CoCQA: Co-Training Over Questions and Answers with an Application to Predicting Question Subjectivity Orientation

Baoli Li
Emory University
csblli@gmail.com

Yandong Liu
Emory University
yliu49@emory.edu

Eugene Agichtein
Emory University
eugene@mathcs.emory.edu

Abstract

An increasingly popular method for finding information online is via the Community Question Answering (CQA) portals such as Yahoo! Answers, Naver, and Baidu Knows. Searching the CQA archives, and ranking, filtering, and evaluating the submitted answers requires intelligent processing of the questions and answers posed by the users. One important task is automatically detecting the question's *subjectivity orientation*: namely, whether a user is searching for subjective or objective information. Unfortunately, real user questions are often vague, ill-posed, poorly stated. Furthermore, there has been little labeled training data available for real user questions. To address these problems, we present *CoCQA*, a co-training system that exploits the association between the questions and contributed answers for question analysis tasks. The co-training approach allows *CoCQA* to use the effectively unlimited amounts of *unlabeled* data readily available in CQA archives. In this paper we study the effectiveness of *CoCQA* for the question subjectivity classification task by experimenting over thousands of real users' questions.

1 Introduction

Automatic question answering (QA) has been one of the long-standing goals of natural language processing, information retrieval, and artificial intelligence research. For a natural language question we would like to respond with a specific, accurate, and complete answer that addresses the question. Although much progress has been made, answering complex, opinion, and even many factual questions automatically is still beyond the current state-of-the-art. At the same time, the rise of popularity in social media and collaborative content creation services provides a promising alternative to web search or completely automated QA. The explicit support for social interactions between participants, such as posting comments, rating content, and responding to questions and comments makes this medium particularly amenable to Question Answering. Some very successful examples of Community Question Answering (CQA) sites are Yahoo! Answers [1] and Naver[2], and Baidu Knows[3]. Yahoo! Answers alone has already amassed hundreds of millions of answers posted by millions of participants on thousands of topics.

The questions posted to such CQA portals are typically complex, subjective, and rely on human interpretation to understand the corresponding information need. At the same time, the questions are also usually ill-phrased, vague, and often *subjective* in nature. Hence, analysis of the questions (and of the corresponding user intent) in this setting is a particularly difficult task. At the same time, CQA content incorporates the relationships between questions and the corresponding answers. Because of the various incentives provided by the CQA sites, answers posted by users tend to be, at least to some degree, responsive to the question. This observation suggests investigating whether the relation-

[1] http://answers.yahoo.com
[2] http://www.naver.com
[3] http://www.baidu.com

Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pages 937–946,
Honolulu, October 2008. ©2008 Association for Computational Linguistics

ship between questions and answers can be exploited to improve automated analysis of the CQA content and the user intent behind the questions posted.

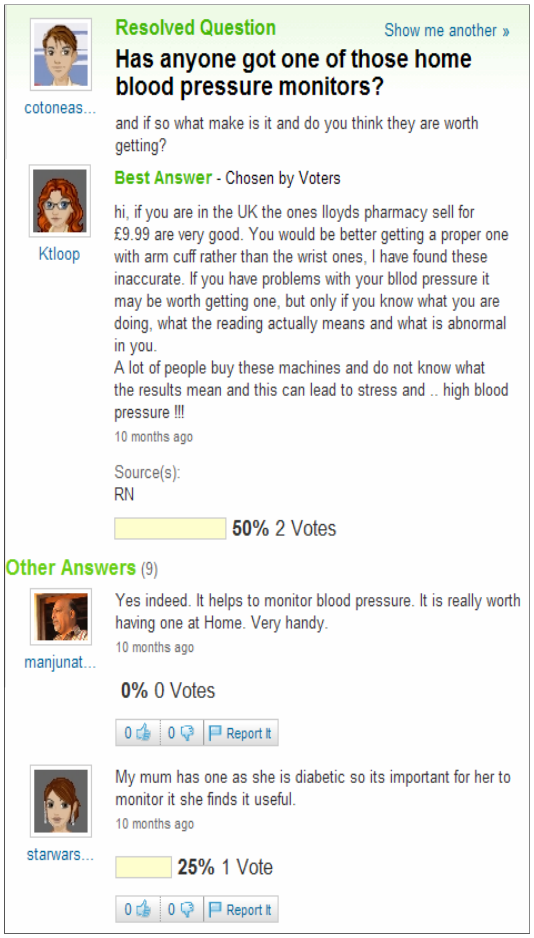
To this end, we exploit the ideas of *co-training*, a general semi-supervised learning approach naturally applicable to cases of complementary views on a domain, for example, web page links and content (Blum and Mitchell, 1998). In our setting, we focus on the complimentary views for a question, namely the text of the question and the text of the associated answers.

As a concrete case-study of our approach we focus on one particularly important aspect of intent detection: the *subjectivity orientation*. We attempt to predict whether a question posted in a CQA site is subjective or objective. Objective questions are expected to be answered with reliable or authoritative information, typically published online and possibly referenced as part of the answer, whereas subjective questions seek answers containing private states, e.g. personal opinions, judgment, experiences. If we could automatically predict the orientation of a question, we would be able to better rank or filter the answers, improve search over the archives, and more accurately identify similar questions. For example, if a question is objective, we could try to find a few highly relevant articles as references, whereas if a question is subjective, useful answers are not expected to be found in authoritative sources and tend to rank low with current question answering and CQA search techniques. Finally, learning how to identify question orientation is a crucial component of inferring user intent, a long-standing problem in web information access settings.

In particular, we focus on the following research questions:

- Can we utilize the inherent structure of the CQA interactions and use the unlimited amounts of *unlabeled* data to improve classification performance, and/or reduce the amount of manual labeling required?

- Can we automatically predict question subjectivity in Community Question Answering – and which features are useful for this task in the real CQA setting?



Resolved Question

Has anyone got one of those home blood pressure monitors?

and if so what make is it and do you think they are worth getting?

Best Answer - Chosen by Voters

hi, if you are in the UK the ones lloyds pharmacy sell for £9.99 are very good. You would be better getting a proper one with arm cuff rather than the wrist ones, I have found these inaccurate. If you have problems with your bllod pressure it may be worth getting one, but only if you know what you are doing, what the reading actually means and what is abnormal in you.
A lot of people buy these machines and do not know what the results mean and this can lead to stress and .. high blood pressure !!!

10 months ago

Source(s):
RN

50% 2 Votes

Other Answers (9)

Yes indeed. It helps to monitor blood pressure. It is really worth having one at Home. Very handy.

10 months ago

0% 0 Votes

My mum has one as she is diabetic so its important for her to monitor it she finds it useful.

10 months ago

25% 1 Vote

Figure 1: Example question (Yahoo! Answers)

The rest of the paper is structured as follows. We first overview the community question answering setting, and state the question orientation classification problem, which we use as the motivating application for our system, more precisely. We then introduce our *CoCQA* system for semi-supervised classification of questions and answers in CQA communities (Section 3). We report the results of our experiments over thousands of real user questions in Section 4, showing the effectiveness of our approach. Finally, we review related work in Section 5, and discuss our conclusions and future work in Section 6.

2 Question Orientation in CQA

We first briefly describe the essential features of question answering communities such as Yahoo! Answers or Naver. Then, we formally state the problem addressed in this paper, and the features used for this setting.

2.1 Community Question Answering

Online social media content and associated services comprise one of the fastest growing segments on the Web. The explicit support for social interactions between participants, such as posting comments, rating content, and responding to questions and comments makes the social media unique. Question answering has been particularly amenable to social media by directly connecting information seekers with the community members willing to share the information. Yahoo! Answers, with millions of users and hundreds of millions of answers for millions of questions is a very successful implementation of CQA.

For example, consider two example user-contributed questions, objective and subjective respectively:

Q1: What's the difference between chemotherapy and radiation treatments?

Q2: Has anyone got one of those home blood pressure monitors? and if so what make is it *and do you think they are worth getting*?

Figure 1 shows an example of community interactions in Yahoo! Answers around the question Q2 above. A user posted the question in the *Health* category of the site, and was able to obtain 10 responses from other users. Eventually, the asker chooses the best answer. Failing that, as shown in the example, the best answer can also be chosen according to the votes from other users. Many of the interactions depend on the perceived goals of the asker: if the participants interpret the question as subjective, they will tend to share their experiences and opinions, and if they interpret the question as objective, they may still share their experiences but may also provide more factual information.

2.2 Problem Definition

We now state our problem of question orientation more precisely. We consider question orientation from the perspective of user goals: authors of objective questions request authoritative, objective information (e.g., published literature or expert opinion), whereas authors of subjective questions seek opinions or judgments of other users in the community. We state our problem as follows.

> **Question Subjectivity Problem:** *Given a question Q in a question answering community, predict whether Q has objective or subjective orientation, based on question and answer text as well as the user and community feedback.*

3 *CoCQA*: A Co-Training Framework over Questions and Answers

In the CQA setting we could easily obtain thousands or millions of unlabeled examples from the online CQA archives. On the other hand, it is difficult to create a labeled dataset with a reasonable size, which could be used to train a perfect classifier to analyze questions from different domains and sub-domains. Therefore, semi-supervised learning (Chapelle et al., 2006) is a natural approach for this setting.

Intuitively, we can consider the text of the question itself or answers to it. In other words, we have multiple (at least two) natural views of the data, which satisfies the conditions of the co-training approach (Blum and Mitchell, 1998). In *co-training*, two separate classifiers are trained on two sets of features, respectively. By automatically labeling the unlabeled examples, these two classifiers iteratively "teach" each other by giving their partners a newly labeled data that they can predict with high confidence. Based on the original co-training algorithm in (Blum and Mitchell, 1998) and other implementations, we develop our algorithm *CoCQA* shown in Figure 2.

At Steps 1 and 2, the *K* examples are coming from different feature spaces, and each category (for example, *Subjective* and *Objective*) has top K_j most confident examples chosen, where K_j corresponds to the distribution of class in the current set of labeled examples L. *CoCQA* will terminate when the increments of both classifiers are less than a specified threshold *X* or the maximum number of iterations are exceeded. Following the co-training approach, we include the most confidently predicted examples as additional "labeled" data. The SVM output margin value was used to estimate confidence; alternative

```
Input:
    • $F_Q$ and $F_A$ are Question and Answer feature views
    • $C_Q$ and $C_A$ are classifiers trained on $F_Q$ and $F_A$ respectively
    • L is a set of labeled training examples
    • U is a set of unlabeled examples
    • $K$: Number of unlabeled examples to choose on each iteration
    • $X$: the threshold for increment
    • $R$: the maximal number of iterations
Algorithm CoCQA
1. Train $C_Q,0$ on L: $F_Q$, and record resulting $ACC_Q,0$
2. Train $C_A,0$ on L: $F_A$, and record resulting $ACC_A,0$
3. for j=1 to $R$ do:
       Use $C_Q,j-1$ to predict labels for U and choose
           top $K$ items with highest confidence $\rightarrow E_{Q,}, j-1$
       Use $C_A,j-1$ to predict labels for U and choose
           top $K$ items with highest confidence $\rightarrow E_{A,}, j-1$
       Move examples $E_{Q,}, j-1 \cup E_{A,}, j-1 \rightarrow$ L
       Train $C_Q,j$ on L: $F_Q$ and record training $ACC_Q,j$
       Train $C_A,j$ on L: $F_A$ and record training $ACC_A,j$
       if $Max(\Delta ACC_Q,j, \Delta ACC_A,j) < X$ break

4. return final classifiers $C_Q,j \rightarrow C_Q$ and $C_A,j \rightarrow C_A$
```

Figure 2: Algorithm CoCQA: A co-training algorithm for exploiting redundant feature sets in community question answering.

methods (including reliability of this confidence prediction) could further improve performance, and we will explore these issues in future work. Finally, the next question is how to estimate classification performance with training data. For each pass, we randomly split the original training data into *N* folds (N=10 in our experiments), and keep one part for validation and the rest, augmented with the newly added examples, as the expanded training set.

After *CoCQA* terminates, we obtain two classifiers. When a new example arrives, we will classify it with these two classifiers based on both of the feature sets, and combine the predictions of these two classifiers. We explored two strategies to make the final decision based on the confidence values given by two classifiers:

- Choose the class with higher confidence
- Multiply the confidence values, and choose the class that has the highest product.

We found the second heuristic to be more effective than the first in our experiments. As the base classifier we use SVM in the current implementation, but other classifiers could be incorporated as well.

4 Experimental Evaluation

We experiment with supervised and semi-supervised methods on a relatively large data set from Yahoo! Answers.

4.1 Datasets

To our knowledge, there is no standard dataset of real questions and answers posted by online users, labeled for subjectivity orientation. Hence, we had to create a dataset ourselves. To create our dataset, we downloaded more than 30,000 resolved questions from each of the following top-level categories of Yahoo! Answers: Arts, Education, Health, Science, and Sports. We randomly chose 200 questions from each category to create a raw dataset with 1,000 questions total. Then, we labeled the dataset with annotators from the Amazon's Mechanical Turk service[4].

For annotation, each question was judged by 5 Mechanical Turk workers who passed a qualification test of 10 questions (labeled by ourselves) with at least 9 of them correctly marked. The qualification test was required to ensure that the raters were sufficiently competent to make reasonable judgments. We grouped the tasks into 25 question batches, where the whole batch was submitted as the Mechanical Turk's Human Intelligence Task (HIT). The batching of questions was done to easily detect the "random" ratings produced by irresponsible workers. That is, each worker rated a batch of 25 questions.

While precise definition of subjectivity is elusive, we decided to take the practical perspective, namely the "majority" interpretation. The annotators were instructed to guess orientation according to how the question would be answered by most people. We did not deal with multi-part questions: if any part of question was subjective, the whole question was labeled as subjective. The gold standard was thus derived with the majority strategy, followed by manual inspection as a "sanity check". At this stage we removed 22 questions with undeterminable meaning, including gems such as "`Upward Soccer`

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

Shorts?"[5] and "1+1=?fdgdgdfg?"[6]. Finally, we create a labeled dataset consisting of 978 resolved questions, available online[7].

	Num. of SUB. Q	*Num. of OBJ. Q*	*Total Num.*	*Annotator agreement*
Arts	137 (70%)	58 (30%)	195	0.841
Education	127 (64%)	70 (36%)	197	0.716
Health	125 (64%)	69 (36%)	194	0.833
Science	103 (52%)	94 (48%)	197	0.618
Sports	154 (79%)	41 (21%)	195	0.877
Total	**646 (66%)**	**332 (34%)**	**978**	**0.777**

Table 1: Labeled dataset statistics.

Table 1 reports the statistics of the annotated dataset. The overall inter-annotator percentage agreement between Mechanical Turk workers and final annotation is 0.777, indicating that the task is difficult, but feasible for humans to annotate manually.

The statistics of our labeled sample show that the vast majority of the questions in all categories except for Science are subjective in nature. The relatively high ratio of subjective questions in the Science category is surprising. However, we find that users often post polemics and statements instead of questions, using CQA as a forum to share their opinions on controversial topics. Overall, we were struck by the expressed need in Subjective information, even for categories such as Health and Education, where objective information would intuitively seem more desirable.

4.2 Features Used in Experiments

For the subjectivied experiments to follow, we attempt to capture the linguistic characteristics identified in previous work (Section 5) in a lightweight and robust manner, due to the informal and noisy nature of CQA. In particular, we use the following feature classes, computed separately over question and answer content:

- Character 3-grams
- Words
- Word with Character 3-grams
- Word n-grams (n<=3, i.e. W_i, W_iW_{i+1}, $W_iW_{i+1}W_{i+2}$)
- Word and POS n-gram (n<=3, i.e. W_i, W_iW_{i+1}, $W_i\ POS_{i+1}$, POS_iW_{i+1} , POS_iPOS_{i+1}, etc.).

We use the character 3-grams features to overcome spelling errors and problems of ill-formatted or ungrammatical questions, and the POS information to capture common patterns across domains, as words, especially the content words, are quite diverse in different topical domains. For word and character 3-gram features, we consider two different versions: case-sensitive and case-insensitive. Case-insensitive features are assumed to be helpful for mitigating negative effects of ill-formatted text.

Moreover, we experimented with three term weighting schemes: Binary, TF, and TF*IDF. Term frequency (TF) exhibited better performance in our development experiments, so we use this weighting scheme for all the experiments in Section 4. Regarding features: both words and structure of the text (e.g., word order) can be used to infer subjectivity. Therefore, the features we employ, such as words and word n-grams, are expected to be useful as a (coarse) proxy to grammatical and phrase features. Unlike traditional work on news-like text, the text of CQA and has poor spelling, grammar, and heavily uses non-standard abbreviations, hence our decision to use character n-grams.

4.3 Experimental Setting

Metrics: Since the prediction on both subjective questions and objective questions is equally important, we use the **macro-averaged F1** measure as the evaluation metric. This is computed as the macro average of F1 measures computed for the *Subjective* and *Objective* classes individually. The **F1** measure for either class is computed as $\frac{2\cdot \text{Precision}\cdot \text{Recall}}{\text{Precision}+\text{Recall}}$.

Methods compared: We compare our approach with both the base supervised learning, as well as *GE*, a state-of-the-art semi-supervised method:

- ***Supervised***: we use the LibSVM implementation (Chang and Lin, 2001) with linear kernel.

[5] http://answers.yahoo.com/question/?qid=20060829074901AADBRJ4

[6] http://answers.yahoo.com/question/?qid=1006012003651

[7] Available at http://ir.mathcs.emory.edu/datasets/.

- ***GE***: This is a state-of-the-art semi-supervised learning algorithm, Generalized Expectation (GE), introduced in (McCallum et al., 2007) that incorporates model expectations into the objective functions for parameter estimation.
- ***CoCQA***: Our method (Section 3).

For semi-supervised learning experiments, we selected a random subset of 2,000 unlabeled questions for each of the topical categories, for the total of 10,000 unlabeled questions.

4.4 Experimental Results

First we report the performance of our *Supervised* baseline system with a variety of features, reporting the average results of 5-fold cross validation. Then we investigate the performance to our new *CoCQA* framework under a variety of settings.

4.4.1 Supervised Learning

Table 2 reports the classification performance for varying units of representation (e.g., question text vs. answer text) and the varying feature sets. We used case-insensitive features and TF (term frequency within the text unit) as feature weights, as these two settings achieved the best results in our development experiments. The rows show performance considering only the question text (**question)**, the best answer (**best_ans**), text of all answers to a question (**all_ans**), the text of the question and the best answer (**q_bestans**), and the text of the question with all answers (**q_allans**), respectively. In particular, using the words in the question alone achieves F1 of 0.717, compared to using words in the question and the best answers text (F1 of 0.695). For comparison, a naïve baseline that always guesses the majority class (*Subjective*) obtains F1 of 0.398.

With character 3-gram, our system achieves performance comparable with words as features, but combining them together does not improve performance. We observe a slight gain with more complicated features, e.g. word n-gram, or word and POS n-grams, but the gain is not significant, and hence not worth the increased complexity of the feature generation. Finally, combining question text with answer text does not improve performance.

Interestingly, the best answer itself is not as effective as the question for subjectivity analysis, nor is using all of the answers submitted. One possible reason is that approximately 40% of the best answers were chosen by the community and not the asker herself, are hence not necessarily representative of the asker intent.

Feature set / Unit	Char 3-gram	Word	Word+ Char 3-gram	Word n-gram (n<=3)	Word POS n-gram (n<=3)
question	0.700	**0.717**	0.694	0.716	0.720
best_ans	0.587	0.597	0.578	0.580	0.565
all_ans	0.603	0.628	0.607	0.648	0.630
q_bestans	0.681	**0.695**	0.662	0.687	0.712
q_allans	0.679	0.677	0.676	0.708	0.689
Naïve (majority class) baseline:				0.398	

Table 2. Performance of predicting question orientation on the mixed dataset with varying feature sets (Supervised).

Table 3 reports the supervised subjectivity classification performance for each question category with word features. The overall classification results are significantly lower compared to training and testing on the mixture of the questions drawn from all categories, likely caused by the small amount of labeled training data for each category. Another possibility is that the subjectivity clues are not topical and hence are not category dependent, with the possible exception of the questions in the Health domain.

Category	Arts	Edu.	Health	Science	Sports
F1	0.448	0.572	**0.711**	0.647	0.441

Table 3. Experiment results on sub-categories with supervised SVM (q_bestans features).

As words are simple and effective features in this experiment, we will use them in the subsequent experiments. Furthermore, the feature set using the words in the question with best answer together (**q_bestans**) exhibit higher performance than question with all answers (**q_allans**). Thus, we will only consider questions and best answers in the following experiments with GE and *CoCQA*.

4.4.2 Semi-Supervised Learning

We now focus on investigating the effectiveness of *CoCQA*, our co-training-based framework for community question answering analysis. Table 4 summarizes the main results of this section. The values for *CoCQA* are derived with the parameter settings: K=100, *X*=0.001. These optimal settings are chosen after comprehensive experiments with different combinations, described later in this section. *GE* does not exhibit a significant improvement over *Supervised*. In contrast, *CoCQA* performs significantly better than the purely supervised method, with F1 of 0.745 compared to the F1 of 0.717 for *Supervised*. While it may seem surprising that a semi-supervised method outperforms a supervised one, note that we use all of the available *labeled* data as provided to the Supervised method, as well as a large amount of *unlabeled* data, that is ultimately responsible for the performance improvement.

Features Method	Question	Question+ Best Answer
Supervised	0.717	0.695
GE	0.712 (-0.7%)	0.717 (+3.2%)
CoCQA	0.731 (+1.9%)	**0.745** (+7.2%)

Table 4. Performance of *CoCQA*, *GE*, and *Supervised* with the same feature and data settings.

As an added advantage, *CoCQA* approach is also practical. In a realistic application, we have two different situations: offline and online. With online processing, we may not have best answers available to predict question's orientation, whereas we can employ information from best answers in offline setting. Co-training is a solution that is applicable to both situations. With *CoCQA*, we have two classifiers using the question text and the best answer text, respectively. We can use both of them to obtain better results in the offline setting, while in online setting, we can use the text of the question alone. In contrast, *GE* may not have this flexibility.

We now analyze the performance of *CoCQA* under a variety of settings to derive optimal parameters and to better understand the performance. Figure 3 reports the performance of *CoCQA* with varying the *K* parameter from 20 to 200. In this experiment, we fix *X* to be 0.001. The combination of question and best answer is superior to that of question and all answers. When *K* is 100, the system obtains the best result, 0.745.

Figure 4 reports the number of co-training iterations needed to converge to optimal performance. After 13 iterations (and 2500 unlabeled examples added), *CoCQA* achieves optimal performance, and eventually terminates after an additional 3 iterations. While a validation set should have been used for *CoCQA* parameter tuning, Figures 3 and 4 indicate that *CoCQA* is not sensitive to the specific parameter settings. In particular, we observe that any *K* is greater than 100, and for any number of iterations *R* is greater than 10, *CoCQA* exhibits in effectively equivalent performance.

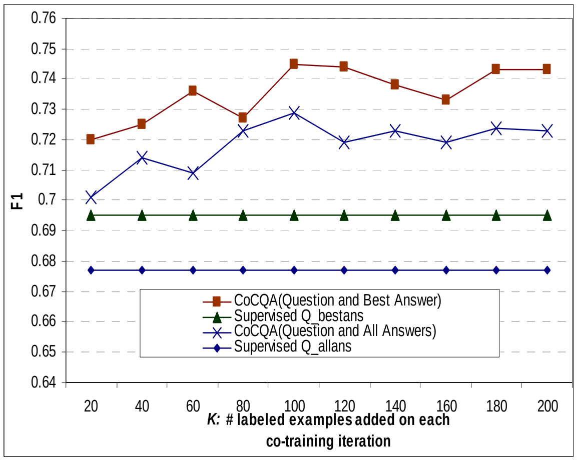


Figure 3: Performance of CoCQA for varying the *K* (number of examples added on each iteration of co-training).

Figure 5 reports the performance of *CoCQA* for varying the number of labeled examples from 50 to 400 (that is, up to 50% of the available labeled training data). Note that for this comparison we use the same feature sets (words in question and best answer text), but using only the (varying) fractions of the manually labeled data. Surprisingly, *CoCQA* exhibits comparable performance of F1=0.685 with only 200 labeled examples are used, compared to the F1=0.695 for *Supervised* with all 800 labeled training examples on this feature set. In other words, *CoCQA* is able to achieve comparable performance to supervised SVM with only one quarter of the labeled training data.

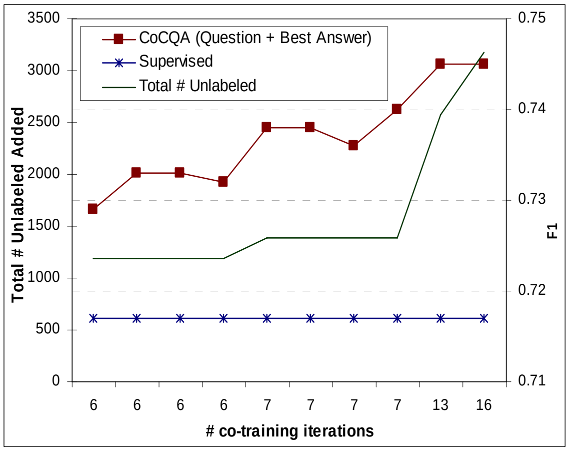


Figure 4: Performance and the *total* number of unlabeled examples added for varying number of co-training iterations (*K*=100, using *q_bestans* features)

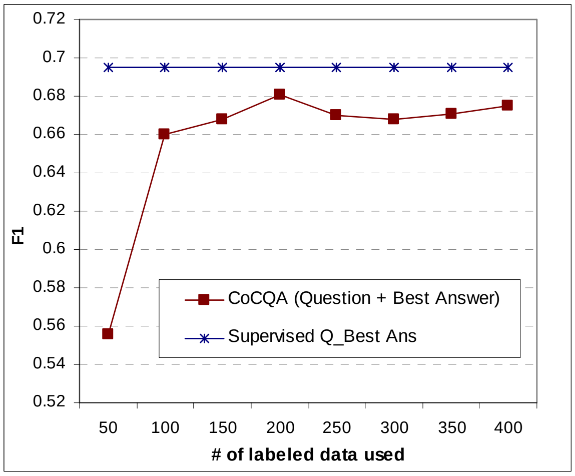


Figure 5: Performance of *CoCQA* with varying number of labeled examples used, compared to *Supervised* method, on same features.

5 Related Work

Question analysis, especially question classification, has been long studied in the question answering research community. However, most of the previous research primarily considered factual questions, with the notable exception of the most recent TREC opinion QA track. Furthermore, the questions were specifically designed for benchmark evaluation. A related thread of research considered deep analysis of the questions (and corresponding sentences) by manually classifying questions along several orientation dimensions, notably (Stoyanov et al., 2005). In contrast, our work focuses on analyzing real user questions posted in a question answering community. These questions are often complex or subjective, and are typically difficult to answer automatically as the question author probably was not able to find satisfactory answers with quick web search.

Automatic complex question answering has been an active area of research, ranging from simple modification to factoid QA techniques (e.g., Soricut and Brill, 2003) to knowledge intensive approaches for specific domains (e.g., Harabagiu et al. 2001, Fushman and Lin 2007). However, the technology does not yet exist to automatically answer open-domain complex and subjective questions. While there has been some recent research (e.g., Agichtein et al. 2008, Bian et al. 2008) on retrieving high quality answers from CQA archives, the subjectivity orientation of the questions has not been considered as a feature for ranking.

A related corresponding problem is complex QA evaluation. Recent efforts at automatic evaluation show that even for well-defined, objective, complex questions, evaluation is extremely labor-intensive and introduces many challenges (Lin and Fushman 2006, Lin and Zhang 2007). As part of our contribution we showed that it is feasible to use the Amazon Mechanical Turk service for evaluation by combining large degree of annotator redundancy (5 annotators per question) with more sparse but higher-quality expert annotation.

The problem of automatic subjective question answering has recently started to be addressed in the question answering community, most recently as the first opinion QA track in (Dang et al., 2007). Unlike the controlled TREC opinion track (introduced in 2007), many of the questions in Yahoo! Answers community are inherently subjective, complex, ill-formed, or all of the above. To our knowledge, this paper is the first large-scale study of subjective/objective orientation of information needs, and certainly the first in the CQA environment.

A closely related research thread is subjectivity analysis at document and sentence level. For example, reference (Yu, H., and Hatzivassiloglou, V. 2003; Somasundaran et al. 2007) attempted to classify sentences into those reporting facts or opinions. Also related is research on sentiment analysis (e.g., Pang et al., 2004) where the goal is to classify a sentence or text fragment as being overall positive or negative. More generally, (Wiebe et al. 2004) and subsequent work focused on the analysis of subjective language in narrative text, primarily news. Our problem is quite different in the sense that we are trying to identify the orientation of a *question*. Nevertheless, our baseline method is similar to the methods and features used for sentiment analysis, and one of our contributions is evaluating the usefulness of the established features and techniques to the new CQA setting.

In order to predict question orientation, we build on co-training, one of the known semi-supervised learning techniques. Many models and techniques have been proposed for classification, including support vector machines, decision tree based techniques, boosting-based techniques, and many others. We use LIBSVM (Chang and Lin, 2001) as a robust implementation of SVM algorithms.

In summary, while we draw on many techniques in question answering, natural language processing, and text classification, our work differs from previous research in that a) develop a novel co-training based algorithm for question and answer classification; b) we address a relatively new problem of *automatic* question subjectivity prediction; c) demonstrate the effectiveness of our techniques in the new CQA setting and d) explore the characteristics unique to CQA – while showing good results for a quite difficult task.

6 Conclusions

We presented *CoCQA*, a co-training framework for modeling the textual interactions in question answer communities. Unlike previous work, we have focused on real user questions (often noisy, ungrammatical, and vague) submitted in Yahoo! Answers, a popular community question answering portal. We demonstrated *CoCQA* for one particularly important task of automatically identifying question subjectivity orientation, showing that *CoCQA* is able to exploit the structure of questions and corresponding answers. Despite the inherent difficulties of subjectivity analysis for real user questions, we have shown that by applying *CoCQA* to this task we can significantly improve prediction performance, and substantially reduce the size of the required training data, while outperforming a general state-of-the-art semi-supervised algorithm that does not take advantage of the CQA characteristics.

In the future we plan to explore more sophisticated features such semantic concepts and relationships (e.g., derived from WordNet or Wikipedia), and richer syntactic and linguistic information. We also plan to explore related variants of semi-supervised learning such as co-boosting methods to further improve classification performance. We will also investigate other applications of our co-training framework to tasks such as sentiment analysis in community question answering and similar social media content.

Acknowledgments

This research was partially supported by the Emory University Research Committee (URC) grant, and by the Emory College Seed grant. We thank the Yahoo! Answers team for providing access to the Answers API, and anonymous reviewers for their excellent suggestions.

References

Agichtein, E., Castillo, C., Donato, D., Gionis, A., and Mishne, G. 2008. Finding High-Quality Content in Social Media with an Application to Community-Based Question Answering. *WSDM2008*

Bian, J., Liu, Y., Agichtein, E., and H. Zha. 2008, to appear. Finding the Right Facts in the Crowd: Factoid Question Answering over Social Media, *Proceedings of the Inter-national World Wide Web Conference (WWW), 2008*

Blum, A., and Mitchell, T. 1998. Combining Labeled and Unlabeled Data with Co-Training. *Proc. of the Annual Conference on Computational Learning Theory*.

Chang, C. C. and Lin, C. J. 2001. LIBSVM : a library for support vector machines. Software available at *http://www.csie.ntu.edu.tw/~cjlin/libsvm.*

Chapelle, O., Scholkopf, B., and Zien, A. 2006. *Semi-supervised Learning*. The MIT Press, Cambridge, Mas-sachusetts.

Dang, H. T., Kelly, D., and Lin, J. 2007. Overview of the TREC 2007 Question Answering track. *In Proceedings of TREC-2007.*

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

Harabagiu, S., Moldovan, D., Pasca, M., Surdeanu, M. , Mihalcea, R., Girju, R., Rusa, V., Lacatusu, F., Morarescu, P., and Bunescu, R. 2001. Answering Complex, List and Context Questions with LCC's Question-Answering Server. *In Proc. of TREC 2001*.

Lin, J. and Demner-Fushman, D. 2006. Methods for automatically evaluating answers to complex questions. *In-formation Retrieval*, 9(5):565–587

Lin, J. and Zhang, P. 2007. Deconstructing nuggets: the stability and reliability of complex question answering evaluation. *In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pages 327–334.*

Mann, G., and McCallum, A. 2007. Simple, Robust, Scalable Semi-supervised Learning via Expectation Regularization. *Proceedings of ICML 2007*.

Pang, B., and Lee, L. 2004. A Sentimental Education: Sen-timent Analysis Using Subjective Summarization Based on Minimum Cuts. *In Proc. of ACL.*

Prager, J. 2006. Open-Domain Question-Answering. *Foundations and Trends in Information Retrieval*.

Sindhwani, V., Keerthi, S. 2006. Large Scale Semi-supervised Linear SVMs. *Proceedings of SIGIR 2006*.

Somasundaran, S., Wilson, T., Wiebe, J. and Stoyanov, V. 2007. QA with Attitude: Exploiting Opinion Type Analysis for Improving Question Answering in Online Discussions and the News. *In proceedings of International Conference on Weblogs and Social Media (ICWSM-2007)*.

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

Stoyanov, V., Cardie, C., and Wiebe, J. 2005. Multi-Perspective question answering using the OpQA corpus. In Proceedings of EMNLP.

Tri, N. T., Le, N. M., and Shimazu, A. 2006. Using Semi-supervised Learning for Question Classification. *In Proceedings of ICCPOL-2006*.

Wiebe, J., Wilson, T., Bruce R., Bell M., and Martin M. 2004. Learning subjective language. *Computational Linguistics*, 30 (3).

Yu, H., and Hatzivassiloglou, V. 2003. Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences. *In Proceedings of EMNLP-2003.*

Zhang, D., and Lee, W.S. 2003. Question Classification Using Support Vector Machines. *Proceedings of the 26th Annual International ACM SIGIR Conference on Re-search and Development in Information Retrieval*.

Zhu, X. 2005. Semi-supervised Learning Literature Survey. *Technical Report 1530*, Computer Sciences, University of Wisconsin-Madison.