A Data-Driven Approach to Question Subjectivity Identification in Community Question Answering

Tom Chao Zhou[1*], **Xiance Si**[2], **Edward Y. Chang**[2], **Irwin King**[3,1], and **Michael R. Lyu**[1]

[1]Department of Computer Science and Engineering,
The Chinese University of Hong Kong, Shatin, N.T., Hong Kong
[2]Google Research, Beijing 100084, China
[3]AT&T Labs Research, San Francisco, CA, USA
[1]{czhou, king, lyu}@cse.cuhk.edu.hk [2]{sxc, edchang}@google.com [3]irwin@research.att.com

Abstract

Automatic Subjective Question Answering (ASQA), which aims at answering users' subjective questions using summaries of multiple opinions, becomes increasingly important. One challenge of ASQA is that expected answers for subjective questions may not readily exist in the Web. The rising and popularity of Community Question Answering (CQA) sites, which provide platforms for people to post and answer questions, provides an alternative to ASQA. One important task of ASQA is question subjectivity identification, which identifies whether a user is asking a subjective question. Unfortunately, there has been little labeled training data available for this task. In this paper, we propose an approach to collect training data automatically by utilizing social signals in CQA sites without involving any manual labeling. Experimental results show that our data-driven approach achieves 9.37% relative improvement over the supervised approach using manually labeled data, and achieves 5.15% relative gain over a state-of-the-art semi-supervised approach. In addition, we propose several heuristic features for question subjectivity identification. By adding these features, we achieve 11.23% relative improvement over word n-gram feature under the same experimental setting.

1 INTRODUCTION

Automatic Question Answering (AQA) has been a long-standing research problem which attracts contributions from the information retrieval and natural language processing communities. AQA ranges from Automatic Subjective Question Answering (ASQA) (Soricut and Brill 2004; Li et al. 2008) to Automatic Factual Question Answering (AFQA) (Harabagiu et al. 2001; Demner-Fushman and Lin 2007; Ferrucci et al. 2010). Although much progress has been made in AFQA, with the notable example of the IBM Watson system (Ferrucci et al. 2010), high quality ASQA is still beyond the state-of-the-art. There are two fundamental differences of ASQA compared with AFQA: firstly, ASQA aims at returning opinions instead of facts; secondly, ASQA aims at returning an answer summarized from different perspectives instead of a fixed answer.

The rising and popularity of Community Question Answering (CQA) sites provides an alternative to ASQA. CQA sites such as Yahoo! Answers[1], Google Confucius (Si et al. 2010), and Baidu Zhidao[2] provide platforms for people to post questions, answer questions, and give feedbacks to the posted items (Adamic et al. 2008; Lou et al. 2011). The structure of QA archives from CQA sites makes these QA pairs extremely valuable to ASQA (Xue, Jeon, and Croft 2008; Zhou et al. 2011; Zhou, Lyu, and King 2012). However, the inherently ill-phrased, vague, and complex nature of questions in CQA sites makes question analysis challenging. In addition, the lack of labeled data hinders the adventure of effective question analysis.

The explicit support of social signals in CQA sites, such as rating content, voting answers, and posting comments, aggregates rich knowledge of community wisdom. Thus, it is worthwhile to investigate whether we can leverage these social signals to advance question analysis. Motivated by Halevy, Norvig and Pereira's argument "Web-scale learning is to use available large-scale data rather than hoping for annotated data that isn't available" (Halevy, Norvig, and Pereira 2009), and inspired by the unreasonable effectiveness of data in statistical speech recognition, statistical machine translation (Halevy, Norvig, and Pereira 2009), and semantic relationship learning (Riezler, Liu, and Vasserman 2008), our approach works towards utilizing social signals to collect training data for question analysis without manual labeling.

As a test case of our study, we focus on one important aspect of question analysis: *question subjectivity identification (QSI)*. The goal is to identify whether a question is a subjective question. The asker of a subjective question expects one or more subjective answers, and the user intent is to collect people's opinions. The asker of an objective question expects an authoritative answer based on common knowledge or universal truth (Aikawa, Sakai, and Yamana 2011). High quality QSI could be used to decide whether the system should try to identify the correct answer (AFQA) or summarize a diversity of opinions (ASQA).

[*]This work was done when Tom Chao Zhou was an intern at Google.

[1]http://answers.yahoo.com
[2]http://zhidao.baidu.com

2 RELATED WORK

Question classification has been long studied in the question answering community. Stoyanov et al. (Stoyanov, Cardie, and Wiebe 2005) conducted deep analysis of the questions by manually classifying questions along several orientation dimensions. Ferrucci et al. (Ferrucci et al. 2010) employed question classification in DeepQA project. However, most previous research works focused on factual questions. In contrast, our work analyzes complex and realistic questions in CQA services. Automatic question answering has been an active area of research (Soricut and Brill 2004). Many existing approaches considered specific domains (Harabagiu et al. 2001; Demner-Fushman and Lin 2007), with the notable exception of the recent IBM Watson system (Ferrucci et al. 2010). Different from previous automatic factual question answering systems, Dang et al. (Dang, Kelly, and Lin 2007) started to address automatic subjective question answering in a controlled TREC opinion track. But questions in CQA services are more complex compared with the controlled TREC track. Li et al. (Li et al. 2008) employed a supervised framework for question subjectivity prediction, and a subsequent work of Li et al. (Li, Liu, and Agichtein 2008) proposed a co-training approach. Aikawa et al. (Aikawa, Sakai, and Yamana 2011) employed a supervised approach in Japanese subjective question classification. While previous approaches relied on manually labeled data, our work utilizes social signals in CQA services to automatically collect training data without manual labeling. There are existing works (Turney 2002; Hu and Liu 2004; Pang and Lee 2008) on classifying sentences or text fragment as being overall positive or negative, but our work focuses on classifying questions as being subjective or objective. There are approaches on utilizing online repositories as training data for some supervised task. Surdeanu et al. (Surdeanu, Ciaramita, and Zaragoza 2008) used publicly available online QA collections to investigate features for answer ranking without the need for costly human evaluations. Mintz et al. (Mintz et al. 2009) used Freebase, a large semantic database, to provide distant supervision for relation extraction. Bernhard et al. (Bernhard and Gurevych 2009) presented three datasets for training statistical word translation models for use in answer finding. Blitzer et al. (Blitzer, Dredze, and Pereira 2007) employed structural correspondence learning on unlabeled data of different domains to perform domain adaptation for sentiment classification. Zhou et al. (Zhou et al. 2009; 2010) employed tags to perform interested-based recommendation. However, different from these approaches, our work utilizes social signals to collect training data effectively.

3 QUESTION SUBJECTIVITY IDENTIFICATION

We treat question subjectivity identification as a classification task. Subjective questions are considered as positive instances, and objective questions are considered as negative instances. In this section, we propose several social signals for collecting training data, and propose several heuristic features for the QSI task.

3.1 Social Signal Investigation

Like (L): in CQA sites, users *like* an answer if they find the answer is useful. Even the best answer of a question has been chosen, users could *like* other answers as well as the best answer. The intuition of the *like* signal is as follows: answers posted to a subjective question are opinions. Due to different tastes of the large community of users, not only the best answer, but also other answers may receive *likes* from users. Thus, if the best answer receives similar number of *likes* with other answers, it is very likely that the question is subjective. If a question is objective, the majority of users would *like* an answer which explains universal truth or common knowledge in the most detailed manner. Thus, the best answer would receive extremely high *likes* than other answers. Equation (1) presents the criteria of selecting positive training data:

$$L(Q_{best\_answer}) \leq \frac{\sum L(Q_{answer})}{AN(Q)}, \quad (1)$$

where $L(\cdot)$ is the number of people *like* this answer, Q_{best_answer} is the best answer of a question Q, Q_{answer} is an answer of a question Q, and $AN(\cdot)$ is the number of answers of a question. Equation (2) presents the criteria of selecting negative training data:

$$L(Q_{best\_answer}) \geq \alpha \times MAX(L(Q_{other\_answer})), \quad (2)$$

where α is a parameter, Q_{other_answer} is an answer except the best answer of a question Q, and $MAX(\cdot)$ is the maximum function. *Like* signal is commonly found in CQA sites, such as *rate* in Yahoo! Answers, *support* in Baidu Zhidao and *like* in AnswerBag[3].

Vote (V): users could vote for the best answer in CQA sites. An answer that receives the most *votes* is chosen as the best answer. The intuition of *vote* signal is as follows: the percentage of *votes* of the best answer of an objective question should be high, since it is relatively easy to identify which answer contains the most thorough universal truth or common knowledge. However, users may *vote* for different answers of a subjective question since they may support different opinions, resulting in a relatively low percentage of *votes* on the best answer. Equation (3) shows the criteria of selecting positive training data:

$$V(Q_{best\_answer}) \leq \beta, \quad (3)$$

where $V(\cdot)$ is the percentage of votes of an answer, and β is a parameter. Equation (4) shows the criteria of selecting negative training data:

$$V(Q_{best\_answer}) \geq \gamma, \quad (4)$$

where γ is a parameter. There is *vote* signal in many popular CQA sites, such as Yahoo! Answers, Baidu Zhidao, and Quora[4].

Source (S): to increase the chance of an answer to be selected as the best answer, users often provide *sources* of their answers. A *source* of an answer is a reference to authoritative resources. The intuition of *source* signal is that

[3] http://answerbag.com

[4] http://quora.com

Table 1: **Social signals investigated to collect training data.**

Name	Description	Training Data
Like	Social signal that captures users' tastes on an answer.	Positive and Negative.
Vote	Social signal that reflects users' judgments on an answer.	Positive and Negative.
Source	Social signal that measures users' confidence on authoritativeness of an answer.	Negative.
Poll and Survey	Social signal that indicates users' intent of a question.	Positive.
Answer Number	Social signal that implies users' willingness to answer a question.	Positive.

source is only available for an objective question that has a fact answer. For an subjective question, users just post their opinions without referencing authorities. In our approach, we collect questions with *source* as negative training data. *Source* signal exists in many popular CQA sites such as Yahoo! Answers, Baidu Zhidao, and AnswerBag.

Poll and Survey (PS): since a large number of community users are brought together in CQA sites, users often post *poll and survey* questions. The intuition of *poll and survey* signal is that the user intent of a *poll and survey* question is to seek opinions on a certain topic. Thus, a *poll and survey* question is very likely to be a subjective question. In addition, CQA sites often have mechanisms to enable users to post *poll and survey* questions. For example, Yahoo! Answers has a dedicated category named *poll and survey*. In our approach, we collect *poll and survey* questions as positive training data.

Answer Number (AN): the number of posted answers to each question in CQA sites varies a lot. The intuition of *answer number* signal is as follows: users may post opinions to a subjective question even they notice there are other answers for the question. Thus, the number of answers of a subjective question may be large. However, users may not post answers to an objective question that has already received other answers since an expected answer is usually fixed. Thus, a large *answer number* may indicate subjectivity of a question, but a small *answer number* may be due to many reasons, such as objectivity, small page views. Equation (5) presents the criteria of collecting positive training data.

$$AN(Q) \geq \theta, \tag{5}$$

where $AN(\cdot)$ is the number of answers of a question, and θ is a parameter. Table 1 summarizes all social signals that are investigated in this study.

3.2 Feature Investigation

Word (word): word feature is shown to be effective in many question answering applications. We also study this feature in this paper. Specifically, each word is represented with its term frequency (tf) value.

Word n-gram (ngram): we utilize word n-gram feature in our approach. Previous supervised (Li et al. 2008) and small scale semi-supervised (Li, Liu, and Agichtein 2008) approaches on QSI observed that the performance gain of word n-gram compared with word feature was not significant, but we conjecture that it may be due to the sparsity of their small amount of labeled training data. We investigate whether word n-gram would have significant gain if we have a large amount of training data. Specifically, each word n-gram is represented with its tf value.

Besides basic features, we also study several light-weight heuristic features in this paper. These heuristic features could be computed efficiently, leading to the scalability of proposed approach.

Question length (qlength): information needs of subjective questions are complex, and users often use descriptions (Wang et al. 2010) to explain their questions, leading to larger question length. We investigate whether question length would help QSI. We divide question length into 10 buckets, and the corresponding bucket number is used as a feature.

Request word (rword): we observe that in CQA sites, users use some particular words to explicitly indicate their request for seeking opinions. We refer to these words as *request words*. Specifically, 9 words are manually selected, i.e. "should", "might", "anyone", "can", "shall", "may", "would", "could", and "please". The total number of request words is used as a feature.

Subjectivity clue (sclue): we investigate whether external lexicons would help QSI. Specifically, in this study, we utilize subjectivity clues from the work of Wilson et al. (Wilson, Wiebe, and Hoffmann 2005), which contain a lexicon of over 8000 subjectivity clues. Subjectivity clues are manually compiled word lists that may be used to express opinions, i.e., they have subjective usages.

Punctuation density (pdensity): punctuation density is measured according to the density of punctuation marks in questions. Equation (6) presents the formulation of calculating punctuation density for a question:

$$PDensity(Q) = \frac{\#\text{ punctuation marks}}{\#\text{ punctuation marks} + \#\text{ words}}. \tag{6}$$

Grammatical modifier (gmodifier): inspired by opinion mining research of using grammatical modifiers on judging users' positive and negative opinions, we investigate the effectiveness of using grammatical modifier as a feature. Specifically, adjective and adverb are considered as grammatical modifiers.

Entity (entity): the expected answer for an objective question is fact or common knowledge, leading to less relationships among entities compared with a complex subjective question. Thus, we conjecture that the number of entities varies between subjective and objective questions. Specifically, we use noun as the surrogate of entity in our study.

Table 2: **Performance of supervised, CoCQA, and combinations of social signals with the word n-gram feature. Value in parenthesis means relative performance gain compared with supervised approach.**

Method	Precision
Supervised	0.6596
CoCQA	0.6861 (+4.20%)
L + V + PS + AN + S	0.6626 (+0.45%)
L	0.5714 (−13.37%)
V + PS + AN + S	0.6981 (+5.84%)
PS + AN + S	0.6915 (+4.84%)
V + PS + AN	**0.7214 (+9.37%)**
V + AN	0.7201 (+9.17%)
AN + S	0.7038 (+6.70%)

4 EXPERIMENTAL EVALUATION

4.1 Experimental Setting

Comparison methods: the baseline approach of question subjectivity identification was supervised learning using labeled training data. In addition, we compared with the state-of-the-art approach CoCQA proposed by Li et al. in (Li, Liu, and Agichtein 2008). CoCQA was a co-training approach that exploits the association between the questions and contributed answers.

Dataset: the raw data that was used to collect training data using social signals was from Yahoo! Answers, and there was $4,375,429$ questions with associated answers and social signals. They were relatively popular questions according to user behaviors, and were actively indexed with high priority in our system. They could be considered as reusable and valuable resources. In Yahoo! Answers data, *rate* function was used as the *like* signal, *vote* function was used as the *vote* signal, *source* field in the best answer was used as the *source* signal, the category *poll and survey* was used as the *poll and survey* signal, and *number of answers* was used as the *answer number* signal. Social signals investigated in this study are quite general, and other CQA sites could be leveraged to collect training data as well. The ground truth data set we used was adapted from Li et al. (Li, Liu, and Agichtein 2008). They created the data set using Amazon's Mechanical Turk service[5]. As suggested in (Provost 2000; Yang et al. 2011), we used a sampling method to deal with the imbalance problem in their data set, i.e. to keep all objective questions and randomly sample the same number of subjective questions. We obtained 687 questions in total, and we referred it as T. We also employed sampling method when using social signals to collect training data. The same with Li et al. (Li, Liu, and Agichtein 2008), we reported the average results of 5-fold cross validation on T for supervised learning and CoCQA. Unlabeled data for CoCQA was from Liu et al. (Liu, Bian, and Agichtein 2008). The results of our approach on T were also reported for comparison. It is worthwhile to point out that our approach did not use any manually labeled data. To tune the parameters for different social signals, 20% of questions in T were randomly selected. This data set was used as the development set, and referred to as D.

Classification method: we employed Naive Bayes with add-one smoothing classification method (Croft, Metzler, and Strohman 2010) in our experiments. Aikawa et al. (Aikawa, Sakai, and Yamana 2011) found Naive Bayes was more effective than Support Vector Machines (Hearst et al. 1998) in classifying subjective and objective questions. In addition, the training process of Naive Bayes was able to be parallelized using MapReduce framework (Dean and Ghemawat 2008).

Metric: precision on subjective questions was used as the evaluation metric in our experiments. The reason was as follows: a user's satisfaction would be increased if he/she receives an answer that summarizes people's opinions for a subjective question, but his/her satisfaction would not be decreased if he/she receives an answer the same with existing CQA sites that are not equipped with subjective question identification component. A user's satisfaction would be decreased if he/she receives a summarized answer that repeats the fact for an objective question. Thus, precision on subjective questions was the appropriate metric.

Parameter tuning: we performed grid search using different parameter values over D. We ran grid search from 1.0 to 2.5 for α in *like* signal, from 0.1 to 1.0 for β and γ in *vote* signal alternatively, and from 10 to 30 for θ in *answer number* signal. The optimal setting was as follows: $\alpha = 2.0$, $\beta = 0.2$, $\gamma = 0.5$ and $\theta = 20$.

4.2 Effectiveness of Social Signals

We employed different combinations of social signals to automatically collect positive and negative training data, and used the trained classifier to identify subjective questions. Table 2 presents the results using word n-gram feature. Specifically, we employed unigram and bigram for word n-gram. By employing co-training over questions and associated answers, CoCQA utilizes some amount of unlabeled data, and achieves better results than supervised approach. However, similar with (Li, Liu, and Agichtein 2008), we found CoCQA achieved optimal performance after adding $3,000$ questions. It means CoCQA could only utilize a small amount of unlabeled data considering the large volume of CQA archives.

In Table 2, it is promising to observe that collecting training data using social signals $V + PS + AN$ achieves the best results. It improves 9.37% and 5.15% relatively over supervised and CoCQA respectively. The results indicate the effectiveness of collecting training data using well-designed social signals for QSI. Selecting training data using $V + AN$ and $AN + S$ achieve the second and third best performance. Both combinations perform better than supervised and CoCQA. In addition, social signals of V, AN, S could be found in almost all CQA sites. Due to the page limit, we report results of several combinations of social signals. Other combinations achieve comparable performances. Collecting training data using $like$ signal does not perform well. We look into the training data, and find that some objective questions are considered as subjective because their best answers receive fewer *likes* than other answers. Considering

[5]http://www.mturk.com

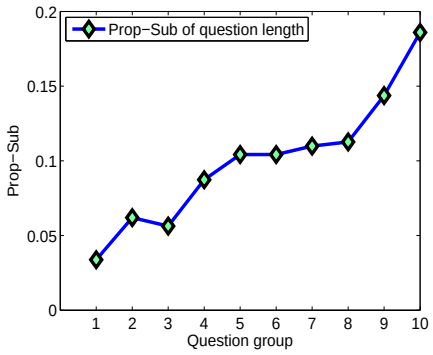

Figure 1: **The question length feature.**

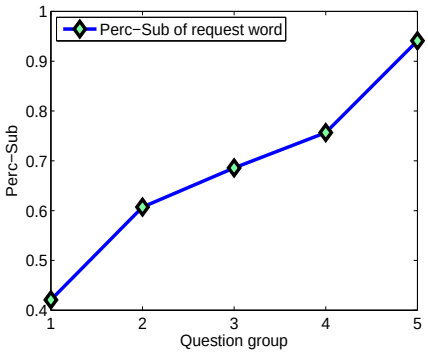

Figure 2: **The request word feature.**

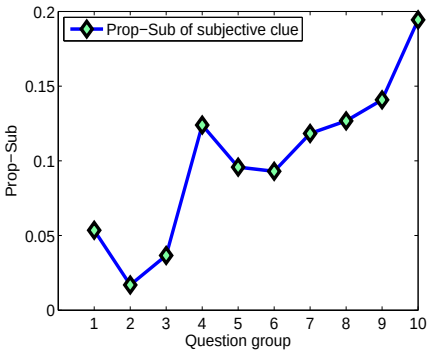

Figure 3: **The subjective clue feature.**

Table 3: **Performance of different approaches using word and word n-gram. Value in parenthesis means relative performance gain of word n-gram compared with word.**

Method/Feature	Word	Word n-gram
Supervised	0.6380	0.6596 (+3.39%)
CoCQA	0.6432	0.6861 (+6.66%)
V + PS + AN	0.6707	0.7214 (+7.56%)
V + AN	0.6265	0.7201 (+14.94%)
AN + S	0.6157	0.7038 (+14.31%)

Table 4: **Performance of three best performing combinations of social signals with varying training data.**

	20%	40%	90%	100%
V + AN	0.6549	0.7004	0.7188	0.7201
AN + S	0.6550	0.6696	0.6842	0.7038
V + PS + AN	0.6640	0.6846	0.7037	0.7214

the fact that many best answers are chosen by the asker, we conjecture that this phenomenon may be due to the complex of best answer selection criteria in CQA sites. Previous work also found socio-emotional factor affected a lot in the best answer selection (Kim, Oh, and Oh 2007). We leave the detailed study of how users choose their best answers to our future work.

Table 3 reports the results of different approaches using word and word n-gram feature. In line with our intuition, all approaches achieve better performance using word n-gram feature compared with word feature. More interestingly, we find that combinations of social signals, $V + PS + AN$, $V + AN$ and $AN + S$ achieve on average 12.27% relative gain of employing word n-gram over word. But supervised approach only achieves 3.39%, and CoCQA achieves 6.66% relative gain of using word n-gram over word. We conjecture the reason is as follows: supervised approach only utilizes manually labeled training data, resulting in the sparsity of employing word n-gram. CoCQA uses several thousand unlabeled data, and tackles the sparsity problem to some extent. Training data collected according to social signals is quite large compared with previous approaches, and data sparsity problem is better solved.

Table 4 reports the performance of three best performing combinations of social signals with varying amount of training data using word n-gram. With the increase of training data, performances of three approaches all improve accordingly. This finding is encouraging because in practical, we may integrate training data from several CQA sites with the same social signal.

4.3 Effectiveness of Heuristic Features

Previously, we discussed results of utilizing social signals to automatically collect training data. In this section, we study the effectiveness of heuristic features. To allow others to repeat our results, experiments investigating heuristic features were conducted on the data set T, which contains 687 questions adapted from Li et al. (Li, Liu, and Agichtein 2008).

Question length: Fig. 1 shows the proportion of subjective questions (denoted as Prop-Sub) with respect to questions' lengths. We rank the questions according to their lengths in ascending order, and equally partition them into 10 groups. Figures 3, 4, 5, and 6 apply similar methods to show Prop-Sub with respect to the corresponding features. Interestingly, we find the proportion of subjective questions increases as the question length increases. To find out the reason, we look into the data, and observe that when a user asks an objective question, he/she just expresses his/her information needs precisely, e.g., "Which player has won the fa cup twice with 2 different teams?" However, when a user asks a subjective question, he/she also shares his/her personal opinion together with the question, e.g., "Has anyone read "Empire" by Orson Scott Card? This is scary. I especially liked the "Afterword" by him. It's amazing how close you can feel today to it coming true."

Request word: Fig. 2 demonstrates the percentage of subjective questions (denoted as Perc-Sub) with respect to the number of request word. Group 1 contains questions that don't have any request word, group 2 contains questions having 1 request word, group 3 contains 2 request words, group 4 contains 3 request words, and group 5 contains at least 4 request words. Perc-Sub measures the percentage of subjective questions among all questions in each group. Quite surprisingly, we find Perc-Sub increases as the number of request words increases. After checking some sample questions, we conclude the reason is that when users ask subjective questions, they also add complicated background

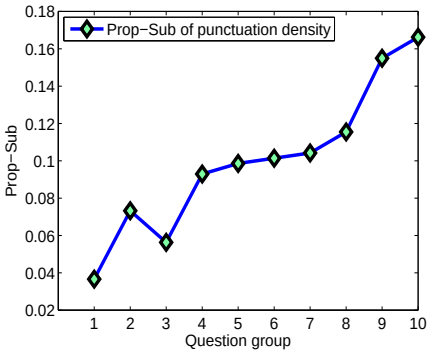

Figure 4: **The punctuation density feature.**

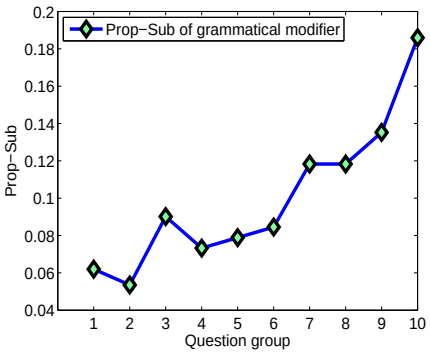

Figure 5: **The grammatical modifier feature.**

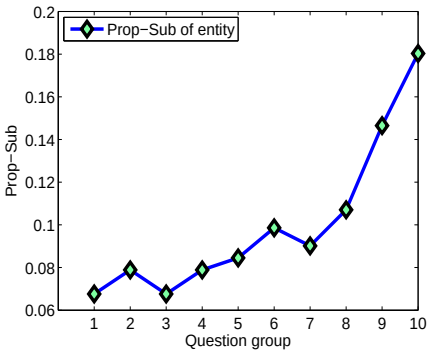

Figure 6: **The entity feature.**

Table 5: **Performance of heuristic features.**

	ngram	ngram + qlength	ngram + rword	ngram + sclue	ngram + pdensity	ngram + gmodifier	ngram + entity	heuristic features	ngram + heuristic
Precision	0.6596	0.6896	0.6834	0.6799	0.7000	0.6950	0.6801	0.6995	$\mathbf{0.7337}(+\mathbf{11.23}\%)$

or detailed opinions, making the question quite long. To attract potential answerers, users add these request words.

Subjective clue: in Fig. 3, we can see a clear trend that the more subjective clues, the larger proportion of subjective questions. This is an interesting finding that although subjective clues used in our experiments are from other documents, such as news, they still help distinguish between subjective and objective questions to some extent.

Punctuation density: in Fig. 4, we observe that the higher punctuation density, the higher proportion of subjective questions. In other words, the punctuation mark density of subjective questions is higher than that of objective questions. After examining some examples, we find that users use short sentence segments when sharing their experiences in subjective questions. In addition, we conjecture that short sentence segments help better express users' feelings and opinions in asking subjective questions.

Grammatical modifier: in Fig. 5, we find the proportion of subjective questions is positively correlated with the number of grammatical modifiers. The reason comes from the observation that grammatical modifiers are commonly used to describe users' feelings, experiences, and opinions in subjective questions. Thus, the more grammatical modifiers used, the larger proportion of subjective questions.

Entity: it is interesting to observe from Fig. 6 that the proportion of subjective question increases as the number of entities increases. After investigating some samples, we find that information needs of objective questions involve fewer entities compared with subjective questions. The reason is that subjective questions involve more descriptions, which also contain entities.

Table 5 shows results of employing heuristic features and word n-gram. We observe that adding any heuristic feature to word n-gram would improve precision to some extent, and employing only heuristic features performs even better than word n-gram. Combining heuristic features and word n-gram achieves 11.23% relative performance gain over employing word n-gram. Table 6 shows examples of questions wrongly classified using n-gram, but correctly classified with the incorporation of heuristic features. These results demonstrate the effectiveness of proposed heuristic features.

Table 6: **Examples of questions wrongly classified using n-gram, but correctly classified with the incorporation of heuristic features.**

Examples
Who is Mugabe?
When and how did Tom Thompson die? He is one of the group of seven.
Was Roy Orbison blind?
How is an echocardiogram done?
Fluon Elastomer material's detail?
What does BCS stand for in college football?

5 CONCLUSIONS

In this paper, we present a data-driven approach for utilizing social signals in CQA sites. We demonstrate our approach for one particular important task of automatically identifying question subjectivity, showing that our approach is able to leverage social interactions in CQA portals. Despite the inherent difficulties of question subjectivity identification for real user questions, we have demonstrated that our approach can significantly improve prediction performance than the supervised approach (Li et al. 2008) and a state-of-the-art semi-supervised approach (Li, Liu, and Agichtein 2008). We also study various heuristic features for QSI, and experimental results confirm the effectiveness of proposed features. In the future we plan to explore more sophisticated features such as semantic analysis using natural language processing techniques. We will investigate characteristics of subjective questions, and study whether we could find popular seman-

tic patterns for subjective questions.

6 ACKNOWLEDGEMENT

The work described in this paper was fully supported by two grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CUHK 413210 and CUHK 415311) and two grants from Google Inc. (one for Focused Grant Project "mobile 2014" and one for Google Research Awards). The authors would like to thank Hiyan Alshawi, Fangtao Li, Decheng Dai and Baichuan Li for their insightful suggestions. The authors would like to thank the anonymous reviewers for their insightful comments and helpful suggestions. The authors would like to thank Baoli Li, Yandong Liu and Eugene Agichtein for sharing the ground truth data set.

References

Adamic, L. A.; Zhang, J.; Bakshy, E.; and Ackerman, M. S. 2008. Knowledge sharing and yahoo answers: everyone knows something. In *Proceedings of WWW*, 665–674.

Aikawa, N.; Sakai, T.; and Yamana, H. 2011. Community qa question classification: Is the asker looking for subjective answers or not? *IPSJ Online Transactions* 160–168.

Bernhard, D., and Gurevych, I. 2009. Combining lexical semantic resources with question & answer archives for translation-based answer finding. In *Proceedings of ACL*, 728–736.

Blitzer, J.; Dredze, M.; and Pereira, F. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of ACL*, 440–447.

Croft, W.; Metzler, D.; and Strohman, T. 2010. *Search Engines: Information Retrieval in Practice*.

Dang, H. T.; Kelly, D.; and Lin, J. 2007. Overview of the trec 2007 question answering track. In *Proceedings of TREC*.

Dean, J., and Ghemawat, S. 2008. Mapreduce: Simplified data processing on large clusters. *Communications of the ACM* 51(1):107–113.

Demner-Fushman, D., and Lin, J. 2007. Answering clinical questions with knowledge-based and statistical techniques. *Computational Linguistics* 33(1):63–103.

Ferrucci, D.; Brown, E.; Chu-Carroll, J.; Fan, J.; Gondek, D.; Kalyanpur, A.; Lally, A.; Murdock, J.; Nyberg, E.; Prager, J.; Schlaefer, N.; and Welty, C. 2010. Building watson: An overview of the deepqa project. *AI Magazine* 59–79.

Halevy, A.; Norvig, P.; and Pereira, F. 2009. The unreasonable effectiveness of data. *IEEE Intelligent Systems* 8–12.

Harabagiu, S.; Moldovan, D.; Pasca, M.; Surdeanu, M.; Mihalcea, R.; Girju, R.; Rus, V.; Lacatusu, F.; Morarescu, P.; and Bunescu, R. 2001. Answering complex, list and context questions with lcc's question-answering server. In *Proceedings of TREC*.

Hearst, M.; Dumais, S.; Osman, E.; Platt, J.; and Scholkopf, B. 1998. Support Vector Machines. *IEEE Intelligent systems* 13(4):18–28.

Hu, M., and Liu, B. 2004. Mining and summarizing customer reviews. In *Proceedings of KDD*, 168–177.

Kim, S.; Oh, J.; and Oh, S. 2007. Best-answer selection criteria in a social q&a site from the user-oriented relevance perspective. *Proceedings of ASIST* 44(1):1–15.

Li, B.; Liu, Y.; Ram, A.; Garcia, E.; and Agichtein, E. 2008. Exploring question subjectivity prediction in community qa. In *Proceedings of SIGIR*, 735–736.

Li, B.; Liu, Y.; and Agichtein, E. 2008. Cocqa: Co-training over questions and answers with an application to predicting question subjectivity orientation. In *Proceedings of EMNLP*, 937–946.

Liu, Y.; Bian, J.; and Agichtein, E. 2008. Predicting information seeker satisfaction in community question answering. In *Proceedings of SIGIR*, 483–490.

Lou, J.; Lim, K.; Fang, Y.; and Peng, Z. 2011. Drivers of knowledge contribution quality and quantity in online question and answering communities. In *Proceedings of PACIS*.

Mintz, M.; Bills, S.; Snow, R.; and Jurafsky, D. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of ACL*, 1003–1011.

Pang, B., and Lee, L. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval* 2(1-2):1–135.

Provost, F. 2000. Machine learning from imbalanced data sets. In *AAAI Workshop on Imbalanced Data Sets*.

Riezler, S.; Liu, Y.; and Vasserman, A. 2008. Translating queries into snippets for improved query expansion. In *Proceedings of COLING*, 737–744.

Si, X.; Chang, E.; Gyöngyi, Z.; and Sun, M. 2010. Confucius and its intelligent disciples: Integrating social with search. volume 3, 1505–1516.

Soricut, R., and Brill, E. 2004. Automatic question answering: Beyond the factoid. In *Proceedings of HLT-NAACL*, 191–206.

Stoyanov, V.; Cardie, C.; and Wiebe, J. 2005. Multi-perspective question answering using the opqa corpus. In *Proceedings of EMNLP*, 923–930.

Surdeanu, M.; Ciaramita, M.; and Zaragoza, H. 2008. Learning to rank answers on large online qa collections. In *Proceedings of ACL*, 719–727.

Turney, P. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of ACL*, 417–424.

Wang, K.; Ming, Z.-Y.; Hu, X.; and Chua, T.-S. 2010. Segmentation of multi-sentence questions: Towards effective question retrieval in cqa services. In *Proceedings of SIGIR*, 387–394.

Wilson, T.; Wiebe, J.; and Hoffmann, P. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of EMNLP*, 347–354.

Xue, X.; Jeon, J.; and Croft, W. B. 2008. Retrieval models for question and answer archives. In *Proceedings of SIGIR*, 475–482.

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 AAAI*, 1273–1278.

Zhou, T. C.; Ma, H.; King, I.; and Lyu, M. R. 2009. Tagrec: Leveraging tagging wisdom for recommendation. In *Proceedings of CSE*, 194–199.

Zhou, T. C.; Ma, H.; Lyu, M. R.; and King, I. 2010. Userrec: A user recommendation framework in social tagging systems. In *Proceedings of AAAI*, 1486–1491.

Zhou, T. C.; Lin, C.-Y.; King, I.; Lyu, M. R.; Song, Y.-I.; and Cao, Y. 2011. Learning to suggest questions in online forums. In *Proceedings of AAAI*, 1298–1303.

Zhou, T. C.; Lyu, M. R.; and King, I. 2012. A classification-based approach to question routing in community question answering. In *Proceedings of CQA*.