2014 47th Hawaii International Conference on System Science

Questioning the Question – Addressing the Answerability of Questions in Community Question-Answering

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

Abstract

In this paper, we investigate question quality among questions posted in Yahoo! Answers to assess what factors contribute to the goodness of a question and determine if we can flag poor quality questions. Using human assessments of whether a question is good or bad and extracted textual features from the questions, we built an SVM classifier that performed with relatively good classification accuracy for both good and bad questions. We then enhanced the performance of this classifier by using additional human assessments of question type as well as additional question features to first separate questions by type and then classify them. This two-step classifier improved the performance of the original classifier in identifying Type II errors and suggests that our model presents a novel approach for identifying bad questions with implications for query revision and routing.

1. Introduction

Studies within community question-answering (cQA) often focus on determining textual features comprising a good answer, indicated by community-based ratings, using mechanical extraction and machine learning approaches [26]. However, few studies have focused on determining question quality. A drawback to cQA studies that only focus on answer quality is their assumption that the question asked is of sufficient quality to receive a good answer [1][9][15][28], when in fact question quality has been shown to have a positive correlation with answer quality [2]. Therefore, this assumption is unrealistic, particularly in light of studies conducted within the library and information science (LIS) field regarding the difficulties faced by an individual when articulating his information need as a question [5][7][22]. These studies find that a disconnect often exits between an expressed information need and how this is interpreted by others, and many librarians devote their careers and training to assisting others in better articulating their questions.[1] Therefore, while studies on answer quality are a valuable area of study, more studies are needed for assessing question quality in order to depict both components – answers and questions – of cQA.

Therefore, the work described here starts with a goal to determine whether machine-based feature extraction can be used to approximate human judgments of question quality, and has implications for question routing, question reformulation, and question suggestions. Specifically, the work presented here assesses question quality by using the following features: (1) mechanically extracted textual features from the question content, (2) ratings provided by human assessors of whether a question is good or bad, and (3) classification of questions into types (Fact, Opinion, Advice, Social), again provided by human assessors.

The rest of the paper will proceed as follows. First, background on answer quality, question quality, and question types within cQA will be discussed. Then, we will discuss the results of three experiments. In Part I, we will describe how we measured the accuracy and validity of mechanically extracted textual features in predicting ratings provided by human assessors of whether a question is good or bad. In Part II we discuss how we built another classifier trained on features hypothesized to identify question type and tested on human assessments of question type. Then, in Part III we used the classifier built in Part II to first divide questions by type and then used the classifier built in Part I to classify these questions as good or bad. Findings indicate that the AUC values for question quality distributed by question type improve upon the original values found in Part I before the questions were divided by type, providing us with a model of both high accuracy and validity in predicting question quality. Following the experiments, we will conclude with a discussion of implications for findings, limitations of the study, and avenues for future work.

[1] See http://bit.ly/mKzaQX for the guidelines created by Reference

978-1-4799-2504-9/14 $31.00 © 2014 IEEE
DOI 10.1109/HICSS.2014.180

2. Background

cQA services provide a popular outlet for obtaining information from the Internet. Users can post a question in natural language and receive answers personalized to their information needs. One of the most popular services is Yahoo! Answers, which has over 200 million users asking over a billion questions at a rate of 90,000 new questions per day [11]. In fact, use of cQA services supersedes web-based search in some markets [2].

Advantages of using cQA include the exchange of personalized content [26], the ability to interact with and receive social support from others [28], and exposure to a large volume of content [2]. However, a relative disadvantage is the variability of content quality [2]. Studies of answer and to a lesser degree, question quality have emerged to address this issue.

2.1. Answer quality in cQA

Studies of answer quality in cQA evaluate textual (e.g. length) and non-textual (e.g. user profile information) features for predicting answer quality; these features are often determined using community feedback measures, such as Best Answer ratings [13][17][21]. In the past, these features have been obtained using mechanical extraction, and a consistency has developed within the literature of features that produce models with high accuracies and validities in predicting answer quality (see [28]). Due these consistencies, current studies focus on building more robust models by experimenting with different classification techniques and methods. For example, learning to rank (LETOR) [20] makes the implicit assumption that answer quality may have gradations beyond the binary values of "Best Answer" and "Not a Best Answer" and ordinally ranks answers, rather than provide binary judgments. Other cQA studies compare the performance of models trained on mechanically extracted features to models trained on judgments of answer quality made by human assessors. One such example is Shah & Pomerantz's work [29] on predicting answer quality within Yahoo! Answers using features derived from judgments of Mechanical Turk (MTurk) assessors.

2.2. Question quality in cQA

Fewer studies have been published on question quality within cQA and we find shortcomings with this work. Ignatova et al. [16] outlined a small list of features that could be mechanically extracted to predict question quality. The suggested features, misspelling, syntactical errors, and ambiguity, are all used within the current work. Agichtein et al. [2] assessed both answer and question quality within Yahoo! Answers, as well as the relationship between answers and questions. Textual features found to have a significant influence in the authors' model that we use in this work include punctuation density, number of words per sentence, number of unique words, and entropy between subject and content [2]. Bian et al. [4] and Li et al. [20] also found that non-textual features such as the profile of the asker also influence question quality.

Yang et al. [32] considered both textual and non-textual features influencing whether or not a question receives an answer, with implications for a system that flags questions likely to not receive an answer and suggests how to revise them. The findings indicate that topics distinguishing non-answered questions, heuristic textual features (e.g., question length) and non-textual features (e.g., time of day posted) all contribute to a predictive model for question quality; however the authors conclude that results are not adequate for practical use.

Our current study differs from these past works in that we use human assessments as a quality baseline, rather than community-driven feedback (e.g. Best Answer ratings, votes, stars, etc.) or Yang et al.'s [32] baseline of whether a question receives an answer or not. We could not the former, since only answers receive such feedback and did not choose the latter since we are determining question quality, not whether the question is likely to receive an answer. Even if a question receives an answer, there is no indication that the answerer adequately understood the asker's information need [28][35]. 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, which, incidentally, is one of the non-textual features Yang et al. [32] used in their prediction model. In addition, human judgments have experienced a high level of agreement regarding what criteria constitute a good versus poor quality answer in past works [28], and we argue that human judgments might provide insights into question quality that would not be determined by mechanical extraction alone.

2.3. Question types in cQA

Most studies on questions within cQA instead focus on type. Since different question types anticipate different types of answers, if one can determine the type of question being asked there are implications for routing the question to a service that best addresses this question type [2][13]. However, most of these studies evaluate question types based on the archival values of their answers, rather than the quality of the actual questions. For example, Harper et al. [15] developed two distinct question types in order to investigate archival value in social Q&A sites: informational questions, which are more likely to gather information, and conversational questions, which stimulate discussion in order to solicit others' opinions. Another study by Harper et al. [15] utilized a rhetorical framework [3] to classify questions in cQA sites and found that factual (31%) questions were most frequently asked, followed by identification (28%), advice (11%), and prescriptive (11%) questions. Rodrigues and Milic-Frayling [25] developed a typology similar to Harper et al.'s [14] and classified question types within Yahoo! Answers as belonging to the following high-level categories of questions seeking: (1) Factual Information, (2) Advice, (3) Opinion, (4) Chatting, (5) Entertainment, and (6) Other. Finally, a recent study by Choi et al. [8] focused on the distributions of frequencies for each type of question among four different types of cQA sites using the following typology: (1) Information seeking questions, (2) Advice-seeking questions, (3) Opinion-seeking questions, and (4) Non-information, or social, seeking questions. The typology for question type used in this study is influenced by these prior works and includes the following types: (1) Fact, (2) Advice, (3) Opinion, and (4) Social. We provide more details about how these four categories were refined in Section 4.1.

Within this study, we take the work discussed above on answer quality, question quality, and question type and attempt to learn from what has already been done, as well as addressing the perceived gaps within these studies. These gaps include lack of a good baseline for determining question quality, overreliance on non-textual features, and the unexplored link between question type and question quality. We will now address these gaps in the following three experiments.

3. Part I: Determining Question Quality

A total of N=5,000 questions were extracted from Yahoo! Answers using the service's API[2]. From this total, n_1=2,000 questions were used in order to train and evaluate the classifier developed in Part I, while a total of $n_{2\&3}$=3,000 questions were used in order to train and evaluate the classifier developed in Parts II and III. To reduce sampling bias, half of the total questions sampled (N=5,000) were unresolved, meaning that either after four days they had not received an answer or were deleted by the user; and half were resolved, meaning the question received an answer. In addition, both resolved and unresolved questions were sampled equally among five categories – Business and Finance, Entertainment and Music, Health, Sports, and Travel.

Questions for Part I were first run through the feature extractor and then given to human assessors from MTurk who rated the question as either good or bad, which provided our baseline for question quality. We chose to use human assessors outside of the Yahoo! Answers community given that, arguably, many of the participants with Yahoo! Answers do not constitute an actual community when measured in amount of interactions exchanged between members. For example, we performed a simple data mining operation using approximately 3.2 million questions collected between 2007 and 2009 to determine that only around 7% (n=230,840) of the askers and answers interacted more than one time to seek and share information within Yahoo! Answers. In addition, most askers within Yahoo! Answers only focus on asking questions and do not participate in other community-based roles such as answering or providing feedback [12]. This is not to say that there is not a concerted user base within the service, however when looking at asking activities, it appears as if many askers within Yahoo! Answers are driven to perform one task – asking a question. For this reason and also given the related difficulties in recruiting a substantial number of Yahoo! Answers users given the relative privacy related elements of the site, we chose MTurkers to provide assessments.

3.1. Feature extraction for Part I

Since most work on feature extraction has focused on answers, we faced a unique challenge of determining which textual features make a significant contribution to question quality. To determine this, we conducted a literature review on content quality within cQA as discussed in Sections 2.1 and 2.2. In addition to a list of common features derived from this review, we also were informed by Shah et al. [30], who developed four main categories that contribute to poor question quality: (1) Unclear, (2) Inappropriate, (3) Broad, and (4) Presence of multiple questions.

The current study extends this research avenue by translating these attributes of poor question quality, along with features commonly used in assessing answer (and to a lesser extent) question quality, into empirical features used to develop a prediction model. The resultant question features

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

were extracted using a Java program created by the authors, unless otherwise noted.

Questions within Yahoo! Answers have a mandatory subject line and an optional area for content. Therefore, some of the features extracted are divided by whether they belong to the subject or content. For those features where this division is not noted, the subject and content (if present) have been combined for analysis. These features are grouped in three categories: Boolean (whether something is present or not), counting-based (simply enumerating certain entities), and derived (based on calculations using certain entities).

Boolean features

Content present: As a measure of question clarity, we hypothesize that presence of content might impair the reader from understanding the question if the content happens to be unclear and extraneous, or unrelated to the question asked in the subject.

Subject starts with an interrogative word: The difference between types of interrogative words used within the subject might influence question quality [11][15]. [15][11]Therefore we distinguished questions based on the interrogative words they started with using a script that recognized from a list of words (e.g. who, what, where, when, why, how, should) which one(s) were present in a given question.

Content starts with an interrogative word: Harper et al. [15] found that whether or not the content section starts with an interrogative word affects the likelihood of a question receiving an answer. Presence of content that starts with an interrogative word could signal multiple questions being asked, which could convolute the question.

Subject / Content has URL: Based on findings that answer quality can be positively correlated to presence of a URL [11], we decided to see whether this might be the same for question quality by writing a script to detect common URL tags.

Presence of taboo words: Questions were identified as taboo by comparing the words used in the question to a dictionary of "taboo" words. Question quality might deteriorate if the question is considered inappropriate, as indicated by the presence of these taboo words.

Counting-based features

Number of Misspelled Words: Misspellings were measured using Jazzy, a Java-based spell checker built on the Aspell algorithm.[3] Spelling contributes to measuring the resultant clarity of a question.

Number of Question Marks: Presence of multiple questions might overwhelm and/or confuse the reader. We identified multiple questions by counting the presence of a question mark at the end of each sentence within the subject and/or content of one posted question. The technique used only counted one distinct question mark at the end of a word to not misidentify cases where multiple question marks were repeated for emphasis.

Question Length / Number of Sentences / Number of Words: These measures represent standard data mining approaches to traditionally assessing answer quality within question and answering forums. It has been found that often the length of an answer has an effect on quality; sometimes answers might be too short and not provide enough information, while in other cases, an answer that is too long might provide superfluous information that ultimately confuses the reader or demands too much of them. We hypothesize that a similar effect might occur based on question length and therefore incorporated these measures.

Number of Complex Words: A Java function available in Fathom library[4] was used to extract the number of complex words. The extractor assigned a related complexity score to a question based on presence of these words. We hypothesize that complexity is related to readability in the sense that if the complexity of a question transcends the cognitive capacity of an average Yahoo! Answers user, the resultant question quality will be poor since the community is not comprised of experts.

Derived features

Edit Distance Between Subject and Content: Difference between subject and content text was measured using Levenshtein (edit) distance [19]. This compares the common distance between words in the subject to the measured distance between words in its related content section, given that the question also contained a content section and is reflective of syntactic appropriateness. It could also contribute to measuring the resultant clarity of a question.

Cosine Similarity Between Subject and Content: As a more sophisticated measure of finding similarity between two strings, cosine similarity measure was computed between vector representations of subject and content.

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

[4] http://www.representqueens.com/fathom/docs/api/

Readability of Subject / Content: Flesch-Kincaid Readability scores [18] were calculated for each question and was used to determine complex, ambiguous questions.

Entropy of Subject and Content: To quantify the clarity of a question, we decided to employ a query clarity measure often used within the IR domain [11]. 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[5] with 131,896 documents containing 66,373,380 terms. The clarity score was computed using the Lemur toolkit. This toolkit has been previously used for measuring clarity (see [6][16][19]), including evaluating high accuracy retrieval [35].

3.2. Rating, revising, and ranking for Parts I and III

After the features discussed in Section 3.1 had been extracted from the dataset, we used Amazon's paid rater service, Mechanical Turk (MTurk)[6] to assess question quality by asking MTurk workers to rate a question as good or bad. Workers were provided with the following guidelines:

You will be given 1,000 questions. Each question has been posted to the social question and answering site Yahoo! Answers. Yahoo! Answers is community-based; anyone can sign up and participate by asking questions, answering questions, voting on the best answer to a question, and other activities.

We would like for you to identify whether a question is good or not. A question is considered good if you think it can be answered. It is ok if you cannot answer the question, however the question should clearly indicate the asker's information need.

These guidelines were purposefully broad since, given our past experience working with MTurk workers [9], we have experienced better levels of inter-coder reliability when providing less detailed guidelines. After being informed of the guidelines, MTurkers viewed the question without any other contextual elements (e.g., asker profile information). This decision was made because the current study was to focus on textual features only in order to determine *goodness* of each question from the dataset.

A test set of 100 questions was posted for three workers to complete in order to assess inter-coder reliability to make sure that the guidelines provided enough consistency to how each individual rated the question. After reliability reached a substantial level ($\varkappa > 0.61$), a total of six workers completed the rating task - three workers for one set of 1,000 questions and three workers for another set of 1,000 questions. A voting system was used to address any disagreement in rating. Using the data collected from human assessors, we analyzed the data using descriptive statistical techniques along with classification techniques, specifically ten-folds cross-validation using Support Vector Machine (SVM). The results generated by these methods will now be further discussed.

3.3. Classification and Cross-Validation on Feature Ratings

With the combined feature extraction, and expert and user-provided ratings, classification of question quality was performed using the Weka[7] framework. Specifically, Support Vector Machine (SVM) was used for ten-fold cross-validation (internal validity and robustness of the model) as well as separate training and testing data (external validity). SVM was chosen as it optimizes the division of features, with a focus on more "difficult" points closer to the decision boundary; a particularly important requirement given the imbalanced proportion between the frequencies of good and bad classes.

If both the classes ('Good' and 'Bad') of the questions had equal likelihood of occurring, a random process can be assumed to have 50% accuracy. However, in our dataset, about 85% of the questions were marked as 'Good', and therefore, a classification process may simply declare every question to be 'Good' and achieve 85% accuracy (100% on 'Good' and 0% on 'Bad'). The other extreme case would be declaring every question to be 'Bad', which will receive 15% accuracy (0% on 'Good' and 100% on 'Bad'). Since the purpose of this study is to identify bad questions that could potentially be flagged in the future and submitted for revision, we chose to revise the baseline of the classifier by training with an equal amount of good and bad questions, as identified by our MTurk assessors.

A revised classifier was built using N=636 questions, half good (n=318) and half bad (n=318) for training. To determine the robustness of this classifier, its performance was assessed using 10-

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

[6] https://www.mturk.com/mturk/welcome

[7] http://www.cs.waikato.ac.nz/ml/weka/

folds cross-validation and tested on the full (n_1=2,000) dataset. The classifier tested on MTurk data was able to identify 84% of bad questions and 21% of good questions, with an overall performance of 32%. This is an improvement over the extreme classifier that identifies bad questions by more than double the accuracy. Therefore, the revised classifier provides more stability in distinguishing between good and bad rankings, which is a difficult task given the sheer imbalance between the classes.

However, more important than looking at the overall accuracy of the classifier is to examine the area under the ROC curve (AUC), which indicates the amount of times the classifier will correctly define an instance of a question being good or bad over every possible iteration within the model. In the ROC curve, the area of 1 indicates that accuracy is measured perfectly, whereas the area of below .7 represents a poor performance of a classifier. The AUC value for the classifier trained on good/bad data (*n*=636) and tested on the rest of the MTurk data have AUC values almost occurring by chance alone (AUC=0.594). For this reason, the importance of reporting AUC values cannot be understated. This finding, combined with the low overall accuracy of the revised classifier suggests that further experiments should be performed in order to improve both the accuracy and validity of the model.

In Parts II and III, we attempt to improve on the performance of the revised classifier by incorporating question type as a determinant of question quality. In order to do this, in Part II, we will review how we developed a classifier to sort questions by type.

Table 1. Rating classification results.*

	Good	Bad	Overall
Extreme case – Good questions biased (n=1,000)	100%	0%	85%
Extreme case – Bad questions biased (n=1,000)	0%	100%	15%
10-fold SVM on equal Split of Good/Bad questions (n=636)	100%	55%	93%
Training (n=636) Testing SVM-MTurk (n=2,000)	21%	84%	32%

*Percentages calculated by distributions in confusion matrix.

4. Part II: Determining Question Type

4.1. Question type assessment

One additional feature we thought might contribute to question quality is the type of question asked - Fact, Advice, Opinion, or Social. Fact-finding questions represent those that usually have one "right" answer, while advice, opinion, and social questions all are subject to multiple interpretations. We hypothesize that different types of questions might have different characteristics. For example, questions that solicit facts might be shorter than those soliciting opinions, advice and/or social engagement, since there is less personal context that might have to be provided.

Another example might be that advice-seeking questions prove to be more complex than other types of questions, since community members might go into more detail and ask more questions about a specific topic if it relates to them personally. It could also be that questions of a specific type are generally of higher quality than those of another type. For this reason, a separate classifier was built to distinguish question types, which could inform the performance of the revised classifier developed in Part I, which could then be tested on questions separated by type.

To assess the validity of these categories and their understanding, we posted a test set of 150 questions on MTurk and asked three different workers to label each question with one of the four categories. We measured intercoder reliability among the three coders and having found it lacking a good level of agreement, revised the definitions and reposted new questions, repeating the same process. Once a reasonable agreement was found ($\varkappa > 0.61$), we proceeded with creating a Human Intelligence Task (HIT) on MTurk where MTurk workers were again asked to provide an assessment of 150 questions. A total of 20 HITs were created for the set of *n*=3,000 questions, which comprised resolved (*n*=1,500) and unresolved (*n*=1,500) questions. Each set of 150 questions was assessed by five different MTurk workers. A voting method for each question was used in order to decide between disagreements of category assignation in the current study. A set of guidelines was developed indicating how to discern between different question types and distributed to MTurk workers:

For each question, please identify the category it best fits into out of the following four choices:

Fact: Asking for only one right answer that is independent of personal views (e.g. What is the capitol of France?).

Advice: Asking for help in making a decision between more than one answer or an opinion / recommendation of how to do something (e.g. Which computer should I buy - Mac or PC? How do I remove a wine stain from a carpet?)

Opinion: Asking for people to share their ideas or thoughts on a specific subject (e.g. Which do you like better - Coke or Pepsi?)

Social: Rephrasing personal thoughts / ideas without necessarily expecting an answer (e.g. Why are some people so negative when you are trying to be positive?)

4.2. Feature extraction based on question type

In order to build a classifier to identify question type based on MTurk assessments, we were informed by literature written on features categorizing textual and non-textual content within cQA [11], as well as by inspecting the data. We hypothesize that question type could be predicted by the following features, which were all obtained by a script written by the authors, which marked the frequency of each word from the feature types listed below. These words were both extracted individually (42 features, 21 features for subject and 21 features for content) and as the clusters denoted below.

Qualifiers: Adverbs that qualify a noun (e.g., good, best, better, favorite, etc.) are hypothesized to be more prevalent in opinion-oriented questions, where a person is expressing personal views.

Opinion words: Verbs that appeal to soliciting the personal views or expressions of others (e.g., recommend, think, feel, believe, etc.) are hypothesized to be more prevalent in opinion-oriented questions.

Personal pronouns: Personal pronouns (e.g., I, me, ours, you, yours, etc.) are hypothesized to appeal to the asker's personal experience, and therefore would more likely require an advice or opinion-oriented answer than a fact-based one.

Indefinite pronouns: Indefinite pronouns (i.e., everybody, everybody's) are also hypothesized to appeal to the asker's personal experience, and therefore would more likely require an advice or opinion-oriented answer than a fact-based one.

Directive words: Directive words (e.g., name, find, compare, describe, etc.) are hypothesized to direct the answerer to perform a specific task. Since advice and opinion-oriented questions often seek answers based on personal experience, such words are hypothesized to indicate fact-finding questions.

Headwords: Combinations of interrogative words with a noun, adverb, or verb (e.g., How do, How long, How can, Where would, What would, Which one) are hypothesized to indicate question type. Headwords were developed from studying the data and noting frequent occurrences of specific clusters.

4.3. Classification and Cross-Validation on Question Type

Next, classification to determine question type from the dataset derived in Section 3.1 ($n_{2\&3}$==3,000) was performed using features derived from the first classifier, with the additional six features from Section 4.1. Since Social questions were not prevalent (less than 2% (n=50) of all questions), this question type was removed from further classification. The distribution of the remaining three question types was: Advice (*n*=2,029, 68%), Opinion (*n*=359, 12%), and Fact (*n*=562, 19%). Due to the imbalance of data distributed within the Advice category, two iterations of bootstrapping sampling were used in order to first separate Opinion-seeking and Fact-finding questions from Advice-seeking ones, and then to differentiate Opinion-seeking questions from Fact-finding ones. Then SVM with ten-folds cross-validation and model weighting was used in order to determine the robustness of the question type classifier. The performance of this model has an overall classification accuracy of 93.08%. Further, when looking at the ROC curves separated by question type, the AUC values separated by question type are all at high, acceptable levels, indicating that the validity the question type classifier holds among Advice (AUC=0.866), Opinion (AUC=0.857), and Fact-finding questions (AUC=0.873).

These findings indicate the impressive ability of the classifier to identify questions by question type. This signifies that if our original classifier from Part I (referred to as the revised classifier) performs better when evaluating datasets first separated by question type using the classifier build in Part II, future studies can be completed using a two-step approach completely based on machine feature extraction. This approach would first use a classifier to identify question type and then feed these results into a classifier separated by question type to determine question rating. For Part III, we then tested the viability of this two-step approach by separating features by question type before classifying them in order to determine whether the classification results using the optimized model from Part I could be improved.

5. Part III: Combining Part I and Part II Classifiers

For Part III, we then tested whether separating features by question type before classifying them will improve the classification results using the optimized model from Part I. Results of this experiment are presented in Table 2.

One interesting observation to note is the features making the most significant contributions to the classifier answer quality (Table 2), as indicated by the chi-square rankings. Results suggest that features making the most contribution to determining question quality using the two-stage approach of determining question type and then question quality, include choice of interrogative words, length of a question and its content, and the clarity of the content written (indicated by readability and number of misspelled words).

The overall accuracy of the Part I classifier when tested on questions divided by type, jumps from 33% to (or nearly to) 100% across all three question types. Further more, the AUC values increased almost +0.3, with Advice (AUC=0.8219), Opinion (AUC=0.817), and Fact (AUC=0.8085). This suggests that dividing questions by question type before classifying them proves an effective means by which to distinguish a small proportion of bad questions from an overarching proportion of good ones. For this reason, we can advocate that the hypothesis that question type contributes to assessing question quality is a valid one that should be tested in future works.

Table 2. Result of classification of each question type on 10-fold cross-validation.

	Advice	Fact	Opinion
Correctly Classified Instances	99.10%	98.81%	100%
Ranked attributes by Chi-squared Ranking filter (top ten)			
Length of the subject+content		124.19193	
Number of words		116.19691	
Edit distance between subject and content		106.85332	
Entropy of subject+content		103.47162	
Number of misspelled words		89.4045	
Number of sentences		46.05922	
Content present? (true/false)		44.5418	
Readability of content		43.40942	
Number of complex words		40.31633	
Subject starts with an interrogative word		20.14454	
Total Number of Instances	1448	336	190

6. Summary and Discussion

6.1. Limitations

Our findings are not without limitations. The first limitation is the small sample size (*N*=5,000) with subsets n_1=2,000 for Part I, and $n_{2\&3}$=3,000 for Parts II & 3. This sample size is smaller than other datasets mostly because we relied on human evaluators to assess question quality and question type in order to provide a standard from which we could train the two models used in Parts I and II. However, we did attempt to accommodate for this limitation via stratified sampling in order to approximate a normalized distribution that would be derived from randomly sampling a larger dataset within Yahoo! Answers. We would have like to have performed more experiments, attempting to further push the ROC curve and therefore account for even more area underneath the curve. However, given the promising results of the models we did test, future work can attempt to improve on this performance, using the two-step model we have developed within this paper.

The second limitation is that we chose not to incorporate non-textual features (e.g., time question was posted, asker profile information, etc.) into our model. Since it was proven that such non-textual features could be another factors that influence question quality [4][23], future work may include these features in order to evaluate the performance of predicting question quality.

6.2. Findings and implications

Our experiments indicate that a classifier with high accuracy and validity can be built using mechanically extracted textual features that is able to classify questions as good or bad, but most importantly can differentiate between good and bad questions, thus limiting the occurrence of Type II errors inherent to the imbalanced dataset.

Within this study, we found that question type makes a significant contribution to improving the ability of our original model to predict question quality. Since the question type classifier also performs with good ability to differentiate between good and bad questions, this suggests that it is possible to classify question type using cheaper and timelier machine extraction methods in lieu of human evaluators. Further, these questions can then be categorized and run through the original model according to type with good results on predicting question quality.

These findings have several implications. One is the creation of a question routing tool that could identify question type, and then using a typology similar to one from the study by Choi et al. [8], find a cQA service that predominately addresses questions of this type. By forwarding a question of a specific type to a service with community members and/or experts that predominately address this question type, the chance of the question receiving a good answer will likely increase. Another implication is the creation of a question reformulation tool. Online Q&A users may lack opportunities of question negotiation [31] or scaffolding [10] that enable them to interact with librarians, teachers, other experts for improving question quality that helps adequately identify the asker's information need in a question. To overcome this shortcoming, a simple model would indicate to an asker that the question she has posted is of low quality, and thus has less of a chance of receiving a good answer; this would allow the asker to reformulate the question how she saw fit in hopes of getting a good answer.

A more complicated model could also be created. This model might take into account expert-generated data regarding reasons for question failure (e.g. too complex, asking multiple questions) and attempt to distinguish textual features that would classify a bad question into one of these reasons. Here, the asker would be presented not only with the information that her question is bad, but also why it is bad, which would give her more information of how to improve it. Although this study did not focus on reasons for question failure, we intend to complete further study in attempts to add a stage to the model that would further subdivide bad questions by predominant reasons for failure.

7. Conclusion

Our study addressed an area within cQA neglected by the IR community: determining question quality. This is an essential area of study within the field, given that it is difficult to formulate a good question, and that a quality question has a greater chance of leading to a quality answer. Although studies on answer quality have been important in deriving textual and non-textual features that may be used to assess quality, these studies make the assumption that is the question asked is good enough already to solicit a quality answer. However, this is often not the case. What is the point in revising an answer if the original question is so poor that it is unclear of the asker's information need?

For this reason, our study attempted to go back one step and start with question quality. Since, to the best of our knowledge, there have been no studies in IR concerned with question quality in cQA; we had to rely on the judgments of human assessors to provide a classification standard. Based on this standard, we performed feature extraction and developed a model, informed by both the extensive research of the IR community on answer quality as well as by other literature within LIS on question type and reference best practices.

One issue we dealt with when developing the model was an imbalance between good to bad questions (about a 5:1 ratio). To lower potentiality for Type II errors, we experimented with changing our baseline. We were able to derive an acceptable model in regard to accuracy and validity, however still experienced difficulties in mitigating Type II errors. We then tested our hypothesis that question type contributes to determining answer quality.

To test this hypothesis, we first had to be able to prove that it could be fairly easy to classify questions mechanically into different question types. We again relied on classifications of questions into types provided by human assessors and trained a model on these classifications. Our end result was a model with both high accuracy and a large AUC value.

Since we had an effective model for separating questions into question type, we could then justify running the original classifier on questions divided by type. The performance of the classifier on all three types of questions was near perfect, with large AUC values, indicating an improvement from the performance of the classifier on questions before they were divided by type. These findings suggest that machine learning can be applied to questions, as well as answers, within cQA, with robust and accurate results. The resultant two-step model then presents novel implications for future research, including research to further improve its performance, as well as to assess its contribution in providing assistance to askers in formulating a better quality question to receive a better quality answer.

8. Acknowledgement

The work reported here is supported by IMLS Early Career Development grant # RE-04-12-0105-12.

9. References

[1] Agichtein, E., Brill, E., and Dumais, S. (2006). Improving web search ranking by incorporating user behavior information. In *Proceedings of ACM SIGIR.*

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

[3] Aristotle. (2007). On rhetoric: A theory of civic discourse. Translated with introduction, notes, and appendices by G. Kennedy. New York: Oxford University Press.

[4] Bian, J., Liu, Y., Agichtein, E. and Zha, H. (2008). Finding the right facts in the crowd: factoid question answering over social media. In *Proceedings of WWW*.

[5] Belkin, N. J. (1980). Anomalous states of knowledge as a basis for information retrieval. *Canadian Journal of Information Science, 5,* 133-143.

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

[7] Belkin, N. J., Oddy, R. N., and Brooks, H. M. (1982). ASK for information retrieval: Part I. Background and theory. *Journal of Documentation. 38*(2), 61-71.

[8] Choi, E., Kitzie, V. and Shah, C. (2012). Developing a Typology of Online Q&A Models and Recommending the Right Model for Each Question Type. Poster in *Proceedings of ASIST Conference*. Baltimore, Maryland.

[9] Choi, E., Kitzie, V. and Shah, C. (2013). A Machine Learning Based Approach to Predicting Success of Questions on Social Question-answering Sites. *Proceedings of iConference 2013*. Fort Worth, Texas.

[10] Choi, I., Land, S.M., and Turgeon, A.J. (2005). Scaffolding peer-questioning strategies to facilitate metacognition during online small group discussion. *Instructional Science, 33*(5/6), 483–511.

[11] Cronen-Townsend, S., Zhou, Y., and Croft, W. B. (2002). Predicting query performance. In *Proceedings of ACM SIGIR Conference*, 299-306.

[12] Gazan, R. (2006). Specialists and synthesists in a question answering community. In *Proceedings of the ASIST Conferece*.

[13] Gyöngyi, Z., Koutrika, G., Pedersen, J., and Garcia-Molina, H. (2008). Questioning Yahoo! Answers. In *Proceedings of the First Workshop on Question Answering on the Web*, held at WWW 2008.

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

[15] Harper, M. F., Raban, D. R., Rafaeli, S., and Konstan, J. K. (2008). Predictors of answer quality in online Q&A sites. In *Proceedings of the ACM ACM CHI*, 865−874.

[16] Harper, F. M., Weinberg, J., Logie, J., and Konstan, J. A. (2010). "Question types in social Q&A sites," First Monday, volume 15, number 7.

[17] Ignatova, K., Bernhard, D., and Gurevych, I. (2008). Generating high quality questions from low quality questions. In *Proceedings of Workshop on the Question Generation Shared Task and Evaluation Challenge*. Arlington, VA, USA.

[18] Kim, S., Oh, J. S., and 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*.

[19] 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.

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

[21] Li, B., Tan, J., Lyu, M.R., King, I., and Mak, B. (2012). Analyzing and predicting question quality in community question answering services. In *Proceedings of the WWW.*

[22] Li, H., Liu, T., and Zhai, C. (2009). Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval, 3*(3), 225-331.

[23] Liu, Y., Bian, J., and Agichtein, E. (2008). Predicting information seeker satisfaction in community question-answering. In S-H. Myaeng, D. W. Oard, F. Sebastiani, T-S Chua, & M-K. Leong (Eds.), In *Proceedings of SIGIR*.

[24] MacMullin, S. D., and Taylor, R. S. (1984). Problem dimensions and information traits. *The Information Society, 3*, 91-111.

[25] Qiu, G., Liu, K., Bu, J., Chen, C., and Kang, Z. (2007). Quantify query ambiguity using ODP metadata. In *Proceedings of ACM SIGIR*, 697–698.

[26] Rodrigues, E.M. and Milic-Frayling, N. (2008). Socializing or knowledge sharing? Characterizing social intent in community question answering. In *Proceedings of ACM CKIM.*

[27] Sang-Hun, C. (2007). To outdo Google, Naver taps into Korea's collective wisdom. *International Herald Tribune*.

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

[29] Shah, C, and Pomerantz, J. (2010). Evaluating and Predicting Answer Quality in Community QA. In *Proceedings of ACM SIGIR 2010 Conference*. Geneva, Switzerland: July 19-23, 2010.

[30] Shah, C., Radford, M., Connaway, L. S., Choi, E., and Kitzie, V. (2012). *"How Much Change Do You Get from 40$?" – Analyzing and Addressing Failed Questions on Social Q&A.* In *Proceedings of ASIST Conference.*

[31] Taylor, R.S. (1968). Question-negotiation and information seeking in libraries. *College & Research Libraries, 29*(3), 178-194.

[32] Yang, L., Bao, S., Lin, Q., Wu, X., Han, D., Su, Z., and Yu, Y. (2011). Analyzing and predicting not-answered questions in community-based question answering services. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligenc*e.