Understanding User Intent in Community Question Answering

Long Chen
DCSIS
Birkbeck, University of London
Malet Street
London WC1E 7HX, UK
long@dcs.bbk.ac.uk

Dell Zhang
DCSIS
Birkbeck, University of London
Malet Street
London WC1E 7HX, UK
dell.z@ieee.org

Mark Levene
DCSIS
Birkbeck, University of London
Malet Street
London WC1E 7HX, UK
mark@dcs.bbk.ac.uk

ABSTRACT

Community Question Answering (CQA) services, such as Yahoo! Answers, are specifically designed to address the innate limitation of Web search engines by helping users obtain information from a community. Understanding the user intent of questions would enable a CQA system identify similar questions, find relevant answers, and recommend potential answerers more effectively and efficiently. In this paper, we propose to classify questions into three categories according to their underlying user intent: subjective, objective, and social. In order to identify the user intent of a new question, we build a predictive model through machine learning based on both text and metadata features. Our investigation reveals that these two types of features are conditionally independent and each of them is sufficient for prediction. Therefore they can be exploited as two views in co-training — a semi-supervised learning framework — to make use of a large amount of unlabelled questions, in addition to the small set of manually labelled questions, for enhanced question classification. The preliminary experimental results show that co-training works significantly better than simply pooling these two types of features together.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*data mining*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; I.2.6 [**Artificial Intelligence**]: Learning; I.5.2 [**Pattern Recognition**]: Design Methodology—*classifier design and evaluation*

General Terms

Algorithms, Experimentation, Performance

Keywords

Community Question Answering, User Intent, Semi-Supervised Learning, Co-Training

1. INTRODUCTION

A number of Community Question Answering (CQA) services have emerged and become popular in the last decade. Typical examples include Yahoo! Answers[1], Baidu Zhidao[2], Quora[3], Facebook Questions[4], and also domain-specific forums like Stack Overflow[5]. These services help users obtain information from a community — a user can post his or her questions which may then be answered by other users. Such a paradigm of information seeking is particularly appealing when the user's information need cannot be satisfied directly by Web search engines due to the unavailability of relevant online content. Furthermore, compared with computers, humans have a better ability of understanding natural language so other users are often able to give more pertinent and comprehensive results to complex information needs expressed as natural language questions than Web search engines are.

Although CQA services have been booming recently, there is still a large margin for improvement [1, 16]. We concentrate on understanding the user intent of questions which would enable a CQA system identify similar questions, find relevant answers, and recommend potential answerers more effectively and efficiently, and thus greatly increase its quality and usability. However, this is a very challenging task, and regarded as a long-term grand goal in Information Retrieval [12].

In this paper, we propose to classify questions into three categories according to their underlying user intent: subjective, objective, and social. In order to identify the user intent of a new question, we build a predictive model through machine learning based on both text and metadata features. Our investigation reveals that these two types of features are conditionally independent, and each of them is sufficient for prediction, therefore they can be exploited as two views in co-training [2] — a semi-supervised learning framework — to make use of a large amount of unlabelled questions, in addition to the small set of manually labelled questions, for enhanced question classification. The preliminary experimental results show that co-training works significantly better than simply pooling these two types of features together.

Our investigation is carried out on the *Yahoo! Answers Comprehensive Questions and Answers (v1.0)* corpus that is kindly provided to research communities by Yahoo! Re-

WWW 2012 Companion, April 16-20, 2012, Lyon, France
ACM 978-1-4503-1230-1/12/04.

[1] http://answer.yahoo.com/
[2] http://zhidao.baidu.com/
[3] http://www.quora.com/
[4] http://www.facebook.com/questions
[5] http://stackoverflow.com/

search through their Webscope[6] programme. The original corpus consists of 4,483,032 questions and their corresponding answers. We randomly selected 1,539 questions from this large corpus, and manually labelled each of them by its user intent (see Section 6).

The rest of this paper is organised as follows. In Section 2, we review the related work. In Section 3, we define our taxonomy of user intent in CQA. In Section 4, we investigate the usefulness of text and metadata features for identifying the user intent of questions. In Section 5, we introduce the co-training approach to question classification according that can make use of both labelled data and unlabelled data. In Section 6, we describe the experimental setup and present the experiment results. In Section 7, we make conclusions and discuss the future work.

2. RELATED WORK

The problem of understanding user intent has been extensively studied in the context of Web search engines, starting from Broder's seminal work [4] that puts the intent of Web search queries into three categories: informational, navigational, and transactional. However, such a taxonomy cannot be directly applied to CQA due to the different expectations within people's mind-sets: in CQA users normally ask natural language questions which are addressed to human beings, whereas in Web search users submit keyword queries which are addressed to automated search engines. More specifically, this leads to the following two major differences between CQA questions and Web search queries. First, many CQA questions are inherently subjective. It has been shown that the proportion of Yahoo! Answers oriented to factual question answering is decreasing while subjective/complex question answering is gradually increasing [11]. This is understandable as Web search engines usually handle the former to a satisfactory degree but have difficulties with the latter. Second, many CQA questions are socially motivated, as users know that the answers to their questions come from other users. Instead of satisfying an information need, such questions are actually about establishing social connections (e.g., finding a date), or about generating some empathy (e.g., complaining), or just for entertainment purposes (e.g. telling jokes).

The subjective intent of questions has been investigated. For example, in TREC competitions, subjective/complex question answering started to be addressed in the opinion QA track from 2007 [6]. The work most similar to ours is [9, 10] in which Li et al. use supervised and semi-supervised machine learning methods to predict the subjectivity orientation of questions, i.e., whether a user is seeking objective or subjective information. However, their proposed approach relies on features extracted from both questions and their corresponding answers, therefore it can only be used to classify questions that have already been answered. In contrast, our approach aims to classify questions instantly once they are asked so only features extracted from questions are used. Thus a CQA system can identify a new question's underlying user intent through our approach and furthermore exploit it to improve the question answering process (e.g., in finding similar questions or relevant answers).

The social content of questions has also received some attention from researchers. Liu et al. have extended Broder's taxonomy of Web search queries to include a social category for CQA questions [12]. However, as mentioned above, that taxonomy is not really suitable for CQA. For example, the navigational category in their study literally contains no questions at all. Rodrigues and Milic-Frayling has analysed the social vs. non-social intent of questions in CQA [15], but their definition for social intent is quite different from ours as they mainly focus on defining measures of social engagement to characterise users' participation and contribution. Harper et al. have proposed to describe the user intent of questions in CQA as informational and conversational [8]. Their conversational category is somewhat similar to our social category.

More generally, a number of machine learning techniques have been applied to automatic question classification according to the expected answer types [19]. There have also been studies on how to motivate users to make more contributions [7] and encourage them to be more responsive [13] in online communities.

3. TAXONOMY

Taking into account the special characteristics of CQA, we propose the following taxonomy that classifies questions into three categories according to their underlying user intent: objective, subjective, and social.

They may want to seek for advices or bring up a discussion, or just complaint about something disappointing.

looking for advices (general/personal opinions).

Objective Questions The intent of such questions is to get factual knowledge about something. For example, the question "Which country in Africa that was colonized by France did assimilation policy succeed?" asks for specific details of a particular event. For another example, the question "how do I find the website for the brick township high school baseball team for this year 2006?" asks for the website address where the user can learn more details about a particular entity afterwards.

Subjective Questions The intent of such questions is to get personal opinions or general advices about something. look at answers on account of personal state, like opinions, advices or experience. For example, the question "Do you believe Canada's flag should be lowered for each soldier that dies in the service of their country?" asks for personal opinions about a topic which could be very different for different people. upbringing and background. For another example, the question "I am a Bangladeshi National girl and I came to USA on B1/B2 visa and now I would like to take admission pls adv?" asks for general advices on a complicated issue.

Social Questions The intent of such questions is not to seek information but to have social interactions with other users. For example, the question "i am 4m kolkata, india. any1 4m here want to be my frnd?gals or guys-no prob with that.betr if a teenagr.i'm 17" and the question "Any1 near Newyork city?" are trying to make friends. For another example, the question "why do people from the USA ask questions as if that is the only country on the web?" is probably trying to get some empathy from people with similar thoughts.

[6] http://webscope.sandbox.yahoo.com

The objective category in the above taxonomy refers to the traditional TREC-style questions, while incorporating both the subjective category and the social category simultaneously distinguishes it from existing taxonomies for CQA questions which only focus on one of them (see Section 2).

Most questions that we encounter in a CQA service can be classified into one of these three categories. However, it is possible to see ambiguous questions. For example, the question "What type of careers are in southeast asia?" could either be interpreted as objective (asking for career facts) or subjective (asking for career advices). After careful inspection of the dataset, we conclude that such questions constitute less than 2% of all questions, so we ignore them in this paper.

Although examining the answers to a question usually helps to infer its user intent accurately, we prefer to utilise the question alone because only by predicting the user intent of a question before it receives answers could we exploit the user intent to enhance the question answering process in CQA services.

4. FEATURES

4.1 Text Features

The text features of a question are extracted from the bag-of-words content of the question title after standard pre-processing steps (tokenization, lower-casing, stopword-removal, and stemming) [14]. Finally each question is represented as a vector of terms weighted by TF×IDF [14].

To have a rough idea about each category of questions, we sort all word/phrase features in terms of information gain for question classification, and show the most discriminative ones in Table 1. It seems that questions with those 5-w words (who, when, where, what, why) are more likely to have an objective intent, whereas questions with polite words and conversational phrases are more likely to have a subjective or social intent. This suggests that text features have relatively more discriminative power to separate objective questions from subjective or social questions.

4.2 Metadata Features

Moreover, we have also identified several metadata features that can be used to help detecting the user intent of questions.

4.2.1 Question Topic

Figure 1 shows the distribution of user intent over the top-10 question topic categories (all questions posted in Yahoo! Answers are annotated by their topic categories). It seems that objective and subjective questions have similar proportion of presence in most topic categories, except for "Arts & Humanities" which contains many questions about history and genealogy etc. The distribution of social questions seems to be quite different: most social questions are about topics like "Family Relationships", "News Events", and "Entertainment & Music" on which people may be more inclined to have social interaction. This suggests that question topic features have relatively more discriminative power to separate social questions from objective or subjective questions.

4.2.2 Question Time

Figure 2 shows the distribution of user intent over the time (hour-of-the-day) when the question was asked on 1st May 2006. It seems that objective and subjective questions do not have apparent differences in terms of question time. In contrast, social questions show interesting patterns: the peak time for social questions is at 18:00 (finishing the daytime work), 15:00 (after lunch), and 03:00 (lonely in the late night). This suggests that question time features have relatively more discriminative power to separate social questions from objective or subjective questions.

Table 1: The most discriminative text features for each category of questions.

intent	text feature	information gain
objective	`anyone`	0.096
	`what's your`	0.087
	`who is`	0.054
	`why is`	0.054
	`what is`	0.044
subjective	`is your`	0.036
	`help`	0.026
	`can I`	0.014
	`favourite`	0.011
	`how do`	0.009
social	`anybody`	0.042
	`is there`	0.035
	`looking for`	0.028
	`do I`	0.028
	`I am`	0.011

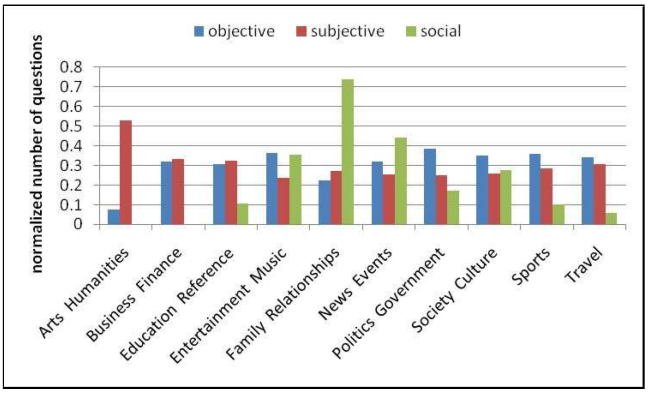

Figure 1: The question topic feature.

4.2.3 Question Asker Experience

Figure 3 shows the distribution of user intent over the question asker's experience, i.e., the number of questions the user has asked before. It seems that subjective and social questions are more likely to come from experienced users than new users, probably because experienced users recognise that the strength of CQA is in subjective or social questions but not objective questions in comparison with Web search engines. This suggests that question asker experience features have relatively more discriminative power to separate objective questions from subjective or social questions.

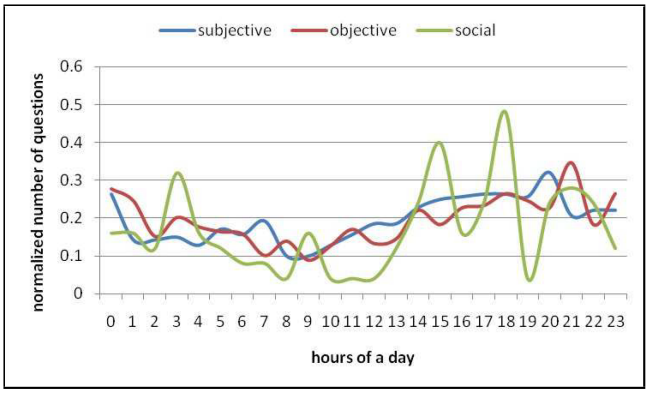

Figure 2: The question time feature.

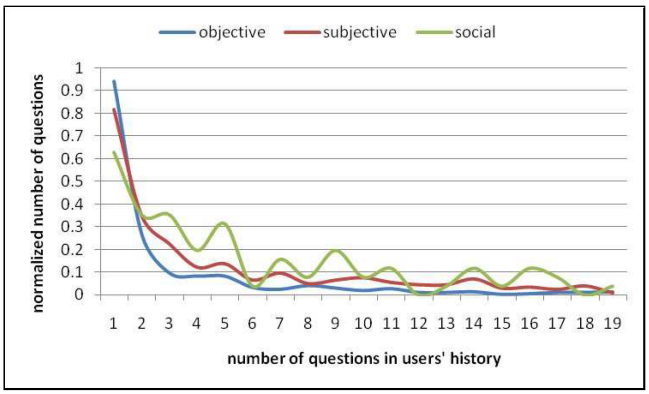

Figure 3: The question asker experience feature.

5. APPROACH

It is time-consuming and error-prone to manually label questions according to their user intent. Usually we can only have a small set of labelled questions, which would seriously limit the success of supervised learning for question classification. However, obtaining unlabelled questions are quite easy and cheap. So it is promising to apply semi-supervised learning [5] here which can make use of a large amount of unlabelled data in addition to the small set of labelled data.

There are many semi-supervised learning techniques available. For this problem of question classification according to user intent, we think the co-training [2] approach is particularly suitable. Basically, co-training is a semi-supervised learning framework that requires two views of the data: each example is described by two different feature sets (views) that provide different, complementary information about it. In the ideal situation, the two views are conditionally independent (given the class) and each view is sufficient (to be used for classification alone). The main steps of co-training are as follows. It first learns a separate classifier for each view from the labelled data, and then the most confident predictions of each classifier on the unlabelled data are used to construct additional labelled training examples. This process iterates again and again until a stopping criterion is met.

As we have pointed out in Section 4, the text and metadata features are both effective in detecting the user intent of questions but with quite different discriminative powers for different question categories, therefore they can be considered as the two views for co-training.

Table 2: The dataset for experiments.

data	objective	subjective	social	*total*
training	503	442	70	1015
testing	259	228	37	524
all	762	670	107	1539

6. EXPERIMENTS

6.1 Dataset

Table 2 shows the statistics about the dataset for experiments. It consists of 1,539 questions that are randomly selected from the original Yahoo! Answers dataset and manually labelled accoring to their user intent. The dataset is split into training and testing sets with a proportion of 2:1.

6.2 Performance Measure

Since the class sizes are imbalanced in this problem, we use the F_1 score [14] instead of accuracy to measure the performance of question classification. The F_1 score is the harmonic mean of precision P and recall R: $F_1 = \frac{2PR}{P+R}$, where $P = \frac{\text{true positive}}{\text{true positive + false positive}}$, $R = \frac{\text{true positive}}{\text{true positive + false negative}}$. Furthermore, both micro-averaged F_1 (miF_1) and macro-averaged F_1 (maF_1) [18] will be reported in the next section. The former carries out averaging over all test questions while the latter over all question categories, therefore the former is dominated by performance on major question categories while the latter treats all question categories equally.

6.3 Results

A number of machine learning algorithms implemented in Weka[7], including C4.5, Random Forest, Naive Bayes, k-Nearest-Neighbours, and Linear Support Vector Machine (SVM), have been tried out for both supervised learning and semi-supervised learning (co-training). The one-vs-rest ensemble scheme was used to achieve multi-class classification. Linear SVM kept to deliver the best classification performance in our experiments, so we only report its results here.

6.3.1 *Supervised Learning*

Table 3 shows the performance (miF_1) of question classification through supervised learning with different sets of features. The Linear SVM parameters are set to their default values except that the class weights are optimised for each question category by 5-fold cross-validation.

It is obvious that using both text features and metadata features works better than using either kind of features alone, for all question categories.

The performance improvement brought by using metadata features in addition to text features for supervised learning is statistically significant ($P < 0.025$), according to the micro sign test (s-test) [18].

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

Table 3: The performance of supervised learning with different sets of features.

features	objective	subjective	social
text	0.693	0.689	0.152
metadata	0.609	0.642	0.378
text+metadata	0.731	0.693	0.412

Table 4: The performance of supervised learning vs semi-supervised learning (co-training).

approach	miF_1	maF_1
supervised (text+metadata)	0.712	0.510
co-training (text+metadata)	0.757	0.534

6.3.2 Semi-Supervised Learning

Table 4 shows the performances (miF_1 and maF_1) of question classification through supervised learning and also semi-supervised learning (co-training) based on both text and metadata features. The Linear SVM parameters are set as in supervised learning, while the co-training algorithm parameters are tuned to their optimal values via 5-fold cross-validation.

It is obvious that the co-training approach that regards text features and metadata features as two views works better than the supervised learning approach that simply pooling these two types of features together. This is probably because co-training, as a semi-supervised learning method, can make use of a large amount of unlabelled questions in addition to the small set of labelled questions.

The performance improvement brought by using unlabelled data in addition to labelled data through co-training rather than simply combining text and metadata features together is statistically significant ($P < 0.005$), according to the micro sign test (s-test) [18].

Figure 4 shows the performance of co-training over iterations with the optimal incremental size. For miF_1, the optimal performance is achieved at the 13th iteration (with 260 unlabelled questions being added to the training set each round). For maF_1, the optimal performance is achieved at the 25th iteration (with 150 unlabelled questions being added to the training set each round). Choosing a smaller incremental size could lead to a better performance, but meanwhile it would require more iterations and thus be less efficient.

Figure 5 shows the performance of co-training vs supervised learning with varying number of labelled questions available. It can be seen that co-training consistently outperforms supervised learning with a substantial gap for miF_1, though there is no clear winner for maF_1. Furthermore, co-training only needs about 30% of labelled questions to reach the same miF_1 performance as supervised learning.

7. CONCLUSIONS

The main contribution of this paper is threefold. First, we propose a taxonomy of user intent in CQA that incorporates both subjective/objective and informational/social perspectives. Second, we identify several metadata features which can be used together with standard text features by machine learning algorithms to classify questions according to their underlying user intent. Third, we demonstrate that it is better to exploit both text features and metadata features through the semi-supervise learning framework, co-training, rather than simply combining them in supervised learning, since the former can make use of a large amount of unlabelled data. For future work, we plan to expand question text using a translational language model learned from question/answer pairs [17], and employ more sophisticated semi-supervised learning algorithms such as co-EM support vector learning [3].

8. REFERENCES

[1] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. Finding high-quality content in social media. In *Proceedings of the 1st ACM International Conference on Web Search and Data Mining (WSDM)*, pages 183–194, Palo Alto, CA, USA, 2008.

[2] A. Blum and T. M. Mitchell. Combining labeled and unlabeled sata with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory (COLT)*, pages 92–100, Madison, WI, USA, 1998.

[3] U. Brefeld and T. Scheffer. Co-em support vector learning. In *Proceedings of the 21st International Conference on Machine Learning (ICML)*, pages 121–128, Banff, Alberta, Canada, 2004.

[4] A. Z. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.

[5] O. Chapelle, B. Scholkopf, and A. Zien, editors. *Semi-Supervised Learning*. MIT Press, 2005.

[6] H. T. Dang, D. Kelly, and J. J. Lin. Overview of the TREC 2007 question answering track. In *Proceedings of The Sixteenth Text REtrieval Conference (TREC)*, Gaithersburg, MD, USA, 2007.

[7] F. M. Harper, S. X. Li, Y. Chen, and J. A. Konstan. Social comparisons to motivate contributions to an online community. In *Proceedings of the 2nd International Conference on Persuasive Technology (PERSUASIVE)*, pages 148–159, Palo Alto, CA, USA, 2007.

[8] F. M. Harper, D. Moy, and J. A. Konstan. Facts or friends?: Distinguishing informational and conversational questions in social Q&A sites. In *Proceedings of the 27th International Conference on Human Factors in Computing Systems (CHI)*, pages 759–768, Boston, MA, USA, 2009.

[9] B. Li, Y. Liu, and E. Agichtein. Cocqa: Co-training over questions and answers with an application to predicting question subjectivity orientation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 937–946, Honolulu, HI, USA, 2008.

[10] B. Li, Y. Liu, A. Ram, E. V. Garcia, and E. Agichtein. Exploring question subjectivity prediction in community qa. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 735–736, Singapore, 2008.

[11] Y. Liu and E. Agichtein. On the evolution of the Yahoo! answers QA community. In *Proceedings of the 31st Annual International ACM SIGIR Conference on*

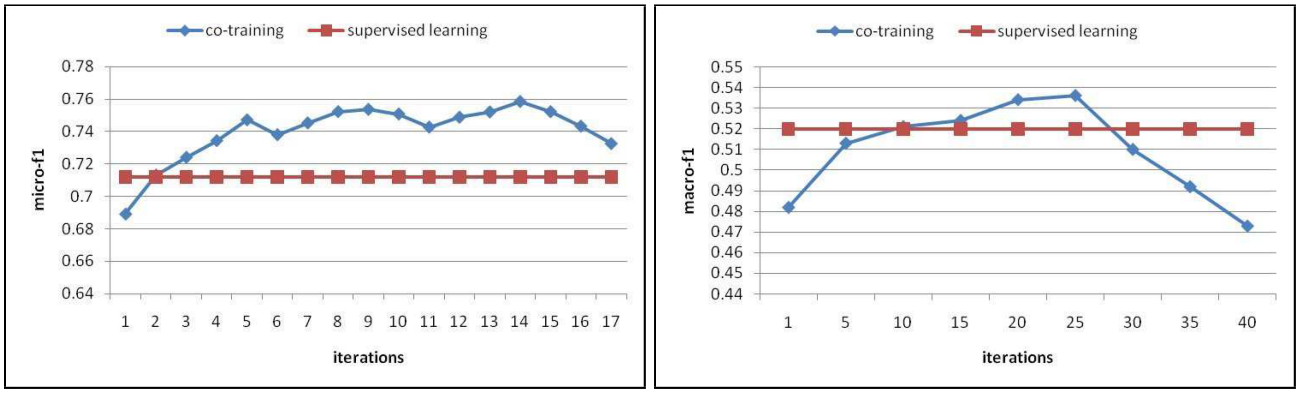

Figure 4: The performance of co-training over iterations with the optimal incremental size.

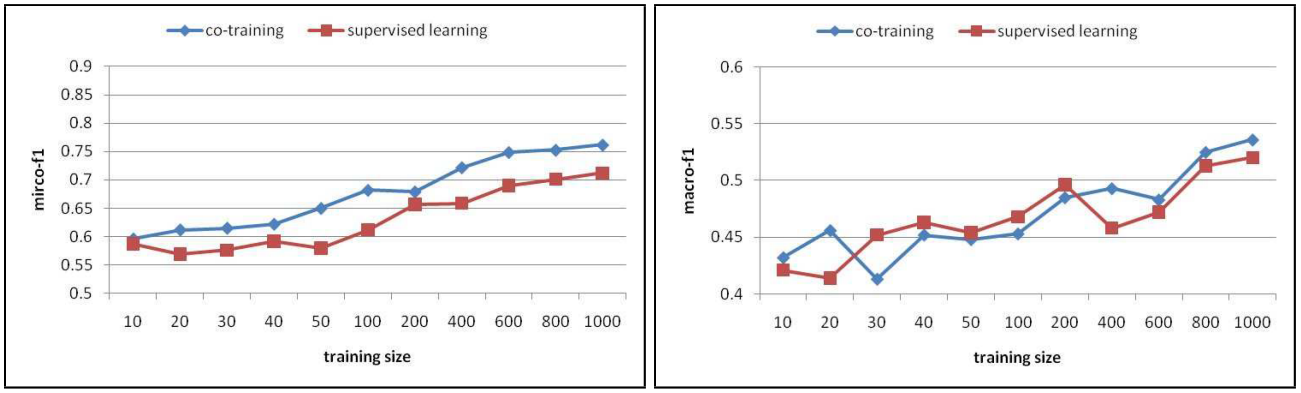

Figure 5: The performance of co-training vs supervised learning with varying number of labelled questions.

Research and Development in Information Retrieval (SIGIR), pages 737–738, Singapore, 2008.

[12] Y. Liu, S. Li, Y. Cao, C.-Y. Lin, D. Han, and Y. Yu. Understanding and summarizing answers in community-based question answering services. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING)*, pages 497–504, Manchester, UK, 2008.

[13] Y. Liu, N. Narasimhan, V. Vasudevan, and E. Agichtein. Is this urgent?: Exploring time-sensitive information needs in collaborative question answering. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 712–713, Boston, MA, USA, 2009.

[14] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[15] E. M. Rodrigues and N. Milic-Frayling. Socializing or knowledge sharing?: Characterizing social intent in community question answering. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM)*, pages 1127–1136, 2009.

[16] X.-J. Wang, X. Tu, D. Feng, and L. Zhang. Ranking community answers by modeling question-answer relationships via analogical reasoning. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 179–186, Boston, MA, USA, 2009.

[17] X. Xue, J. Jeon, and W. B. Croft. Retrieval models for question and answer archives. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 475–482, Singapore, 2008.

[18] Y. Yang and X. Liu. A re-examination of text categorization methods. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, pages 42–49, New York, NY, USA, 1999. ACM.

[19] D. Zhang and W. S. Lee. Question classification using support vector machines. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 26–32, Toronto, Canada, 2003.