Density Based Active Self-training for Cross-Lingual Sentiment Classification

Mohammad Sadegh Hajmohammadi*, Roliana Ibrahim, and Ali Selamat

Software Engineering Research Group, Faculty of Computing, Universiti Teknologi Malaysia, 81300 UTM Skudai, Johor, Malaysia
shmohammad2@live.utm.my, {roliana,aselamat}@utm.my

Abstract. Cross-lingual sentiment classification aims to utilize annotated sentiment resources in one language (typically English) for sentiment classification in another language. Most existing research works rely on automatic machine translation services to directly project information from one language to another. However, since machine translation quality is still far from satisfactory and also term distribution across languages may be dissimilar, these techniques cannot reach the performance of monolingual approaches. To overcome these limitations, we propose a novel learning model based on active learning and self-training to incorporate unlabeled data from the target language into the learning process. Further, in this model, we consider the density of unlabeled data to avoid outlier selection in active learning. The proposed model was applied to book review datasets in two different languages. Experiments showed that the proposed model could effectively reduce labeling efforts in comparison with some baseline methods.

Keywords: Sentiment Classification, Self-training, Active Learning, Density.

1 Introduction

Text sentiment classification is the process of automatically predicting the sentiment polarity of a given text document[1]. Although traditional classification algorithms can be used to train sentiment classifiers from labeled text data, construction of manually labeled data is a very expensive and time-consuming task. However, since most labeled sentiment resources are in English, there are not enough labeled sentiment data in other languages [2]. Therefore, the challenge is how to utilize labeled sentiment resources in one language (source language) for sentiment classification in another language (target language) and leads to an exciting research area called cross-lingual sentiment classification (CLSC).

Most existing works employed machine translation to directly project the data from the target language into the source language [3] and then treated the problem as mono-lingual sentiment classification in the source language. However, since machine translation quality is still far from satisfactory and also term distribution across

* Corresponding author.

H.-Y. Jeong et al. (eds.), *Advances in Computer Science and Its Applications*,
Lecture Notes in Electrical Engineering 279,
DOI: 10.1007/978-3-642-41674-3_146,

languages may be dissimilar due to the difference in cultures and writing styles, these methods cannot reach the performance of monolingual methods. To solve this problem, making use of unlabeled data from the target language can be helpful because they are always easy to obtain and have the same term distribution and writing style with the target language. Active learning (AL) and semi-supervised learning (SSL) are two well-known techniques that make use of unlabeled data to improve classification performance. In this paper, we propose a new model based on a combination of Active learning and self-training in order to incorporate unlabeled data from the target language into the learning process.

The rest of this paper is organized as follows. The next section presents related work on CLSC. The proposed model is described in Section 3 while evaluation and experimental results are given in Section 4. Finally, Section 5 concludes this paper.

2 Related Works

Cross-lingual sentiment analysis has been extensively studied in recent years. These research studies are based on the use of annotated data in the source language (always English) to compensate for the lack of labeled data in the target language. Most approaches focus on resource adaptation from one language to another language with few sentiment resources. For example, Mihalcea, Banea [4] generate subjectivity analysis resources into a new language from English sentiment resources by using a bilingual dictionary. In other works [5, 6], automatic machine translation engines were used to translate the English resources for subjectivity analysis. In [6], the authors showed that automatic machine translation is a viable alternative for the construction of resources for subjectivity analysis in a new language. Pan et al. [7] designed a bi-view non-negative matrix tri-factorization (BNMTF) model to solve the problem of cross-lingual sentiment classification. Another approach is that of cross-lingual classification, that is translating the features extracted from labeled documents [8]. It can, however, suffer from the inaccuracies of dictionary translation, in that words may have different meanings in different contexts. In another work, Wan [3] used the co-training method to overcome the problem of cross-lingual sentiment classification. The author exploited a bilingual co-training approach to leverage annotated English resources to sentiment classification in Chinese reviews.

3 The Proposed Model

As mentioned before, because translated data in cross-lingual sentiment classification cannot cover all vocabularies used in test data, the performance of sentiment classifier in this case is limited. To increase the performance, making use of unlabeled data from the target language can be helpful since these data are always easy to obtain and have the same term distribution as test documents. However, manually labeling unlabeled data is a hard and time-consuming task. To reduce the labeling effort, we propose a new model based on the combination of active learning and self-training.

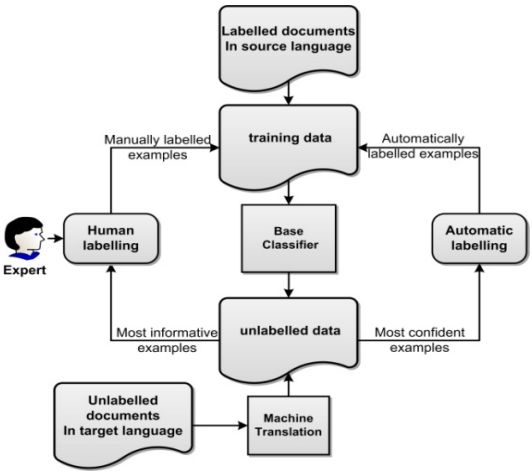

Fig. 1. Framework of the proposed approach

This model attempts to enrich initial training data through manually (AL) and automatically (self-training) labeling of some unlabeled data from the target language in an iterative process. The framework of the proposed model is illustrated in figure 1.

The query function is essential in the active learning process. The simplest query function is uncertainty sampling [9] in which unlabeled examples with the maximum uncertainty are selected for manual labeling in each learning cycle. Entropy is a popular uncertainty measurement widely used in recent researches [10]. Formula (1) shows the uncertainty function calculated based on the entropy estimation. $P(.)$ is the posterior probability of the classifier and $H(.)$ is the uncertainty function.

$$H(x) = \sum_{y \in Y} P(y|x) log P(y|x) \tag{1}$$

As reported in [11, 12], many unlabeled examples selected by the uncertainty sampling cannot help the learner since they are outliers. It means that a good selected example for manual labeling should not only be the most informative, but also the most representative one. Jingbo, Huizhen [12] proposed a density based technique to select the most informative and representative example to solve this problem. To determine the density degree of an unlabeled example, they used a novel method called k-nearest neighbor based density (*k*NN density). In this measure, the density degree of an example is computed by average similarity between this example and k most similar unlabeled examples in the unlabeled pool. Suppose $S(x)=\{s_1,s_2,s_3,\ldots,s_k\}$ is a set of k most similar unlabeled examples to the x. Therefore, average similarity for x ($A(x)$) can be computed based on the following formula:

$$A(x)=\frac{\sum_{s_i \in S(x)} \text{Similarity}(x,s_i)}{k} \tag{2}$$

We employ this density degree to avoid selecting outlier example in active learning. We use cosine measure as the similarity function to compute the pair-wise similarity value between two examples. In this model, an unlabeled example with the maximum uncertainty and density is selected based on the following formula for manually labeling.

$$u = \arg\max_{x \in U} \big(H(x) \times A(x)\big) \tag{3}$$

On the other hand, the self-training algorithm is used to label the most confident examples and generate new training examples along with active learning. These most confident classified documents are selected and added to training data with corresponding predicted labels in each step (automatic labeling). Confidence in each newly classified example is computed based on the distance of each example from the current decision boundary. *p* positive and *n* negative the most confident examples are selected as auto labeled examples for the next iteration. These two groups of selected examples are then added to the training data and removed from the unlabeled data. We called this model density based active self-training (DBAST).

4 Evaluation

In this section, we evaluate the proposed approach in CLSC on two different languages in the book review domains and compare it with some baseline methods.

4.1 Datasets

Two different evaluation datasets have been used in this paper.

1. English-French dataset (En-Fr): This dataset contains Amazon book review documents in English and French languages. This dataset was used by Prettenhofer and Stein [13].
2. English-Chinese dataset (En-Ch): This dataset was selected from Pan reviews dataset [7]. It contains book review documents in English and Chinese languages.

All review documents in target languages are translated into the source language (English) using the Google translate engine[1]. In the pre-processing step, all English reviews are converted into lowercase. Special symbols, words with one character length and other unnecessary characters are eliminated from each document. Unigram and bi-gram patterns were extracted as sentimental patterns. To reduce computational complexity, we performed feature selection using the information gain (IG) technique. We selected 5000 high score unigrams and bi-grams as final features. Term presence was used as feature weights because this method has been confirmed as the most efficient feature weighting method in sentiment classification [14].

[1] `http://translate.google.com/`

4.2 Based Lines Methods

The following baseline methods are implemented in order to evaluate the effectiveness of proposed models.

- Active Self-Training model (AST): this model is similar to DBAST but without considering the density measure of uncertain examples.
- Active learning (AL): this model is based on the simple uncertainty sampling.
- Random Sampling (RS): In random sampling approach, in each cycle, one example is randomly selected from unlabeled data for manually labeling.

4.3 Experimental Setup

In all experiments, SVM^{light} (http://svmlight.joachims.org/) is used as the base classifier with all parameters set to their default values. However, SVM does not directly output the posterior probabilities of predicted labels. Therefore, we use a strategy that introduced in [15] to compute the probabilities. In the experiments, we used the 5-fold cross validation to obtain the results. In this setting, translated documents are split into five groups. In each cycle of cross validation, the text documents from 4 groups are considered as unlabeled data and the remaining group being used as test data.

In order to compare the proposed active learning methods, we used the deficiency metric [16] that has been employed in recent papers [12]. The deficiency metric between two methods BASE and *ALG* is defined by:

$$Def_n(ALG, BASE) = \frac{\sum_{t=1}^{n}(Acc_n(BASE) - Acc_t(ALG))}{\sum_{t=1}^{n}(Acc_n(BASE) - Acc_t(BASE))} \quad (4)$$

Where *BASE* is the baseline method (in our experiment, uncertainty sampling) and *ALG* is the proposed methods such as DBAST and AST. $Acc_t(.)$ refers the accuracy of active learning method in *t*th learning cycle and $Acc_n(.)$ denotes the accuracy of active learning at the end of the learning process. This metric is always non-negative measure, and smaller values (i.e., < 1.0) indicate that *ALG* is better than *BASE* method.

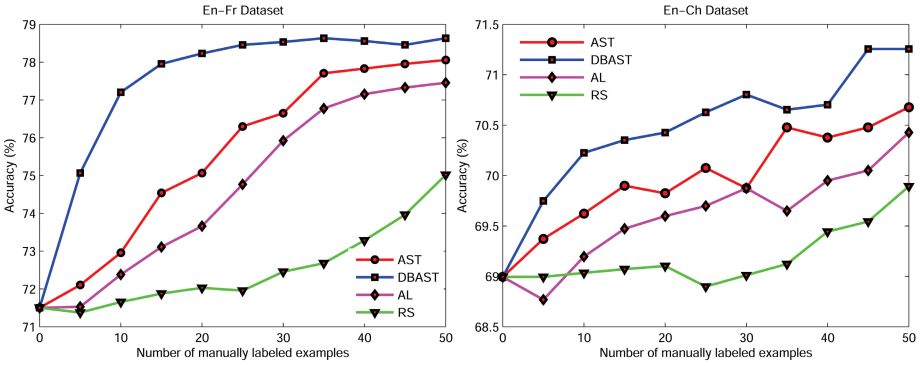

Fig. 2. The classification accuracy over the number of manually labeled examples

4.4 Results and Discussions

In this section, the proposed method is compared with three baseline methods. We set $k=20$ in the kNN density measure. We also used $p=n=5$ for the self-training algorithm. The total number of iterations is set to 50 iterations for all algorithms. After full learning process, test data is presented into learned classifier for evaluation.

Fig. 2 shows the classification accuracy of various methods on two evaluation datasets. As shown in this figure, by comparing the proposed method (DBAST) with the AST model, the classification accuracy of the proposed model improves very quickly in the first few cycles (specially in French language). This is due to the examples, selected based on density and uncertainty, are more representative than examples, selected only based on uncertainty in active learning. This figure also shows that combining active learning with self-training helps to obtain better accuracy. This is most likely due to the augmentation of most confident automatic classified examples, along with manually labeled examples, into training data during the learning process.

Table 1 shows the deficiency metric of DBAST and AST method in compare with uncertainty sampling active learning (AL). DBAST achieves smallest deficiency in all datasets, which indicates better performance than AST and AL method.

Table 1. Deficiency metric - compared with uncertainty sampling (AL)

Dataset	Methods	
	DBAST	AST
En-Fr	**0.0248**	0.7010
En-Ch	**0.0384**	0.5571

5 Conclusion

In this paper, we have proposed a new model by combining active learning and self-training in order to reduce the human labeling effort in CLSC. We also considered a density measure to avoid selecting outlier examples from unlabeled data to increase the representativeness of selected examples for manual labeling in the active learning algorithm. We applied this method to cross-lingual sentiment classification datasets in two different languages and compared the performance on the proposed model with some baseline methods. The experimental results show that the proposed model outperforms the baseline methods in all datasets.

Acknowledgement. This work is supported by the Ministry of Higher Education (MOHE) and Research Management Centre (RMC) at the Universiti Teknologi Malaysia (UTM) under Research University Grant Scheme (Vote No. Q.J130000.2628.07J52).

References

1. Liu, B.: Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies, vol. 5, p. 167. Morgan & Claypool Publishers (2012)
2. Montoyo, A., Martínez-Barco, P., Balahur, A.: Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments. Decision Support Systems 53(4), 675–679 (2012)
3. Wan, X.: Bilingual co-training for sentiment classification of chinese product reviews. Comput. Linguist. 37(3), 587–616 (2011)
4. Mihalcea, R., Banea, C., Wiebe, J.: Learning multilingual subjective language via cross-lingual projections. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (2007)
5. Banea, C., Mihalcea, R., Wiebe, J.: Multilingual subjectivity: are more languages better? In: Proceedings of the 23rd International Conference on Computational Linguistics 2010, pp. 28–36. Association for Computational Linguistics, Beijing (2010)
6. Banea, C., et al.: Multilingual subjectivity analysis using machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing 2008, pp. 127–135. Association for Computational Linguistics, Honolulu (2008)
7. Pan, J., Xue, G.-R., Yu, Y., Wang, Y.: Cross-Lingual Sentiment Classification via Bi-view Non-negative Matrix Tri-Factorization. In: Huang, J.Z., Cao, L., Srivastava, J. (eds.) PAKDD 2011, Part I. LNCS, vol. 6634, pp. 289–300. Springer, Heidelberg (2011)
8. Moh, T.-S., Zhang, Z.: Cross-lingual text classification with model translation and document translation. In: Proceedings of the 50th Annual Southeast Regional Conference 2012, pp. 71–76. ACM, Tuscaloosa (2012)
9. Lewis, D.D., Gale, W.A.: A sequential algorithm for training text classifiers. In: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval 1994, pp. 3–12. Springer-Verlag New York, Inc., Dublin (1994)
10. Zhu, J., Ma, M.: Uncertainty-based active learning with instability estimation for text classification. ACM Trans. Speech Lang. 8(4), 1–21 (2012)
11. Tang, M., Luo, X., Roukos, S.: Active learning for statistical natural language parsing. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics 2002, pp. 120–127. Association for Computational Linguistics, Philadelphia (2002)
12. Jingbo, Z., et al.: Active Learning With Sampling by Uncertainty and Density for Data Annotations. IEEE Transactions on Audio, Speech, and Language Processing 18(6), 1323–1331 (2010)
13. Prettenhofer, P., Stein, B.: Cross-Lingual Adaptation Using Structural Correspondence Learning. ACM Trans. Intell. Syst. Technol. 3(1), 1–22 (2011)
14. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of the ACL 2002 Conference on Empirical Methods in Natural Language Processing, vol. 10, pp. 79–86. Association for Computational Linguistics (2002)
15. Brefeld, U., Scheffer, T.: Co-EM support vector learning. In: Proceedings of the Twenty-First International Conference on Machine Learning 2004, p. 16. ACM, Canada (2004)
16. Baram, Y., El-Yaniv, R., Luz, K.: Online Choice of Active Learning Algorithms. J. Mach. Learn. Res. 5, 255–291 (2004)