Clarifications and Question Specificity in Synchronous Social Q&A

Makoto P. Kato
Kyoto University
Yoshida Honmachi, Sakyo
Kyoto 6068501 Japan
kato@dl.kuis.kyoto-u.ac.jp

Jaime Teevan
Microsoft Research
One Microsoft Way,
Redmond, WA 98052, USA
teevan@microsoft.com

Ryen W. White
Microsoft Research
One Microsoft Way,
Redmond, WA 98052, USA
ryenw@microsoft.com

Susan T. Dumais
Microsoft Research
One Microsoft Way,
Redmond, WA 98052, USA
sdumais@microsoft.com

CHI 2013 Extended Abstracts, April 27–May 2, 2013, Paris, France.
ACM 978-1-4503-1952-2/13/04.

Abstract
Synchronous social question-and-answer (Q&A) systems help people find answer by connecting them with others via instant messaging. To understand how such systems can quickly and effectively establish fruitful connections, we analyze conversations collected from a working enterprise social Q&A system. We show that when askers start with underspecified questions (e.g., "I need help with mail access"), they receive clarification requests, extended dialogs, and poor responses. To address this we are implementing and deploying support within a Q&A system to foster more complete questions, reduce the need for clarification, and benefit both askers and answerers.

Author Keywords
Synchronous social Q&A; dialog analysis; clarification

ACM Classification Keywords
H.5.3 [Information interfaces and presentation]: Group and organization interfaces

General Terms
Human Factors, Measurement

Introduction and Background
Synchronous social Q&A systems, like Aardvark [4] and IBM Community Tools [13], let people ask questions of others via synchronous communication channels like instant messaging (IM). Researchers have studied question features that affect answer quality in asynchronous social Q&A. For example, Teevan et al.

[12] conducted a controlled study where people broadcasted variants of the same question to their social network, and found that question phrasing impacts the quantity, quality, and speed of the responses received. Other research has focused on the quality of answers as evaluated by the asker in Web forums and synchronous Q&A settings [6], [10]. For example, Richardson and White [10] developed a model to predict the subjective quality of answers in synchronous social Q&A. They reported that features of the dialog, the question, and the historical information of users were effective for prediction.

Synchronous Q&A systems allow for rapid interaction between askers and answerers. An important feature of such systems is that answerers can probe the asker's question, and askers can request additional details regarding the provided answers. Clarifications play a key role in establishing *common ground* (i.e., mutual knowledge, beliefs, and assumptions) in conversations [2]. Purver et al. [9] describe possible forms such requests take, including reprise sentences (where a previous utterance is repeated in full) and gaps (where part of a previous utterance is repeated). Schlangen [11] summarized possible reasons for clarification requests, ranging from concrete (e.g., not understanding the meaning of a word) to more abstract (e.g., ambiguous intentions). Establishing common ground in synchronous Q&A corresponds to the process by which an answerer understands the question, and the asker understands the answer provided to them.

Several studies have been conducted focusing on how the establishment of common ground can vary by communication medium. McCarthy et al. [7] show, through a task-based user study, that common ground is difficult to achieve via text alone. Their experimental results support a prediction by Clark and Brennan [1] that common ground is difficult to establish without audio or visual support. This can present challenges for the users of real-time Q&A systems, who typically rely on text-based IM to communicate. Nonetheless, establishing a common frame of reference is important for these dialogs to be successful, in part because expertise differences exist between participants [5]. By design, Q&A systems aim to bring together people with different levels of domain knowledge. Human-human dialogs can be hampered by vocabulary differences between participants and the tendency of novices to underspecify their goals [8].

In this paper, we analyze how people interact with an existing synchronous social Q&A system, focusing on how clarification requests by the answerer relate to dialog outcomes and the relationship between question specificity and clarification requests. In showing that question specificity impacts the nature and success of dialogs in an open domain Q&A system, we lay the ground work for improved system design. We also describe a work-in-progress system that was designed for requesting a clarification to an unspecific question so that askers and answerers can have a smooth and fruitful Q&A dialog.

Data Analyzed

We studied conversations collected from a field trial of an enterprise social Q&A system, IM-an-Expert [10]. The system receives questions via IM, locates and contacts potential answerers within the enterprise with expertise or interest in the question topic, and mediates an IM dialog between the asker and answerer. The system is deployed within Microsoft Corporation and used by over 4,000 employees. A screenshot of a dialog is shown in Figure 1, which includes an example of a

Figure 1. Screenshot of an IM-an-Expert dialog. Q, A, and S correspond to query, answer, and social labels. Subscripts represent pointers from answers to questions.

clarification request from the answerer ("you are asking about autoarchive in outlook 2010, i assume?") The asker omitted the version of Outlook in her question because people in the company used Outlook 2010 when the question was posted. The answerer assumed she was asking about Outlook 2010, but tried confirming this assumption by the clarification request.

The degree of match between a posted question and user-generated profiles (measured by BM25, a document ranking function) is used to rank candidate experts. The top-ranked experts who are also available at question time (according to their presence information from the IM client) are contacted via IM to determine whether they can answer. If one of the contacted people accepts, the other pending requests are cancelled. If everyone declines or does not respond in time, the system sends additional requests, until a maximum of 45 people have been contacted. Approximately 58% of the questions asked via IM-an-Expert are accepted by an answerer. Once an answerer is identified, IM-an-Expert connects the parties for a free-form IM conversation. When the conversation ends, the system asks the asker to optionally judge, "How helpful was the answer?" (penultimate utterance in Figure 1), on a scale from one (not helpful) to five (very helpful). Most (79%) answered questions receive a rating, which is recorded along with the dialog.

To understand how common ground is established in the IM-an-Expert conversations, we randomly selected 350 conversations with answers from the IM-an-Expert logs. Of these, 50 received no rating, 50 received a low (1 or 2) rating, and 250 received a high (3 to 5) rating. Although the distribution of ratings closely matched that of all rated questions, we biased towards selecting questions with ratings so we could evaluate outcome quality. We manually removed demonstration questions (e.g., "test question" and "help?"), resulting in a set of 333 questions asked by 212 unique askers and answered by 245 unique answerers. On average dialogs lasted around 8 minutes and consisted of 16.3 utterances (i.e., a single IM message).

Understanding Clarification Requests

Utterances were manually labeled by one of the paper authors as *question*, *answer*, or *social*, with the utterance being labeled *question* if it asked something that could potentially be answered by the other party, *answer* if it answered a *question* utterance, and *social* if it was made for social purposes (e.g., "hi," "thanks," or "no problem"). The conversation in Figure 1 show several examples. The total number of labeled utterances is 3,406 (63%), distributed as 1,069 (31%) *questions*, 1,375 (40%) *answers*, and 962 (28%) *social*. Unlabeled utterances (37%) typically provided additional information on the success or failure of solutions attempted by the asker, or progress updates (e.g., "just a sec," or "let me find it").

Not surprisingly, the distribution of utterance type is different for askers and answerers. Table 1 shows the average number of utterances of each type per dialog. Askers ask more questions, and answerers answer more questions. However, we also observe answerers asking questions (average 1.27 per dialog) and askers providing answers (average 1.14 per dialog).

To understand these role reversals, we identified the first answer provided to the asker's initial question and automatically extracted all questions from the answerer prior to that point. Visual inspection revealed that these 234 questions were almost all intended to clarify aspects of the initial question. For this reason, we refer

	Asker	Answerer
Question	1.94 (1.36)	1.27 (2.05)
Answer	1.14 (2.00)	2.99 (2.18)
Social	1.75 (1.28)	1.14 (1.07)
Total	4.83 (3.45)	5.40 (4.61)

Table 1. Average number of utterances per dialog. Standard deviation values are parenthesized.

Check (23%)
"You are trying to create shortcut programmatically?"

More Info (21%)
"What version of Visio are you using?"

General (14%)
"What kind of help do you need?"

Selection (14%)
"Are you in Windows or outside Windows?"

Confirmation (10%)
"You're trying to install on Win 7 x86, right?"

Experience (9%)
"Did you try creating a firewall exception?"

Other (9%)
"If you pick a port over 1024, do you still have the problem?"

Figure 2. Breakdown of clarification request types observed, with examples.

to them as *clarification requests* (CR). Clarification requests occurred in 111 of the 333 dialogs, and utterances related to clarification (including the questions and answers) comprised 8.2% of all utterances in our data set, and they comprise 9.1% of the utterances in the dialogs that contain clarifications.

Types of Clarification Requests

To understand the types of clarification requests used, we classified them into the six user intents shown in Figure 2, similar to Ginzburg and Cooper [3]. The most common clarification type *check* (representing 23% of all requests), aimed to test an answerer's hypothesis. Also popular (21%) were requests for *more information*, which often started with one of the following six interrogative words: who, what, where, when, why, and how. The remaining four types (*general* requests that do not require knowledge on the question, *selection* requests where hypotheses are posed, *confirmation* requests that repeat a part of the previous utterance, and *experience* requests that ask about the asker's experience) represent between 9% and 14% of the total clarification requests. At least 72% of the requests (*check*, *more information*, *general*, *selection*) might have been avoided if the asker had provided more detail in their initial question. This suggests a substantial opportunity to improve Q&A systems.

Question Specificity and Clarification Requests

In examining the Q&A dialogs with clarification requests, we often observed poorly specified initial questions. This prompted us to investigate the relationship between question specificity and clarification requests. The initial questions in each dialog were labeled as *low*, *medium*, or *high* specificity. *Low* specificity questions do not specify the question objective and do not provide sufficient detail to allow an answerer to understand the question. *Medium* specificity questions lack an objective or sufficient detail. *High* specificity questions are specific enough to be answered. For example, questions like, "email web access" and "help with [product] beta" were labeled *low*; "how to free up hard disk space" and "how can I find out when my password will expire?" were labeled *medium*; and "what is the square root of pi?" and "when will [product] be released?" were labeled *high*. Three of the authors labeled question specificity, resolving disagreements via discussion and changing 4.5% of the labels.

Question specificity	Low	Medium	High
N	39	80	214
% of dialogs with CRs	**71.7**%	37.5%	24.7%
# of CRs per dialog	**1.51** (1.66)	0.73 (1.52)	0.55 (1.36)
Expert match	.029 (.031)	.043 (.088)	**.049** (.091)
Answer rating	3.72 (1.00)	3.89 (0.95)	**4.00** (0.83)

Table 2. Dialogs with clarification requests (CRs) by specificity, as well as the average match scores from the expert finding algorithm and average answer rating. Bold = highest value in row. Parenthesized = standard deviation.

Table 2 shows the relationship between clarification requests and question specificity. An ANOVA shows a significant difference in the number of clarification requests across *low*-, *medium*-, and *high*-specificity questions: $F(2,331) = 7.41$, $p < .001$. Tukey post-hoc tests showed significant differences between *low*- and *medium*-specificity questions, and between *low*- and *high*-specificity questions. *Low*-specificity questions were significantly more likely to lead to clarifications than *high*- and *medium*-specificity questions. In

addition, a chi-square test showed significant differences in the fraction of dialogs with clarifications for questions with different specificity: $\chi^2(2) = 33.65$, $p < .001$. Tukey tests on proportions showed significant differences in the percentage of clarification requests for all specificity pairs.

Our findings show that *low*-specificity questions were more likely to be followed by clarification requests than *high*- or *medium*-specificity questions. Question difficulty could also potentially contribute to the presence of clarification requests, but our experience with the data suggests that underspecified questions are of comparable difficulty.

Dialogs with Clarification Requests

To understand whether dialogs with clarification requests are different from those without them, we examined several different metrics, including the average answer rating, total dialog duration, and the number of utterances (Table 3). Our analysis revealed a relationship between clarification requests and lower answer ratings, and longer dialogs.

Answer ratings were significantly higher in dialogs without clarification requests, compared to those with such requests ($t(331) = 2.59$, $p < .01$). One explanation for this is that if the asker provided more specific information in the original question the expert finding algorithm could better find expert answerers. This hypothesis is supported by the expert match and ratings in Table 2, which show that *high*-specificity questions find more qualified answerers than *low*-specificity questions ($F(2,331) = 3.96$, $p = .02$; Tukey test: $p = .021$), and that *high*-specificity questions also receive better answers ($F(2,331) = 4.72$, $p = .009$; Tukey test: all $p < .01$).

	Clarification Requests in Dialog	
	Absent	***Present***
N	222	111
Answer rating	**3.98** (0.89)	3.77 (1.01)
Dialog time (secs)	416 (367)	**659** (460)
Num. utterances	13.0 (12.4)	**23.0** (18.4)

Table 3. Metrics of dialogs, broken down by whether a clarification request was present or not.

Although we expect dialogs with clarification requests to take longer than those without, the differences in duration were surprisingly large. Dialogs with clarification requests are 58% longer than those without such requests ($t(331) = 5.21$, $p < .001$). Similarly, the number of utterances in dialogs with clarification requests is almost double that in those without them ($t(331) = 5.81$, $p < .001$). The magnitude of these differences is remarkable considering that clarification requests comprised just 9.1% of utterances in dialogs with clarifications.

Supporting the Clarification Process

We implemented a clarification request module as part of the existing IM-an-Expert system. The module requests a clarification to an unspecific question before the real-time Q&A system starts to locate answerers (Figure 3). To automatically assess question specificity we learned a logistic regression classifier. Using several characteristics of the asker, the question (e.g., length, form, topic and when it was asked), and potential answerers (e.g., the quality of match) as input features, the learned model output a predicted specificity score. More information was requested for around 10% of questions. Given a clarification by the asker, the module appends the clarification to the original

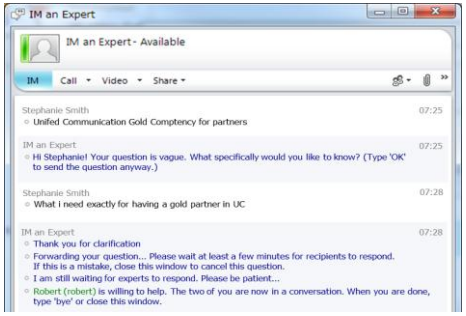

Figure 3. A screenshot of IM-an-Expert with the clarification request module.

question, and then processes the question in the same way we described earlier to find matching experts.

Preliminary analysis suggests that the clarification request and its use need to be carefully designed. We observed that when people clarify, they often add very little content (e.g., "I need help" became "I have a question"). Askers may need more support in how to clarify (e.g., requesting specifics such as knowledge and motivation). Surprisingly, when people did provide meaningful clarification requests we observed that their questions were less likely to receive an answer. Clarifications make the question lengthy, which may be less attractive to answerers. One solution could be to use the question plus clarification in the matching and only show the original question to prospective answerers, assuming that people who better match the full question would be able to quickly decide whether they will answer.

Conclusion

We have presented an investigation of clarification requests using the IM dialogs of an operational Q&A system. This gave us access to real questions and rich dialogs. We showed that unspecific questions may lead to clarification requests, dialogs with clarification requests tend to be longer, and importantly, unspecific questions find less qualified answerers and lead to lower answer ratings. These findings suggest at least two design implications for synchronous social Q&A systems: (1) they need to estimate question specificity, and (2) they should leverage these estimations to prompt askers to provide additional detail on unspecific questions before it is distributed to candidate answerers. To this end, we also described work-in-progress on developing clarification support and early findings from its deployment within our enterprise.

References

[1] Clark, H.H. and Brennan, S.E. Grounding in communication. In: *Perspectives on Socially Shared Cognition*. American Psychological Association, 1991.

[2] Clark, H.H. and Schaefer, E.F. Contributing to discourse. *Cognitive Science*, 13(2): 259-294, 1989.

[3] Ginzburg, J. and Cooper, R. Resolving ellipsis in clarification. *ACL*, 236-243, 2001.

[4] Horowitz, D. and Kamvar, S.D. The anatomy of a large-scale social search engine. *WWW*, 431-440, 2010.

[5] Isaacs, E.A. and Clark, H.H. References in conversations between experts and novices. *Journal of Experimental Psychology*, 116(1): 26-37, 1987.

[6] Liu, Y., Bian, J., and Agichtein, E. Predicting information seeker satisfaction in community question answering. *SIGIR*, 483-490, 2008.

[7] McCarthy, J.C., Miles, V.C., and Monk, A.F. An experimental study of common ground in text-based communication. *SIGCHI*, 209-215, 1991.

[8] Pollack, M.E. Information sought and information pro-vided: an empirical study of user/expert dialogues. *SIGCHI*, 155-159, 1985.

[9] Purver, M., Ginzburg, J., and Healey, P. On the means for clarification in dialogue. *SIGDIAL*, 1-10, 2001.

[10] Richardson, M. and White, R.W. Supporting synchronous social Q&A throughout the question lifecycle. *WWW*, 755-764, 2011.

[11] Schlangen, D. Causes and strategies for requesting clarification in dialogue. *SIGDIAL*, 136-143, 2004.

[12] Teevan, J., Morris, M.R., and Panovich, K. Factors affecting response quantity, quality and speed in questions asked via online social networks. *ICWSM*, 630-33, 2011.

[13] Weisz, J.D., Erickson, T., and Kellogg, W.A. Synchronous broadcast messaging: the use of ICT. *SIGCHI*, 1293-1302, 2006.