Question Quality in Community Question Answering Forums: a survey

Antoaneta Baltadzhieva

Tilburg University
PO Box 90153
5000 LE Tilburg
The Netherlands

a_baltadzhieva@yahoo.de

+31 (0)6 151 01 557

Grzegorz Chrupała

Tilburg University
PO Box 90153
5000 LE Tilburg
The Netherlands

g.chrupala@uvt.nl

+31 (0)13 466 3106

ABSTRACT

Community Question Answering websites (CQA) offer a new opportunity for users to provide, search and share knowledge. Although the idea of receiving a direct, targeted response to a question sounds very attractive, the quality of the question itself can have an important effect on the likelihood of getting useful answers. High quality questions improve the CQA experience and therefore it is essential for CQA forums to better understand what characterizes questions that are more appealing for the forum community. In this survey, we review existing research on question quality in CQA websites. We discuss the possible measures of question quality and the question features that have been shown to influence question quality.

Keywords

Question quality, Community Question Answering, question features

1. INTRODUCTION

The web has changed the way people provide, search and share information and knowledge. It has become straightforward to submit keywords in a search engine to express a need, and the search engine immediately lists a large number of more or less relevant webpages from which the user can choose. However, the search results may not provide an exact solution to the user's problem and it may be time-consuming to review all of them, without having a guarantee of finding the desired answer. Community Question Answering websites offer a new opportunity to obtain the desired knowledge in a more rapid and efficient way.

Community Question Answering (CQA) websites provide an interface for users to exchange and share knowledge. The users in such a forum can be divided in three groups: 1) users who only ask questions, 2) users who only answer questions and 3) users who ask and answer questions [1]. The user asking a question lacks knowledge of a specific topic and searches for an expert on the same topic to provide the desired knowledge. In this way, an asker is querying a topic and the experts providing the knowledge about this topic are the source of information, thus replacing other sources like documents or databases. Although the idea of receiving a direct response to a certain information need sounds very appealing, CQA websites also involve risk because the quality of the provided information is not guaranteed. An important difference between user-generated content and traditional content is the range of the content quality: user-generated content shows a higher variance in quality than traditional content [2], [4]. The quality distribution varies from very high to very low.

As there is a large number of CQA websites, it is important for a CQA website to provide high-quality content to distinguish itself from other websites. The importance of high-quality content in community-driven question and answering websites has been recognized and investigated in several studies. Importantly, in [2] is has been shown that there is a correlation between the question quality and answer quality- good answers are more likely to be given in response to good questions. Similarly, bad answers appeared in response to bad questions. According to the definition in [14], high quality questions are expected to draw greater user attention, to have more answer attempts and to obtain the best answer within a short period of time. High-quality questions thus help to improve the CQA website's popularity as they, on the one side, contribute to efficient problem solving, and on the other side, enrich the community knowledge.

With the increase in popularity of CQA websites, not only the number of questions and the number of new members increased, but also the number of unanswered questions became high. For example, in Stack Overflow, a programming CQA forum, approximately 45 questions per month remained unanswered, according to statistics from 2012 [6]. By March 20, 2014, the number of unanswered questions was 752,533 out of 6,912,743 (approximately 10.9%). In Yahoo! Answers also approximately 15% of the incoming English questions remain unanswered [22]. Interestingly, the fact that those questions are not answered is not caused by users not having seen them. In fact, unanswered questions are seen 139 times on average [6]. A number of studies have attempted to predict whether a certain question will receive an answer or not and to determine the features that define question quality. In this survey, we will discuss the different measures used to evaluate question quality in CQA websites and will present the features of questions used in previous work that influence the question quality.

2. PROBLEM DESCRIPTION

The ability to recognize high quality questions is of high importance for both asker and answerer in a CQA website. In a good forum there should be a supply and demand balance – the aim of an asker is to receive a good answer to her question and the aim of a person providing an answer is to give a satisfying answer in order to increase her reputation. Therefore, it is essential for CQA forums to better understand what characterizes questions that are more appealing to be answered. A well-formulated question will increase the answerer's willingness to answer and will help her to give an appropriate answer which, at the same time, will increase the asker's satisfaction. It can also increase the question score and, respectively, the user's reputation which would motivate her to increase her participation and privileges on the programming CQA website Stack Overflow[1]. Seen from the asker's side, not receiving an answer, or only a limited number of answers, is not only disappointing, but may also be of educational or professional disadvantage. What is more, even if a question finally receives an answer, the longer it takes to get an answer, the more likely that the answer quality will not be satisfactory [5].

Question quality is important not only for the personal use but also for question and answering platforms as a whole because appropriate question-answer pairs attract users and improve platform traffic [16]. The aim of the asker is to obtain the information she is searching for and the aim of the person providing an answer is to give a satisfying answer to increase her reputation. For example, in StackOverflow there are a lot of questions with differing numbers of answers with a large variation, and it is observed that low-quality questions receive low-quality answers [2]. Low-quality questions also affect the user's experience on question and answering website. Furthermore, high-quality questions improve the entire question and answering platform as these questions are more appealing to users to share their knowledge and in this way to improve the overall platform knowledge. Finally, high-quality questions improve question retrieval and question recommendation in question and answering websites [2].

Several studies have focused on the quality of *answers* on CQA websites [15, 12, 23]. More recently, there has been increasing interest in the quality of questions. A variety of metrics of question quality has been used, and accordingly, conclusions about which features influence question quality also differ. For example, in [27] the authors find that short and long questions are more likely to be answered, whereas other studies state that too short questions have a low probability of obtaining an answer [6]. Therefore, the aim of this survey is to provide an overview of 1) the different measures to predict question quality, and 2) the different features used in previous work that determine question quality.

3. QUESTION QUALITY

Different studies employ different definitions of question quality. In this section we describe the measures most commonly used in research on define question quality - the number of answers and the question score. Finally, additional measures for question quality used in some research will be summarized.

The number of answers is one of the clearest reflections of how interesting and useful a posted question is to the CQA community. In a domain-specific QA website, users share expertise knowledge, mostly in the form of question-answer pairs. These question-answer pairs are saved on the website and usually ranked by search engines which makes them retrievable and valuable for future information needs [5]. A question has long term value when it draws the users' attention long into the future after it was posted. Research has shown that the number of answers is the most significant feature to predict the long term value of a question together with its answers set [5]. The number of answer is direct feedback on the usefulness/quality of the question. If the users assess it as being off topic or for some other reason inappropriate for the question answering community, they will be less likely to provide an answer.

Other research focuses not on the number of answers but on whether a question received at least one answer [27]. Similarly, in [6] a qualitative study is conducted on unanswered questions. Although one might think that questions remain unanswered because they were for some reason not discovered, the authors found that unanswered questions were on average seen 139 times.

A second measure of the community response, indicating whether a question has been valuable for the question and answering community, is the question score. The question score consists of the sum of the upvote count and the (negative) downvote count. In forums like Stack Overflow, Quandora[2] and Askbot[3], a question can be rewarded by voting it up. Equivalently, if the question was of low quality or useless to the community, the asker can be punished by having it voted down or even deleting it from the website. In general, answered questions on Stack Overflow have higher scores compared to unanswered questions [21]. In [3] it has been shown that prior *up* and *down* votes correlate with the asker's current question score. Hence, the question score is an appropriate feature to measure how the community users assess the question quality.

As both the question score and the number of answers are considered quality determinants, one would expect that a question with a high score receives many answers. That would be the case when a question is found very interesting and valuable to the community and if there were enough experts to answer it. Also if a question was not appropriate for the CQA community, it may not receive an answer and get a lot downvotes. However, the question score and the number of answers may not necessarily correlate. A question may address a new development or topic that is very interesting to the community but at the same time also very difficult to answer as there may be not enough experts familiar with it. Such a question may receive no answers but a lot of upvotes. If however a question was too easy or posted previously it may receive answers, but may not be evaluated highly as it does not contribute to the question answering website.

A number of other measures of question quality have been used in the literature. In [13] a survey is conducted among professional software developers to identify code-related questions they find difficult to answer. In [10] the focus is also on the content quality of Stack Overflow, but instead of investigating the features characterizing well-formulated questions, they concentrate on the features that describe low-quality questions. In Stack Overflow, if questions are off topic or of poor quality, they can be deleted by Stack Overflow moderators, experienced users with high reputation, or by the user who posted the question. As a deletion of a question is direct feedback regarding its quality, the aim in [10] is to find out what defines a question that is considered bad

1 http://stackoverflow.com/

2 http://www.quandora.com/

3 https://askbot.com/

enough to be deleted as well as how long it takes to remove it. The statistics revealed that for most of the deleted questions it took a long period of time to receive the first delete vote - approximately eighty percent of questions after one month, and half of the questions after six months. They also find that eighty percent of deleted questions received one delete vote and 14 percent received 3 delete votes. If the question was deleted by the asker herself, it was removed much faster than by a moderator. The feature that has the biggest influence on the deletion decision is the question score and eighty percent of the deleted questions have a zero score [10].

4. FEATURES DETERMINING QUESTION QUALITY

The features which have influence on the question quality can be divided in two groups: question-related and asker-related attributes. The group of the question-related features is represented by the features *tags, terms, question title and question body length* and *the presence of an example, in the case of Stack Overflow - a code snippet*. Regarding asker-related features, the reputation of the user can be taken into consideration. Since a user would better understand the question quality at the moment of posting, we focus on features that relate to information that is available at the moment a question is posted, i.e. those features that contain information that only becomes available once it is already known whether and how many answers a question received, are left out. The reasoning behind this choice is that features which are not available at the moment of the posting cannot help the asker to improve his or her question [9, 10].

4.1.1 Question Related Features

4.1.1.1 Tags and Terms

In many QA forums, the asker can add tags to her question to indicate to which topic(s) the question is related. Intuitively, one would expect that some question topics will elicit more answers than others, just because more people might be working on a certain topic, i.e. there will be more potential answerers available. Although tags may potentially differentiate between the number of answers, the large number of unanswered questions cannot be explained by a lack of sufficient experts for certain topics [21]. The assigned tags are considered as representative topics and investigated tags used for unanswered questions but not for answered questions. The authors of this study found 274 unanswered topics linked to only 378 questions in total. The number of questions with these specific tags is very small compared to the total number of unanswered questions which would indicate that there is at least one expert for each tag/topic. However, as users mostly assign several tags to a question, covering very general to specific tags, the large number of unanswered questions cannot be explained by a lack of experts [21]. In [10] tags are also analyzed to investigate the topics of questions in Stack Overflow. They found that approximately ten percent of the tags found in deleted questions were not present in closed or regular questions. These questions, tagged for example as homework, job-hunting and polls, are beyond the interests of the programmer community. In both, [21] and [10], the authors assume that tags are representative of the actual question topics. According to [6], however, incorrect tagging is one of the characteristics of unanswered questions. In [18] an investigation is performed of the, among others, relationship between characteristics and the question type in Stack Overflow data. The question types are described based on two dimensions – the question topic and the main concern of the asker. The former is described by the technology or construct the user is asking about, and the latter dimension concerns the problem the asker wants to solve. In [18] the following question types are considered based on the problem of the asker: debug/corrective, need-to-know, how-to-do-it, seeking different-solution. The authors found that the answer attributes are likely to be determined only by the second question dimension, the main concern of the asker. In [10] the authors observed that a high percentage of author-deleted questions are marked as too localized and off topic, and that a high percentage of moderator-deleted questions are marked as subjective and not a real question. These results indicate that question topics, i.e. tags, may either be incorrect and/or may not be fully informative of the likelihood of receiving an answer, the number of answers, or question score.

A number of recent studies tried to infer question topics from the natural language used to formulate the questions. In [25] the contents of thousands of questions and answers on Stack Overflow are analyzed. The authors assume the number of the topics to be equal to five and use Latent Dirichlet Allocation (LDA) [78] to find latent topics. They manually label each of these topics to: user interface, stack trace, large code snippet, web document, miscellaneous. Table 1 shows the topics with the representative key words. The results if the study showed that the category of miscellaneous topics which consists of many different kinds of questions, hold the largest number of questions. The second largest category is the web document topic, followed by large code snippet, stack trace and user interface.

Similarly, in [27] question topics from natural language are inferred using supervised latent Dirichlet allocation (SLDA) for classification [7]. They focus on questions in Yahoo! Answers and set the number of topics to 50. They discovered that, unsurprisingly, the topic with the lowest probability of remaining unanswered is a seasonal topic (questions were crawled on April 1st and the most answered questions were about "April Fools"). The second and third most answered questions are related to pets and food, respectively.

Topic	Words
User Interface	view, image, button, etc.
Stack Trace	java, error, org, server, etc.
Large Code Snippet	code, string, new, object, class, etc.
Web Document	href, page, html, php, etc.

Table 1. Topics and related words (obtained from [25])

4.1.1.2 Length of the Question

In [27] the authors find that the top 10% shortest and the top 10% longest questions have the highest probability of obtaining an answer, while the medium length questions were less likely to be answered. They explain this phenomenon by noting that reading and answering a short question can be accomplished in a very short time. Long questions are mostly expertise-related and need more explanation. These questions attract more users with the same interest and are therefore more appealing to be answered. The authors of this study assume that medium length questions are less interesting and unnecessarily long which makes them less likely to receive an answer. In contrast, according to [6] too short questions are very likely to remain unanswered. Those questions may miss important information, be too vague or unclear. Also,

too time-consuming questions are not very attractive for answerers. In another research on the effect of question length on question quality it was found that for answered questions, the minimum length is 5 characters and the maximum length is 48,258 characters (M = 1,079, SD = 1,389); for unanswered questions, these numbers are 19 and 35,588, respectively (M = 1,300; SD = 1,845) [21]. Their results show that although unanswered questions have longer length, both answered and unanswered questions, have the same probability of receiving an answer. In the ranking list of importance for differentiating unanswered from answered questions the attribute "question length" gets the same place for both classes. Finally, in [10] the authors found that compared to closed questions, deleted questions had a slightly higher number of characters in the question body. Existing literature thus does not provide a consistent answer to the question of whether and to what extent question length influences question quality. Further, it is not clear whether users mainly look at the length of the question title or the question body in deciding whether to answer the question or not, since question length and question body length are never analyzed separately.

4.1.1.3 Presence of an Example

Providing the audience with an example may help the asker to clarify or specify the information need she is seeking for. However, multiple examples may result in an overload of information and overwhelm the potential answerers making them less willing to answer the question. In programming CQA platforms, users can provide a code example to clarify their information need. Although code snippets are very specific for programming CQA platforms, they can be perceived in a more general way and are therefore included in the current survey. In [24] a few hundred questions were manually analyzed and assigned to ten different categories. Their goal was to investigate which questions are answered well and which remain unanswered. Using qualitative and quantitative data from Stack Overflow, they distinguished ten question categories – how-to, environment, discrepancy, error, decision help, conceptual, review, non-functional, novice and noise. They found that review questions had a high answer ratio as they often contain a code snippet and may have more than one possible good answer. The presence of a code snippet will not only elicit more answers to review questions; also other types of questions may benefit from the extra information provided by a code fragment. According to [6], for example, program specific questions are very hard to answer if no code fragment or detailed explanation is included. In [10] the authors analyze deleted and closed questions. Question that are extremely off topic, have very poor quality or have no activity after a long period of time are deleted. A user can delete her own question when it has not received any answers or upvotes[4]. Closed questions also indicate low quality and are questions which are considered duplicate, subjective, off topic, too localized or not a real question [11]. Similarly to the research of [6], the authors in [11] found that deleted questions had a lower percentage of code blocks compared to closed questions. Interestingly, the presence of a code snippet may have adverse effects as well. A user may not receive an answer if the code is hard to follow or if other users are able to understand the code but cannot see the problem [6]. The possible adverse effects may explain why in [21] the presence of code ranks only ninth in terms of importance for differentiating between answered and unanswered questions.

4 http://stackoverflow.com/help/deleted-questions

4.1.2 Asker-related Features

4.1.2.1 User Reputation

Among asker-related features, asker's reputation has received the most attention. The users' reputation scores are built on their participation on the question answering website. It has been shown that the experts, i.e. users with high reputations, do not only provide an essential contribution to answering websites in general, but they also provide the most helpful answers [26, 19, 17]. For example, on Stack Overflow, reputation can be gained when a user's question or answer is upvoted, when an answer is accepted as the best answer or by receiving a bounty. A bounty can be assigned to a question when the question is not receiving answers and for that reason a user wants to draw attention to this question. After the bounty period ends, the user placing the bounty, can award it to the user providing the best answer[5]. To encourage users to submit and maintain high quality content on the website, Stack Overflow rewards upvotes on answers more than upvotes on questions. In addition, high reputation users have more privileges in site management and receive more bonuses than regular users. The most reputation points are scored when a user's answer is accepted as the best answer, when it is upvoted or when the answer has received a bounty. Also in Askubuntu, a user's reputation increases when other users vote up her questions, answers or edits[6]. Unity Answers gives users that post a good question or helpful answer 15 reputation points and removes reputation points when a user posts something that is off topic or incorrect[7]. The authors in [5] show that users build their reputation mainly by receiving upvotes for their answers and not by asking questions themselves. Nevertheless, when high-reputation users post a question, these questions are expected to be of higher quality, because high-reputation users have more experience in answering questions themselves, i.e. they would be better skilled in what topics are more popular among the community and how to formulate their questions in order to receive the required answer. New users would be, on the contrary, less experienced in what to ask and how to ask it. Indeed, in [21] it was shown that the asker's reputation is one of the most dominant attributes to distinguish between answered and unanswered questions. The minimum reputation score for answered questions is 1, the maximum score is 465,166 (M = 1,886; SD = 7,005). For unanswered questions the minimum score is 1, the maximum score is 223,117 (M = 579; SD = 2,586). This result is in line with [27] where the asker's reputation is defined by the number of resolved questions posted, the number of answered questions and the number of answers that were indicated as best answer. The authors of this study showed that users with a larger question and answering history were more likely to receive an answer than new users. Similarly, as found in [6], the number of unanswered questions decreases with an increase in the asker's reputation. This statement is supported by [20] who estimated a linear quality function-based model to predict high and low quality questions. They showed that the user's reputation provides insights in the quality of the question that they will post in the future, which again emphasizes the importance of the interaction between the users and the community for the prediction of the question quality.

5 http://stackoverflow.com/help/bounty

6 http://askubuntu.com/tour

7 http://answers.unity3d.com/page/faq.html

5. CONCLUSION

The increasing popularity and usage of Community Question and Answering platforms calls for an investigation into factors influencing content quality and ways to encourage improving the quality. The CQA content can be roughly divided in questions and their answers: in this survey we address the quality of the questions. The number of answers and the question score have been shown to be a good measure of question quality on CQA websites. The question features most frequently used in research on predicting question quality were tags and terms, length of the question, the presence of an example and user reputation. Certain content-related features of the questions have a negative influence on the question quality - questions containing incorrect tags, or that are too localized, subjective or off topic are considered of bad quality. Results on the influence of the question length are more mixed and further research should provide better insights in its importance for the number of answers and the question score. On the other hand, the presence of an example has a positive effect on the question score and the number of answers. Such insights about the effect of these features are useful when engineering forums so as to guide users to improve their questions and to make it more likely they obtain the information they are searching for.

6. REFERENCES

[1] Adamic, L. A., Zhang, J., Bakshy, E., & Ackerman, M. S. (2008). Knowledge sharing and yahoo answers: Everyone knows something. *Proceedings of the 17th international conference on the World Wide Web*

[2] Agichtein, E., Castillo, C., Donato, D., Gionis, A., & Mishne, G. (2008). Finding high-quality content in social media. *Proceedings of WSDM*.

[3] Ahn, J., Butler, B.S., Weng, C. & Webster, S. (2013). Learning to be a Better Q'er in Social Q&A Sites: Social Norms and Information Artifacts. *Proceedings of the Association for Information Science & Technology*, 1-10

[4] Anderson, C. (2006). *The long tail: Why the future of Business is selling less of more.* Hyperion

[5] Anderson, A., Huttenlocher, D. P., Kleinberg, J. M., & Leskovec, J. (2012). Discovering value from community activity on focused question answering sites: a case study of Stack Overflow. *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 850–858.

[6] Asaduzzaman, M., Mashiyat, A. S., Roy, C. K. & Schneider, K. (2013) A. Answering Questions about Unanswered Questions of Stack Overflow. *Proceedings of Mining Software Repositories*, 97–100.

[7] Blei, D., & McAuliffe, J. (2007). Supervised topic models. Proceedings of Neural Information Processing Systems, 121-128.

[8] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of Machine Learning Research, 3.

[9] Chang, D., Schiff, M., & Wu, W. (2013). Eliciting Answers on StackOverflow. Working Paper

[10] Correa, D., & Sureka, A. (2014). Chaff from the Wheat: Categorization and Modeling of Deleted Question on StackOverflow. Working Paper.

[11] Correa, D., & Sureka, A. (2013). Fit or unfit: Analysis and prediction of 'closed questions' on stack overflow. In: *Proceedings of the ACM Conference on Online Social Networks*, Boston, MA, USA.

[12] Jeon, J., Croft, W.B., Lee, J.H., & Park, S. (2006). A framework to predict the quality of answers with non-textual features. *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval,* 228–235. ACM Press.

[13] LaToza, T. D., & Myers, B. A.(2010). Hard-to-answer questions about code. *Proceedings of the 2nd Workshop on the Evaluation and Usability of Programming Languages and Tools*. 81-86.

[14] Li, B. , Jin, T., Lyu, M. R., King, I., & Mak, B. (2012) Analyzing and predicting question quality in community question answering services. *Proceedings of the international conference on the World Wide Web*, 775–782.

[15] Lou, J. , Lim, K., Fang, Y., & Peng, Z. (2011). Drivers of knowledge contribution quality and quantity in online question and answering communities. *Proceedings of the 15th Pacific Conference on Information Systems.*

[16] Mamykina, L., Manoim, B., Mittal, M., Hripcsak, G., & Hartmann., B. (2011). Design lessons from the fastest q&a site in the west. *Proceedings of the 2011 annual conference on Human factors in computing systems*, 2857–2866. ACM Press.

[17] Movshovitz-Attias, D., Movshovitz-Attias, Y, Steenkiste, P., & Faloutsos, Ch. (2013). Analysis of the reputation system and user contributions on a question answering website: Stackoverflow. *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 886–893.

[18] Nasehi, S. M., Sillito, J., Maurer, F., & Burns, C. (2012). What Makes a Good Code Example? A Study of Programming Q&A in StackOverflow. *Proceedings of the 28th IEEE International Conference on Software Maintenance*, 25-34.

[19] Pal, A., Harper, F. M., & Konstan, J. A. (2012). Exploring question selection bias to identify experts and potential experts in community question answering. *ACM Transactions on Information Systems,* 30.

[20] Ponzanelli, L., Mocci, A., Bacchelli, A., & Lanza, M. (2014). *Understanding and Classifying the Quality of Technical Forum Questions. Proceedings of QSIC 2014 (14th International Conference on Quality Software)*. IEEE CS Press

[21] Saha, R.K., Saha, A.K., & Perry, D.E. (2013). Toward Understanding the Causes of Unanswered Questions in Software Information Sites: A Case Study of Stack Overflow. *Proceedings of the 9th Joint Meeting of the European Software Engineering Conference and the ACM SIGSOFT Symposium on the Foundations of Software Engineering,* 663-666.

[22] Shtok, A., Dror, G., Maarelk, Y., & Szpektor, I. (2012).Learning from the past: answering new questions with past answers. *Proceedings of the International Conference on World Wide Web*

[23] Suryanto, M. A. , Lim, E. P., Sun, A., & Chiang, R. H. L. (2009). Quality-aware collaborative question answering: methods and evaluation. *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, 142-151.

[24] Treude, C., Barzilay, O., & Storey, M. A. (2011). How do programmers ask and answer questions on the web? *Proceedings of the 33rd ACM/IEEE International Conference on Software Engineering,* 2, 804-807.

[25] Wang, S., Lo, D., & Jiang, L. (2013). An empirical study on developer interaction in StackOverflow. *Proceedings of the 28th Annual ACM Symposium on Applied Computing.* 18-22.

[26] Welser, H. T., Gleave, E., Fisher, D., & Smith, M. (2007). Visualizing the signatures of social roles in online discussion groups. *Journal of Social Structure*, 8.

[27] Yang, L. Bao, Lin, Q., Wu, X., Han, D. , Su, Z., & Yu, Y. (2011). Analyzing and predicting not-answered questions in community-based question answering services. *Proceedings, Association for the Advancement of Artificial Intelligence*. 1273–1278.