Incorporate Credibility into Context for the Best Social Media Answers

Qi Su[a,b], Helen Kai-yun Chen[a], and Chu-Ren Huang[a]
[a]Department of Chinese & Bilingual Studies,
The Hong Kong Polytechnic University, Hong Kong, China
sukia@pku.edu.cn
[b]Key Laboratory of Computational Linguistics,
Peking University, Beijing, China
{helenkychen, churenhuang}@gmail.com

Abstract. In this paper, we focus on the task of identifying the best answer for a user-generated question in Collaborative Question Answering (CQA) services. Given that most existing research on CQA has focused on non-textual features such as click-through counts which are relatively difficult to access, we examine the effectiveness of diverse content-based features for the task. Specially, we propose to explore how the information of *evidentiality* can contribute to the task. By the comparison of diverse textual features and their combinations, the current study provides useful insight into the issues of detecting the best answer to a given question in CQA without user features or system specific link structures.

Keywords: collaborative question answering, answer assessment, credibility, quality

1 Introduction

The technology of web search has provided a powerful platform for seeking and accessing information on the Web. However, due to the limitations of keyword based retrieval, people often find it hard to locate their desired content through ad hoc searching. Therefore, it is no surprise that Web 2.0 system is trying to seek improvement in information foraging. Given that a general-purpose, fully-automated question answering is still beyond the state-of-the-art, the human-powered Collaborative Question Answering (CQA) services start to draw new attentions. Through CQA services, people could post their questions concerning any subject, even those the ad hoc searching may find difficult to answer, on a community portal, and get answers directly provided by community contributors. Due to the advantage of human interaction, CQA has become a popular question-answering platform and an active research area up till now. It turns out to be an effective supplement to the ad hoc searching and question answering (Banerjee and Han, 2009; Weerkamp and Rijke, 2008).

Although CQA tries to help users come up with the best answers with its human-oriented strategy, the quality of user-generated answers is sometimes difficult to control (Gyongyi et al, 2008). Among the answers received, there might be only few knowledgeable responses and helpful information. Some postings could be vague or even purposely misleading. Thus seeking the best answers to questions asked within the CQA community has been a challenge to the CQA system.

The existing researches on best answer seeking in CQA can roughly be categorized by textual feature based and non-textual feature based approaches. Usually, best answers could be decoded as high-quality and authoritative. To date, there is no much attention on the content-based answer quality judgment in CQA research. Instead, non-textual features are utilized extensively in estimating the quality of answers (Jeon et al., 2006). Much of the related work is based on user authority analysis by the user’s interaction network. In some related area, researchers have also proposed textual feature based answer quality judgment. However, they usually equal it to the writing quality, and other features such as spelling errors, lacking of leading capitals and so on (Weerkamp & Rijke, 2008).

Since most of the existing researchers have focused on non-textual features, in this paper, we aim to explore the effectiveness of diverse textual features on the task of best answer detection. The problem of best answer detection is formulated as a text classification task. We will verify the effectiveness of different textual features, including n-gram, bag-of-word, and the relationship between questions and answers, also the combination of these features. Specially, we propose to model the textual features based on the linguistic theory of evidentiality. Evidentiality is concerned with linguistic expression reflecting users' degree of certainty, or commitment to the truth. It explicitly expresses information providers' specification for the information sources and their attitudes toward the information by the use of evidential. Thus evidentiality could be an explicit cue for the information quality in CQA answers.

The following answer phrases are derived from the CQA service of Yahoo! Answers .

- *I doubt this is true but it's a neat thing to think about...*
- *im not sure i was always told never to look directly at the sun cos its bad for ur eyes.*

The evidentials (such as *doubt*, *not sure*) in the context signals the uncertainty of the answer by the contributor himself and thus it is less possible to be considered as the best answer. In other words, evidentials could play the role as an important context clue that provides us with further insights in locating the best answers. In the current study, we model evidentiality within the framework of machine learning based text classification. The result from this study would contribute to both CQA and other applications in making judgments on textual information quality.

This paper is organized as follows. We discuss related work in section 2. The proposed framework for best answer detection which utilizes several textual features is presented in section 3. In section 4, we focus on one of the textual features in the framework, namely cognitive evidentiality. Using these textual features, we report the results of a large scale evaluation over Yahoo! answers in section 5. Lastly, in section 6, we discuss further our findings from the experiments and conclude this paper.

2 Related Work

Although extensive researches have been done on many aspects of question answering, the research on collaborative question answering (CQA) has not been the center of focus until recently. The purpose of CQA is quite similar to the traditional QA in that both aim to get the most exact answer for a given question. Comparing with the traditional QA solutions, which are generally content based, the previous works on CQA utilize more features that are related to CQA link structures and characteristics (Agichtein et al., 2008).

To model the best answers, several factors have been considered in CQA researches, including: the quality of answers (Agichtein et al., 2008), the users' authority (Jurczyk and Agichtein, 2007), and the relationships between questions and answers (Wang et al, 2009). These factors are not logically independent of one another. Agichtein et al. (2008) have shown that, in social media, high quality content is usually generated by highly authoritative authors. To score the authority of users, a common approach is using a graph-based ranking algorithm such as HITS and PageRank (Zhang et al., 2007; Bouguessa et al., 2008).

Jeon et al. (2006) showed a successful incorporation of quality measure into a language modeling-based retrieval model for the CQA task. However, the framework of answer quality predication which they proposed utilized mainly non-textual features such as click counts, answer's activity level, copy counts, etc, which is relatively hard to be accessed.

In CQA services, community members are usually incented to vote for best answers, and typically use the simple plurality voting to select best answers. They are expected to vote based on their conviction about the quality of the answer. However, the results of voting would not be revealed to the community until the voting period is over. Therefore, waiting for the voting results is sometimes a long process. In some cases, there could even be no voting by any user. This is the reason why we conducted the task of best answer detection by learning from the existing dataset. Although some researches have questioned the approach of plurality voting (Lee et al., 2009), here we accepted the user-voted best answers as the gold standard in our experiments.

Representing text with salient features is an important part of text processing tasks. In the field of CQA, non-textual features have drawn much attention. There hadn't been, however, as many works focusing on textual features for the CQA tasks, especially lexical semantic related features. Some related researches which involved answer quality predication incorporated only secondary textual features such as spelling errors, the lacking of leading capitals, the large number of exclamation markers, personal pronouns and text length (Weerkamp and Rijke, 2008). These researches usually treated the writing quality of documents as a cue of best answer identification. Yet there has not been any attempt to directly evaluate inherent linguistic cues in reflecting the credibility of the information.

Evidentiality is presented as a type of subjective information available in texts. Statements usually bring with explicit evidentiality markers (evidentials). Here we want to check the credibility of text content by evidentials in the CQA answers, and other textual features, as well as their combination to form a content-based framework of best answer detection.

3 A Content-based Framework for Best Answer Detection

We propose a content-based framework to detect the best answers in CQA applications. In the framework, we encode the process of best answer detection into a machine learning based classification problem. We aim to explore the effectiveness of several textual features (instead of structural link features) for such task. The features we used could be divided into two categories. The former one takes the relevance between questions and answers into account; the latter considers the characteristics of the answer content, including n-grams and answer credibility.

3.1 Question-Answer Relevance

We adopt the Query Likelihood Model (Language Model) to calculate the likelihood of how an answer would be relevant to the question, and take the likelihood score as a feature in the feature vector. In the model, the similarity between a query (Q) and a document (D) is given by the probability of generating the query from the document language model, as shown in following (Manning et al., 2008):

$$\mathit{sim}(Q,D) = \mathrm{P}(D\,|\,Q) = \mathrm{P}(D)\mathrm{P}(Q\,|\,D)\,/\,\mathrm{P}(Q)$$

In the equation, $\mathrm{P}(Q)$ is the same for all answer documents, and thus could be ignored. The prior probability of $\mathrm{P}(D)$ could also be ignored since it is often treated as uniform across all documents. We estimate $\mathrm{P}(Q|D)$ as following:

$$\mathrm{P}(Q\,|\,D) \propto \mathrm{P}(D)\prod_{w\in Q}((1-l)\hat{\mathrm{P}}(w\,|\,\mathrm{M_C}) + l\,\hat{\mathrm{P}}(w\,|\,\mathrm{M_D}))$$

Here l is a smoothing parameter. $\mathrm{M_C}$ is a language model built from the entire answer document collection; $\mathrm{M_D}$ is built from each answer document in the collection.

3.2 Answer N-grams

Bag-of-words (BOW) is the most frequently used features in text classification. Other N-gram models have also been shown to be very effective for many text processing applications. Therefore, we would like to check the effect of these N-gram models for our CQA task. For the experiments, we use the following 3 experimental settings:

- unigram

- bigram
- unigram + bigram

3.3 Answer Credibility

Moreover, we come up with an answer representation based on the credibility which is encoded by evidentials in the text of CQA answers. The information of text credibility has proved to be helpful in many other natural language processing applications. For example, Banerjee and Han (2009) modulated answer score in their question answering research by using this answer credibility: score'= (1-λ)*score + λ*AnswerCredibility, a weighted combination of the original score and answer credibility evaluation.

Credibility is a board definition which incorporates many aspects, such as the credibility of the information providers and of the text content. From the latter perspective, we consider the feature of evidentiality, which involves the expression of the users' degree of certainty, or commitment to the truth. Our goal is to explore the contribution of evidentiality in the text content, which will be discussed in detail in the next section.

4 Evidentiality as the Feature for CQA Answer Detection

Evidentiality is information providers' self specification for the information sources and their attitudes toward the information. Aikhenvald (2003) observed that every language has some ways of making reference to the source of the information. Once language is being used, it is always imprinted with the subjective relationship from the speakers towards the information. As a linguistic phenomenon, evidentiality could be expressed at different linguistic levels, and most commonly it is marked on the lexical level (such as in English, Chinese and many other languages). Sometimes evidentiality could be a label for the verbal category indicating the alleged source of information about the narrated information, that is, the evidence through which information is acquired (e.g. hear, reportedly, see, deduce, recall) (DeLancey, 2001). Meanwhile, evidentiality could also be characterised as expressions of speaker's attitude toward information, typically expressed by the so-call epistemic modalities (e.g. surely, ought to, may) (Chafe, 1986; Mushin, 2000).

The linguistic expressions of evidentiality are named as evidentials or evidential markers. Mushin (2000) defines evidentials as a marker which qualifies the reliability of information. It is an explicit expression of the speaker's attitudes toward the trustworthiness of the information source. For instance,

a). *It's probably raining.*
b). *It must be raining.*
c). *It sounds like it's raining.*
d). *I think/guess/suppose it's raining.*
e). *I can hear/see/feel/smell it raining.*

From the above examples, as can be seen, the information provided is based on a subjective viewpoint. The information conveyed would bear the personal experience or attitudes, which at the same time reflects the speakers' estimation toward the trustworthiness of the statements by the information providers.

Although evidentiality seems to be an obvious and straightforward evidence for text trustworthiness detection, it has not attracted much attention that it merits within the natural language processing society. A preliminary theoretical framework has been proposed for manual categorization of explicit certainty information by Rubin et al. (2005). However, as mentioned in (Rubin et al., 2005), the fields of information retrieval and natural language processing have not yet considered in detail the task of certainty identification.

In this paper, we focus on detecting the lexical semantic feature of evidentiality within a machine learning based text classification framework for best CQA answer detection. The items of evidentials in text expression form a relatively closed set, which is consist of these categories: attributive/modal adverb, lexical verb, auxiliary verb, and epistemic adjective. We extract the evidentials from the dataset manually. The list of the extracted evidentials is presented in table 1.

Table 1: The Extracted Evidentials as Features

Category	Evidential
Attributive/ modal adverb	*certainly, sure, of course, definitely, absolutely, undoubtedly, clearly, obviously, apparently, really, always, seemingly, probably, maybe, personally, perhaps, possibly, presumably*
Lexical verb	*report, certain, believe, see, seem, think, sound, doubt, wish, wonder, infer, assume, forecast, fell, heard*
Auxiliary verb	*must, ought, should, would, could, can, may, might*
Epistemic adjective	*definite, possible, likely, unlikely, probable, positive, potential, not sure, doubtful*

5 Experiment

5.1 Dataset

We experiment with a dataset extracted from Yahoo! Answers, which is distributed by Emory University. The dataset consist of the collections of questions, answers, user data and question categories. In this research, we only take into account the information of textual content, without the consideration of both user information and question category. For the dataset of Yahoo! Answers, a question only has one best answer and thus all the other answers will be marked as non-best answers. As a result, the set of best answers contains much fewer documents than the set of non-best answers. We extracted 10,000 questions and the corresponding 83,586 answers (including both best answers and non-best answers) to form our experimental dataset. Typically, the total numbers of best answers and non-best answers form a skew distribution. So we reduce the set of non-best answers to a comparable scale as best answers.

We adopt support vector machine (SVM) as the machine learning model to classify best answers from non-best answers, and use the SVMlight package (http://svmlight.joachims. org) as the classifier with the default parameters and a linear kernel. For the evaluation, we conducted a 10-fold cross validation, and used the metrics of precision (Prec. as in table 3), recall, accuracy (Accu. as in table 3) and F1-measure (F1: the harmonic mean of the precision and recall).

5.2 Experimental Settings and Results

We conduct binary classification experiments using different textual features, as well as the combinations of those features. The features which we used individually and the dimensionality of each kind of feature are summarized in Table 2.

Table 2: The Individual Features Used in the Experiments

Feature	Abbrev.	Dimensionality
Query Likelihood Model	LM	1
Unigrams	UG	145,454
Bigrams	BG	631,534
Evidential	EV	57

Note that the dimensionality of evidentials shown in Table 2 is bigger than the overall sizes of the evidential collection shown in Table1. This is mainly because we also include the cases of morphological changes (e.g. possible contractions of word forms).

To gain insight into the performance of individual feature, we first examine the experimental results using single features, as shown in Table 3.

Table 3: Experimental Results Using Single Features

Feature	Prec.	Recall	F1	Accu.
UG	0.5030	0.3400	0.4057	0.5020
BG	0.5116	0.2420	0.3286	0.5055
UG+BG	0.5072	0.3180	0.3909	0.5045
EV	0.5656	0.2890	0.3825	0.5335

As seen in Table 3, using evidential (EV) as the feature achieved the best performance among all the single features. Although more complex N-gram measures (e.g. BG, UG+BG) outperform the standard unigram measures (UG) in many cases, in this research, it did not show much advantage. Meanwhile, the utilization of N-grams also suffers from the problem of high feature dimensionality. The experimental results suggest that evidentiality could be an essential role in the prediction of answer quality, and therefore contribute to the task of best answer detection in CQA applications.

We then further experiment on the combinations of these single features. The experimental results based on the combined feature vectors are provided in Table 4.

Table 4: Experimental Results Using Combined Features

Feature	Prec.	Recall	F1	Accu.
LM+BG	0.6100	0.4890	0.5882	0.5420
LM+EV	0.6345	0.4301	0.5912	0.5127
LM+BG+EV	0.5942	0.6159	0.5976	0.6049

We combine the Query-Likelihood Model, which indicates the relevance of an answer to a question, with the feature of answer content as N-grams, as well as the evidentials which encode information credibility to form a combined feature vector. From the above results, we note that the best performance can be achieved by incorporating LM features, N-gram features and the evidentiality based text representation for our task.

6 Conclusion

In this paper, we propose a content based framework to predict the best answer for CQA applications. By the comparison of diverse textual features and their combinations, the current study provides a useful insight into the task of detecting the best answer to a given question in CQA applications. Specifically, we incorporate the linguistic knowledge of evidentiality into the text representation framework for best answer detection. We try to explore how the information of evidentiality can contribute to the task. As evidentiality is an integral and inherent part of any statement and explicitly expresses information about the trustworthiness of this statement, it should provide the most robust and direct model for predicting the quality of answer documents. Our experimental results also show an improvement of the evidential feature over other textual features such as N-grams. By combining evidentiality with other textual features, we show a better overall performance. In the future works, we plan to further examine other textual features and their weighted combination on both feature-level and classifier-level. Also, we will further explore the contribution of evidentiality on information quality prediction.

References

Agichtein E, Castillo C, and etc. 2008. Finding high-quality content in social media. In Proceedings of WSDM'08.

Aikhenvald A and Dixon, ed. 2003. Studies in evidentiality. Amsterdam/Philadelphia: John Benjamins Publishing Company

Banerjee P, Han H. 2009. Credibility: A Language Modeling Approach to Answer Validation, In Proceedings of NAACL HLT 2009, Boulder, Bolorado, US

Bouguessa M, Dumoulin B, Wang S. 2008. Identifying Authoritative Actors in Question-Answering Forums - The Case of Yahoo! Answers, In Proceedings of KDD'08, Las Vegas, Nevada, USA

Chafe W. 1986. Evidentiality: The Linguistic Coding of Epistemology, Evidentiality in English Conversation and Academic Writing. In Chafe and Nichols, (ed.). Evidentiality: The Linguistic Coding of Epistemology. Norwood, NJ: Ablex

DeLancey S. 2001. The mirative and evidentiality. In Journal of Pragmatic, 33

Forman G. 2003. An Extensive Empirical Study of Feature Selection Metrics for Text Classification, In Journal of Machine Learning Research, 3

Gyongyi Z, Koutrika G, Pedersen J. 2008, Garcia-Molina H. Questioning Yahoo! Answers. In First Workshop on Question Answering on the Web, held at WWW2008

Jeon J, Croft W, Lee J and Park S. 2006. A Framework to Predict the Quality of Answers with Non-textual Features, In Proceedings of SIGIR'06, Seattle, Washington, USA

Jurczyk P and Agichtein E. 2006. Discovering Authorities in Question Answer Communities by Using Link Analysis, In Proceedings of 16th ACM Conf. on Information and Knowledge Management (CIKM'07)

Lee C et al., 2007. Model for Voter Scoring and Best Answer Selection in Community Q&A Services, In Proceedings of International Conference on Web Intelligence and Intelligent Agent Technology-workshops

Leopold E, Kindermann J. 2002. Text Categorization with Support Vector Machines. How to Represent Texts in Input Space?, In Machine Learning, 46, 423-444

Manning C, Raghavan P, and Schütze H. 2008. Introduction to Information Retrieval, Cambridge University Press

Mushin I. 2000. Evidentiality and Deixis in Retelling, In Journal of Pragmatics, 32

Rubin V, Liddy E, and Kando N. 2005. Certainty Identification in Texts: Categorization Model and Manual Tagging Result. In J. Wiebe (Ed.), Computing Attitude and Affect in Text: Theory and Applications (The Information Retrieval Series), Springer-Verlag New York, Inc.

Wang X, Tu X, Feng D and Zhang L. 2009. Ranking Community Answers by Modeling Question-answer Relationships via Analogical Reasoning, In Proceedings of SIGIR'09, Boston, Massachusetts, USA

Weerkamp W, Rijke M. 2008. Credibility Improves Topical Blog Post Retrieval. In Proceedings of ACL08: HLT

Zhang J, Ackerman M, Adamic L. 2007. Expertise Networks in Online Communities: Structure and Algorithms. In Proceedings of the 16th ACM International World Wide Web Conference (WWW'07)