From Bad to Good: An Investigation of Question Quality and Transformation

Vanessa Kitzie, Erik Choi, Chirag Shah
School of Communication & Information (SC&I), Rutgers, The State University of New Jersey
4 Huntington St, New Brunswick, NJ 08901
vkitzie@gmail.com, erikchoi@gmail.com, chirags@rutgers.edu

ABSTRACT

Social question answering (SQA) services are a popular way for people to exchange information. Unfortunately, the quality of information exchanged can be variable and few studies focus on the quality of questions asked. To address this, we explored the influence of textual features on question quality based on 126 questions taken from five different categories of Yahoo! Answers labeled as “Bad” by human assessors and then revised to be “Good” by them. Findings indicate significant differences between the means of each feature before and after revision, suggesting the potential for an automated system that could flag questions of poor quality. In addition, by exploring the relationship between features contributing to good quality questions, we suggest a simple set of strategies askers can take when writing a question in order to improve its chances of receiving a satisfactory answer.

Keywords
Social question answering; question quality; feature extraction

INTRODUCTION

Social question answering (SQA) services provide an outlet for information exchange. These types of services are community-based; members can ask a question and receive personalized answers tailored to their specific information needs from other members (Shah, Oh, & Oh, 2009). In addition, members interact with one another by providing feedback to preexisting content by voting, assigning points, etc. The relative popularity of SQA services can be attributed in part to these advantages of personalization and social engagement. For example, Yahoo! Answers, a popular SQA service, has over 200 million users, with over one billion questions asked at a rate of 90,000 new questions per day (Harper, Moy, & Konstan, 2009). One element of SQA services that can present a disadvantage as compared to other forms of search (e.g. consulting a reference librarian) is their variable content quality. Studies of answer, and to a lesser degree, question quality have emerged in order to determine what constitutes a good question or answer within SQA, and often employ machine-learning methods (see Harper et al., 2008; Li et al., 2012; Shah & Pomerantz, 2010 for examples). However, these studies have shortcomings indicated by their disproportionate focus on answer quality and the narrow scope employed by studies that address question quality. These shortcomings will be addressed by studying question quality within Yahoo! Answers with regard to the following research questions:

RQ1. Do the textual features of a question labeled as “Bad” differ from the textual features of the same question after it is revised to be “Good”?
RQ2. Which question features change the most from before to after revision?
RQ3. Which relationships between features change the most from before to after revision?

Findings will indicate elements of a question that can make a difference in question quality and suggest avenues toward the development of an automated system that can make suggestions for improving a question.

The rest of the poster will proceed as follows. First, we will review background on question-answering and determining answer and question quality within SQA, followed by the method taken to address the research questions. Then, we will present findings followed by a discussion and conclusion.

BACKGROUND

Question-Answering

Studies within Library and Information Science (LIS) question the assumption often made within Information Retrieval (IR) that an asker can specify his or her information need. In identifying four stages of articulating an information need, Taylor (1968) stressed the role of the mediator (e.g. a reference librarian) who interprets the asker’s information need and expresses this need to a system in order to retrieve relevant documents. Belkin (1980) also argued that question-answering systems and services cannot accept a query at face value. Instead, these systems and services should solicit feedback from the asker during the search and use this feedback to inform what sort of results should be derived (Belkin, 1980).
In order to solicit such feedback, we find it imperative not only to extract textual features from questions to assess their quality, but also use human assessments as a baseline for quality. Since the average asker may not be able to properly articulate his or her information need (Taylor, 1968), we decided to employ trained reference librarians to provide human assessments.

**Question Quality in SQA**

Studies of question quality in SQA look at both textual and non-textual features. Examples of such works include Agichtein et al.’s (2008) assessment of answer and question quality within Yahoo! Answers, as well as the relationship between answers and questions. Textual features found to have a significant influence on the authors’ model also used in this work include punctuation density, number of words per sentence, number of unique words, and entropy. Bian et al. (2008) and Li et al. (2012) also found that non-textual features, such as the profile of the asker, influence question quality. Non-textual features were not considered within this study since our focus was to investigate revision of textual features by human assessors.

Yang et al. (2011) had a similar objective to this study, specifically using findings to inform the development of a system that flags questions for revision; however the authors focused on unanswered questions, which are not synonymous with question quality. Even if a question receives an answer, there is no indication that the answerer understood the asker’s information need. Alternatively, a question that clearly states the asker’s information need might not receive an answer based on variable factors, such as time of day the question was posted, a non-textual feature Yang et al. (2011) used in their prediction model. Therefore, within this study we sampled both questions that received answers and those that did not.

**METHOD**

**Human Assessors**

Using the Yahoo! Answers Application Programming Interface (API)
http://developer.yahoo.com/answers/, a total of N=2,000 questions were extracted. To reduce sampling bias, half of the questions were resolved, meaning they received an answer chosen as the “Best Answer” by the asker or community, and the other half were unresolved, meaning that they did not receive an answer or were removed from the server after a period of four days. Additionally, questions were equally sampled from five categories: Business and Finance, Entertainment and Music, Health, Sports, and Travel.

Four experts, defined as having a Master’s in Library and Information Science and currently working as reference librarians, were directly recruited using convenience sampling by one of the authors. In the course of two days, this author met with these experts, two-on-one. The author explained to the pair of experts the purpose of the study and general guidelines affecting question quality adapted from a typology for question quality developed by Shah et al. (2012), who utilized a grounded theoretical approach (Strauss & Glaser, 1967) to classify unresolved questions seeking a fact-based answer asked within Yahoo! Answers. The typology includes the following characteristics:

1. **Not clear:** The question cannot be comprehended by the average person. An example would be a question in which it is difficult to understand what the person is trying to ask. Another example would be a question that uses expert terminology or complicated vocabulary words.

2. **Too complex:** The question demands expert knowledge and/or an unreasonable amount of time to be answered.

3. **Asking multiple questions:** When posting a question, users have the option of elaborating their information need in the content section. Too many follow-up questions asked within the content section might entail an unreasonable amount of time to answer all of them. In addition, there may be little to no relationship between the ideas being expressed within the subject as compared to the content, which might confuse the reader as to the actual information need of the asker.

4. **Lack of information:** You do not feel you have enough information from the asker to answer the question.

Following this explanation, the author and the two experts worked together to rate ten questions and provide reasons why the question was bad/revise if necessary. After this, the experts worked together privately on fifty questions. Inter-coder reliability (ICR) was then measured, and when at an acceptable level (κ, > 0.61), the experts were given the remainder of the N=2,000 questions, divided between them. Out of these questions, a subset of n=129 (6.45%) questions were labeled as “Bad” and then revised by the experts. These revisions were both informed by the prior guidelines and discussion of question quality as well as the experts’ background and training in reference-based services, which involves helping patrons articulate questions to reflect their information needs (Taylor, 1968). These questions labeled as “Bad” and then revised by the experts comprised the dataset used for the feature extraction and analysis portion of this study.

**Feature Extraction**

We chose the textual features to extract based on those commonly used within similar mechanical extraction problems and the typology for question failure by Shah et al. (2012). Each textual feature will now be identified. Unless otherwise noted, a Java-based extractor created by the authors was used to derive these features.

**Number of complex words.** A dictionary was created with a list of complex words and the extractor assigned a complexity score to a question based on the presence of these words.
**Reading level.** Flesch-Kincaid Readability scores (Kincaid, 1975) were calculated to determine the reading ease of a question based on the amount of syllables contained within a word. The higher the score, the more easily understood the piece of content.

**Amount of unique information.** This measure indicates the amount of novel information communicated within the question, which may assist an answerer in interpreting an asker’s information need with a higher level of specificity, improving the overall answer quality. This is calculated by counting the number of distinct words over the total number of words used in a question.

**Number of questions.** Content containing multiple questions might confuse the answerer in interpreting what information the asker is looking for. A script identified the number of unique question marks in order to assign a related score. Question marks directly adjacent to one another were only counted once.

**Number of misspelled words.** A dictionary was created that measured misspellings using Jazzy, a Java based spell checker built on the Aspell algorithm. Questions containing many misspelled words may be unclear to the reader.

**Entropy.** Entropy measures the similarity between the language used within questions asked on Yahoo! Answers and the language used in the baseline comparison exemplar data corpus, in this case the LA Times collection available from TREC. The higher the entropy score, the less clear the question.

**Number of characters / Number of words / Number of sentences.** Often the length of an answer affects its quality (Shah & Pomerantz, 2010). Sometimes answers might be too short and not provide enough information, while in other cases, answers might be too long and provide superfluous information that ultimately confuses the reader or demands too much of them in answering the question.

### FINDINGS

A paired sample t-test was performed in order to determine whether significant differences (at p<0.001) exist between each of the nine features identified in the Feature Extraction section before and after revision. Results indicate that significant differences existed among all of the variables. The highest t-values, which indicate the greatest differences, were present for entropy (t=11.98), number of misspelled words (t=9.73), number of characters (t=9.72), and number of words (t=9.36). Full test results are given in Table 1.

When looking at the descriptive statistics (as presented in Table 2) for these nine features both before and after revision, it appears that the following values decreased after revision – number of complex words, number of questions, number of misspelled words, entropy, number of words, number of sentences, and number of characters. Values that increased are reading level (as indicated by the Flesch Kincaid score) and amount of unique information. It should be noted that a higher reading score indicates a converse reading level, i.e. content is easier to understand.

### Table 1. Results of paired sample t-test

<table>
<thead>
<tr>
<th>Pair</th>
<th>t</th>
<th>df</th>
<th>Sig.*</th>
</tr>
</thead>
<tbody>
<tr>
<td>numComplexWords - numComplexWords_rev</td>
<td>7.04</td>
<td>12</td>
<td>.000</td>
</tr>
<tr>
<td>FleschKincaidScore - FleschKincaidScore_rev</td>
<td>-3.90</td>
<td>12</td>
<td>.000</td>
</tr>
<tr>
<td>uniqueInfo - uniqueInfo_rev</td>
<td>-6.36</td>
<td>12</td>
<td>.000</td>
</tr>
<tr>
<td>numQuestions - numQuestions_rev</td>
<td>-6.78</td>
<td>12</td>
<td>.000</td>
</tr>
<tr>
<td>numMisspelledWords - numMisspelledWords_rev</td>
<td>9.73</td>
<td>12</td>
<td>.000</td>
</tr>
<tr>
<td>entropy - entropy_rev</td>
<td>11.98</td>
<td>12</td>
<td>.000</td>
</tr>
<tr>
<td>numWords - numWords_rev</td>
<td>9.36</td>
<td>12</td>
<td>.000</td>
</tr>
<tr>
<td>numSentences - numSentences_rev</td>
<td>7.61</td>
<td>12</td>
<td>.000</td>
</tr>
<tr>
<td>numCharacters - numCharacters_rev</td>
<td>9.72</td>
<td>12</td>
<td>.000</td>
</tr>
</tbody>
</table>

* Two-tailed

### Table 2. Descriptive statistics of paired sample t-test

<table>
<thead>
<tr>
<th>Features</th>
<th>Mean</th>
<th>S.D.*</th>
<th>S.E**</th>
</tr>
</thead>
<tbody>
<tr>
<td>numComplexWords</td>
<td>5.52</td>
<td>6.86</td>
<td>.61</td>
</tr>
<tr>
<td>numComplexWords_rev</td>
<td>1.27</td>
<td>1.14</td>
<td>.10</td>
</tr>
<tr>
<td>FleschKincaidScore_content</td>
<td>74.88</td>
<td>19.81</td>
<td>3.52</td>
</tr>
<tr>
<td>FleschKincaidScore_rev</td>
<td>89.60</td>
<td>17.19</td>
<td>1.52</td>
</tr>
<tr>
<td>uniqueInfo</td>
<td>0.85</td>
<td>0.19</td>
<td>.02</td>
</tr>
<tr>
<td>uniqueInfo_rev</td>
<td>0.96</td>
<td>0.07</td>
<td>.01</td>
</tr>
<tr>
<td>numQuestions</td>
<td>2.98</td>
<td>1.71</td>
<td>.15</td>
</tr>
<tr>
<td>numQuestions_rev</td>
<td>2.08</td>
<td>0.42</td>
<td>.04</td>
</tr>
<tr>
<td>numMisspelledWords</td>
<td>99.95</td>
<td>95.38</td>
<td>8.43</td>
</tr>
<tr>
<td>numMisspelledWords_rev</td>
<td>15.90</td>
<td>9.70</td>
<td>.86</td>
</tr>
<tr>
<td>entropy</td>
<td>5.06</td>
<td>1.06</td>
<td>.09</td>
</tr>
<tr>
<td>entropy_rev</td>
<td>3.60</td>
<td>0.76</td>
<td>.07</td>
</tr>
<tr>
<td>numWords</td>
<td>82.43</td>
<td>79.91</td>
<td>7.06</td>
</tr>
<tr>
<td>numWords_rev</td>
<td>14.75</td>
<td>8.27</td>
<td>.73</td>
</tr>
<tr>
<td>numSentences</td>
<td>5.13</td>
<td>5.49</td>
<td>.49</td>
</tr>
</tbody>
</table>

2 http://jazzy.sourceforge.net/

3 http://trec.nist.gov/data/docs_eng.html

Correlations between the variables before revision indicate high positive correlations (r=0.5, p<0.001) between the number of complex words, number of misspelled words, entropy, number of words, number of sentences, and number of characters. In addition, there is also a high positive correlation between entropy and reading score (r=0.617). On the other hand, high negative correlations are present between the amount of unique information and the number of complex words (r=-0.525), the reading score (r=-0.590), number of misspelled words (r=-0.685), entropy (r=-0.677), number of words (r=-0.721), number of sentences (r=-0.517), and number of characters (r=-0.652).
After revision, there is a high negative correlation between reading score and number of complex words ($r=-0.610$). High positive correlations ($r>0.5$, $p<0.001$) remain between number of misspelled words, entropy, number of words, number of sentences, and number of characters. On the other hand, high negative correlations remain between amount of unique information and number of misspelled words ($r=-0.517$), number of words ($r=-0.617$), and number of characters ($r=-0.559$).

**DISCUSSION**

The results indicate a few key findings. In relation to $RQ1$, there are significant differences between the means of each feature before and after revision. This indicates that changing textual features affects question quality, suggesting the validity of Shah et al.’s (2012) typology, as well as features derived from the literature review. In regard to $RQ2$, questions were mostly revised for clarity as indicated by entropy, and length as measured by number of words, sentences, and characters. In regard to $RQ3$, high positive correlations between clarity, as indicated by number of misspelled words and entropy, and length, as indicated by number of words and number of characters, both before and after revision indicate that these features are likely interrelated. When comparing the content of the questions sampled before and after revision, questions were often rendered clearer by shortening the overall length of the question. This suggests that before revision, the asker often provides extraneous information that could ultimately confuse potential respondents. At the same time, when referring to the mean values before and after revision, the amount of unique information has increased and negatively correlates to answer length, specifically number of words and number of characters. Therefore, removing extraneous information and only leaving content that articulated the information need can contribute to better questions.

However, the generalizability of our findings is limited by sample size ($n=129$). For this reason, the merit of this work is in presenting an exploratory approach to determining question quality and future work should be performed on larger data sets in order to determine whether these findings are consistent.

**CONCLUSION**

Our study extracted textual features from questions sampled within Yahoo! Answers before and after revision from “Bad” to “Good,” using human assessors to both label and revise the questions. Based on results from tests for mean differences, all of our features relate to question quality. When reviewing some of the relationships between these features, we found that within our sample ($n=129$) questions were improved by being shortened in length so that the information need of the asker could be clearly communicated. Although the small size of this sample does not allow for generalizability of these results, the framework for assessment established within this study can be lent to future work.

**ACKNOWLEDGEMENT**

The work reported here is partly supported by IMLS Early Career Development grant # RE-04-12-0105-12, as well as NSF award # BCS-1244704.

**REFERENCES**


