Active Learning for Cross-Lingual Sentiment Classification

Shoushan Li[1,2], Rong Wang[1], Huanhuan Liu[1], and Chu-Ren Huang[2]

[1] Natural Language Processing Lab,
School of Computer Science and Technology, Soochow University, China
[2] CBS, The Hong Kong Polytechnic University, Hong Kong
{shoushan.li,wangrong2022,huanhuanliu.suda,
churenhuang}@gmail.com

Abstract. Cross-lingual sentiment classification aims to predict the sentiment orientation of a text in a language (named as the target language) with the help of the resources from another language (named as the source language). However, current cross-lingual performance is normally far away from satisfaction due to the huge difference in linguistic expression and social culture. In this paper, we suggest to perform active learning for cross-lingual sentiment classification, where only a small scale of samples are actively selected and manually annotated to achieve reasonable performance in a short time for the target language. The challenge therein is that there are normally much more labeled samples in the source language than those in the target language. This makes the small amount of labeled samples from the target language flooded in the aboundance of labeled samples from the source language, which largely reduces their impact on cross-lingual sentiment classification. To address this issue, we propose a data quality controlling approach in the source language to select high-quality samples from the source language. Specifically, we propose two kinds of data quality measurements, intra- and extra-quality measurements, from the certainty and similarity perspectives. Empirical studies verify the appropriateness of our active learning approach to cross-lingual sentiment classification.

1 Introduction

Sentiment classification is a task of predicting the sentimental orientation (e.g., positive or negative) for a certain text (Pang et al., 2002; Turney, 2002). This task has drawn much attention in the natural language processing (NLP) community due to its wide applications (Pang and Lee 2008; Liu, 2012). Up to now, extensive studies have been conducted on this task and various kinds of resources are available, such as polarity lexicons and labeled corpora. However, these resources are rather imbalanced across different languages. For example, due to dominant studies on English sentiment classification, the labeled data in English is often in a large scale while the labeled data in some other languages is much limited. This motives the research on cross-lingual sentiment classification, which aims to perform sentiment classification

G. Zhou et al. (Eds.): NLPCC 2013, CCIS 400, pp. 236–246, 2013.

in a resource-scarce language (named as the target language) with the help of labeled data from another resource-rich language (named as the source language). Representative studies include Wan (2008, 2009), Wei and Pal (2010), Lu et al. (2011), and Meng et al. (2012).

Although existing studies have yielded certain progress in cross-lingual sentiment classification, the classification performance of only using the labeled data in the source language remains far away from satisfaction due to the huge difference in linguistic expression and social culture. For example, in Wan (2009) where English is considered as the source language and Chinese is considered as the target language, only using the labeled data from English yields the performance of around 0.75 in accuracy which is much lower than 0.92 that achieved by using 1000 labeled samples from the target language (Chinese). Even when the unlabeled data from the target language is employed via co-training, the obtained performance can be only improved to around 0.82 in accuracy (Wan, 2009).

One possible solution to handle this dilemma is to deploy active learning, where a small scale of samples (called newly-added data) are actively selected and manually annotated to quickly improve the classification performance for the target language. However, one challenge in active learning-based cross-lingual sentiment classification lies in the much imbalanced labeled data from the source and target languages. For example, in Wan (2009), the labeled samples in the source language can be around 8000 while the labeled samples in the target language are generally as less as 200, a reasonable number one can expected to be manually annotated in a fast deploying application. Such huge imbalance in the labeled data easily floods the small amount of the labeled target data in the abundance of labeled source data and largely reduces the contribution of the labeled data in the target language.

In this paper, we address above challenge by proposing a data quality controlling approach to select high-quality samples in the source language instead of using all the samples. Consequently, the data imbalance can be much reduced when only a small partition of labeled samples in the source language is employed. We believe that using a partition of them could be as useful as (or even possibly better than) using all of them for cross-lingual sentiment classification. For example, consider following three reviews from the product-review corpora, introduced in Blitzer et al. (2007):

E1: *This book is not worth wasting your money on. To the novice, this book may appear to represent the art of cabales serrada escrima, but it does not. More than half of the book is unrelated to the system of serrada. ……*

E2: *This fourth installment of becky's trying tribulations is the worst. I don't understand how kinsella's editor didn't draw the line (and the red pencil) at the litany of shopping expeditions. I am not making this up. ……*

E3: *This is one of the worst books ever, it is not worth wasting your money on. Don't buy it.*

While **E1** has a strong sentimental expression of "*not worth wasting*" and **E2** has another strong sentimental expression of "*the worst*", **E3** has both of them. Therefore, once **E3** is selected, we can safely throw away **E1** and **E2**.

Accordingly, we propose a certainty-based quality measurement, together with cross-validation to select high-quality samples in the source language. Besides, we propose a similarity measurement to select the samples in the source language that are similar to those in the target language. In this paper, we call the former the intra-quality measurement because it only employs the data in the source language to measure the quality of the samples in the source language, and the latter the extra-quality measurement due to the consideration of the samples in the target language. For a particular data in the target language, these two kinds of measurements are integrated to select high-quality samples in the source language. After obtaining the high-quality samples in the source language, we employ standard uncertainty sampling for active learning-based cross-lingual sentiment classification.

The remainder of this paper is organized as follows. Section 2 overviews the related work on cross-lingual sentiment classification. Section 3 presents our approach to data quality controlling. Section 4 applies the data quality controlling to active learning-based cross-lingual sentiment classification. Section 5 evaluates the proposed approaches. Finally, Section 6 gives the conclusion and future work.

2 Related Work

Although sentiment classification have been extensively studied in the last decade (Pang et al., 2002; Turney, 2002), cross-lingual sentiment classification only merges in recent years (Wan, 2008; Wan 2009; Pan et al., 2011; Prettenhofer and Stein, 2011; Lu et al., 2011; Meng et al., 2012).

Wan (2008) proposes an ensemble method to combine one classifier trained with labeled data from the source language and another classifier trained with their translated data. Subsequently, Wan (2009) incorporates the unlabeled data in the target language with co-training to improve the classification performance.

Wei and Pal (2010) regard cross-lingual sentiment classification as a domain adaptation task and apply a structural correspondence learning approach (SCL) to tackle this problem. Their approach is shown to more effective than the co-training algorithm.

More recently, Lu et al. (2011) perform cross-lingual sentiment classification from a different perspective. Instead of using machine translation engines, they use a parallel corpus to help perform semi-supervised learning in both English and Chinese sentence-level sentiment classification.

Unlike all of them, this study suggests to use only those high-quality samples instead of all of them to perform cross-lingual sentiment classification. As a result, the data imbalance between the labeled data in the source and target languages can be largely reduced. This largely eliminates obstacles towards active learning to cross-lingual sentiment classification. To the best of our knowledge, this is the first attempt to consider data quality, active learning and integrate them in cross-lingual sentiment classification.

3 Data Quality Controlling in the Source Language

Let X_S be the set of the labeled samples in the source language and X_T the set of the unlabeled samples (testing data) in the target language. The objective of cross-lingual sentiment classification is to estimate a hypothesis h: $X_S \to C$ which classifies the samples in X_T into C, the predefined set of class labels, i.e., *negative* and *positive*.

In contrast to traditional sentiment classification, where the training and testing data are from the same language, it is not possible to directly train a hypothesis h: $X_S \to C$ to classify X_T because the training and test samples have different feature spaces due to the language difference. Therefore, the feature spaces for the training and test data need to be unified. One common way to achieve this is to translate the samples in the source (or target) language into the target (or source) language. Let X_S^t be the set of the translated samples in the source language and X_T^t the set of the translated samples in the target language. Then, the objective of cross-lingual sentiment classification is changed into estimating the hypothesis h: $X_S^t \to C$ which classifies the samples in X_T or the hypothesis h: $X_S \to C$ which classifies the samples in X_T^t. For simplicity, in the following, we only focus on the solution of translating the labeled data in the source language into the target language. Note that our research is certainly suitable for the case of translating the test data in the target language into the source language.

As stated in Introduction, the task of data quality controlling in cross-lingual sentiment classification is first to measure the quality of the samples in X_S^t and then select a subset of X_S^t (i.e. those high-quality samples, denoted as X_{S-sub}^t) to train a classifier rather than using all the labeled samples in the source language. In the following, we describe two measurements to evaluate the quality of a translated sample in the source language.

3.1 Intra-quality Measurement with Certainty and Cross-validation

The quality measurement that measured only through the resource from the source language is called intra-quality measurement.

To obtain a high-quality sample that representing some other samples, we first split the labeled data from the source language into two different parts. One is severed as the training data and the other is severed as the validation data. Then, we use the training data to train a classifier which is used to predict the samples in the validation data. After the prediction process, all posterior possibilities of the validation samples are provided and we assume that the samples with high posterior possibilities are

capable of representing the classification knowledge in the training data. Formally, the certainty measurement is employed to rank the validation samples, which is defined as follows:

$$Cer(x) = \max_{y \in \{pos, neg\}} P(y \mid x) \tag{1}$$

Where x is a sample in the validation data and $P(y \mid x)$ is its posterior possibility estimated by the classifier trained with the training data.

To represent all the data in the source language, the cross-validation strategy is applied (Kohavi, 1995). In k-fold cross-validation, X_S^t is randomly partitioned into k equal size subsamples. Of the k subsamples, a single subsample is used as the validation data, and the remaining $k - 1$ subsamples are used as the training data. The cross-validation process is then repeated k times (the *folds*). In this way, each of the k subsamples used exactly once as the validation data to find the high quality samples.

3.2 Extra-Quality Measurement with Similarity

Instinctively, the quality of the samples in the source language is also related to the testing samples in the target language. We name the quality measurement measured with the resource from the target language as the extra-quality measurement. In this study, the samples with higher similarity to the target language are thought to be of higher quality.

Suppose the labeled data in the source language contains n samples, i.e., $X_S^t = (x_{S1}, x_{S2}, ..., x_{Sn})$ and the testing data in the target language contains m samples, i.e., $X_T = (x_{T1}, x_{T2}, ..., x_{Tm})$. The similarity between one sample x_{Si} in the source language and the target language is defined as following:

$$SIM(x_{Si}, X_T) = \frac{1}{m} \sum_{j=1}^{m} sim(x_{Si}, x_{Tj}) \tag{2}$$

Where $sim(x_{Si}, x_{Tj})$ is the similarity between the sample x_{Si} and x_{Tj}. In this study, the standard cosine method is applied to compute the similarity between two samples.

3.3 Integrating Intra- and Extra-Quality Measurements

One straightforward way to integrate the two quality measurements is to linearly combine the certainty and similarity scores. However, in fact, the similarity measurement, as the extra-quality measurement in this study, is not a good way to select high-quality samples. In contrast, it performs even worse than the random selection strategy. This is mainly because the similarity measurement does not take the sentimental information into account and thus the selected samples are not useful for sentiment classification.

Input:

Translated training data from the source language X_S^t

Testing data from the target language X_T

Output:

The selected data set X_{S-sub}^t

Procedure:

(1) Initialize the selected data set: $X_{S-sub}^t = \varnothing$

(2) Compute the similarity between each sample in X_S^t and X_T with formula (2)

(3) Repeat until the predefined stop criterion is met

a) Perform *k*-fold cross-validation in X_S^t

b) Rank the samples in each validation data sets according to their certainty values computed with formula (1).

c) Select top-*N* certainty samples that take the higher similarities to X_T than σ in each validation data, which is denoted as $X_l^{Cer}(l = 1,2,...,k)$

d) $X_{S-sub}^t = X_{S-sub}^t + \sum_{l=1}^{k} X_l^{Cer}$

e) $X_S^t = X_S^t - \sum_{l=1}^{k} X_l^{Cer}$

Fig. 1. Algorithm of data quality controlling in the source language

Therefore, we consider the certainty measurement as the main ranking factor and leave the similarity measurement as a supplementary one when designing the way to integrate them. Specifically, we select high-certainty samples that take the similarities to the target language higher than a threshold σ. In this way, only the samples that are similar to the source language are possibly be selected as the high-quality candidates.

Our algorithm of data quality controlling in the source language is shown in Figure 1. This algorithm integrates the intra- and extra-quality measurements in the steps of b) and c) respectively.

4 Active Learning-Based Cross-Lingual Sentiment Classification

As mentioned in Introduction, the performance of cross-lingual sentiment classification usually remains very limited and unsatisfactory. To quickly improve the performance, a small amount of informative samples in the target language are encouraged to be annotated and leveraged. This is a typical active learning task.

Input:

Translated training data from the source language X_S^t

Unlabeled data U_T

Testing data from the target language X_T

Output:

The classifier for cross-lingual sentiment classification

Procedure:

(1) Obtain the high-quality data set X_{S-sub}^t from X_S^t

(2) Initialize the labeled data $L_T = X_{S-sub}^t$

(3) Loop for *M* iterations

a) Learn a classifier using L_T

b) Use the current classifier to label all samples in U_T

c) Use the uncertainty measurement to select *n* most uncertainty samples for manual annotation

d) Move the newly-annotated sample from U_T to L_T

(4) Learn the classifier with L_T for cross-lingual sentiment classification

Fig. 2. Algorithm of active learning for cross-lingual sentiment classification

However, different from traditional active learning-based sentiment classification, the initial labeled data in active learning-based cross-lingual sentiment classification is from a different language and in a large amount. This makes the small amount of informative samples in the target language submersed and thus difficult to well affect the classification decision.

Our solution to the above challenge is to use only those high-quality samples from the source language as the initial labeled data instead of using all the data. Then, the standard uncertainty sampling method is employed to add the informative samples from the target language for manual annotation, with the uncertainty measurement defined as follows:

$$Uncer(x) = \min_{y \in \{pos, neg\}} P(y \mid x) \tag{3}$$

Figure 2 illustrates the detailed algorithm.

5 Experimentation

5.1 Experimental Settings

Labeled Data in the Source Language: The labeled data from the source language contains English reviews from four domains: Book (B), DVD (D), Electronics (E) and

Kitchen (K)[1] (Blitzer et al., 2007). Each domain contains 1000 positive and 1000 negative reviews. All together, 8000 labeled samples are available in the source language. All these labeled samples are translated into Chinese ones with *Google Translate*[2].

Testing Data in the Target Language: The testing data from the target language contains Chinese reviews from two domains. they are from the data collection by Wan (2011): Chinese reviews from IT168 (451 positive and 435 negative reviews) and Chinese reviews from 360BUY (560 positive and 370 negative reviews) [3], together with 2000 unlabeled reviews.

Unlabeled Data in the Target Language: We manually annotate the unlabeled reviews collected by Wan (2011) and select 500 positive and 500 negative as the unlabeled samples for active learning.

Feature Space: Each review text is treated as a bag-of-words and transformed into binary vectors encoding the presence or absence of word unigrams.

Classification Algorithm: The maximum entropy (ME) classifier implemented with the public tool, Mallet Toolkits[4] is employed in all our experiments. The posterior probabilities belonging to the categories are also provided in this tool.

5.2 Experimental Results on Active Learning-Based Cross-Lingual Sentiment Classification

In this section, we compare following approaches to active learning in cross-lingual sentiment classification.

Random+No_source: Perform active learning in the target language by randomly selecting samples in the target language and no samples in the source language are used. We perform 5 runs of such approaches and report the average results.

Uncertainty+No_source: Perform active learning in the target language with the uncertainty selection strategy and no samples in the source language are used. 20 samples are randomly selected as the initial labeled data.

Uncertainty+All_source: Perform active learning in the target language with the uncertainty selection strategy. All the translated samples in the source language are served as the initial labeled data.

Uncertainty+Selected_source: Perform active learning in the target language with the uncertainty selection strategy. 500 high-quality translated samples selected by our quality controlling approach in the source language are served as the initial labeled data. In the implementation of sample selecting in the source domain, the fold number is set to 10 (k=10) and top 10 certainty samples are selected in each validation data (N=10). As for the parameter of σ, we set it to 0.27, 0.14 in the domains of IT168 and 360BUY respectively. These values are referred to the average similarity between each sample and all the other samples in the target language.

---

[1] http://www.seas.upenn.edu/~mdredze/datasets/sentiment/
[2] http://translate.google.com/translate_t
[3] http:// google.com/site/wanxiaojun1979/
[4] http://mallet.cs.umass.edu/

Table 1. The classification performance by using all 8000 samples in the source domain

Domain	IT168	360BUY
Accuracy	0.756	0.754

Table 1 shows the classification performance by using all the samples in the source domain. From the results, we can see that only using the labeled samples (even the scale of data is big) from the target domain, the obtained performances are very limited (less than 0.8 in both domains).

Figure 3 shows the performances of different active learning approaches for cross-lingual sentiment classification. From this figure, we can see that:

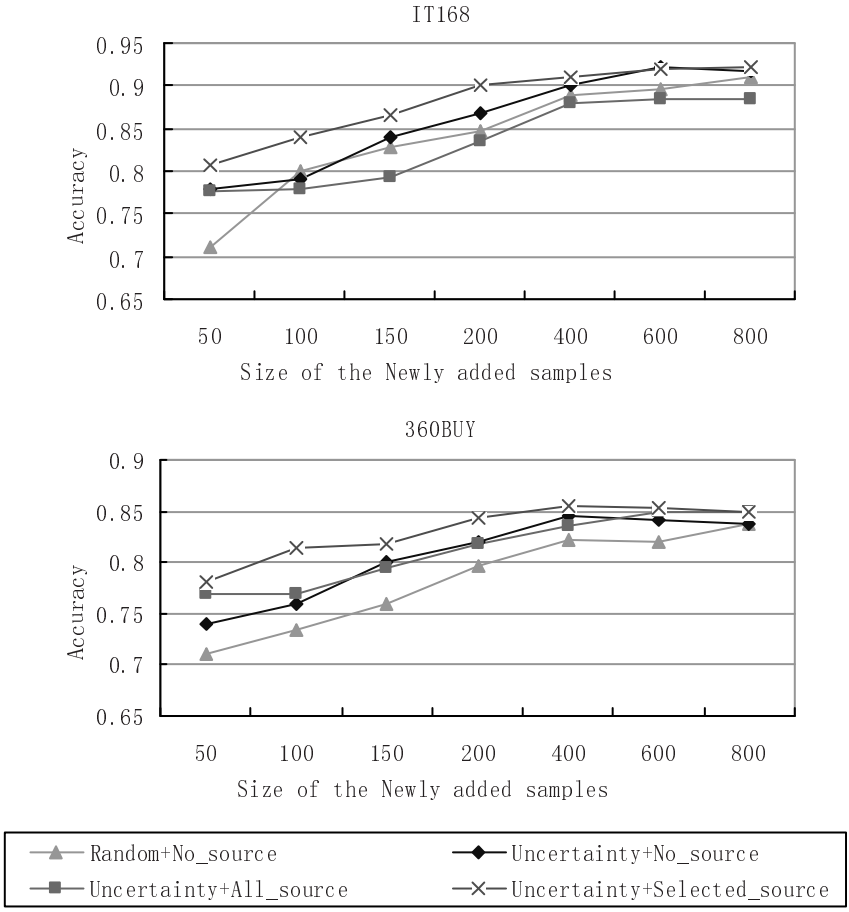

Fig. 3. Performances of different approaches to active learning-based cross-lingual sentiment classification

Employing labeled samples from the target language is indeed effective for sentiment classification in the target language. For example, as shown in Table 1, using all 8000 samples from the source language yields the accuracy of 0.756 in IT168. In contrast, as shown in Figure 4, using only 100 randomly-selected samples from the target language could yield a higher accuracy, i.e., 0.8.

Uncertainty+No_source generally performs better than **Uncertainty+All_source** when the labeled samples in the target language are more than 100. This result demonstrates that the labeled samples in the source language become unhelpful when a certainty number of labeled data in the target language is available.

Uncertainty+Selected_source performs best. The high quality samples, together with only 100-200 samples, achieve a comparable performance to that of using more than 800 samples in the target language. This result verifies the necessity of data quality controlling in the source language when performing active learning in cross-lingual sentiment classification.

6 Conclusion

In this paper, we propose an active learning approach for cross-lingual sentiment classification and address the huge challenge of the data imbalance by controlling data quality in the source language. Specifically, we design a certainty measurement, together with a similarity measurement, to select high quality samples in the source language. Experimentation verifies the appropriateness of active learning for cross-lingual sentiment classification. Specifically, the results show that with the selected samples in the source language, manually annotating only 100-200 samples in the target language can achieve a comparable performance to that of using more than 800 samples only in the target language.

Acknowledgments. This research work has been partially supported by two NSFC grants, No.61003155, and No.61273320, one National High-tech Research and Development Program of China No.2012AA011102, one General Research Fund (GRF) sponsored by the Research Grants Council of Hong Kong No.543810.

References

1. Balahur, A., Turchi, M.: Multilingual Sentiment Analysis using Machine Translation? In: Proceedings of the 3rd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, pp. 52–60 (2012)
2. Blitzer, J., Dredze, M., Pereira, F.: Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. In: Proceedings of ACL 2007, pp. 440–447 (2007)
3. Boyd-Graber, J., Resnik, P.: Holistic Sentiment Analysis across Languages Multilingual Supervised Latent Dirichlet Allocation. In: Proceedings of ACL 2010, pp. 45–55 (2010)
4. Kohavi, R.: A Study of Cross-validation and Bootstrp for Accuracy Estimation and Model Selection. In: Proceedings of IJCAI, pp. 1137–1143 (1995)

5. Liu, B.: Sentiment Analysis and Opinion Mining (Introduction and Survey). Morgan & Claypool Publishers (May 2012)
6. Lu, B., Tan, C., Cardie, C., Tsou, B.: Joint Bilingual Sentiment Classification with Unlabeled Parallel Corpora. In: Proceedings of ACL 2011, pp. 320–330 (2011)
7. Pang, B., Lee, L.: Opinion Mining and Sentiment Analysis: Foundations and Trends. Information Retrieval 2(12), 1–135 (2008)
8. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment Classification using Machine Learning Techniques. In: Proceedings of EMNLP 2002, pp. 79–86 (2002)
9. Prettenhofer, P., Stein, B.: Cross Language Text Classification Using Structural Correspondence Learning. In: Proceedings of ACL 2010, pp. 1118–1127 (2010)
10. Turney, P.: Thumbs up or Thumbs down? Semantic Orientation Applied to Unsupervised Classification of reviews. In: Proceedings of ACL 2002, pp. 417–424 (2002)
11. Wan, X.: Using Bilingual Knowledge and Ensemble Techniques for Unsupervised Chinese Sentiment Analysis. In: Proceedings of ACL 2008, pp. 553–561 (2008)
12. Wan, X.: Co-Training for Cross-Lingual Sentiment Classification. In: Proceedings of ACL 2009, pp. 235–243 (2009)
13. Wan, X.: Bilingual Co-Training for Sentiment Classification of Chinese Product Reviews. Computational Linguistics 37, 587–616 (2011)
14. Wei, B., Pal, C.: Cross Lingual Adaptation An Experiment on Sentiment Classifications. In: Proceedings of ACL 2010, pp. 258–262 (2010)