Data Quality Controlling for Cross-Lingual Sentiment Classification

Shoushan Li[†‡] Yunxia Xue[†] Zhongqing Wang[†] Sophia Yat Mei Lee[‡] Chu-Ren Huang[‡]

[†]Natural Language Processing Lab
Soochow University
Suzhou, China
{shoushan.li, yunxia.xue, wangzq870305}@gmail.com

[‡]Department of Chinese and Bilingual Studies
The Hong Kong Polytechnic University
Hong Kong
{churenhuang, sophiaym}@gmail.com

***Abstract*—Cross-lingual sentiment classification aims to perform sentiment classification in a language (named as the target language) with the help of the resources from another language (named as the source language). Previous studies are prone to using all available data in the source language while using all data is observed to perform no better or even worse than using a partion of *good* data. In this paper, we propose a novel task called data quality controlling in the source language to select high quality samples from the source language. To tackle this task, we propose two kinds of data quality measurements: intra- and extra-quality measurements which are implemented with the certainty and similarity measurements respectively. The empirical studies demonstrate the effectiveness of the proposed approach to data quality controlling in the source language.**

I. INTRODUCTION

Sentiment classification is a task for predicting the sentimental orientation (e.g., positive or negative) of 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 related resources are available, such as polarity lexicon and large amount of labeled corpora. However, these resources are rather imbalanced across different languages. For example, due to many previous work on English sentiment classification, the labeled data for sentiment classification in English is often in a large scale while the labeled data in Chinese is rather limited. This motivates the research on cross-lingual sentiment classification which aims to perform sentiment classification in a resource-scarce language (named as the target language) with the help of the rich resources from another language (named as the source language). Recently, such studies have become more and more popular in NLP (Wan, 2009).

Although the resources from the source language are rich, not all the labeled samples in the resources are helpful. Using a partition of them could sometimes be more useful for cross-lingual sentiment classification than using all of them. This is because the resources usually contain multiple data sets collected from different domains or different sources and their classification abilities in the target language possibly differ a lot. Let us consider the case of cross-lingual sentiment classification in Wan (2009) where the source-language data consists of the labeled English reviews from four domains: Book (B), DVD (D), Electronic (E) and Kitchen (K) and the target-language data contains the Chinese reviews from the IT products. As a pilot test, we select two domains, i.e., B and D, from the source language and use the translated text in each single domain or both domains to train three classifiers which are applied to predict the Chinese testing samples. The classification results are shown in Table 1. We observe that using all translated samples from both domains performs even worse than using the ones from DVD only. Therefore, how to select a partition of good samples remains a challenging problem in cross-lingual sentiment classification.

Domain	B	D	B&D
Accuracy	0.677	0.75	0.711

Table 1: Performances on the target language when using the classifiers trained with the samples from Book, DVD, and both of them

In this paper, we address the above challenging issue by proposing a data quality controlling approach to select high-quality samples in the source language. The high-quality samples could achieve a better performance than using all data. Our approach to data quality controlling integrates two kinds of quality measurements: intra- and extra-quality measurements. The former employs the labeled data in the source language to measure the quality of a sample in the source language while the latter employs the unlabeled data in the target language to measure the quality of a sample in the source language. More specifically, a certain measurement is proposed as the intra-quality measurement together with the cross-validation technology, and the similarity measurement is proposed as the extra-quality measurement. For a particular data set in the target language, these two kinds of measurements are integrated to select high-quality samples in the source language.

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 5 evaluates our approach. Finally, Section 6 gives the conclusion and future work.

II. RELATED WORK

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

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

More recently, Lu et al. (2011) perform cross-lingual sentiment classification from a different perspective. Instead of using the machine translation engines, they use a parallel corpus to help perform semi-supervised learning in both English and Chinese sentence-level sentiment classifications. Similar to Lu et al. (2011), Meng et al. (2012) also employ the parallel corpus to help cross-lingual sentiment classification. They also explore the case when no labeled data is available in the parallel corpus.

Unlike all of them, we aim to measure the quality of the samples in the source language and suggest using only the high-quality samples rather than all of them. To the best of our knowledge, this is the first attempt to consider the data quality issue in cross-lingual sentiment classification.

III. Data Quality Controlling

A. Problem Formulation

Let X_S be the set of the labeled samples in the source language and X_T be 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: $X_S \rightarrow C$ which classifies the samples in X_T into C, the predefined class label set including two categories, i.e., negative and positive.

In contrast to traditional sentiment classification, where the training and testing data is from the same language, it is not possible to directly train a hypothesis h: $X_S \rightarrow C$ to classify X_T because they possess a totally different feature space. Therefore, the feature spaces for the training and testing 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 ${X_T}^t$ be the set of the translated samples in the target language. In this case, the objective of cross-lingual sentiment classification becomes different, i.e. estimating the hypothesis *h*: ${X_S}^t \rightarrow C$ which classifies the samples in X_T or the hypothesis *h:* $X_S \rightarrow C$ which classifies the samples in ${X_T}^t$. For simplicity, in this paper, we only focus on the solution that translating the labeled data in the source language into the target language. Note that our approach is also suitable for the case of translating the testing data in the target language into the source language.

The task of data quality controlling in cross-lingual sentiment classification is first to measure the qualities of the samples in X_S^t and then select a subset of X_S^t with high-quality samples, denoted as X_{S-sub}^t to train the classifier rather than using all the labeled samples in the source language. The main challenge in this task is how to design the evaluation of the quality of each sample.

B. Intra-Quality Measurement with Certainty and Cross-validation

The quality measurement based on the resource from the source language is called intra-quality measurement. Let us consider the following three reviews from the product-review corpora (Blizer 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.*

E1 contains a strong sentimental expression of "not worth wasting" and E2 has another strong sentimental expression of "the worst". Except these expressions, the other expressions in the two reviews are more likely to express facts and thus are not helpful for sentiment classification. In contrast, E3 contains the two strong sentimental expressions and fewer factual expressions. Among the three reviews, E3 is believed to be of higher quality because it represents the sentimental information on both E1 and E2 and contains fewer noisy information. Therefore, a sample with more representative opinions and less noisy expressions is intuitively thought to be of higher quality.

To obtain a high-quality sample that representing some other samples, we could split the labeled data from the source language into two different parts and use one part as the training data to train a classifier. Then, the classifier is used to predict the samples in the other part, denoted as the validation data. After the prediction process, all posterior possibilities of the validation samples are provided. We assume that the samples with high posterior possibilities are capable of representing the classification knowledge in the training data. Meanwhile, these samples are believed to contain less noisy information. Otherwise, the posterior probabilities could not be so high. Formally, the certainty measurement is employed to rank the validation samples, which is defined as follows:

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

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.

The cross-validation strategy is applied to obtain the samples representing all the data in the source language (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 selected 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.

C. 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 based on the resource from the target language the extra-quality measurement. 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 langue and the whole sample set in the target langue is defined as follows:

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

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 the two samples.

D. Integrating Intra- and Extra-Quality Measurements

```
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 bear the higher
     similarity to $X_T$ than $\sigma$ in each validation data,
     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) $X_S^t = X_S^t - \sum_{l-1}^{k} X_l^{Cer}$
```

Figure 1: The algorithm of selecting samples in the source language according to the data quality controlling with both intra- and extra- measurements

One straightforward way to integrate the two quality measurements is to linearly combine the certainty and similarity scores. However, the similarity measurement, as the extra-quality measurement in this study, is not a good way to select high-quality samples. It in fact performs even worse than the random selection strategy (These results will be shown in Section IV). This is mainly because the similarity measurement does not take the sentimental information into account. Therefore, we consider the certainty measurement as the key ranking factor. Specifically, we select high-certainty samples that bear higher similarities to the target language than a threshold σ. In this way, only the samples that are similar to the source language are possibly selected as the high-quality samples. 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).

IV. Experimentation

A. Experimental Setting

The Data from the Source Language: The 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 and thus 8000 labeled samples are available in the source language.

The Data from the Target Language: The data from the target language contains Chinese reviews from three domains. Among them, the data of the two domains are taken from Wan (2011): Chinese reviews from IT168 (451 positive + 435 negative reviews) and Chinese reviews from 360BUY (560 positive and 370 negative reviews)[2], together with 2000 unlabeled reviews. Another domain, i.e., named Beauty, is collected by ourselves from AMAZON (400 positive and 400 negative reviews)[3].

Features: 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 is implemented by the public tool, Mallet Toolkits[4]. The posterior probabilities belonging to the categories are also provided in this tool.

B. Results on Data Quality Controlling

In this section, we will compare the following classifiers trained with the samples selected with different approaches of data quality controlling in cross-lingual sentiment classification.

Random: Train the ME classifier with the labeled samples randomly selected from the source language.

Certainty: Train the ME classifier with the labeled samples selected from the source language with the certainty measurement only (ranking the samples according to their certainty scores).

Similarity: Train the ME classifier with the labeled samples selected from the source language with the similarity measurement only (ranking the samples according to their similarity scores to the target language).

Certainty+Similarity: Train the ME classifier with the labeled samples selected by our approach as shown in Figure 1 In the implementation, 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, 0.19 in the domains of IT168, 360BUY,

[1] http://www.seas.upenn.edu/~mdredze/datasets/sentiment/

[2] http:// google.com/site/wanxiaojun1979/

[3] http://www.amazon.cn/

[4] http://mallet.cs.umass.edu/

and Beauty respectively. These values refer to the average similarity between each sample and all the other samples in the target language.

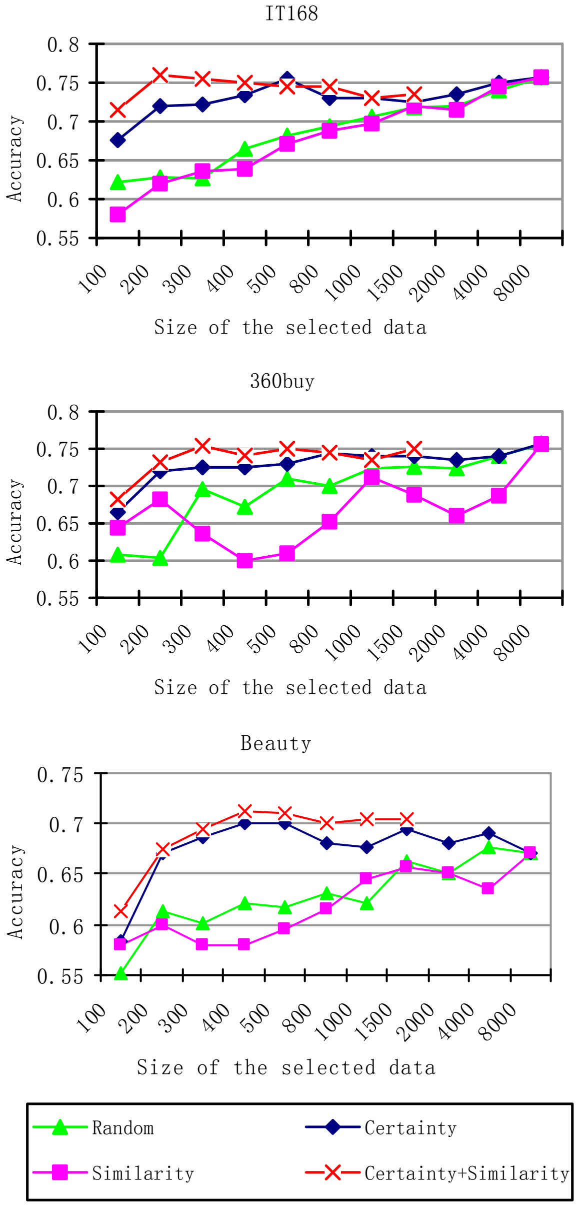

Figure 2: Performances of cross-lingual sentiment classification with different approaches to data quality controlling

Figure 2 shows the performances of these approaches on cross-lingual sentiment classification. Since the threshold value is used in our approach, many samples with the lower similarity scores than the threshold are filtered out. Thus we only report the results when less than 2000 samples are selected. We can see, in Figure 2, that the certainty measurement always performs much better than the random selection. However, the similarity measurement performs worse than the certainty one. This is mainly due to the fact that the similarity measurement does not take the sentimental information into account. This is exactly why we use the similarity as an extra restrictive condition rather than a major condition for selecting samples when we integrate the two kinds of quality measurements.

Our approach that integrating both the intra- and extra-quality measurements is the most effective one. Selecting 300-500 samples by our approach achieves comparative performance or even better performance than using all 8000 samples. For example, in Beauty, using only 400 samples obtain the accuracy of 0.71 which is higher than that (0.671 in accuracy) of using all 8000 samples.

V. Conclusion

In this paper, we address a novel task called data quality controlling in the source language. Specifically, we design a certainty measurement together with a similarity measurement to select high quality samples. Empirical studies show that using a small amount of high quality samples achieves a comparable performance to or even better performance than using all the data.

In future work, we would like to enhance the effectiveness of the extra-quality measurement in selecting high quality samples. We will also apply our data quality controlling approach to other NLP tasks.

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] Blitzer J., M. Dredze and F. Pereira. 2007. Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. In Proceedings of ACL-07, pp.440-447.

[2] Kohavi, R. 1995. A Study of Cross-validation and Bootstrp for Accuracy Estimation and Model Selection. In Proceedings of IJCAI, pp.1137–1143.

[3] Liu B. 2012. Sentiment Analysis and Opinion Mining (Introduction and Survey). Morgan & Claypool Publishers, May 2012.

[4] Lu B., C. Tan, C. Cardie and B. K. Tsou. 2011. Joint Bilingual Sentiment Classification with Unlabeled Parallel Corpora. In Proceedings of ACL-11, pp.320-330.

[5] Meng X., F. Wei, X. Liu, M. Zhou, G. Xu, H. Wang. Cross-Lingual Mixture Model for Sentiment Classification. In Proceedings of ACL-12, pp.572-581.

[6] Pang B. and L. Lee. 2008. Opinion Mining and Sentiment Analysis: Foundations and Trends. Information Retrieval, vol.2(12): 1-135.

[7] Pang B., L. Lee and S. Vaithyanathan. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. In Proceedings of EMNLP-02, pp.79-86.

[8] Turney P. 2002. Thumbs up or Thumbs down? Semantic Orientation Applied to Unsupervised Classification of reviews. In Proceedings of ACL-02, pp.417-424.

[9] Wan X. 2007. Co-Training for Cross-Lingual Sentiment Classification. In Proceedings of ACL-09, pp.235-243.

[10] Wan X. 2008. Using Bilingual Knowledge and Ensemble Techniques for Unsupervised Chinese Sentiment Analysis. In Proceedings of ACL-08, pp.553-561.

[11] Wan X. 2011. Bilingual Co-Training for Sentiment Classification of Chinese Product Reviews. Computational Linguistics, 37: 587-616.

[12] Wei B. and C. Pal. 2010. Cross Lingual Adaptation An Experiment on Sentiment Classifications. In Proceedings of ACL-10, pp.258-262.