ORIGINAL ARTICLE

Analysing Clarification in Asynchronous Information-Seeking Conversations

Leila Tavakoli[1] | Hamed Zamani[2] | Falk Scholer[3] | W. Bruce Croft[4] | Mark Sanderson[5]

[1]PhD candidate, STEM college, School of Computing Technologies, RMIT University, Melbourne, Australia, leila.tavakoli@rmit.edu.au

[2]Assistant Professor, College of Information and Computer Sciences, University of Massachusetts, Amherst, MA, United States, zamani@cs.umass.edu

[3]Professor, STEM college, School of Computing Technologies, RMIT University, GPO Box 2476, Melbourne, 3001, Victoria, Australia, falk.scholer@rmit.edu.au

[4]Professor, College of Information and Computer Sciences, University of Massachusetts, Amherst, MA, United States, croft@cs.umass.edu

[5]Professor, STEM college, School of Computing Technologies, RMIT University, GPO Box 2476, Melbourne, 3001, Victoria, Australia, mark.sanderson@rmit.edu.au

Correspondence
Leila Tavakoli, RMIT University, Melbourne, Australia
Email: leila.tavakoli@rmit.edu.au

Funding information

This research analyses human-generated clarification questions to provide insights into how they are used to disambiguate and provide a better understanding of information needs. A set of clarification questions is extracted from posts on the *Stack Exchange* platform. Novel taxonomy is defined for the annotation of the questions and their responses. We investigate the clarification questions in terms of whether they add any information to the post (the initial question posted by the asker) and the accepted answer, which is the answer chosen by the asker. After identifying which clarification questions are more useful, we investigate the characteristics of these questions in terms of their types and patterns. Non-useful clarification questions are identified, and their patterns are compared with useful clarifications. Our analysis indicates that the most useful clarification questions have similar patterns, regardless of topic. This research contributes to an understanding of clarification in conversations and can provide insight for clarification dialogues in conversational search scenarios and for the possible system generation of clarification requests in information seeking conversations.

1 | INTRODUCTION

In order to answer complex natural language information needs, conversational search systems may need to ask clarification questions to better determine the intent of the user who submitted the question (Christakopoulou et al., 2016; Zhang et al., 2018; Zou and Kanoulas, 2019). Past work has examined generating clarification questions (Zamani et al., 2020a; Rao and Daumé III, 2018), offline evaluation of clarification (Aliannejadi et al., 2019; Zamani et al., 2020b), and interactions with clarification (Zamani et al., 2020c). What is less explored is an understanding of the characteristics of useful clarification.

Question Answering (QA) forums record asynchronous exchanges between people seeking to solve complex information needs, including those that a web search may have failed to answer (Liu et al., 2012; Shah et al., 2009). This is because human-driven online QA services allow users to seek different forms of information, ranging from factual information, to personal opinion or advice through human-to-human interactions (Choi and Shah, 2016).

Information seeking forums such as *Stack Exchange*, *Quora*, or *Yahoo! Answers* serve needs covering diverse topics. Although an interaction is asynchronous on a forum compared to the synchronous conversations of a search system, following Anand et al. (2020), we contend that it is reasonable to assume that users of such forums have similar expectations and behaviours as when engaging with a conversational search system. An initial investigation of information seeking forums shows that a high percentage of clarification questions are left unanswered, suggesting that not all clarifications are necessarily useful. This motivated us to investigate what the characteristics of a useful clarification question are. Therefore, we ask the following research questions:

- **RQ. 1**: What clarification questions are more useful (in terms of helping the Asker to get a correct answer)?
- **RQ. 2**: What are the characteristics of useful clarification questions?

To address the research questions, we focus on conversations in *Stack Exchange*, a popular QA forum. Figure 1, shows an initial question posted by an *Asker*, followed by a clarification question from a *Responder* (any user other than the Asker). The interaction led to an *accepted answer* chosen by the Asker. We use this terminology throughout the paper.

In this paper, we describe how we classify clarification questions with respect to the type of *answerer*: Asker, Responder, or when clarification questions are left unanswered. We then investigate the usefulness of clarifications based on manual annotation. The analysis shows that if Askers interact with a clarification question, they commonly provide an informative answer. Perhaps unsurprisingly, we find that the Asker tends to answer more clarification questions. We propose new definitions for useful and non-useful clarification questions based on the findings in this study and distinguish them accordingly. We show that useful clarification questions is a clarification question which is answered by the Asker, has an informative answer, and is valuable for the post and the accepted answer. While a non-useful clarification question is a clarification question when it is left unanswered and is not valuable for the post but, the post still receives an accepted answer. Comparing the patterns of useful and non-useful clarification questions shows us some specific patterns, which have higher chances to engage the Asker. We notice useful clarifications often target *Ambiguity/Incompleteness* or *Confirmation* in the post. Our key contributions are:

- Presenting new taxonomy to investigate the usefulness of clarification questions (Section 3).
- Examining the relationship between posts with accepted answers and different types of answerers (Section 4.1).
- Detecting useful and non-useful clarification questions (Section 4.2).
- Extracting and identifying the types and patterns of useful and non-useful clarification questions (Section 4.3).

FIGURE 1 A question posted on *Stack Exchange*.

2 | RELATED WORK

We describe related work from four perspectives: the study of human interactions with clarification systems, the design of clarification systems, the design of clarification datasets, and finally, the study of human to human interactions when tackling complex information-seeking tasks.

2.1 | Human Interaction With Systems

Insights about the importance of clarification questions in conversational search systems have led to models to generate clarification in conversational search to understand user intent better. Kiyota et al. (2002) proposed a dialogue-based QA system utilising a large text knowledge base. The system was designed to navigate users to the desired answer by asking clarification questions using dialogue cards and description extraction of each retrieved text. When a user asked a vague question, the system asked the user a dialogue card. This process continued until the question was clarified. Their study showed that clarification questions are required for any dialogue with a searching system. Kim and Oh (2009) investigated the criteria which questioners used to choose the best answers in *Yahoo! Answers*, a social QA site. They found that the best-answer selection criteria were very similar to traditional relevance criteria; however, socioemotional-, content-, and utility-related criteria were dominant, and there were patterns between topic categories and selection criteria. In another study on *Yahoo! Answers*, Shah (2011) found out that most of the questions received at least one answer within a few minutes, although it took longer to obtain a satisfactory answer. They also noticed that the sooner a question received an answer, the chances of that answer being selected as the best answer by the Asker increased.

In 2016, Bodoff and Raban (2016) attempted to understand what the answerer in online QA services must do to provide a high-quality answer. They believed that different question types might require taking into account different aspects of the user situation and showed that answerers extract more conceptual clarifications for closed questions than for open questions. Jeng et al. (2017) focused on *ResearchGate*, an academic social networking site. They examined how scholars exchanged information across three distinct disciplines of Library and Information Services, History of Art, and Astrophysics. They concluded that scholars took a long time to respond to their peers in information-typed questions compared to discussion ones. Moreover, users in the field of Astrophysics were more likely to provide factual information and write longer responses.

Human behaviour in conversational systems was investigated by Vtyurina et al. (2017). They studied participants conversing with three different conversational agents; an existing commercial intelligent assistant, a human expert and a human disguised as an automatic system (Wizard agent). The last two agents could ask clarification questions to help participants complete an allocated information-seeking task. The researchers noticed that participants were happy to interact with the agents as long as system accuracy was acceptable. Qu et al. (2018) analysed the distribution, co-occurrence, and flow pattern of user intent in information seeking conversations. They classified twelve classes of intent, including the original question, repeat the question, clarification question, further details, follow up question, information request, potential answer, positive feedback, negative feedback, greetings/gratitude, junk and others. Their research led to finding some frequently occurring intent patterns during information seeking. In another empirical study, Zou et al. (2020) investigated the domain of online *Amazon* retail to quantify whether and to what extent users were willing or able to answer clarification questions. The users interacted by the system with a “Yes”, a “No”, or a “Not Sure” in order to find a target product to buy. They performed an online experiment collecting both implicit interaction data and explicit feedback from users. They found that users answered 11-21 clarification questions on average. Most users answered clarifications until they reached their target product. However, 21% of users stopped answering either due to fatigue or being asked irrelevant questions. Krasakis et al. (2020) studied the effect of user feedback in mixed-initiative conversations and analysed the performance of a lexical ranking model on a conversational search dataset with clarification questions. The researchers showed that effectively understanding and incorporating explicit conversational feedback is important, and a more fine-grained treatment of the conversations is crucial.

Clarification in voice queries was investigated by Kiesel et al. (2018) who examined the impact of clarification on user satisfaction, finding that satisfaction depends on language proficiency levels. They also noticed that the effectiveness of query clarification depends on the number and lengths of the possible answers. In a follow-up, Kiesel et al. (2019) showed that clarifications usually improve user satisfaction and that satisfaction is significantly impacted by the tone of voiced clarifications.

2.2 | Design of Clarification Systems

Zhang et al. (2018) proposed the *System Ask - User Respond* paradigm for conversational search. They developed a multi-memory network architecture and trained their model on a large scale dataset in e-commerce. The system was capable of asking clarification questions from users directly to understand user needs. They performed their experiments on the Amazon e-commerce scenario based on real-world user purchase datasets and found out that their approach outperformed state-of-the-art product search and recommendation baselines. Rao and Daumé III (2019) proposed an initial model for generating clarification questions. They found that their model produced more useful and specific questions compared to previous models including models trained using maximum likelihood objective and trained using utility reward-based reinforcement learning. This research inspired other research groups such as Zamani et al. (2020a), Hashemi et al. (2020), Shwartz et al. (2020) and Dhole (2020) to focus on design of clarification systems. Zamani et al. (2020a) focused on the task of generating clarification for open-domain search. They proposed three models of a Template-based approach, Question Likelihood Maximisation and Query Clarification Maximisation for asking clarification questions, and they examined their models using human annotation. They showed that the Query Clarification Maximisation approach performed better than others. Hashemi et al. (2020) proposed a multi-source attention network and applied it to conversational search tasks for utilising user responses to clarification questions. They focused on both conversations with only one clarification question and multi-turn setting. Their evaluation showed that their models, which implemented guided transformer, substantially outperformed state-of-the-art baselines. Shwartz et al. (2020) proposed an unsupervised framework based on self-talk to generate natural language clarification questions and their corresponding answers. They generated multiple clarification questions considering a) concatenating one of several question prefixes, curated for each task and b) generating five questions for each prefix using Nucleus sampling. They noticed several shortcomings of using pre-trained learning models as knowledge providers, including (i) insufficient coverage, (ii) insufficient precision, and (iii) limited reasoning capabilities. Their empirical results demonstrated that the self-talk procedure they proposed substantially improved the performance of zero-shot language model baselines and outperformed models that obtaining knowledge from external knowledge bases Dhole (2020) presented a method to disambiguate queries that are ambiguous between two intents. They stated that the proposed method could take advantage of any question generating system with no need for annotated data of clarification questions.

In other studies on clarification systems, Wu et al. (2020) proposed the Predicting, Explaining, and Rectifying Failed Questions (PERQ) framework and Kumar et al. (2020) conducted ranking clarification questions using natural language inference. Wu et al. (2020) developed the PERQ framework to improve Knowledge-based Question Answering systems' accuracy over simple questions. They designed an interactive system that identified ambiguities in failed questions and requested minimal clarification actions from users. Kumar et al. (2020) ranked clarification questions using natural language inference and defined a clarification question as good when the answer to the clarification led to a resolution of the underspecification in the question posed in the original post from the Asker.

2.3 | Clarification Question Datasets

The study of clarification requires datasets on which test can be run. In the absence of standardised datasets, some researchers created their own. Rao and Daumé III (2018) drew data from *Stack Exchange* and a formed model for learning to rank clarification questions. They evaluated their model against expert human judgments and demonstrated significant improvements over baselines. Aliannejadi et al. (2019) constructed an open-domain clarification question dataset using crowdsourcing. Their experiments demonstrated that asking only one good clarification question improves retrieval performance. Penha et al. (2019) created a dataset that focused on the interaction between an agent and a user, including clarification questions. The researchers presented a conceptual model and provided baseline results for conversation response ranking and user intent prediction tasks. Zamani et al. (2020b) introduced a collection of search clarification data drawn from real web search queries. The collection included two datasets of user interactions with search clarification and one dataset based on manual annotations of clarification panes in the *Bing* search engine. Each clarification was generated by a *Bing* production algorithm and contained a clarification question and up to five candidate answers. Datasets included (1) *MIMICS-Click* included unique queries, their associated clarification panes, and the corresponding user interaction signals, (2) *MIMICS-ClickExplore* was an exploration data including aggregated user interaction signals for unique queries with multiple clarification panes and (3) *MIMICS-Manual* with unique real search queries with clarification pairs which were manually labelled by at least three trained annotators. Kumar and Black (2020) proposed a large-scale dataset for the task of clarification question generation. They proposed a bootstrapping framework to employ a neural network for classifying clarification questions. Min et al. (2020) introduced an open-domain QA task to find every possible answer and then to rewrite the question and to clarify the ambiguity which led to each answer. They constructed a dataset from an open-domain QA benchmark. The dataset contained diverse types of ambiguity, which are not normally visible from the prompt question alone.

2.4 | Human to Human Interaction and Clarification

In studies conducted to understand the characteristics of clarification questions, Kato et al. (2013) and Braslavski et al. (2017) investigated human-generated clarification questions in Social Q&A and *Stack Exchange* sites, respectively. Kato et al. (2013) investigated the relationship between clarification questions and dialogue outcomes with respect to the specificity of the posted question and the requested clarification in a social QA system. They classified the clarification questions into seven types of *Check*, *More Info*, *General*, *Selection*, *Confirmation*, *Experience* and *Other* and found out that the most common clarification question type is *Check*. They showed that about one-third of all dialogues had clarification requests. They also developed a question classifier to provide clarifications in the case of underspecified requests. Braslavski et al. (2017) analysed user behaviour and types of clarification questions in *Stack Exchange* to explore the problem of predicting the specific subject of a clarification question. They divided the clarification questions into seven types of *More Info*, *Check*, *Reason*, *General*, *Selection*, *Experience* and *Not a ClarQ*. They also investigated three-word question starting patterns of common clarification questions. They found that clarification questions vary in topic and format and mainly depend on the content and individual characteristics of users.

2.5 | Summary of Related Work

Literature review shows that previous works studied human interaction with different QA systems and the selection and generation of clarifications in the context of web search and conversational search. Studies on generating clarification questions in search engines and also in conversational search systems are in the early stages with limited success. While generating and asking clarification question is important, getting an answer for the clarification question is also important to help the system to respond better. Therefore, any system which generates, ranks and asks clarification questions requires understanding what clarification question is more engaging (useful clarification). Even this needs to be considered in developing any clarification question dataset. However, the characteristic of such clarifications is challenging and relatively less explored. User engagements with clarification can be used as a signal for identifying useful clarifications. Among all discussed studies, the research conducted by Braslavski et al. (2017) which focused on human-generated clarification questions, was the closest research to our work. In contrast, we first classify clarification questions based on the type of the answerer to focus on those which engage the Asker more. We then define new taxonomy to investigate the usefulness of clarification questions from different aspects. The findings of this research help us to determine useful and non-useful clarification questions. This is a more detailed analysis compared to Braslavski et al. (2017). They classified clarification questions into several types and presented common patterns in general, while we define different types regardless of their answers, identify popular patterns of the useful clarification questions, and compare useful with non-useful clarification questions.

TABLE 1 The analysed sites of *Stack Exchange*.

Site	Category	# Posts
Quantitative Finance, (QF)	Business	13,187
English Language and Usage, (EL)	Culture/Recreation	107,266
Science Fiction and Fantasy, (SF)	Life/Arts	55,959

3 | METHODOLOGY

To investigate our research questions and better understand what clarification questions are useful in terms of helping the Asker of a post obtain an accepted answer, we investigate publicly available data from *Stack Exchange*[1] covering a period from July 2009 to September 2019. We investigate three sites that had the highest number of posts in each of three different *Stack Exchange* categories: Business (which holds three sites), Culture/Recreation (46 sites) and Life/Arts (26 sites). Table 1 reports the details of the three chosen sites: Quantitative Finance (*QF*), English Language and Usage (*EL*), and Science Fiction and Fantasy (*SF*). As stated by Shah et al. (2009) and Liu et al. (2011), such popularity may indicate that users cannot satisfy their information needs using web search engines.

The data is first processed by identifying posts, potential clarification questions, and their answers within the data (Sec. 3.1). Second, questions are classified with respect to the type of answerer (Sec. 3.2). Third, the relationship between the type of answerer and the post is investigated (Sec. 4.1).

3.1 | Identifying Potential Clarifications

To identify *potential clarification questions*, we collect comments from within posts that contain at least one sentence ending with a question mark, regardless of the question content. Sentences containing question marks that appear in the form of a quotation are ignored (e.g. *Note swiss German might say: "Wie isch ire Name?"*). A question is also disregarded if it is part of a hyperlink (e.g. *"I've voted to close, per our top answer here [Are stories that only appear to contain fantastical elements on topic?]"*). We exclude any questions submitted by the Asker, assuming that the person who submitted a post would not ask for clarifications. In addition, if the question starts with "@username", it should be the Asker's username. This is to ensure that the question is directed to the Asker and is not part of a conversation between other users.

In order to identify the *answer* to a clarification question, the following criteria have to be met:

- The comment starts with "@username", which is the name of the user who asked the clarification question.
- The comment is submitted after the clarification question, based on timestamps.
- The user who asks the clarification question did not comment between the clarification question and the provided answer to that clarification question (this maximises the likelihood of the comment being a response to the clarification question).

[1] `https://archive.org/details/stackexchange`

3.2 | Annotation and Data Sampling

To investigate the characteristics of clarification questions based on the type of answerer – the Asker, a Responder, or unanswered questions – we conduct a manual annotation to answer the following questions:

- Is a *potential* clarification question as defined above an *actual* clarification question?
- Does the clarification question have an informative answer?
- Does answering the clarification question add any value to the post overall? (The focus of this attribute is the clarification question itself, regardless it has an answer or not.)
- Does answering the clarification question add any information to the accepted answer of the post?

Three annotators, one who also acted as *coordinator*, carried out the labelling. Annotators included one man and two women, one with knowledge in finance. They were all proficient in English and conducted the annotation process in three months. The coordinator met the other two annotators to explain the labelling strategy and the annotation procedure (the *guidelines*) to provide a common ground for everyone. The coordinator collected all annotations, aggregated them, and identified any disagreements. Next, the coordinator met the annotators to get their feedback and discuss any challenges that they encountered. The annotators discussed the labels with disagreement. As a result of the discussions, the guidelines were sometimes refined, and clarification questions with disagreements were re-labelled, and those new labels were aggregated once again. 363 posts out of a total of 557 sampled posts were initially agreed. When the guideline was discussed and amended, in a few cases (17 posts) where there was still disagreement, majority voting was used to obtain the final label. This means the coordinator recorded the related label if the agreement score was greater than or equal to 66.67%. The developed annotation guidelines are summarised below:

- ***Actual clarification question***: To determine the usefulness of a clarification question, it is a prerequisite to ensure that a potential clarification question is an actual clarification question. A potential question is considered as an actual question (hereafter, simply called a clarification question) if it is on the topic of the post, if it appears to be clear, and does not contain:
  - Sarcastic/humorous questions and rhetorical questions (Braslavski et al., 2017)
  - Comments which provide a solution or give a hint for the post in the form of a question (e.g. *"Why don't you just try a backtest ...?"*). This type of question does not generally look for an answer. In contrast to Kato et al. (2013), we do not consider such a question as a clarification question.

  To identify a clarification question accurately, investigation of potential clarification questions, any accompanying sentences, and the post submitted by the Asker is required. For those potential clarification questions that are actual clarification questions, the following three attributes are assessed.
- ***Informative answer to the clarification question***: At this stage, we classify an answer to a clarification question as informative or non-informative. This part of the study is essential because a clarification question needs to have an informative answer to help the Askers with their posts. An answer to a clarification question is informative when it responds to the clarification question or a portion of it (Figure 2). There are some conditions when the answer to the clarification question is not informative:
  - The clarification question has accompanying sentences, and the Asker responds to these sentences rather than the clarification question itself (Figure 3).
  - The clarification question receives a relevant but incorrect answer when the Asker misunderstands the clarifi-

cation question (Figure 4).

- ***Valuable for the post***: A clarification question can be relevant to a post, but it does not necessarily add value to it. Here, we consider a clarification question as valuable for the post if it attempts to resolve ambiguity or to eliminate any incompleteness in the post (Figure 5). In contrast, Figure 6 indicates a clarification question asking something that is not about the post, and therefore does not add value to the post. To evaluate this attribute, the clarification question and the post need to be considered together.
- ***Valuable for the accepted answer***: We consider a clarification question valuable for the accepted answer if it improves an accepted answer or if answering the clarification question is necessary to produce an accepted answer for the post. To be considered valuable for the accepted answer, the clarification question needs to meet the following criteria:
  - The clarification question has an informative answer;
  - The clarification question is labelled as valuable for the post. This is because the category of *valuable for the accepted answer* is a subclass of *valuable for the post*; and
  - The post has an accepted answer.

  The clarification question and its answer, the post and the accepted answer need to be considered together to label this attribute. Since a post usually has some introductory parts or details, in order for a clarification question to be valuable for an accepted answer, the question needs to address the main focus of the post, in contrast to being valuable for the post, which can target any aspect of the post. Moreover, the answer to the clarification questions needs to improve the accepted answer. Figure 7 shows an example of a clarification question that is valuable for both the post and the accepted answer.

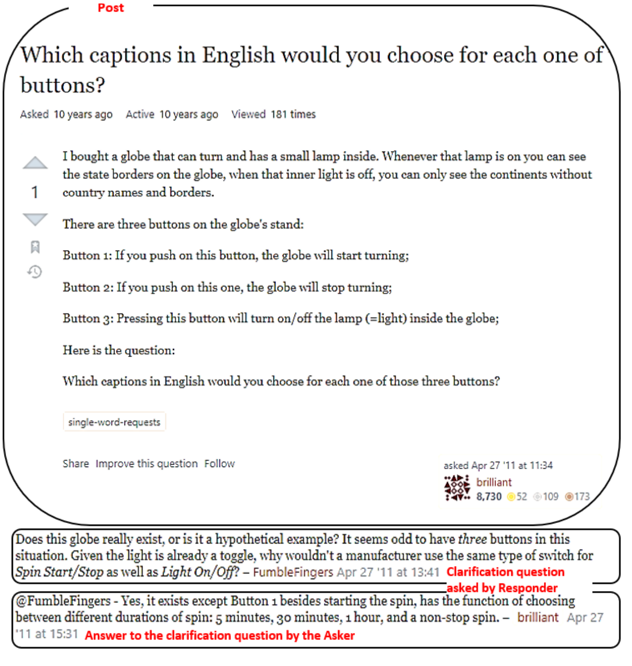

FIGURE 2 Clarification receiving an informative answer.

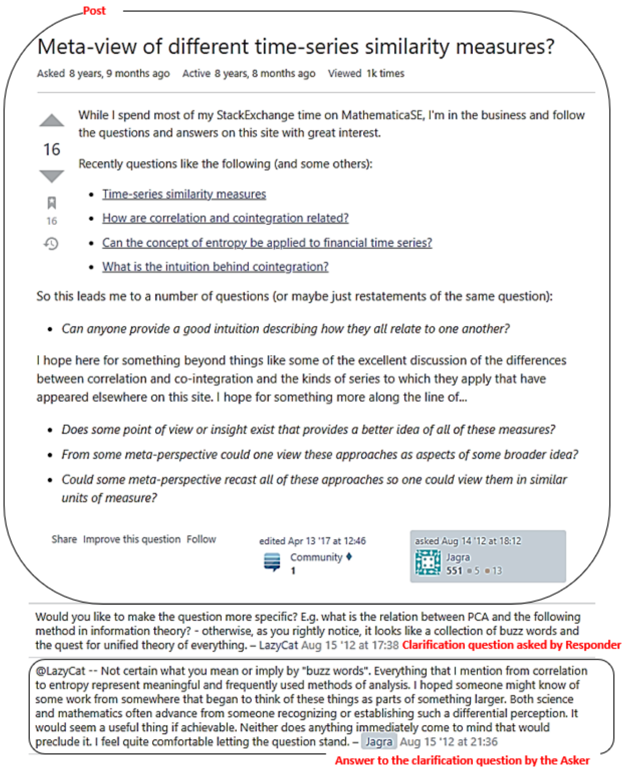

FIGURE 3 Clarification receiving an uninformative answer.

TABLE 2 Sample Size (Number of investigated Potential clarification question).

Type of Answerer	*QF*	*EL*	*SF*
Asker	58	67	67
Responder	30	66	66
No one	67	68	68

When the annotation guidelines were finalised, all clarification questions in the three sites were randomly sampled to assess the labels with respect to the type of answerer. To ensure that the samples were representative of their constituent site, each sample size was estimated based on a finite population with a confidence level of 90% and an error margin of 10%. To give every potential clarification question an equal chance of being selected, the simple random sampling approach with a random number generator is used. In total, 557 potential clarification questions were sampled across the three sites. Table 2 indicates the samples of potential clarification questions (taken from each domain), which were used to assess four different taxonomies based on the type of answerer.

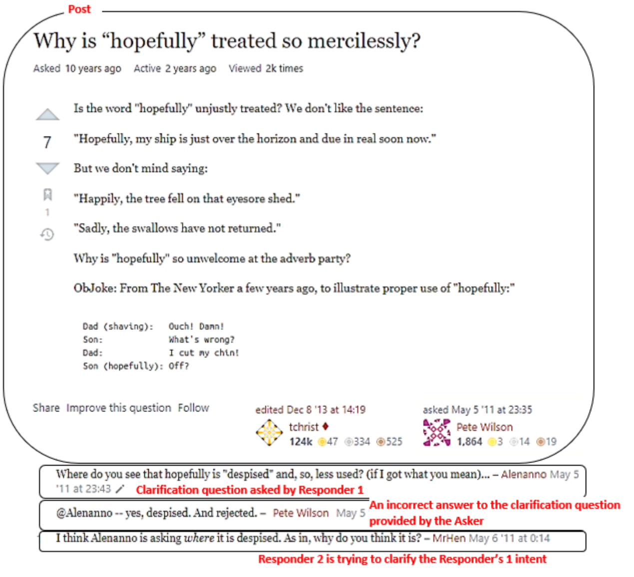

FIGURE 4 A useless answer due to a misunderstanding.

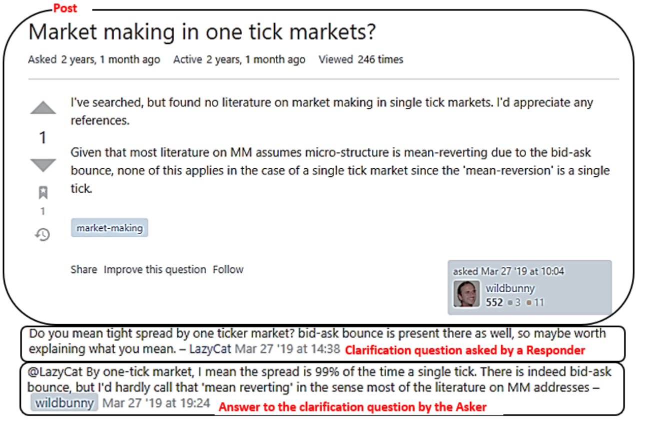

FIGURE 5 A clarification asked to eliminate ambiguity.

4 | RESULTS AND ANALYSIS

In this section, we present the results of this research and analyses on user engagement with clarification questions, the usefulness of clarification questions and clarification types and patterns.

4.1 | User Engagement and Clarification

To understand the utility of clarification questions, we analyse questions with respect to the answerer. Table 3 shows the majority of clarifications (around 90%) are unanswered. Of those that are answered, the Asker is the most likely to reply. The high percentage of clarifications with no answer confirms the importance of investigating the properties of the clarification questions that engage Askers. Results shown in Table 3 are all statistically significant for these three sites and also for the types of the answerer (two-tailed z-test, $p < 0.00001$).

Considering those posts where clarification questions are answered by only one type of answerer, we find that when clarification questions are answered by the Asker, there is a higher chance of the post gaining an accepted answer (Figure 8). The percentage is significantly higher than the other two answerer types across the three sites (two-tailed z-test, $p = 0.05614$ for *QF*; $p < 0.00001$ for *EL*; $p < 0.00001$ for *SF*). Next, we examine the relationship between the percentage of clarification questions answered by the Asker per post with the number of posts with an accepted answer (Figure 9). We divide the percentage of the clarification questions answered by the Asker into four bins. As can be seen, there are similar trends across all three sites: the greater the fraction of clarification questions answered by the Asker, the more posts obtain an accepted answer.

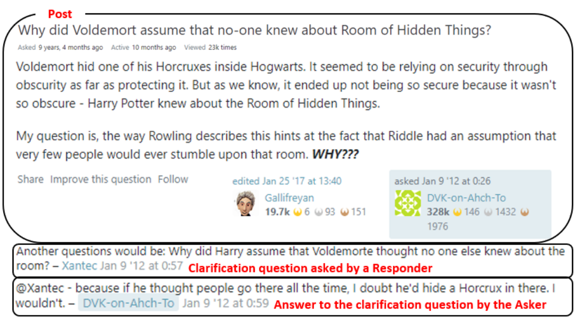

FIGURE 6 A non-valuable clarification question.

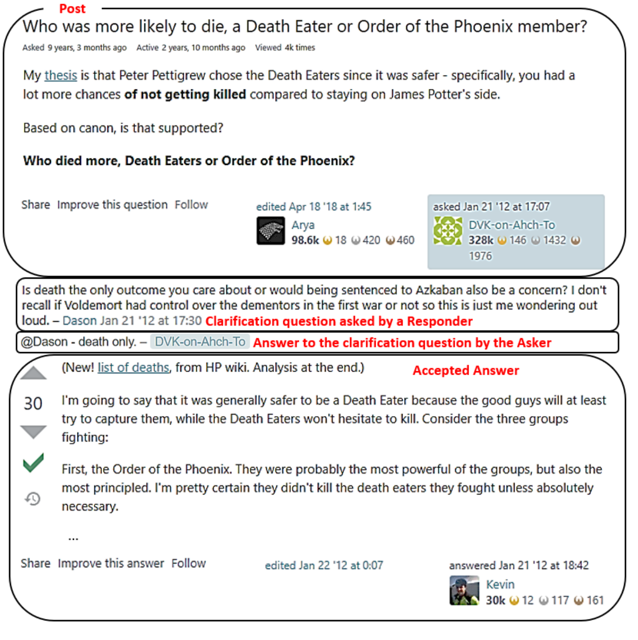

FIGURE 7 A clarification that enhances the accepted answer.

TABLE 3 Who answers clarification questions.

Answerer	*QF*	*EL*	*SF*
Asker	376 (8.7%)	4065 (6.46%)	3144 (9.04%)
Responder	42 (0.97%)	2027 (3.22%)	2100 (6.04%)
Asker & Responder	3 (0.07%)	100 (0.16%)	167 (0.48%)
Unanswered	3905 (90.39%)	56971 (90.48%)	29707 (85.4%)

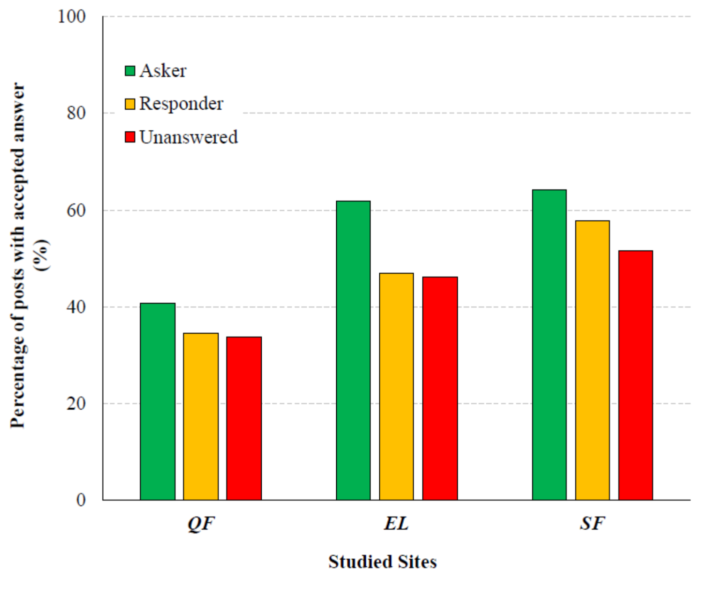

FIGURE 8 Percentage of posts with an accepted answer, grouped by answerer.

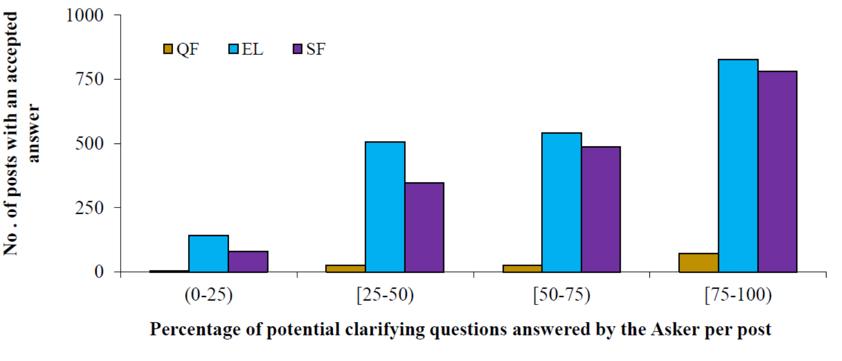

FIGURE 9 Number of posts with an accepted answer grouped by the % of questions answered by the Asker.

Next, for posts with an accepted answer, we examine the relationship between the number of asked clarification questions and the rate at which the Asker answers those questions. Figure 10 shows similar trends across the three sites. We can see the maximum asked potential clarification questions is different in investigated sites (a maximum of 8, 20 and 12 potential clarification questions are asked per post in the *QF*, *EL* and *SF* sites, respectively).

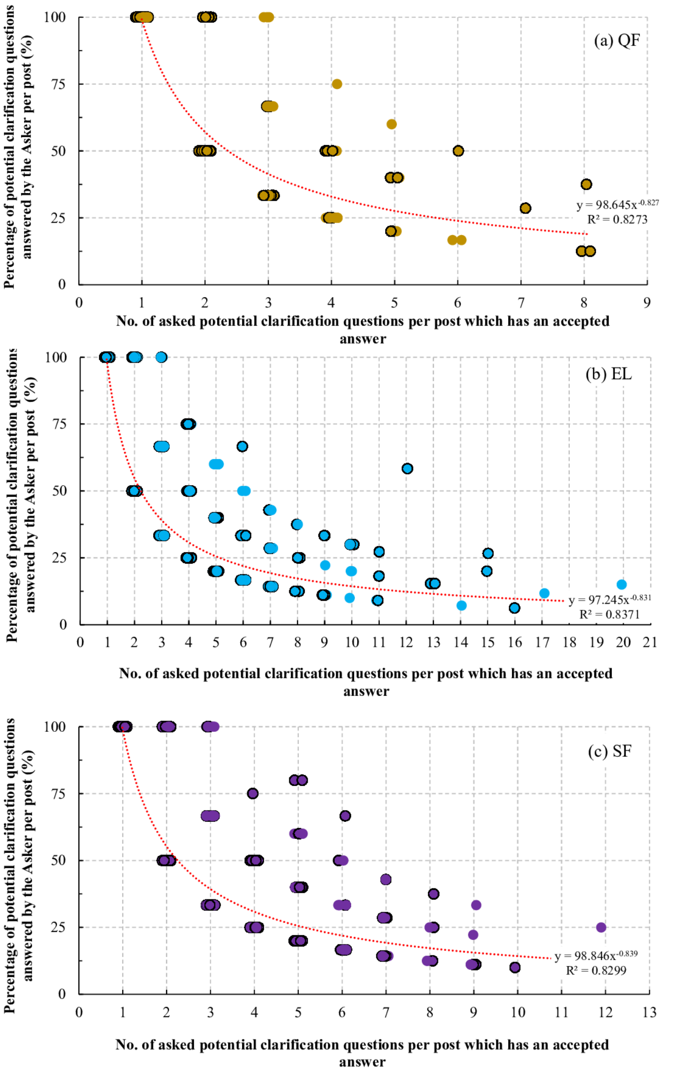

FIGURE 10 Fraction of answered clarifications per question.

Our analysis indicates that more than 79% of posts with an accepted answer have three or fewer potential clarification questions answered by the Asker, regardless of the topic. The mean number of answered potential clarification question per post is 1.12-1.14, with a mode value of 1 for all sites.

Our findings showed that potential clarification questions answered by the Asker increase the chance of a post obtaining an accepted answer. Therefore, for the rest of this analysis, we consider a potential clarification question as useful if it is answered by the Asker and the post receives an accepted answer. We consider a potential clarification question as non-useful if it is left unanswered, but the post still obtained an accepted answer; i.e. answering the question was unnecessary to address Asker's information need.

We also examined the elapsed time between posting a clarification question, the clarification question received an answer, and the post obtaining an accepted answer to investigate if there is any relationship between them. However, we have not observed any noticeable trend. This is perhaps because *Stack Exchange* is an online platform, and the users are located worldwide, so periods of activity can vary substantially.

4.2 | Useful Clarifications

In the previous section, we started to investigate what potential clarification questions are useful in terms of helping a post obtain an accepted answer. In this section, we study clarification usefulness using the manual annotation described in Sec 3.2. We study what percentage of potential clarifications are actual clarification questions. We further investigate the potential clarification questions that have informative answers or help the post to achieve an accepted answer.

Table 4 indicates which potential clarification questions were labelled as actual clarifications. We can see that those answered by Askers are more likely to be actual clarification questions compared to those answered either by a Responder or left unanswered. Table 4 also shows that the average number of actual clarification questions varies across sites (71.6% in *QF* and 59.7% in *SF*, and 46.2% in *EL*); the differences are statistically significant (two-tailed z-test, $p < 0.00001$ for *SF* and *EL* sites; $p = 0.0198$ for *QF* and *SF* sites; and $p = 0.0067$ for *EL* and *SF* sites). The low average number of actual clarification questions in the *EL* site in comparison to the other two suggests that most of the clarification questions in this site might be a comment containing either sarcastic/humorous questions, rhetorical questions, off-topic questions as explained in Section 3.2. This again suggests that the nature of a topic could influence user behaviour regarding asking and responding to clarification questions.

TABLE 4 Percentage of actual clarification questions.

Answerer	*QF*	*EL*	*SF*
Asker	79.3	55.2	64.2
Responder	66.7	40.9	54.5
Unanswered	68.7	42.6	60.3
Average	71.6	46.2	59.7

We next investigate the usefulness of clarification questions in terms of their having an informative answer and being valuable for the post and the accepted answer. In all four of the graphs in Figure 11 we see that more than 90% of clarification questions answered either by the Asker or a Responder have an informative answer. This shows that clarification questions normally receive informative answers regardless of the topic. Figure 11 also demonstrates that the clarification questions answered by Askers are more valuable for the posts compared to clarifications answered either by a Responder or left unanswered. Moreover, such clarification questions are more valuable for the accepted answer in comparison to those answered by a Responder. This highlights the importance of those clarification questions which are answered by Askers. Findings from Table 3 in addition to the result presented in Figure 11 show that although Responders contribute more to answer potential clarification questions in the *EL* and *SF* sites compared to the *QF* site, their contribution in answering the clarification questions are less valuable for the accepted answers.

The results from the manual annotation helped us to get a better understanding of how a clarification question can be considered as useful in obtaining an accepted answer or non-useful as below:

- **Useful Clarification Questions:** We consider a clarification question useful if it is answered by the Asker, has an informative answer, and is valuable for the post and the accepted answer.
- **Non-Useful Clarification Questions:** We consider a clarification question as non-useful if it is left unanswered and is not valuable for the post, but the post still receives an accepted answer. This is because such clarification questions are not only lacking in value for the post but moreover, answering them is not necessary to address the information need originally posted by the Asker.

The definition of useful and non-useful clarification questions can be applied in other studies on clarification questions and on other platforms as the usefulness is defined based on the successful interaction with the user who submits the post and the answer to the clarification in this study whether it is useful for the post to obtain an accepted answer or not.

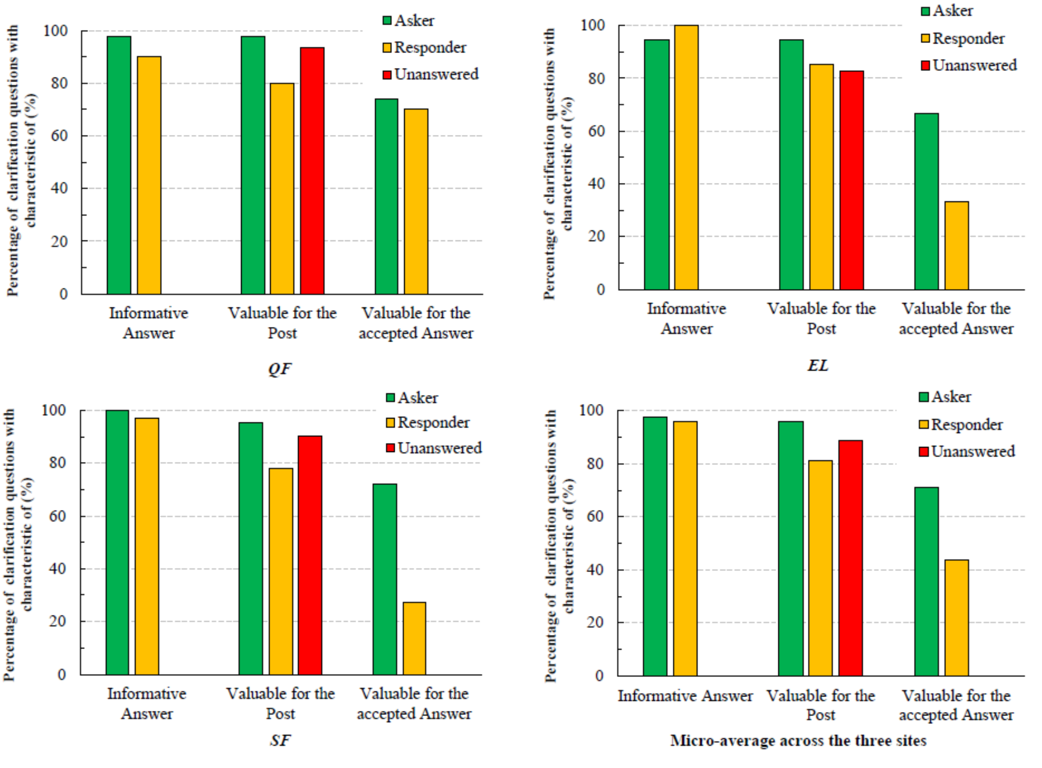

FIGURE 11 Characteristics of clarification questions answered either by the Asker or a Responder.

4.3 | Clarification Types and Patterns

To investigate what are the characteristics of useful clarification questions, we now analyse the type (based on user intent) and the pattern (represented as a trigram of words) of clarification questions. We compare the most common patterns of useful clarification questions with the patterns of non-useful clarification questions. This analysis provides insight for developing models for generating clarification questions in conversational search.

To recognise the patterns and types of clarification questions, we first performed a manual annotation on 376 comments which were answered by the Askers in the *QF* site and contained at least one clarification question. The annotation followed the same procedure described in Section 3.1: two annotators performed the classification to reduce the impact of personal judgement; disagreements were discussed between the annotators and a coordinator to make the final decision, and we iteratively updated the annotation guideline based on the discussions between the annotators and the coordinator.

As the large size of the complete dataset made complete manual labelling infeasible, the remaining clarifications (answered by Responders or left unanswered) in the *QF* site, and all clarifications in the two other sites, were labelled automatically using a Random Forest classifier. We used TF-IDF weighted bag-of-word features which included n-grams ($n \in \{1, 2, 3, 4, 5\}$). To provide more meaningful representation some patterns such as *could you please give*, *can you please give*, *could you give* and *can you give* are compiled. The classifier was trained on 80% of the aforementioned annotated dataset collected from the *QF* site. Each data point includes a clarification question and its type and pattern. We also noticed that there were some questions that did not follow any specific patterns (e.g. *"Perhaps question an example of one of your long sentences?"*). Therefore, we put some clarifications which did not follow any patterns in our training data to detect such clarification questions. We excluded these clarification questions from our analysis. To foster this research, we open-source our implementation and annotated data.[2]

To measure the accuracy of the classifier, we used the remaining 20% of the annotated data as test data. Our model achieved an accuracy of 78.21% on the held-out test set. To further verify the quality of the automatic annotation process, 10% of the results were cross-checked by the annotators. The accuracy of the labelling reported by the annotators was found to be 72.32%. Annotators were also asked to edit any pattern which was labelled mistakenly and to add these to the list of detected patterns and types.

After classification, we investigated if there was any relationship between clarification questions answered by the Asker, posts with an accepted answer, and the type of clarification questions (Section 4.3.1). We then compared and analysed the highest frequency patterns for the useful and non-useful clarification questions (Section 4.3.2).

4.3.1 | Clarification Types

The six identified types of clarification questions are presented below:

- ***Ambiguity/Incompleteness***: The clarification question asks about an unclear part of the post. The post either is ambiguous or some information is missing, which means further clarification or more details are required. In such cases, asking a clarification question may lead to the post being revised (e.g. *"How much money did you assume to start with?"*)
- ***Confirmation***: The Responder may ask a clarification question to confirm her/his perception about a piece of information in the post. In some other cases, the Responder may want to emphasise important information in the post and confirm it with the Asker (e.g. *"Does it have to be a single word?"*).

[2] `github.com/Leila-Ta/Clarification_CQA/`

- ***General***: The clarification question is a general question that does not refer to any specific part of the post. Such clarification questions can often be asked from all posts (e.g. *"Would you like to make the question more specific?"*).
- ***Incorrectness***: When the Responder thinks there is wrong information in the post, this type of clarification question is asked by the Responder to resolve the problem before providing an answer (e.g. *"Are you sure it is 62 and not 66?"*).
- ***Paraphrasing***: The Responder attempts to paraphrase the post by asking clarification questions to make it more digestible and to understand the post correctly (e.g. *"Are you asking how to write an exchange simulator?"*).
- ***Suggestion***: The Responder asks clarification questions to draw the Asker's attention to a specific point, which can sometimes be a solution in the form of a suggestion, a reference, or an example (e.g. *"Can the book "Monte Carlo simulation in financial engineering" by Glasserman help you?"*).

Figure 12 shows the percentage of the type of the asked clarification questions in each site. It is evident that all three sites show similar trends for all types except *Suggestion* in the *SF* site. We can see that *Ambiguity* and *Confirmation* are the most common followed by *General* and *Suggestion*. It is not unusual to find out that the type of *Ambiguity* is the most common type. This is because the clarification questions are mainly asked to eliminate any ambiguity or lack of information in the post.

To answer whether all clarification questions are useful (**RQ1**), we found out in sections 4.1 and 4.2 that a clarification question is useful if it is answered by the Asker and helps the post to obtain an accepted answer. Now, as a first step toward answering **RQ2** which is characterising useful clarification questions, we investigate if there is a relationship between the type of clarification question, Asker interaction, and the likelihood of a post obtaining an accepted answer. We define a series of measures as follows:

- $ClarQ_{Type}$: The type of clarification question.
- $ClarQ_{Asker}$: A clarification question is answered by the Asker.
- $P_{ClarQ_{Type}}$: A post contains a particular clarification type.
- P_{AccAns}: A post contains an accepted answer.

We graph the following conditional probabilities:

- P($ClarQ_{Type} \mid ClarQ_{Asker}$): The conditional probabilities in Figure 13 show the *Ambiguity* type is most answered by Askers, however, *Ambiguity* is the commonest type of clarification question.
- P($ClarQ_{Asker} \mid ClarQ_{Type}$): Figure 14 indicates that relative to the number of clarifications of each type, the probability of the Asker answering clarifications is generally even across each site. There are some site-specific variations, however.
- P($P_{ClarQ_{Type}} \mid P_{AccAns}, ClarQ_{Asker}$): Figure 15 shows that given that a post with an accepted answer has a clarification question, which has been answered by the Asker, the clarification has a high chance to be of type *Ambiguity*. However, this is because clarification questions of this type are asked more.
- P($P_{AccAns} \mid P_{ClarQ_{Type}}, ClarQ_{Asker}$): The graph of this conditional probability (Figure 16) shows that clarification type is less important to a post having an accepted answer if the Asker answers a clarification question. We also see that when clarification questions are answered in the *EL* and *SF* sites, regardless of the type, the post gets an accepted answer almost always.

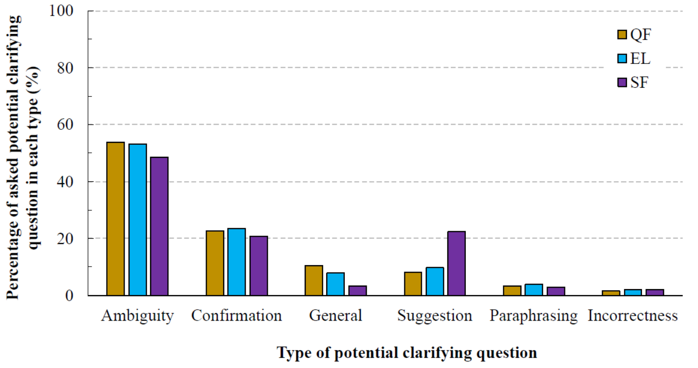

FIGURE 12 Distribution of clarifications by type.

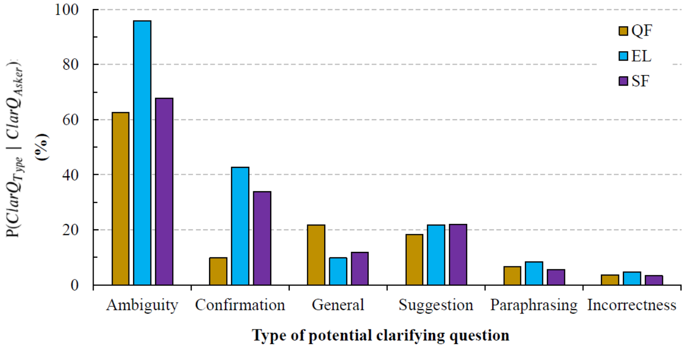

FIGURE 13 The probability of a clarification question which has a certain type, given that the clarification question is answered by the Asker.

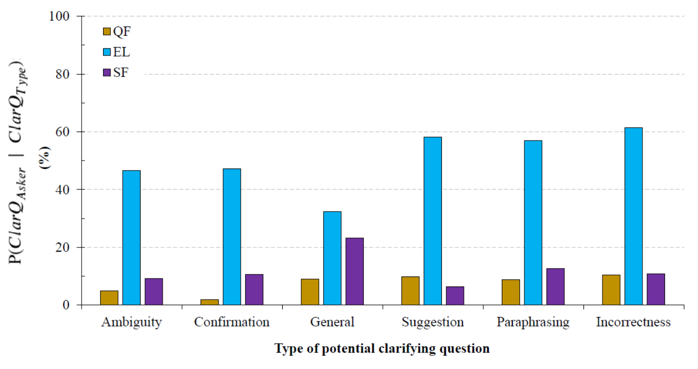

FIGURE 14 The probability of a clarification question being answered by the Asker, given a particular clarification question type.

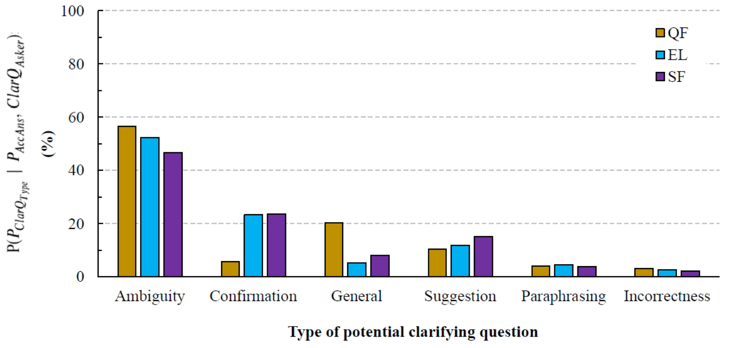

FIGURE 15 The probability of a post has a particular clarification question type, given that the post has an accepted answer.

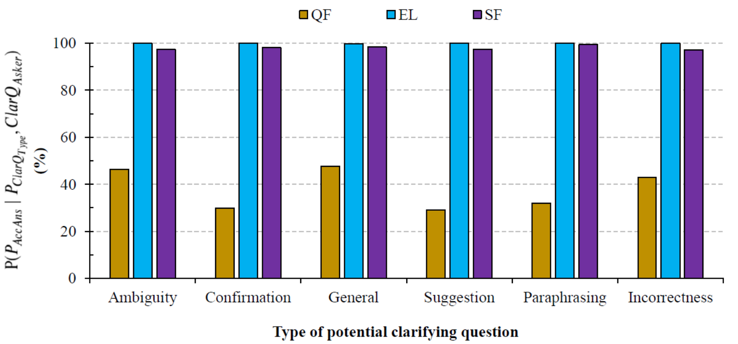

FIGURE 16 The probability of a post having an accepted answer, given the presence of a clarification question of a specific type.

4.3.2 | Clarification Patterns

In this section, we detect the patterns of useful clarification questions. Comparing the detected patterns of useful clarification questions with the detected top 20 patterns and related types of useful potential clarification questions showed the same patterns. This indicates that we have recognised the patterns of useful clarification questions without being impacted by the size of the sampled dataset. We also notice there are 65%, 65%, and 85% similarity in the most popular patterns (top 20 patterns) of the two sites of *QF* and *EL*; the two sites of *QF* and *SF*; and the two sites of *EL* and *SF*, respectively. We present the most popular patterns (the share and sequence of terms) of useful clarification questions with respect to the type of clarification questions. These patterns are collectively extracted from three sites and shown in Figure 17. The gaps in this figure are patterns with a low frequency, which are not shown. The patterns *Is it (noun)* and *Is there a/any* in the type of *Ambiguity/Incomp.*, the pattern of *Do you mean* in the type of *Confirmation*, the pattern of *Can you give* in the type of *General*, the pattern of *Are you asking/looking* in the type of *Paraphrasing*, the pattern of *Are you sure* in the type of *Incorrectness* and the pattern of *Do you know* in the type of *Suggestion* are the patterns with the highest frequency in the useful clarification questions. We also observe that apart from the pattern of *Is it (noun)* which exists in both types of *Ambiguity/Incomp.* and *Confirmation*, there are no other similar patterns in different types of useful clarification questions. Our pattern analysis also shows that more than 80% of useful clarification questions can be generated with 25 patterns. This finding suggests that the identified patterns can be used in asking/generating useful clarification questions in conversational search systems.

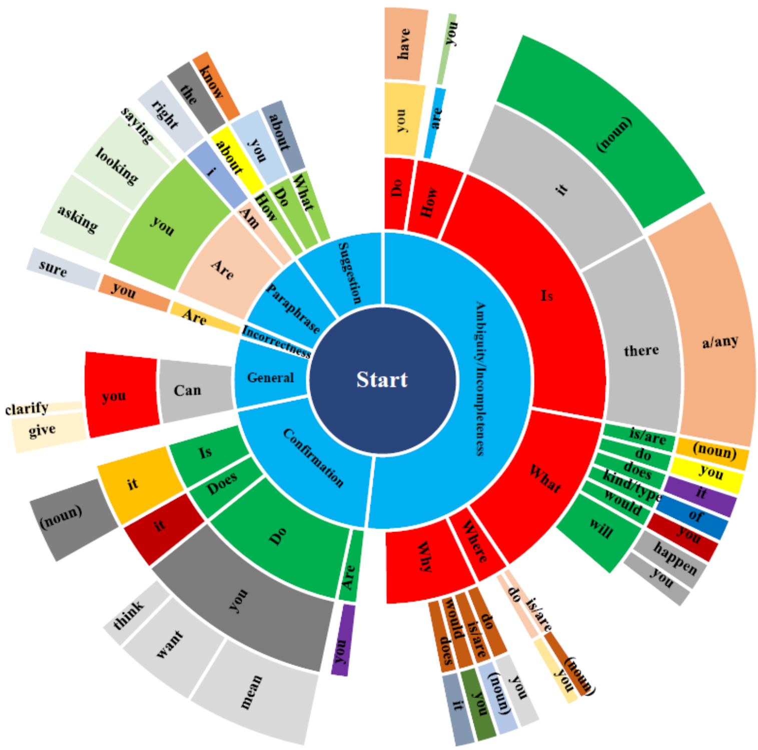

FIGURE 17 Popular clarification patterns grouped by type.

After studying the characteristics, types and patterns of useful clarification questions, we investigate the differences between the pattern of useful and non-useful clarifications. We compare popularity distributions of the patterns of the useful (P(x)) and non-useful clarification questions (Q(x)) by computing point-wise Kullback-Leibler Divergence (D_{KL}). The popularity of distribution is based on the frequency of patterns. Figure 18 shows the top ten and bottom ten patterns by D_{KL} score. Positive scored patterns are used more in useful clarification questions, while negative scores are patterns used more in non-useful clarifications. Comparing Figures 17 and 18 shows that there are some popular patterns that are shared in both useful and non-useful clarification questions. However, some of them are more common in one group. For example, the patterns *Do you mean* and *What is/are (noun)* are in popular patterns of useful clarification questions presented in Figure 17. However, the pattern of *Do you mean* is more likely to be asked in useful clarifications. In contrast, the pattern *What is/are (noun)* is common with non-useful clarifications.

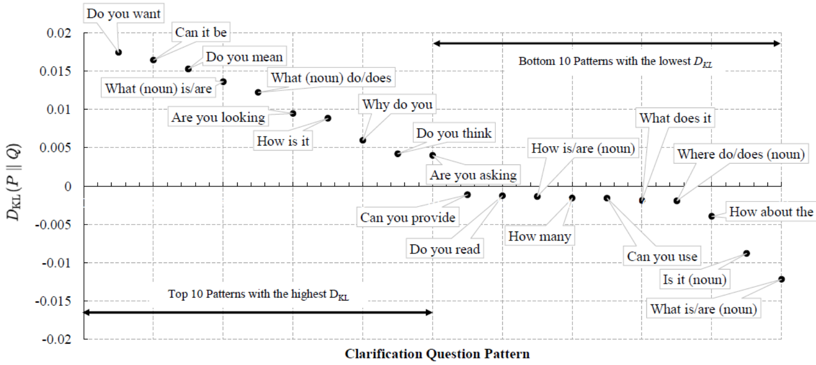

FIGURE 18 The top ten and bottom ten patterns by point-wise *KL-Divergence*, (*P(x)*) is popularity distributions of the patterns of the useful clarification questions and (*Q(x)*) is popularity distributions of the non-useful clarification questions.

5 | DISCUSSION

Information seeking community forums have been studied widely over the last two decades (e.g. (Liu et al., 2008; Li and King, 2010; Riahi et al., 2012; Zhang et al., 2014; Liu et al., 2015; Le and Shah, 2018)). However, human-to-human interaction in such platforms in terms of interaction with clarification questions is a new area of interest. This is particularly helpful as it can enhance our knowledge of developing conversational search systems with higher performance. We showed that there is some research on generating and asking clarification questions and some studies on clarification questions in QA forums. While generating and asking a clarification question is important, getting an answer for the clarification question can even be more important (Tavakoli, 2020). This means a clarification question helps the search system only when the user provides an informative answer. To improve the performance of conversational search systems by generating clarification questions that engage the users more actively (useful clarification question), we attempted to find characteristics of such clarifications in terms of types and patterns in this study.

The aim of our research was to understand why some clarification questions lead to more engagement from Askers. This is because we discovered that many clarification questions are left unanswered, and such clarification questions are generally not useful in helping to better understand an Asker's information need. We could not have understood this without investigating the answer to the clarification questions at the same time. Therefore, we classified clarification questions with respect to the type of *answerer*. This is in contrast to Braslavski et al. (2017) who only studied the clarification question itself.

We understood from Table 3 that there are notably fewer Responders answering in the *QF* site compared to the other sites. We speculate that the nature of a site's topic influences the behaviour of users in engaging with clarification questions. In domains where expert knowledge is required, e.g. *QF*, the questions are more specific where the Asker is the only person who can answer.

We also observed similar trends in three domains in terms of the number of asked potential clarification questions per post with an accepted answer versus the percentage of potential clarification questions which was answered by the Asker in each post (Figure 10). This observation contrasts with the findings of Zou et al. (2020) who investigated clarification question-based systems and showed that Askers normally answer 11-21 clarification questions on average. We conjecture that underlying differences between the analysed platforms account for this divergence: the architecture and the nature of Zou et al. 's system was different from a typical QA forum: while we investigate natural language questions on QA forums, Zou et al. (2020) performed their experiments in the domain of the *Amazon* retail platform, where a user answered questions generated by the system with a "Yes", a "No" or a "Not Sure", to find a target product to buy.

The usefulness of clarifications in this study was investigated using manual annotation, and to enhance the accuracy of the analysis, we sampled 557 potential clarification questions with their answers for labelling, which was more than double the sample size of the previous study conducted by Braslavski et al. (2017). Our annotation showed that those clarification questions answered by the Asker are more valuable in terms of adding value to the post and the accepted answer. The findings led to proposing new definitions to distinguish useful and non-useful clarification questions.

As the first step towards understanding how to generate useful clarification questions in information seeking conversations, we attempted to find the type and patterns of such clarifications. When we detected the type and patterns of the useful and non-useful clarification questions, we applied them to all potential clarification questions in three investigated sites and found similar trends. While the top 20 patterns and related types of useful clarification questions extracted from manual annotation and useful potential clarification questions extracted by the automatic classifier found to be all the same, in type recognition analysis presented in Figures 12 to 16 we only focused on potential clarification questions. However, comparing the results with manually labelled clarification questions, we found that the results are consistent.

Our type classification was also more detailed compared to previous studies (e.g. (Kato et al., 2013)). The type of *Reason* proposed by Kato et al. (2013) questioned a general aspect of the post (*General* type) or an unknown aspect (*Ambiguity* type). Therefore, it was decomposed into the *General* and *Ambiguity* types. In addition, the types of *Selection* and *Check* had noticeable similarity and we merged them into the type *Confirmation*. The types *More Info* and *Experience* in previous studies were the same as our types of *Ambiguity/Incompleteness* and *Suggestion*, respectively. In this work, we also introduced two new types: *Paraphrasing* and *Incorrectness*.

We also noticed that there are some potential clarification questions in each type that are not actual questions. For example, in the type of *Suggestion*, the Responder normally did not look for any answer (e.g. *"Monte Carlo simulation in financial engineering" by Glasserman help you?*). This case was found to be dominant in the type of *Suggestion*. Nevertheless, even in our study, types may overlap in some cases, which means a clarification question may belong to more than one type. For example, the clarification question *"Have you considered using the Fama/French market factor as a reference?"* can fall into either the *Confirmation* or *Suggestion* types; in such cases, it was assigned to both types.

Comparing our results with the previous study carried out by Braslavski et al. (2017) showed us the most frequent patterns are different when we investigate all clarification questions (regardless of their answers) compared with those useful clarification questions that are answered by the Asker and help the post to get an accepted answer. Even some of the most frequent patterns of general clarification questions suggested by Braslavski et al. (2017) fall into non-useful clarification questions (Figure 18). This highlights the importance of distinguishing useful clarification questions from non-useful ones. The same thing was observed in the types of clarification questions. The type of *Ambiguity/Incompleteness* was found to be well above others in terms of frequency in the useful clarification questions, which is different from the previous finding. It is worth noting that three completely different sites of *Quantitative Finance* (from Business domain), *English Language and Usage* (from Culture/Recreation domain) and *Science Fiction and Fantasy* (from Life/Arts domain), were chosen for this study. While the nature of these domains is very different, we observed similar trends in terms of users interaction and type and patterns of clarification questions. We also mentioned earlier that there are 65%, 65%, and 85% similarity in the most popular patterns (top 20 patterns) of the two sites of *QF* and *EL*; the two sites of *QF* and *SF*; and the two sites of *EL* and *SF*, respectively. This suggests that the findings can be applicable cross-domain.

6 | CONCLUSIONS

Asking clarification questions has been found to be an effective way to disambiguate and better understand user information needs. In this paper, we analysed the role of clarification in information seeking conversations. In particular, we explored asynchronous conversations extracted from multiple *Stack Exchange* sites, studied clarification questions based on their answers, and provided insights into useful clarifications–those whose answers help reach accepted answers of user information needs. We examined and answered the following research questions:

- **RQ. 1**: What clarification questions are more useful? (in terms of helping the Asker to get a correct answer)?
  - We found that many clarification questions are left with no answers, and some of them do not add value to the post, i.e., answering the user information needs is independent of those clarifications. This shows that it is important to identify and characterise useful clarifications. Therefore, we discerned useful clarification questions from non-useful ones based upon the results.
- **RQ. 2**: What are the characteristics of useful clarification questions?
  - We classified useful clarification questions into different types based on user intents and extracted their patterns. Our analysis showed that the type of *Ambiguity/Incompleteness* is the most frequent compared to the other types, regardless of the topic. Moreover, we showed that three other types, *Confirmation*, *General* and *Suggestion*, are also useful for the post to obtain an accepted answer as they can lead to about 41.27% more successful resolution of information needs. Investigation of useful and non-useful clarification questions showed that there are specific patterns, which are not only the most common for useful clarification questions but also are less asked for non-useful clarification questions. Such patterns can be employed by conversational search systems for generating clarification questions that are more likely to result in user engagement with a conversational system.

In the future, we intend to expand this study by focusing on information seeking user-agent conversations. The synchronous nature of such conversations and the impact of system mistakes are likely to make the work more challenging. We are also interested in applying our findings to clarification question generation models for various conversational information seeking scenarios (e.g. when we have several generated clarification questions for a query, the common patterns detected in this study can be considered to select the most engaging clarification question). Studying the cost and benefit of asking clarifications in conversational search systems, in addition, to accurately utilising the user responses to clarification for satisfying the user information need, are also left for future work.

references

Aliannejadi, M., Zamani, H., Crestani, F., and Croft, W. B. (2019). Asking clarifying questions in open-domain information-seeking conversations. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '19, pages 475–484.

Anand, A., Cavedon, L., Joho, H., Sanderson, M., and Stein, B. (2020). Conversational Search (Dagstuhl Seminar 19461). *Dagstuhl Reports*, 9(11):34–83.

Bodoff, D. and Raban, D. (2016). Question types and intermediary elicitations. *Journal of the Association for Information Science and Technology*, 67(2):289–304.

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.

Choi, E. and Shah, C. (2016). User motivations for asking questions in online q & a services. *Journal of the Association for Information Science and Technology*, 67(5):1182–1197.

Christakopoulou, K., Radlinski, F., and Hofmann, K. (2016). Towards conversational recommender systems. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 815–824.

Dhole, K. D. (2020). Resolving intent ambiguities by retrieving discriminative clarifying questions. *arXiv preprint arXiv:2008.07559*.

Hashemi, H., Zamani, H., and Croft, W. B. (2020). Guided transformer: Leveraging multiple external sources for representation learning in conversational search. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, pages 1131–1140.

Jeng, W., DesAutels, S., He, D., and Li, L. (2017). Information exchange on an academic social networking site: a multidiscipline comparison on researchgate q&a. *Journal of the Association for Information Science and Technology*, 68(3):638–652.

Kato, M. P., White, R. W., Teevan, J., and Dumais, S. T. (2013). Clarifications and question specificity in synchronous social q&a. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, pages 913–918.

Kiesel, J., Bahrami, A., Stein, B., Anand, A., and Hagen, M. (2018). Toward voice query clarification. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1257–1260.

Kiesel, J., Bahrami, A., Stein, B., Anand, A., and Hagen, M. (2019). Clarifying false memories in voice-based search. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, pages 331–335.

Kim, S. and Oh, S. (2009). Users' relevance criteria for evaluating answers in a social q&a site. *Journal of the American society for information science and technology*, 60(4):716–727.

Kiyota, Y., Kurohashi, S., and Kido, F. (2002). Dialog navigator: A question answering system based on large text knowledge base. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics.

Krasakis, A. M., Aliannejadi, M., Voskarides, N., and Kanoulas, E. (2020). Analysing the effect of clarifying questions on document ranking in conversational search. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*, pages 129–132.

Kumar, V. and Black, A. W. (2020). ClarQ: A large-scale and diverse dataset for clarification question generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7296–7301, Online. Association for Computational Linguistics.

Kumar, V., Raunak, V., and Callan, J. (2020). Ranking clarification questions via natural language inference. In *29th ACM International Conference on Information and Knowledge Management (CIKM)*.

Le, L. T. and Shah, C. (2018). Retrieving people: Identifying potential answerers in community question-answering. *Journal of the association for information science and technology*, 69(10):1246–1258.

Li, B. and King, I. (2010). Routing questions to appropriate answerers in community question answering services. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1585–1588.

Liu, D.-R., Chen, Y.-H., Shen, M., and Lu, P.-J. (2015). Complementary qa network analysis for qa retrieval in social question-answering websites. *Journal of the Association for Information Science and Technology*, 66(1):99–116.

Liu, Q., Agichtein, E., Dror, G., Gabrilovich, E., Maarek, Y., Pelleg, D., and Szpektor, I. (2011). Predicting web searcher satisfaction with existing community-based answers. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 415–424.

Liu, Q., Agichtein, E., Dror, G., Maarek, Y., and Szpektor, I. (2012). When web search fails, searchers become askers: understanding the transition. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 801–810.

Liu, Y., Bian, J., and Agichtein, E. (2008). Predicting information seeker satisfaction in community question answering. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 483–490.

Min, S., Michael, J., Hajishirzi, H., and Zettlemoyer, L. (2020). Ambigqa: Answering ambiguous open-domain questions. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Penha, G., Balan, A., and Hauff, C. (2019). Introducing mantis: a novel multi-domain information seeking dialogues dataset. *arXiv preprint arXiv:1912.04639*.

Qu, C., Yang, L., Croft, W. B., Trippas, J. R., Zhang, Y., and Qiu, M. (2018). Analyzing and characterizing user intent in information-seeking conversations. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, pages 989–992.

Rao, S. and Daumé III, H. (2018). Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2737–2746.

Rao, S. and Daumé III, H. (2019). Answer-based adversarial training for generating clarification questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 143–155.

Riahi, F., Zolaktaf, Z., Shafiei, M., and Milios, E. (2012). Finding expert users in community question answering. In *Proceedings of the 21st International Conference on World Wide Web*, pages 791–798.

Shah, C. (2011). Measuring effectiveness and user satisfaction in yahoo! answers. *First Monday*.

Shah, C., Oh, S., and Oh, J. S. (2009). Research agenda for social q&a. *Library & Information Science Research*, 31(4):205–209.

Shwartz, V., West, P., Bras, R. L., Bhagavatula, C., and Choi, Y. (2020). Unsupervised commonsense question answering with self-talk. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Tavakoli, L. (2020). Generating clarifying questions in conversational search systems. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 3253–3256.

Vtyurina, A., Savenkov, D., Agichtein, E., and Clarke, C. L. (2017). Exploring conversational search with humans, assistants, and wizards. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 2187–2193.

Wu, Z., Kao, B., Wu, T.-H., Yin, P., and Liu, Q. (2020). Perq: Predicting, explaining, and rectifying failed questions in kb-qa systems. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 663–671.

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*, WWW '20, page 418–428, New York, NY, USA. Association for Computing Machinery.

Zamani, H., Lueck, G., Chen, E., Quispe, R., Luu, F., and Craswell, N. (2020b). Mimics: A large-scale data collection for search clarification. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, CIKM '20.

Zamani, H., Mitra, B., Chen, E., Lueck, G., Diaz, F., Bennett, P. N., Craswell, N., and Dumais, S. T. (2020c). Analyzing and learning from user interactions for search clarification. In *Proceedings of the 43rd international ACM SIGIR conference on Research and development in Information Retrieval*, SIGIR '20, New York, NY, USA. Association for Computing Machinery.

Zhang, Y., Chen, X., Ai, Q., Yang, L., and Croft, W. B. (2018). Towards conversational search and recommendation: System ask, user respond. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 177–186.

Zhang, Z., Li, Q., Zeng, D., and Gao, H. (2014). Extracting evolutionary communities in community question answering. *Journal of the Association for Information Science and Technology*, 65(6):1170–1186.

Zou, J. and Kanoulas, E. (2019). Learning to ask: Question-based sequential bayesian product search. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 369–378.

Zou, J., Kanoulas, E., and Liu, Y. (2020). An empirical study of clarifying question-based systems. In *29th ACM International Conference on Information and Knowledge Management (CIKM)*.