] J. Allan, B. Croft, A. Moffat, and M. Sanderson. Frontiers, challenges, and opportunities for information retrieval. *SIGIR Forum*, 46(1):2–32, 2012.

[2] A. Anderson et al. Discovering value from community activity on focused question answering sites: a case study of stack overflow. In *KDD'2012*.

[3] R. Gangadharaiah and B. Narayanaswamy. Natural language query refinement for problem resolution from crowdsourced semi-structured data. In *IJCNLP'2013*.

[4] M. A. Hearst. 'Natural' search user interfaces. *CACM*, 54(11), 2011.

[5] M. P. Kato, R. W. White, J. Teevan, and S. T. Dumais. Clarifications and question specificity in synchronous social Q&A. In *CHI'2013 Extended Abstracts*.

[6] A. Kotov and C. Zhai. Towards natural question guided search. In *WWW'2010*.

[7] C. D. Manning et al. The Stanford CoreNLP natural language processing toolkit. In *ACL System Demonstrations'2014*.

[8] S. Quarteroni and S. Manandhar. Designing an interactive open-domain question answering system. *Natural Language Engineering*, 15(01):73–95, 2009.

[9] F. Radlinski and N. Craswell. A theoretical framework for conversational search. In *CHIIR'2017*.

[10] H. Sajjad, P. Pantel, and M. Gamon. Underspecified query refinement via natural language question generation. In *COLING'2012*.

[11] S. Stoyanchev, A. Liu, and J. Hirschberg. Modelling human clarification strategies. In *SIGDIAL'2013*.

[12] Y. Tang, F. Bu, Z. Zheng, and X. Zhu. Towards interactive qa: suggesting refinement for questions. In *SIGIR'2011 Workshop on "entertain me": Supporting Complex Search Tasks*.