A Machine Learning-Based Approach to Predicting Success of Questions on Social Question-Answering

Erik Choi
erikchoi@gmail.com

Vanessa Kitzie
vkitzie@gmail.com

Chirag Shah
chirags@rutgers.edu

School of Communication & Information (SC&I)
Rutgers, The State University of New Jersey

Abstract

While social question-answering (SQA) services are becoming increasingly popular, there is often an issue of unsatisfactory or missing information for a question posed by an information seeker. This study creates a model to predict question failure, or a question that does not receive an answer, within the social Q&A site Yahoo! Answers. To do so, observed shared characteristics of failed questions were translated into empirical features, both textual and non-textual in nature, and measured using machine extraction methods. A classifier was then trained using these features and tested on a data set of 400 questions—half of them successful, half not—to determine the accuracy of the classifier in identifying failed questions. The results show the substantial ability of the approach to correctly identify the likelihood of success or failure of a question, resulting in a promising tool to automatically identify ill-formed questions and/or questions that are likely to fail and make suggestions on how to revise them.

Keywords: social Q&A; fact-based questions; machine learning; question success prediction

Introduction

In the recent past, a substantial transformation has occurred regarding people's information seeking behaviors, especially within online environments. One behavioral pattern that has developed on account of this transformation is the use of web-based question-answering (Q&A) services along with, and often instead of, web search engines. A popular example is Yahoo! Answers, which has over 200 million users and over a billion questions asked, an average of 90,000 new questions per day (Harper, Moy, & Konstan, 2009). These Q&A services typically provide a web-based interface for asking and answering questions in a variety of categories. Questions can be posted and answered by almost anyone, and often there is little to no monitoring or control over users' activities or quality of content. Such crowd-based Q&A services are often referred to as social Q&A (SQA).[1] Unlike virtual reference (VR) services, which constitute expert based reference interviews conducted by trained librarians via an electronic medium, SQA sites offer very little or no opportunity of interactions between an asker and an answerer to frame the question appropriately. This may result in poor quality of answers or even receiving no answers for a question. For example, Shah et al. (2012) found that within a period of five months, 13,867 questions across the 25 Yahoo! Answers categories were still open to receive a best answer ranking from the original asker, which could be indicative of dissatisfaction with the answers provided,

[1] For a more comprehensive treatment of terminology and typology for online Q&A services, see Choi, Kitzie, & Shah (2012).

Acknowledgements: The work reported here is done under research project *Cyber Synergy: Seeking Sustainability through Collaboration between Virtual Reference and Social Q&A Sites*, funded by the Institute of Museum and Library Services (IMLS), Rutgers, The State University of New Jersey, and OCLC, Inc. (http://www.oclc.org/research/activities/synergy.html). We also acknowledge Xiao Qin for his assistance with prediction model developments and data analysis.

Choi, E., Kitzie, V., & Shah, C. (2013). A machine learning-based approach to predicting success of questions on social question-answering. *iConference 2013 Proceedings* (pp. 409-421). doi:10.9776/13224

and 4,638 (about 33%) of them did not receive any answers. Since people specify an information need in natural language to others within an SQA site, it is important to investigate how the information need was structured and/or expressed to understand how others interpreted what the original asker intended to look for as compared to the true information-seeking goal. Predicting the likelihood a question failing by determining whether it contains any overarching features of past questions that have failed will help an asker to reconstruct his/her question and increase its potential for success, promoting more effective information seeking behaviors within the SQA context.

The goal of the work is to investigate what makes a question in SQA likely to succeed, defined here as a question that receives at least one answer, or to fail, defined here as a question that does not receive an answer. By looking at questions that fail, examining their shared characteristics and using a quantitative approach to determine the empirical influence these variables might have on question failure, the authors hope to provide a more concrete and robust way to not only identify questions that are likely to fail, but also to provide suggestions and other means for which to increase the propensity for success. In order to accomplish this, an examination of existing works focusing on content-based studies within SQA will be provided in the next section, followed by a method for extracting various features from SQA questions collected from Yahoo! Answers and a technique to build a model that predicts if a question is likely to succeed or not. The model will then be tested for robustness and accuracy, with results being discussed in terms of implications for improvement of SQA services.

Background

Within the past few years, various types of social Q&A (SQA) services have been introduced to the public and researchers have begun to evidence interest in information seeking behaviors within these contexts. People ask questions to the community and expect to receive answers from anyone who knows something related to the questions, allowing everyone to benefit from the collective wisdom of many. These services often supplant search engine use, allowing askers to pose a question in natural language rather than submitting a few keywords to a search engine and to receive personalized answers from other people, as opposed to a list of results. Due to the intrinsic humanistic aspect of the site interactions, SQA outlets pose a benefit to those who may not be finding satisfactory search results using a search engine result page (SERP), and also offer specific social benefits such as the opportunity to solicit and provide opinion and advice-based information, as well as the ability to foster social expression by encouraging users to participate in various support activities, including commenting on questions and answers, rating the quality of answers, and voting on the best answers.

Adamic et al. (2008) found that knowledge resources within SQA comprise a broad range of topics, however are not very deep since many questions asked solicit opinion and advice, while a very small proportion seek fact based knowledge. This observation has been continually made, most recently by Shah et al. (2012), which observed a minor amount (around 5%) of information seeking questions versus advice, opinion or social expression based ones. Further, Agichtein et al. (2008) found that as many SQA sites continue to grow, overall performance in answering fact based questions using traditional relevance measures wanes. This suggests that further studies, such as the one reported here, prove valuable to the field by improving performance on a previously identified weaker facet of the SQA environment and could potentially impact both the types of questions posed in the future, as well as overall community participation and use.

Research on SQA can be divided into two distinct areas of study - user-based and content-based (Shah, Oh, & Oh, 2009). The former examines the factors that comprise interactions within Q&A communities. Shachaf (2010) suggested that while these communities may differ in scope and means of operation, they all operate under the pretense that interaction within an SQA model is multi-dimensional and collaborative, hinging on assessment, motivation, identity formation, and communicative norms unique to this platform. Gazan (2007) performed a content analysis using Yahoo! Answers, dividing askers into seekers and sloths, and concluding that the more active seekers group received a larger proportion of responses than the sloth counterpart. Oh (2012) studied answerer motivations within Health Q&A sites, finding that altruism was the leading factor in answerer participation.

Content-based studies attempt to characterize the components of the actual questions and answers posted to the site. Shah and Pomerantz (2010) identified several textual criteria that comprise a good answer using human evaluators to rank a question on each criteria, while those in the information retrieval (IR) community use machine extraction methods of textual and non-textual features to predict answer quality (e.g., Text REtrieval Conference (TREC),[2] held annually). One of the overarching conclusions from these studies was that relevance, answer length, presence of outside sources, and time it took to deliver an answer all constitute significant factors in predicting a best answer.

To the best of the authors' knowledge, similar criteria to evaluate the quality of questions asked within an SQA environment have not yet been developed. Instead, most research focusing on questions within this context attempts to classify all questions based on type (e.g. information seeking, advice seeking, opinion seeking, etc.) in order to examine which questions have the best archival value (Harper, et al., 2009). Harper et al. (2009) also distinguish *informational* questions and *conversational* questions in order to investigate the level of archival value by exploring the use of machine learning techniques to automatically classify questions. The authors argue that informational questions seeking factual knowledge or objective data in which there exists a "right" answer, are more likely to solicit information that the asker may learn or use, whereas conversational questions, which do not have a "right" or "wrong" answer, stimulate discussion to obtain other people's opinions or to perform acts of self-expression. Kim, Oh, and Oh (2007) have investigated criteria that questioners may employ in selecting the best answer to their given question. They also studied how types of questions that users ask correlate to these criteria using a data corpus from Yahoo! Answers and found that affective characteristics, such as answerer politeness, tend to matter more for conversational questions, while traditional relevance theory-based characteristics, such as quality and topicality apply more to informational questions (Kim, Oh, and Oh 2007). Their study of 465 queries found opinion seeking questions (39%) to be most frequent, followed by information seeking questions (35%), and suggestion seeking questions (23%). This finding indicates that conversational questions seeking opinions or suggestions are generated more than informational questions within Yahoo! Answers.

Further studies have touched on how examining question types might improve question dissemination among services, predominately within the realm of virtual referencing (VR) (Duff & Johnson, 2001; Pomerantz, 2005; Arnold & Kaske, 2005), however these studies do not directly address specific practical applications for services yielded from the development of such typologies. A typology for classification of failed fact-based questions was reported in Shah et al. (2012) and summarized in Table 1. The authors defined failed questions as those that did not receive a response after three months, a time period by which most SQA community members reported "giving up," in seeking an answer from the original posted thread. A randomized set of 200 information-seeking questions, defined as questions soliciting a fact-based response, constituted the data corpus.

Findings from the study (Shah et al., 2012) indicate that main characteristics for the 200 failed questions were spread across the categories with significant concentrations in the too complex, overly broad sub-category (68, 34%), followed by lack of information (28, 14%), relatedness (26, 13%), and ambiguity (21, 10.5%) while socially awkward (8, 4%), excessive information (4, 2%), and poor syntax (2, 1%) exhibited a less likely primary influence on failure. Based on these findings, it appears that questions falling within the broader categories of unclear, complex, and multiple questions represent a higher proportion of those that fail in comparison to inappropriate ones, which intuitively suggests that features measuring this latter characteristic may make less of a contribution to the accuracy of the classifier developed within this study.

Prediction Model Using Automatically Extracted fFeatures

Although a large number of content based studies within SQA focus on answer quality, as identified by the previous section, there exists a lack of studies examining its counterpart - question quality. Shah et al. (2012) began to address this area by developing a set of characteristics to describe what types of questions fail within an information-seeking context. The current study extends this research avenue by translating these attributes of question failure into empirical features used to develop a prediction model for question failure. In this section, the authors describe a set of experiments that approximate these empirical translations, construct a classifier trained on these features, and test the predictive accuracy of the subsequent model.

[2] http://trec.nist.gov/

Table 1
Typology for failed informational questions developed by Shah et al. (2012)

Category	Definition
1. Unclear	
Ambiguity	Question is too vague or too broad, and for this reason, is misunderstood or causes multiple interpretations.
Lack of information	Not enough information exists to identify the asker's intended information-seeking goal.
Poor syntax	Question syntax is ill formed, has typos, or has Internet slang that hampers understanding.
2. Complex	
Too complex and/or overly broad	Question is too complicated and a few people have the ability and/or the resources necessary to provide answers, even though enough details are provided to identify the asker's intended information-seeking goal.
Excessive information	Question contains an excessive amount of information that may lose people's attention to (or interest in) answering it.
3. Inappropriate	
Socially awkward	Question is inappropriate, too personal, or socially taboo.
Prank	Question is posed as a joke or to get attention.
Sloths	Question is homework related and often reflects a perceived "laziness" of the askers to obtain an answer themselves or to actively participate in the SQA community outside of posting questions.
4. Multiple Questions	
Relatedness	Title and/or content poses more than one question (although they are related), so the answerers may be confused in interpreting the asker's intended information-seeking goal.
Un-relatedness	There is more than one question posed and subsequent questions are unrelated, causing potential respondents to be confused in interpreting the asker's intended information-seeking goal.

Data

A total of 400 questions posed in Yahoo! Answers were used to develop a classifier for this study. This study investigated two sets of questions from Yahoo! Answers - 200 failed, information-seeking questions used in the previous study by Shah et al. (2012), as well as 200 resolved information-seeking questions. Questions defined as resolved were ones in which the asker of a given question selected any answer provided as the best answer that satisfied his/her information need. Both question sets were selected across the 25 Yahoo! Answers categories and collected via the Yahoo! Search Application Programming Interface (API)[3].

Extracting Question Features

The current study assumes that the main characteristics of question failure have been identified by the previous study (Shah et al., 2012) and provide several necessary measures that can be translated empirically to construct a model that identifies failed questions. A set of features was selected for extraction in order to address each of the characteristics of question failure developed by the typology, as empirical translations of hypothesized critical variables that influence a question's likelihood for failure within Yahoo! Answers. Derived from standard data mining approaches, the resulting features identified

[3] http://developer.yahoo.com/answers/

best represent the original characteristics developed within the typology, and will now be further discussed.

Clarity score (ClarityScore). To quantify the clarity of a question, we decided to employ a query clarity measure often used within the IR domain (Cronen-Townsend, et al., 2002). This measure computes the relative entropy between the query/question language model and the corresponding collection language model. We used the LA Times collection available from TREC with 131,896 documents containing 66,373,380 terms. The clarity score was computed using the Lemur[4] toolkit. This toolkit has been previously used for measuring clarity (see Belkin et al., 2004; Diaz & Jones, 2004; Qiu et al., 2007), including evaluating high accuracy retrieval (Shah & Croft, 2004).

Syntax (TypoNumber). Edit distance (Levenshtein, 1966), which compares the common distance between words to the measured distance of the data corpus, as well as spelling, were measured to determine the syntactical appropriateness, and implied resultant clarity, of a question. Misspellings were detected by Jazzy[5], a Java-based spell checker built on the Aspell algorithm.

Readability (FleschKincaidReadingEase). Flesch-Kincaid Readability scores (Kincaid, 1975) were calculated for each question with the hypothesis that a question with an implied higher cognitive load would attract less potential answers, since less community members would be able to understand the information need of the asker. This measure was used to determine complex, ambiguous questions.

Inverse Document Frequency (iDFCharLength). Inverse document frequency (IDF) measures were used to determine questions that might be too broad. The authors hypothesized that the more novel terms within the data corpus in relationship to the amount of words contained in a question, the more direct the question was in stating the asker's information need, and thus, the increased likelihood that the question would be resolved.

Presence of taboo words (TabooNumber). Questions were identified as inappropriate by using a dictionary of "taboo" words and assessing whether an identified question within the corpus had any of these defined words. While this measure identifies the theoretical sub-characteristic of taboo and/or socially awkward questions, it does not measure questions that might seek homework help. Therefore future work might look to include a measure that determines whether or not a question directly solicits homework help, perhaps by flagging key words and phrases from questions defined as such. However, this would take time to identify and build a corpus of questions, and as to the best of the authors' knowledge this corpus is currently nonexistent, so it was not included as a feature for this study.

Punctuation (QuestionMarkCount). We identified multiple questions posed as a single information need in a question by counting the presence of a question mark at the end of each sentence within a question posed to Yahoo! Answers, containing a title and/or content. To not misidentify a single question that might have been punctuated with more than one question mark at the end of a sentence in order to emphasize an information need, the technique used only counted one distinct question mark at the end of a word. In order to not confound variables due to the exploratory nature of this study, related versus unrelated content were combined into one categorization.

Question length (CharLength) (WordCount) (Sentence Count). Question length constituted a measure of complexity, in which a longer question was hypothesized to correlate positively with question failure since the longer the question, the more cognitive effort needed to process the information need. In addition, a short question might indicate a lack of information provided, which might in turn make it also unclear. The authors measured question length by the number of characters used, the amount of words in the question, and the number of sentences in the content section (if applicable).

Content (Content). When posing a question in Yahoo! Answers, there are two fields - question title, where the actual question is posted, and content, where the asker has the opportunity to describe his/her information need further. A question title is required to pose a question, whereas the question content section is optional allows an asker to supply additional information to provide readers with a better understanding of the information need. As the authors hypothesized that presence of content material could be useful in supplying additional contextual information to certain questions, the significance of whether or not a question has content was measured to determine if a relationship existed between presence of such information and whether or not the related question was likely to fail.

[4] http://www.lemurproject.org

[5] http://jazzy.sourceforge.net/

Additional Features

Additional textual measures utilized in other works identifying features of questions and/or answers within SQA that affected either question and/or answer performance, were also included to build a more representative model.

Interrogative words (StartWith). It is hypothesized that question type might influence likelihood of failure. For example, perhaps informational questions experience more failure than conversational ones. Harper et al. (2009) identified a series of interrogative words (i.e. “who,” “what,” “where,” “when,” “why,” and “how”) that might differ in proportion among informational, or fact-finding questions, versus more conversational ones. The authors found that words such as “where” and “how” were used more frequently in informational questions and “why” in conversational questions. Extending this observation, certain interrogative words might also play a role in contributing to characteristics of failed questions. For this reason, the presence of common interrogative words was used as an additional variable.

Number of external links (URLCount). Gazan (2006) divided answerers within the SQA site, Answerbag into two types, *specialists* who provided answers based on self-identified expertise and therefore did not provide references, and *synthesists* who provided external sources. He found that members of the Answerbag community rated answers provided by *synthesists*, containing external sources, higher than those provided by *specialists*. Based on this observation, the authors decided include number of external links to determine whether or not this factor influenced a question’s likelihood to fail.

The overall results of numeric features (e.g., clarity score, syntax, readability score, inverse document frequency, and presence of taboo words, punctuation, question length, and number of external links) found from failed questions are described in Table 3; the results indicate that inverse document frequency and clarity score are the most significant features of question failure. Additionally, Table 4 illustrates other nominal features (e.g., presence of content for additional information, interrogative words) of failed questions. The results show that more than half of failed questions contain an interrogative word, “what”, followed by “how” and “is”.

Table 3
Summary of numeric features used from questions.

Feature	Min Value	Max Value	Mean	Std. Deviation
TabooNumber	0	1	0.023	0.148
TypoNumber	0	49	3.268	3.938
QuestionMarkCount	0	5	1.408	0.814
URLCount	0	2	0.035	0.209
CharLength	19	2840	168.953	205.274
iDFCharLength	0.204	1.224	0.533	0.134
ClarityScore	7.284	17.621	11.25	1.307
WordCount	4	482	31.118	36.936
SentenceCount	1	81	2.91	4.461
FleschKincaidReadingEase	-18.2	118.2	76.13	21.087

Table 4
Summary of nominal features used from questions.

Feature	N
Content	
Yes	231
No	169
StartWith	
what	240
where	8
when	8
which	0
who	0
why	1
how	35
is	14
are	6
do	8
does	3
other	77

Constructing a model with SVM

Methods

Features were extracted from each question using a variety of tools, including Lemur for question clarity; LingPipe[6], a java natural language processing (NLP) tool (e.g., tokenization, stopwords removal, etc.); and Jazzy, a java-based spell checker. After the features were extracted, a Support Vector Machine classifier was built with split-sample validation and cross-validation using Weka[7]. For further features evaluation, χ^2 feature selection method and correlation feature selection method were applied on the entire dataset to weight features and later to reduce the feature vector. More specifically, 66% of the data was used for model training with split-sample validation and a 10-fold cross-validation was performed for robustness evaluation. K-fold cross validation was one way to improve over the split-sample method. The data set was divided into k subsets, and the split-sample method was repeated k times.

Results

Table 5 summarizes the various outcome measures. The measures performed relatively the same, with the highest percentage of accuracy at 77.94% for SVM. Although these percentages are not indicative of a strong model, they represent enough of a difference from the chance levels as can be illustrated by the Kappa statistic, which measures the agreement of predictions with the actual class. The first value of Relative Information Score and Information Score (Kononenko and Bratko, 1991) corresponds to the cumulative information score and the second one corresponds to previous value divided by the number of instances.

Six of the twelve features make contributions to the model with the highest percentage of accuracy at 76.50%; the largest contribution by far made by the feature *StartWith*, which represents interrogative words as shown in Table 6. The other two features that make significant contributions to the model are *IDFCarLength*, which represents the number of unique words in the question, and *ClarityScore*, which represents the complexity of the question (see Table 7). It is also interesting to note

[6] http://alias-i.com/lingpipe/
[7] http://www.cs.waikato.ac.nz/ml/weka/

that questions identified as inappropriate questions, measured here by presence of taboo words, were not found to be prevalent by Shah et al. (2012), yet had a fairly large effect on the performance of the model within this study. This might be due to the methodology employed by the authors (Shah et al., 2012), which coded the corpus using the developed typology on the perceived main characteristic viewed to have the most significant effect on question failure, while potential secondary features were not included.

Table 5
Result of classification on test split.

Correctly Classified Instances	106	77.94%
Incorrectly Classified Instances	30	22.06%
Kappa statistic	0.558	
K&B Relative Info Score	7602.7895	
K&B Information Score	76.0248 bits	0.53 bits/instance
Class complexity | order 0	136.0272 bits	1 bits/instance
Class complexity | scheme	32220 bits	252.39 bits/instance
Complexity improvement	-32083.9728 bits	-251.39 bits/instance
Mean absolute error	0.2206	
Root mean squared error	0.4697	
Relative absolute error	44.1128%	
Root relative squared error	93.9206 %	
Total Number of Instances	*136*	

Table 6
Result of classification on stratified cross-validation.

Correctly Classified Instances	306	76.50%
Incorrectly Classified Instances	94	23.50%
Kappa statistic	0.53	
K&B Relative Info Score	21200	
K&B Information Score	212 bits	0.53 bits/instance
Class complexity | order 0	400 bits	1 bits/instance
Class complexity | scheme	100956 bits	252.39 bits/instance
Complexity improvement	-100556 bits	-251.39 bits/instance
Mean absolute error	0.235	
Root mean squared error	0.4848	
Relative absolute error	47.00%	
Root relative squared error	96.95%	
Total Number of Instances	*400*	

Table 7
Ranked attributes by Chi-squared Ranking Filter.

StartWith	122.429
iDFCharLength	68.428
ClarityScore	20.253
TabooNumber	9.207
QuestionMarkCount	8.591
Content	2.305
TypoNumber	0
FleschKincaidReadingEase	0
WordCount	0
URLCount	0
CharLength	0
SentenceCount	0

Interrogative words might play a role in determining whether a question fails or is resolved by indicating question type, as hypothesized above. In addition, interrogative words could almost representing a clarity measure in the sense that when one of these common interrogative words (e.g. “what”) is used at the beginning of a question, it immediately indicates to the reader something about the nature of what the asker is looking for and how to frame an answer. For example, the word “what” might indicate the asker is searching for a noun (e.g. What is the capital of France?), whereas the word “how” might indicate that the asker is searching for an opinion and/or directions (e.g. How do you assemble a computer from scratch?). Number of unique words in a question also represents a measure of clarity and also questions identified as too broad, since presence of novel words indicates a question that is more specific and therefore has a clearer identified information need. Finally the clarity score indicates a direct measure of the sub-characteristic clarity. It is interesting to note that these three top features all fit under the “Unclear” major characteristic developed within Shah et al.’s (2012) typology. A secondary feature as indicated by the IDF measure could also be “Too Complex,” although the other feature measuring this characteristic, Reading Level, did not make a significant contribution to the model, suggesting this measure might be representative of clarity. All of the classification results reported in the present section are summarized in Table 8. Classification accuracies of both the model constructed with all question features and the one with selected features are the same. This result shows that six of the twelve features – TypoNumber, FleschKincaidReadingEase, WordCount, URLCount, CharLength, and SentenceCount have no significant attributes to predict the likelihood of failure for fact-based questions in SQA.

Table 8
Summary of various classification models

Model	Training	Testing	Accuracy
SVM	400 samples	Split-sample	77.94%
SVM	400 samples	Cross-validation	76.50%
SVM with selected features	400 samples	Cross-validation	76.50%

Discussion

Limitations

One main limitation of this study lay with the imperfect translation from a theoretical model, exemplified via the characteristics within the typology for failed questions, and the empirical model, or translations of these characteristics into factors that could be measured using text extraction. For example, how can a machine provide a representation of a complex question, given the nuances embodying the concept? The best researchers can hope to do is break apart the characteristics by key facets and identify the appropriate methods and tools by which to define a corresponding feature. For this reason, future study might focus on using human coders to classify a set of failed questions, using the typology definitions. It should be then be determined whether there exists a significant difference between how questions were classified by humans versus machine classification.

This study is also limited in its generalizability, both because it only samples from one SQA community, albeit a popular one, and since it only samples a small subset of the types of questions that comprise the corpus. However, it can be argued that information-seeking questions have a greater likelihood of being addressed since they actually have an answer. Future study could focus on why questions soliciting more open ended answers may fail, however unless failure is due to question features rather than their actual content, there would be no overarching classifier that could be built to identify the propensity for these types of questions to fail in the first place.

Although the purpose of this study was to determine whether certain observed features could be used to classify failed questions, it did not consider user attributes. Previous work on SQA answer quality has indicated that user attributes make a significant contribution to predicting whether an answer receives a best answer rating or not, perhaps presence of these attributes could also contribute to whether or a question fails (Bian et al., 2008). Since initial findings indicate that the classifier is weak, adding user attributes as a variable in future studies has the potential to improve the performance of the classifier.

Further study should be done in order to improve on the accuracy of the model, as the current accuracy with ten-fold validation is 76.50%. While the classifier developed here included only textual features identified in the previous study (Shah et al., 2012), other research includes such non-textual features of a question and/or an asker to investigate how those features might be related to earning a response. For example, Shah and Pomerantz (2010) included, in order to evaluate and predict answer quality, information from the answerer's profile, and reciprocal rank of the answer in the list of answers for the given question, and Teevan, Morris, and Panovich (2011) included both properties of the asker including social network use, social network makeup, profile picture, and time of day that the question was posed for investigating factors affecting response quantity, quality, and speed. These works point out that it would be possible for some non-textual features of a question and/or an asker to affect the likelihood of failure for fact-based questions, and therefore a more comprehensive prediction model with both textual and non-textual features may outperform the current prediction model with regard to the likelihood of fact-based questions get resolved in SQA.

Implications

Within an online Q&A platform, machine extraction could be used to measure the degree of existence for variables, as identified in the previous section, which were found to influence variability in a question's likelihood to fail. Based on the measured identification of these variables, the machine could then employ a pre-identified approach to assist the asker in increasing the likelihood for success. Such approaches could include referring the asker to a different SQA outlet, in which the question has a better chance of getting answered; referring the asker to a VR site where a reference interview can be performed to better elicit and conceptualize an information need; employing an iterative feedback system; and employing an automated measure such as query expansion or syntax correction.

Referring to a different SQA outlet. Different types of SQA outlets exist. Shah, Choi & Kitzie (2012) developed a typology of these outlets into four types: community based, where people exchange information within an online community (e.g. Yahoo! Answers); collaborative, where users can edit the question and/or answer over time to improve it (e.g. WikiAnswers); expert-based, where users receive answers from experts within a specific topic area (e.g. Google Answers); and social Q&A, where people exchange information using their own personal social networks (e.g. Facebook Questions). Presumably each outlet has different strengths and weaknesses in dealing with certain criteria that might contribute to a failed answer. Future study could look at whether the presence of a certain factor predominately factoring in to a question's likelihood to fail within the context of one type of SQA site could be migrated to another site where this same factor presents less of a likelihood for the question posed to fail.

Referring the asker to a VR site. Expert-based SQA services would provide a viable option here. Aside from addressing questions too complex, VR services could also address questions that may not contain a fully articulated information need since professionally trained librarians could conduct a reference interview with the asker to assist him in fully articulating it (Taylor, 1968). Such questions probably have not received an answer since it might require someone with expertise in the field to properly understand and address the stated information need.

Employing an iterative feedback system. Iterative feedback systems, such as query reformulation in interactive information retrieval (Belkin et al., 2001), have been shown to assist users in articulating their information need depending on what information they initially deem relevant, and how this information is processed by the system to provide better results. Within an SQA context, perhaps archived answers of similar questions could be shown to the user, and the user could pick the answer that is most relevant to his question. If the answer chosen still does not satisfy the query, one of the other suggested methods could then be applied.

Employing an automated measure. This would be useful for questions that might lack information necessary to provide a good answer. The system could be trained using simple measures, such as the IDF measure used within this study, to identify questions that might be lacking information and suggest simple techniques such as query expansion or using a thesaurus to suggest more unique terms that might better convey the question to others. Another way to address the problem within the SQA platform is to monitor for syntax and spelling and make suggestions; much like in all commonly used word processing documents, which could go a long way in improving the overall clarity of the question.

Conclusion

Predicting the likelihood of failure for fact-based questions in SQA offers a way to assist information seekers in constructing question with an increased likelihood of being answered. The previous study by Shah et al. (2012) attempted to identify several characteristics of why fact-based questions fail in Yahoo! Answers and revealed that the characteristics unclear, complex, inappropriate, and multiple questions are major attributes of fact-based questions that failed within Yahoo! Answers. However, helping an asker revise his/her question might be the first step in making information seeking behaviors more effective in SQA. To do so, it is important to first identify attributes suspected to affect question failure and suggest solutions to improve ways of constructing a question.

Based on a typology for classification of failed fact-based questions (Shah et al., 2012), the study extracted a variety of textual features in order to build a prediction model for determining the likelihood for question resolution. The study found that a question starting with “what”, the number of unique words in the question as measured by Inverse Document Frequency, the level of clarity, presence of taboo words, punctuation, and significance of whether or not a question has content for additional information, are the most significant features for prediction. These findings shed light on the ways in which an asker poses a question and suggests various applications (i.e., incorporating relevance feedback, enabling question routing) that could address how to revise the question in order to increase its likelihood of being answered. Since SQA enables people to seek and share information to fill the knowledge gaps that might not be addressed by other services, providing an appropriate specification and/or structuration of an information need in natural language constitutes a fundamental step toward conceptualizing an effective method for seeking and sharing information. Identifying attributes for why some questions fail and predicting the likelihood of having a question answered presented in the current study will play a significant role in clarifying and revising an asker’s question for a better question-answering process in SQA, and future studies will also benefit from these findings in order to evaluate the quality of questions asked within an SQA environment.

References

Adamic, L. A., Zhang, J., Bakshy, E., & Ackermen, M. S. (2008). Knowledge Sharing and Yahoo Answers: Everyone Know Something. *In Proceedings of WWW Conference,* 665-674.

Agichtein, E., Castillo, C., Donato, D., Gionis, A., & Mishne. G. (2008). Finding High-Quality Content in Social Media. In *WSDM ’08*, 183–194.

Arnold, J., & Kaske, N. (2005). Evaluating the quality of a chat service. *Portal: Libraries and the Academy*, *5(2)*, 177 - 193.

Belkin, N.J., Chaleva, I., Cole, M., Li, Y.-L., Liu, Y.-H., Muresan, G., Smith, C.L., Sun, Y., Yuan, X.-J., & hang, X.-M. (2004). Rutgers' HARD Track Experiences at TREC 2004. In *Proceedings of TREC 2004*.

Belkin, N., Cool, C., Kelly, D., Lin, S-J., Park, S. Y., Perez-Carballo, J. and Sikora, C. (2001). Iterative exploration, design and evaluation support for query reformulation in interactive information retrieval. *Information Processing and Management, 37(3)*, 403-434.

Bian, J., Liu,Y., Agichtein, E., & Zha, H. (2008). Finding the Right Facts in the Crowd: Factoid Question Answering over Social Media. In *Proceedings of WWW Conference*, 467–476.

Choi, E., Kitzie, V., & Shah, C (2012, October). *Developing a Typology of Online Q&A Models and Recommending the Right Model for Each Question Type*. Poster Proceedings of American Society of Information Science & Technology (ASIST) Annual Meeting. October 26-30 Baltimore, Maryland.

Cronen-Townsend, S., Zhou, Y., and Croft, W. B. (2002). Predicting query performance. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval* (SIGIR '02). ACM, New York, NY, USA, 299-306. DOI=10.1145/564376.564429 http://doi.acm.org/10.1145/564376.564429

Diaz, F. & Jones, R. (2004). Using temporal profiles of queries for precision prediction. In SIGIR '04: Proceedings of the 27th annual international conference on Research and development in information retrieval, 18–24.

Duff, W. M., Johnson, C. A. (2001). A virtual expression of need: An analysis of e-mail reference questions. *The American Archivist*, *64(1)*, 43 - 60.

Harper, F. M., Moy, D., & Konstan, J. A. (2009). Facts or friends? Distinguishing informational and conversational in social Q&A. In *Proceedings of ACM CHI Conference*, 759-768.

Hitwise. (2008). U.S Visits to question and answer websites increased 118 percent year-over-year. Retrieved from http://www.hitwise.com/news/us200803.html

Gazan, R. (2007). Seekers, sloths and social reference: Homework questions submitted to a question-answering community. *New Review of Hypermedia and Multimedia*, *13*(2), 239-248. doi:10.1080/13614560701711917.

Kim, S., Oh, J. S., & Oh, S. (2007). Best-Answer Selection Criteria in a Social Q&A site from the User-Oriented Relevance Perspective. In *Proceedings of the ASIST Conference,* 1-15.

Kincaid, J. P. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel, in National Technical Information Service, 8–75.

Kononenko, I., Bratko, I. (1991) Information based evaluation criterion for classifier's performance. *Machine Learning, 6*, 67-80.

Kwok, C., Etzioni, O., & Weld. D. S. (2001). Scaling question answering to the web. *ACM Transactions on Information Systems (TOIS), 19*(3), 242 – 262.

Levenshtein, I. V. (1966). Binary Codes capable of correcting deletions, insertions, and reversals. *Cybernetics and Control Theory*, 10(8), 707–710.

Morris. M.R., Teevan, J., and Panovich, K. (2010). What do people ask their social networks, and why? A survey study of status message Q&A behavior. In *Proceedings of the ACM CHI 2010 Conference*, 1739-1748.

Qiu, G., Liu, K., Bu, J., Chen, C., & Kang, Z. (2007). Quantify query ambiguity using ODP metadata. In SIGIR '07, 697–698.

Oh, S. (2012). The characteristics and motivations of health answerers for sharing information, knowledge, and experiences in online environments. *Journal of the American Society for Information Science and Technology, 63*(3)*,* 543- 557.

Pomerantz, J. (2005). A conceptual framework and open research questions for chat-based reference service. *Journal of the American Society for Information Science and Technology*, *56(12)*, 1288 - 1302.

Shah, C., & Croft, W. B. (2004). Evaluating high accuracy retrieval techniques. In *Proceedings of ACM SIGIR 2004*, 2–9.

Shachaf, P. (2010). Social reference: A unifying theory. *Library & Information Science Research, 32(1),* 66–76.

Shah, C., Oh, S., & Oh, J. S. (2009). Research agenda for social Q&A. *Library & Information* Science Research, 31(4), 205-209.

Shah, C., Oh, S., & Oh, J. S. (2009). Research agenda for social Q&A. *Library & Information* Science Research, 31(4), 205-209.

Shah, C., & Pomerantz, J. (2010). Evaluating and predicting answer quality in community QA. *SIGIR*, 411–418.

Shah, C., Radford, M., Connaway, L. S., Choi, E., & Kitzie, V. (2012). *"How Much Change Do You Get from 40$?" – Analyzing and Addressing Failed Questions on Social Q&A.* Paper to be presented to at the annual convention of American Society of Information Science & Technology (ASIST). Baltimore, Maryland.

Taylor, R. S. (1968). Question negotiation and information seeking in libraries. *College and Research Libraries, 29*, 178-194.

Teevan, J., Morris, M.R., Panovich, K. (2011). Factors Affecting Response Quantity, Quality, and Speed for Questions Asked Via Social Network Status Messages. In *Proceedings of ICWSM*. 2011.