Active Learning for Cross Language Text Categorization

Yue Liu, Lin Dai*, Weitao Zhou, and Heyan Huang

School of Computer Science and Technology
Beijing Institute of Technology, Beijing 100081, China
{kerwin,dailiu,zhouwt,hhy63}@bit.edu.cn

Abstract. Cross Language Text Categorization (CLTC) is the task of assigning class labels to documents written in a target language (e.g. Chinese) while the system is trained using labeled examples in a source language (e.g. English). With the technique of CLTC, we can build classifiers for multiple languages employing the existing training data in only one language, therefore avoid the cost of preparing training data for each individual language. One challenge for CLTC is the culture differences between languages, which causes the classifier trained on the source language doesn't perform well on the target language. In this paper, we propose an active learning algorithm for CLTC, which takes full advantage of both labeled data in the source language and unlabeled data in the target language. The classifier first learns the classification knowledge from the source language, and then learns the cultural dependent knowledge from the target language. In addition, we extend our algorithm to double viewed form by considering the source and target language as two views of the classification problem. Experiments show that our algorithm can effectively improve the cross language classification performance.

Keywords: Cross Language Text Categorization, Active Learning.

1 Introduction

Due to the explosive growth of electronic documents in different languages, there's an urgent need for effective multilingual text organizing techniques. Cross Language Text Categorization (CLTC) is the task of assigning class labels to documents written in a target language (e.g. Chinese) while the system is trained using labeled examples in a source language (e.g. English). With the technique of CLTC, we can build classifiers for multiple languages employing the existing training data in only one language, thereby avoiding the cost of preparing training data for each individual language.

The basic idea under CLTC is the documents in different languages may share the same semantic information [14], although they're in different representations. Previous works on CLTC have tried several methods to erase the language barrier and show promising results. However, despite language barrier, there's another problem for CLTC. That is the differences between cultures, which may cause

* Corresponding author.

P.-N. Tan et al. (Eds.): PAKDD 2012, Part I, LNAI 7301, pp. 195–206, 2012.

topic drift between languages. For example, news of the category *sports* from China (in Chinese) and US (in English) may concern different topics. The former may talk more about table tennis and Liu Xiang while the later may prefer NBA and NFL. As a result, even if the language barrier is perfectly erased, some knowledge of the target language still can't be learned from the training data in the source language. This will inevitably affect the performance of categorization. To solve this problem, making use of the unlabeled data in the target language will be helpful. Because these data is often easy to obtain and contains knowledge of the target language. If we can provide techniques to learn from it, the resulting classifier is expected to get more fit for the target language, thereby give better categorization performance.

In this paper, we propose an active learning algorithm for cross language text categorization. Our algorithm makes use of both labeled data in the source language and unlabeled data in the target language. The classifier first learns the classification knowledge from the source language, and then learns the cultural dependent knowledge from the target language. In addition, we extend our algorithm to double viewed form by considering the source and target language as two views of the classification problem. Experiments show that our algorithm can effectively improve the cross language classification performance. To the best of our knowledge, this is the first study of applying active learning to CLTC.

The rest of the paper is organized as follows. First, related works are reviewed in Section 2. Then, our active learning approach for CLTC is presented in Section 3 and its extension to double viewed form is introduced in Section 4. Section 5 presents the experimental results and analysis. Finally, Section 6 gives conclusions and future work.

2 Related Work

Several previous works have addressed the task of CLTC. [2] proposes practical approaches based on machine translation. In their work, two translation strategies are considered. The first strategy translates the training documents into the target language and the second strategy translates the unlabeled documents into the source language. After translation, monolingual text categorization is performed. [12] introduces a model translation method, which transfers classification knowledge across languages by translating the model features and takes into account the ambiguity associated with each word. Besides translation, in some other studies multilingual models are learned and used for CLTC, such as the multilingual domain kernels learned from comparable corpora [5] and the multilingual topic models mined from Wikipedia [8]. Moreover, there are also some studies of using lexical databases (e.g. WordNet) for CLTC [1].

All the previous methods have somehow solved the language barrier between training documents and unlabeled documents. But only a few have considered the culture differences between languages. In these works, authors try to solve this problem by employing some semi-supervised learning techniques. [12] employs a self-training process after the model translation. This process applies

the translated model to predict a set of unlabeled documents in target language and iteratively chooses the most confident classified documents to retrain the model. As a result, the model is adapted to better fit the target language. [9] proposes an EM based training algorithm. It consists of an initialization step that trains a classifier using translated labeled documents and an iteration step that repeats the E and M phases. In the E phase, the classifier predicts the unknown labels of a collection of documents in the target language. In the M phase, these documents with labels obtained in E phase are used to estimate the parameters of the new classifier. [16] investigates the use of co-training in cross language sentiment categorization. In their work, Chinese and English features are considered as two independent views of the categorization problem.

The common idea of above methods is to automatically label and use the documents in target language. To reduce the noises introduced by classification errors, the documents with low prediction certainty are usually underutilized. In this paper, we consider such documents to contain important information and will explore them through our active learning algorithm.

3 Active Learning for CLTC

3.1 Cross Language Text Categorization

Given a collection TR_e of labeled documents in language E and a collection TS_c of unlabeled documents in language C, in the scenario of CLTC, we would like to train a classifier using TR_e to organize the documents in TS_c. E and C is usually referred as the source language and target language respectively. In this paper, we suppose E is English and C is Chinese. In practice, they can be replaced with any other language pairs.

To solve the language barrier between training and test documents, we can employ a machine translation tool. The translation can be performed in two directions: the first direction translates all training documents into Chinese and the second direction translates all test documents into English. Both approaches convert the cross language problem into monolingual one. In this section, we choose the training set translation approach. First, we translate TR_e into Chinese, denoting it by $TR_{e\text{-}c}$. Then, we learn a classifier $C_{e\text{-}c}$ based on $TR_{e\text{-}c}$. Suppose the translation process gives accurate enough results, $C_{e\text{-}c}$ obtains the classification knowledge transferred from English.

$C_{e\text{-}c}$ can be applied to the unlabeled Chinese documents directly. Since the documents of same topic in different languages may share some common semantics, $C_{e\text{-}c}$ may be able to make very certain predications for some Chinese documents by the classification knowledge transferred from English. However, for some other documents, $C_{e\text{-}c}$ may get confused, as the class-discriminative information of these documents can't be detected. The latter case is usually caused by culture differences. For instance, a classifier trained using English *sports* samples may not be able to recognize Liu Xiang, a famous Chinese hurdle athlete, in a Chinese document. We can then make an observation that Chinese documents, which are uncertain to be classified by $C_{e\text{-}c}$, usually contain culture dependent

classification knowledge that can't be learnt from the translated training data. From this observation, we derive the active learning algorithm to improve $C_{e\text{-}c}$.

3.2 Apply Active Learning to CLTC

Active learning [11] is a form of learning algorithm for situations in which unlabeled data is abundant but labeling data is expensive. In such a scenario, the learner actively selects examples from a pool of unlabeled data and asks the teacher to label.

In the context of CLTC, we can assume an additional collection U_c of unlabeled documents in target language (Chinese in this paper) is available, since the unlabeled data is usually easy to obtain. Our algorithm consists of two steps. In the first step, we train a classifier using the translated training set $TR_{e\text{-}c}$, this classifier can be considered as an initial learner which has learnt the classification knowledge transferred from the source language. In the second step, we apply this classifier to the documents in U_c and select out the documents with lowest classification certainty. Such documents are expected to contain most culture dependent classification knowledge. We label them and put them into the training set. Consequently, the classifier is re-trained. The second step is repeated for several iterations, in order to let the classifier learn the culture dependent knowledge from the target language. Figure 1 illustrates the whole process.

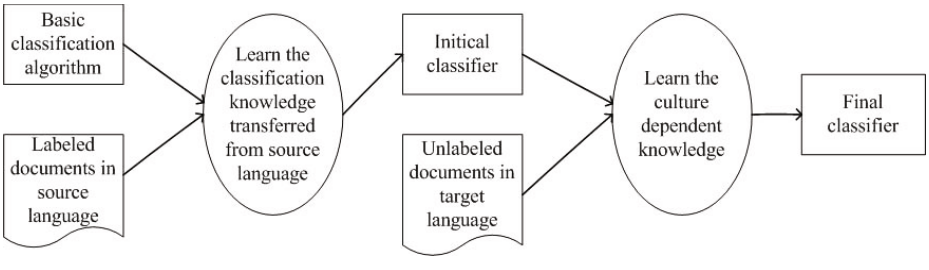

Fig. 1. The active learning process

In our approach, a basic classification algorithm is required to train the initial classifier. We employ support vector machine, as it has been well studied in previous work for active learning [15,6,11]. Note that our algorithm is independent on specific classification techniques.

Given the translated labeled set $TR_{e\text{-}c}$, each example can be represented as (x, y), where $x \in R^p$ is the feature vector and $y \in \{1, 2, \ldots k\}$ is the corresponding class label. A classifier learnt from $TR_{e\text{-}c}$ can predict the unknown class label for a document d in U_c. To measure the prediction certainty, we can refer to the membership probabilities of all possible classes.

However, SVM can't give probabilistic outputs directly. Some tricks have been proposed in [7]. For binary-class SVM, given the feature vector $x \in R^p$, and the

label $y \in \{-1, 1\}$, the membership probability $p(y = 1|x)$ can be approximated using a sigmoid function,

$$P(y = 1|x) = 1/(1 + exp(Af(x) + B)), \tag{1}$$

where $f(x)$ is the decision function of SVM, A and B are parameters to be estimated. Maximum likelihood estimation is used to solve for the parameters,

$$\begin{aligned} &\min_{(A,B)} -\sum_{i=1}^{l}(t_i \log p_i + (1 - t_i) \log (1 - p_i)), \\ &\qquad\qquad\qquad\qquad \text{where,} \\ &p_i = \frac{1}{1 + exp(Af(x_i) + B)}, \\ &t_i = \begin{cases} \frac{N_+ + 1}{N_+ + 2} \text{ if } y_i = 1; \\ \frac{1}{N_- + 2} \text{ if } y_i = -1. \end{cases} \end{aligned} \tag{2}$$

N_+ and N_- are the number of positive and negative examples in the training set. Newton's method with backtracking line search can be used to solve this optimization problem [7]. For multi-class SVM, we can obtain the probabilities through pair coupling [18]. Suppose that r_{ij} is the binary probability estimate of $P(y = i|y = i \text{ or } j, x)$, and p_i is the probability $P(y = i|x)$, the problem can be formulated as

$$\begin{aligned} &min_p \frac{1}{2} \sum_{i=1}^{k} \sum_{j, j \neq i} (r_{ji} p_i - r_{ij} p_j)^2, \\ &subject\ to \sum_{i=1}^{k} p_i = 1 \text{ and } p_i \geq 0, \forall i, \end{aligned} \tag{3}$$

where k denotes the number of classes. This optimization problem can be solved using a direct method such as Gaussian elimination, or a simple iterative algorithm [18].

In practice, we employ the toolbox $LibSVM$ [4], which is widely used in data mining tasks [13]. It implements the above methods for multi-class probability estimation. After obtaining the class membership probabilities of a document, we use the best against second best (BVSB) approach [6] to estimate the classification certainty. This approach has been demonstrated to be effective for multi class active learning task [6]. It measures the certainty by the difference between the probability values of the two classes having the highest estimated probabilities. The larger the difference, the higher the certainty is. Suppose c is the classifier, d is the document to be classified, i and j are the two classes with highest probabilities, then we calculate the certainty score using

$$Certainty(d, c) = P(y = i|d, c) - P(y = j|d, c). \tag{4}$$

Based on the discussions above, we describe the proposed algorithm in Algorithm 1.

Algorithm 1. Active learning algorithm for CLTC

Input:
The labeled set in the source language, TR_e;
The unlabeled set in the target language, U_c;
Output:
Classifier C;
1: Translate examples in TR_e into the target language, and denote the translated set by $TR_{\text{e-c}}$
2: Let $TrainingSet = TR_{\text{e-c}}$
3: Repeat I times:
4: Train classifier C using $TrainingSet$
5: Classify documents in U_c by C and measure the prediction certainty using Equation 4
6: Let S be a set of n documents with the lowest prediction certainty
7: Remove S from U_c
8: Label documents in S by the teacher
9: Add the newly labeled examples to $TrainingSet$
10: Return C

4 Double Viewed Active Learning

In this section, we extend our algorithm to double viewed form. In chief, the source and target language are considered as two views of the classification problem. The same idea was utilized in [16].

4.1 Two Views of the Problem

In Section 3, we convert the cross language problem into monolingual one with the help of a machine translation tool. As illustrated in Figure 2, the translation can be performed in two directions. The first direction translates the training set TR_e into Chinese and the second direction translates both the test set TS_c and additional unlabeled set U_c into English. Each direction gives us a monolingual view of the problem. In Section 3, we apply our active learning algorithm based on the Chinese view. In this section, we will show how to take advantage of both views and extend our algorithm to double viewed active learning.

First, we perform the translation following both directions. As a result, each document is associated with two views: the Chinese view and the English view. We denote the double viewed training set, test set and additional unlabeled set by TR, TS and U respectively. Then, two initial classifiers are trained using TR based on Chinese view and English view individually. We apply both of them to the unlabeled set U.

Since predictions made by the two classifiers are based on individual views of one document, they may have different certainties. As illustrated in Figure 3, the pool of unlabeled documents is split into four regions. In region C, the English classifier is certain on the documents, while the Chinese classifier is not. In region D, it's the opposite. For these scenarios, we can employ a co-training

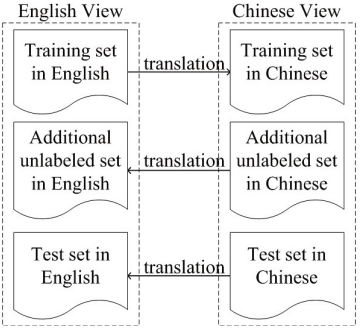

Fig. 2. Two directions of translation

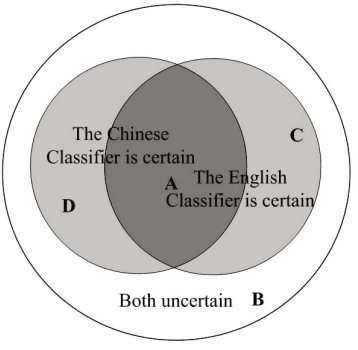

Fig. 3. Certainty distribution over the unlabeled documents

[3] approach, which labels documents according to the confident classifier and generate new training examples for the unconfident one. In other words, the two learners can teach each other in some times, needn't always ask the teacher. Based on this idea, we present the double viewed active learning algorithm in the next section.

4.2 Double Viewed Active Learning

Given a document d and two classifiers C_e and C_c, we measure whether both classifiers are certain about its prediction by the average certainty,

$$Average\_Certainty(d, C_e, C_t) = (Certainty(d, C_e) + Certainty(d, C_c))/2. \quad (5)$$

To measure whether a classifier is more certain than the other, we refer to the difference between their certainties,

$$Certainty\_Difference(d, C_e, C_c) = Certainty(d, C_e) - Certainty(d, C_c). \quad (6)$$

Our double viewed active learning algorithm is described by Algorithm 2.

After the learning phase, we get two classifiers C_e and C_c. As a result, in the classification phase we can obtain two predictions for a document. Since both classifiers output class membership probabilities, they can be combined in the following way to give the overall prediction,

$$P(y = i|x) = (P(y = i|x, C_c) + P(y = i|x, C_e))/2. \quad (7)$$

5 Evaluation

5.1 Experimental Setup

We choose English-Chinese as our experimental language pair. English is regarded as the source language while Chinese is regarded as the target language.

Algorithm 2. Double viewed active learning algorithm for CLTC

Input:

The labeled set in the source language, TR_e;
The unlabeled set in the target language, U_c;

Output:

Classifiers C_e and C_c;

Generate two-view labeled set TR by translate TR_e into the target language
Generate two-view unlabeled set U by translate U_c into the source language
Let $TrainingSet = TR$
Repeat I times:
    Train classifier C_e using $TrainingSet$ based on the source language view
    Train classifier C_c using $TrainingSet$ based on the target language view
    Classify U by C_e and C_c respectively
    Let S be the n documents from U having lowest $Average\_Certainty(d)$
    Let L be the documents from U having $Certainty(d, C_c) > h$ or $Certainty(d, C_e) > h$, where h is the certainty threshold
    Let E_e and E_c be m documents from L having highest $Certainty\_Difference(d, C_e, C_c)$ and $Certainty\_Difference(d, C_c, C_e)$ respectively
    Remove E_e, E_c and S from U
    Label E_e and E_c according to C_e and C_c respectively; Label S by the teacher
    Add E_e, E_c, S to $TrainingSet$
Return C_e and C_c

Since there is not a standard evaluation benchmark available for cross language text categorization, we build a data set from the Internet. This data set contains 42610 Chinese and English news pages during the year 2008 and 2009, which fall into eight categories: Sports, Military, Tourism, Economy, Information Technology, Health, Autos and Education. The main content of each page is extracted and saved in plain text.

In our experiments, we select 1000 English documents and 2000 Chinese documents from each class. The set of English documents is treated as the training set TR_e. For the Chinese documents, we first randomly select 1000 documents from each class to form the test set TS_c, and leave the remaining documents as the additional unlabeled set U_c.

As we will use the two views of each document in our algorithm, we employ *Google Translate* [1] to translate all Chinese documents into English and all English documents into Chinese. Then, for all Chinese or Chinese translated documents, we segment the text with the tool *ICTCLAS* [2], afterwards remove the common words. For all English or English translated documents, the *EuropeanLanguageLemmatizer* [3] is applied to restore each word in the text to its base form. Then we use a stop words list to eliminate common words.

Each document is transformed into an English feature vector and a Chinese feature vector with *TF-IDF* format. The *LibSVM* package is employed for the

[1] http://translate.google.com
[2] http://ictclas.org/
[3] http://lemmatizer.org/

basic classifier. We choose linear kernel due to its good performance in text classification task. Since we need probabilistic outputs, the b option of $LibSVM$ is selected for both training and classification. The cost parameter c is set to 1.0 as default. We use Micro-Average F1 score as the evaluation measure, as it's a standard evaluation used in most previous categorization research [10,17].

5.2 Results and Discussions

In this section, we present and discuss the experimental results of the proposed algorithms.

Single Viewed Active Learning. In the first experiment, we would like to verify the effectiveness of our active learning algorithm described in Algorithm 1. An initial classifier is trained using the translated labeled set $TR_{\text{e-c}}$ and then applied to the Chinese unlabeled set U_c. In each iteration, 10 documents with the lowest prediction certainty are selected and labeled by the teacher. To validate this selecting strategy, we also implement another strategy which selects 10 documents randomly for comparison. In each iteration, a new classifier is retrained on the expanded labeled set and its performance is evaluated on the testing set TS_c. The corresponding micro average F1 curves are plotted in Figure 4.

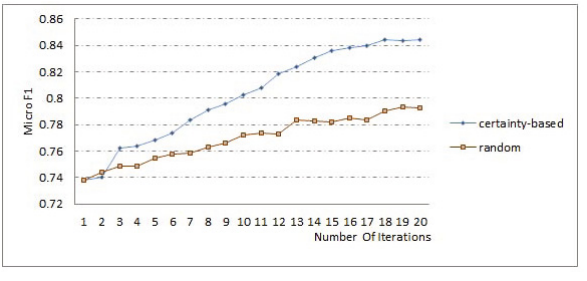

Fig. 4. Micro-F1 curves of single viewed algorithm

We can observe that, the initial classifier doesn't perform well on the Chinese test set. As the number of iterations increases, the performance is significantly improved. The certainty-based strategy shows an obvious advantage over the random strategy. This verify our assumption that documents with low prediction certainty usually contain culture dependent classification knowledge and therefore are most informative for the learner. After 20 iterations, the Micro average F1 measure on the 8000 test documents is increased by about 11 percents while the additional cost is to label 200 selected examples.

Double Viewed Active Learning. In the following experiments, we verify the double viewed algorithm described in Algorithm 2. First, two initial classifiers are trained using the labeled set based on English and Chinese view individually. Then the active learning process is performed. We set the parameter n to 10,

which means in each iteration 10 examples having lowest average certainty are selected and labeled by the teacher; and we set m to 5, which means each classifier labels 5 examples for the other. The certainty threshold h is set to 0.8, in order to reduce the error introduced by automatically labeled examples. In each iteration, the two classifiers are retrained and applied to the test set. We combine their predictions based on each view to get the overall prediction. Figure 5 shows the micro average F1 curves of the Chinese, English and overall classifiers. The curve of the single viewed algorithm is plotted as well for comparison.

We can observe that, the English classifier generally has better performance than the Chinese one, a possible reason is that more noises are introduced in Chinese view due to the text segmentation process. The overall classifier has highest accuracy, as it combines the information from both views. All the three classifiers generated by double viewed algorithm outperform the one of the single viewed algorithm. Because in each iteration they get 10 more labeled examples (each classifier automatically labels 5 examples for the other).

In our double viewed algorithm, the classifiers learn from each other and the teacher. We would like to investigate the effect of the two approaches individually. This can be done by set the parameter n and m in Algorithm 2. We first set n to 10 and m to 0, then set n to 0 and m to 5. The corresponding curves are showed in Figure 6.

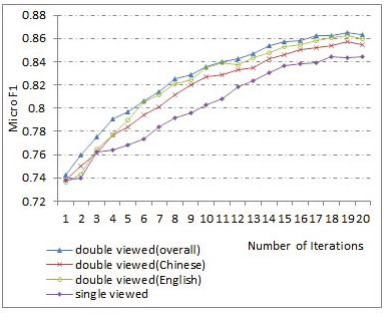

Fig. 5. Micro-F1 curves of double viewed algorithm

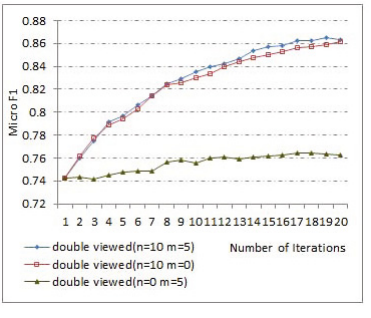

Fig. 6. Compare effects of the two learning approaches

As we can see, learning from the teacher makes a significant contribution to the improvement of the performance, while the effect of learning from the partner is weaker. The latter maybe caused by two reasons: first, there may be some errors introduced by the automatically labeled examples; second, since the Chinese and English views of one document are not completely independent, the C and D region illustrated in Figure 2 may be very limited. However, learning from the partner is still helpful, and it reduces the labor of the teacher to achieve the same performance.

Table 1. Comparison of different methods

Category	ML			MTE			MTC			Single Viewed (n=10, 20 iterations)			Double Viewed (n=10,m=5,20 iterations, overall)		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
sports	94.0	89.3	91.6	86.2	72.1	78.5	83.6	73.8	78.4	86.5	85.5	86	84.3	91.2	87.6
military	92.9	94.2	93.5	68.7	85.2	76.1	67.1	84.7	74.9	78.5	93.7	85.4	82.0	95.8	88.4
tourism	87.4	91.8	89.5	72.3	68.2	70.2	75.7	71.2	73.4	83.2	85.4	84.3	85.8	83.9	84.8
economy	87.5	87.2	87.3	88.4	56.2	68.7	85.2	62.7	72.2	84.2	80.4	82.3	83.9	87.8	85.8
IT	90.4	90.3	90.3	72.1	80.2	75.9	71.6	75.2	73.4	86.9	85.7	86.3	91.7	83.3	87.3
health	92.1	91.2	91.6	69.1	81.4	74.7	73.2	79.7	76.3	85.0	87.3	86.1	87.6	85.6	86.6
autos	93.7	91.4	92.5	65.3	88.5	75.2	62.5	89.8	73.7	85.1	92.6	88.7	89.7	92.3	91.0
education	88.4	90.1	89.2	87.9	58.1	70.0	88.9	53.9	67.1	87.8	64.7	74.5	86.3	70.5	77.6

Comparison. In Table 1, we present the detailed classification results of our algorithms, comparing with two basic machine translation based methods. The first one, denoted as **MTC**, translates the training set TR_e into Chinese and trains a classifier; the second one, denoted as **MTE**, trains a classifier in English and translates the test set TS_c into English. In addition, we also build a monolingual classifier (**ML**) by using all documents in U_c as training data. The **ML** method plays the role of an upper-bound, since the best classification results are expected when monolingual training data is available.

We can observe that, the **ML** classifier has the best performance as expected, since it's trained on the labeled data in the target language, so that there's no drawback caused by language barrier or cultural differences. Comparing with the two basic machine translation methods **MTE** and **MTC**, our active learning algorithms, both single viewed and double viewed, significantly improve the classification performance of each class. The double viewed algorithm has better performance than the single viewed one, as it combines the information from both views and makes use of the automatically labeled examples.

6 Conclusions and Future Works

In this paper, we proposed the active learning algorithm for cross language text categorization. The proposed method can effectively improve the cross language classification performance by learning from unlabeled data in the target language. For the future work, we will incorporate more metrics in the selecting strategy of active learning. For instance, can we detect the scenario in which the classifier is pretty certain but actually wrong? If such examples can be detected and labeled for retraining, the classifier will be further adaptable for the target language.

Acknowledgments. This work is supported by National Natural Science Foundation of China (60803050, 61132009)and BIT Team of Innovation.

References

1. Amine, B.M., Mimoun, M.: Wordnet based cross-language text categorization. In: 2007 IEEE/ACS International Conference on Computer Systems and Applications, pp. 848–855. IEEE (2007)
2. Bel, N., Koster, C.H.A., Villegas, M.: Cross-Lingual Text Categorization. In: Koch, T., Sølvberg, I.T. (eds.) ECDL 2003. LNCS, vol. 2769, pp. 126–139. Springer, Heidelberg (2003)
3. Blum, A., Mitchell, T.: Combining labeled and unlabeled data with co-training. In: Proceedings of the Eleventh Annual Conference on Computational Learning Theory, pp. 92–100. ACM (1998)
4. Chang, C.-C., Lin, C.-J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 27:1–27:27 (2011), software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`
5. Gliozzo, A., Strapparava, C.: Cross language text categorization by acquiring multilingual domain models from comparable corpora. In: Proceedings of the ACL Workshop on Building and Using Parallel Texts, pp. 9–16. Association for Computational Linguistics (2005)
6. Joshi, A.J., Porikli, F., Papanikolopoulos, N.: Multi-class active learning for image classification. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 2372–2379. IEEE (2008)
7. Lin, H.T., Lin, C.J., Weng, R.C.: A note on platt's probabilistic outputs for support vector machines. Machine Learning 68(3), 267–276 (2007)
8. Ni, X., Sun, J.T., Hu, J., Chen, Z.: Mining multilingual topics from wikipedia. In: Proceedings of the 18th International Conference on World Wide Web, pp. 1155–1156. ACM (2009)
9. Rigutini, L., Maggini, M., Liu, B.: An EM based training algorithm for cross-language text categorization. In: Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence, pp. 529–535. IEEE (2005)
10. Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys (CSUR) 34(1), 1–47 (2002)
11. Settles, B.: Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison (2009)
12. Shi, L., Mihalcea, R., Tian, M.: Cross language text classification by model translation and semi-supervised learning. In: Proc. EMNLP, pp. 1057–1067. Association for Computational Linguistics, Cambridge (2010)
13. Tang, J., Liu, H.: Feature selection with linked data in social media. In: SIAM International Conference on Data Mining (2012)
14. Tang, J., Wang, X., Gao, H., Hu, X., Liu, H.: Enriching short texts representation in microblog for clustering. Frontiers of Computer Science (2012)
15. Tong, S., Koller, D.: Support vector machine active learning with applications to text classification. The Journal of Machine Learning Research 2, 45–66 (2002)
16. Wan, X.: Co-training for cross-lingual sentiment classification. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, vol. 1, pp. 235–243. Association for Computational Linguistics (2009)
17. Wang, X., Tang, J., Liu, H.: Document clustering via matrix representation. In: The 11th IEEE International Conference on Data Mining, ICDM 2011 (2011)
18. Wu, T.F., Lin, C.J., Weng, R.C.: Probability estimates for multi-class classification by pairwise coupling. The Journal of Machine Learning Research 5, 975–1005 (2004)