Cross Lingual Adaptation: An Experiment on Sentiment Classifications

Bin Wei
University of Rochester
Rochester, NY, USA.
bwei@cs.rochester.edu

Christopher Pal
École Polytechnique de Montréal
Montréal, QC, Canada.
christopher.pal@polymtl.ca

Abstract

In this paper, we study the problem of using an annotated corpus in English for the same natural language processing task in another language. While various machine translation systems are available, automated translation is still far from perfect. To minimize the noise introduced by translations, we propose to use only key 'reliable" parts from the translations and apply structural correspondence learning (SCL) to find a low dimensional representation shared by the two languages. We perform experiments on an English-Chinese sentiment classification task and compare our results with a previous co-training approach. To alleviate the problem of data sparseness, we create extra pseudo-examples for SCL by making queries to a search engine. Experiments on real-world on-line review data demonstrate the two techniques can effectively improve the performance compared to previous work.

1 Introduction

In this paper we are interested in the problem of transferring knowledge gained from data gathered in one language to another language. A simple and straightforward solution for this problem might be to use automatic machine translations. However, while machine translation has been the subject of a great deal of development in recent years, many of the recent gains in performance manifest as syntactically as opposed to semantically correct sentences. For example, "PIANYI" is a word mainly used in positive comments in Chinese but its translation from the online Google translator is always "cheap", a word typically used in a negative context in English. To reduce this kind of error introduced by the translator, Wan in (Wan, 2009) applied a co-training scheme. In this setting classifiers are trained in both languages and the two classifiers teach each other for the unlabeled examples. The co-training approach manages to boost the performance as it allows the text similarity in the target language to compete with the "fake" similarity from the translated texts. However, the translated texts are still used as training data and thus can potentially mislead the classifier. As we are not really interested in predicting something on the language created by the translator, but rather on the real one, it may be better to further diminish the role of the translated texts in the learning process. Motivated by this observation, we suggest here to view this problem as a special case of domain adaptation, in the source domain, we mainly observe English features, while in the other domain mostly features from Chinese. The problem we address is how to associate the features under a unified setting.

There has been a lot of work in domain adaption for NLP (Dai et al., 2007)(Jiang and Zhai, 2007) and one suitable choice for our problem is the approach based on structural correspondence learning (SCL) as in (Blitzer et al., 2006) and (Blitzer et al., 2007b). The key idea of SCL is to identify a low-dimensional representations that capture correspondence between features from both domains (x_s and x_t in our case) by modeling their correlations with some special pivot features. The SCL approach is a good fit for our problem as it performs knowledge transfer through identifying important features. In the cross-lingual setting, we can restrict the translated texts by using them only through the pivot features. We believe this form is more robust to errors in the language produced by the translator.

Adapting language resources and knowledge to a new language was first studied for general text categorization and information retrieval as in (Bel

Proceedings of the ACL 2010 Conference Short Papers, pages 258–262,
Uppsala, Sweden, 11-16 July 2010.

et al., 2003), where the authors translate a keyword lexicon to perform cross-lingual text categorization. In (Mihalcea et al., 2007), different shortcomings of lexicon-based translation scheme was discussed for the more semantic-oriented task subjective analysis, instead the authors proposed to use a parallel-corpus, apply the classifier in the source language and use the corresponding sentences in the target language to train a new classifier. With the rapid development of automatic machine translations, translating the whole corpus becomes a plausible option. One can either choose to translate a corpus in the target language and apply the classifier in the source language to obtain labeled data, or directly translated the existing data set to the new language. Various experiments of the first strategy are performed in (Banea et al., 2008) for the subjective analysis task and an average 65 F1 score was reported. In (Wan, 2008), the authors propose to combine both strategies with ensemble learning and train a bi-lingual classifier.

In this paper, we are also interested in exploring whether a search engine can be used to improve the performance of NLP systems through reducing the effect of data sparseness. As the SCL algorithm we use here is based on co-occurrence statistics, we adopt a simple approach of creating pseudo-examples from the query counts returned by Google.

2 Our Approach

To begin, we give a formal definition of the problem we are considering. Assume we have two languages l_s and l_t and denote features in these two languages as x_s and x_t respectively. We also have text-level translations and we use $x_{t'}$ for features in the translations from l_s to l_t and $x_{s'}$ for the other direction. Let y be the output variable we want to predict, we have labeled examples (y, x_s) and some unlabeled examples (x_t). Our task is to train a classifier for (y, x_t). In this paper, we consider the binary sentiment classification (positive or negative) problem where l_s and l_t correspond to English and Chinese (for general sentiment analysis, we refer the readers to the various previous studies as in (Turney, 2002),(Pang et al., 2002),and (McDonald et al., 2007)). With these definitions in place, we now describe our approach in further detail.

2.1 Structural Correspondence Learning(SCL)

Due to space limitations, we give a very brief overview of the SCL framework here. For a detailed illustration, please refer to (Ando and Zhang, 2005). When SCL is used in a domain adaptation problem, one first needs to find a set of pivot features x_p. These pivot features should behave in a similar manner in both domains, and can be used as "references" to estimate how much other features may contribute when used in a classifier to predict a target variable. These features can either be identified with heuristics (Blitzer et al., 2006) or by automatic selection (Blitzer et al., 2007b). Take sentiment classification as an example, "very good" and "awful" are good pivot features, if a certain feature in the target domain co-occurs often with "very good" but infrequently with "awful", we could expect this feature will play a similar role as "very good" in the final classifier but a different role from "awful". We can make this observation purely based on the co-occurrence between these features. No hand-labeling is required and this specific feature doesn't need to be present in our labeled training data of the source domain.

The SCL approach of (Ando and Zhang, 2005) formulates the above idea by constructing a set of linear predictors for each of the pivot features. Each of these linear predictor is binary like whether "very good" occurs in the text and we have a set of training instances $(1|0, \{x_i\})$. The weight matrix of these linear predictors will encode the co-occurrence statistics between an ordinary feature and the pivot features. As the co-occurrence data are generally very sparse for a typical NLP task, we usually compress the weight matrix using the singular vector decomposition and only selects the top k eigenvectors v_k. This matrix w of the k vectors $\{v_k\}$ gives a mapping from the original feature space to a lower dimensional representation and is shown in (Ando and Zhang, 2005) to be the optimal choice of dimension k under common loss functions. In the next step we can then train a classifier on the extended feature $(x, w * x)$ in the source domain. As w groups the features from different domains with similar behavior relative to the pivot features together, if such a classifier has good performance on the source domain, it will likely do well on the target domain as well.

2.2 SCL for the Cross-lingual Adaptation

Viewing our task as a domain adaptation problem. The source domain correspond to English reviews and the target domain for Chinese ones. The full feature vector is (x_s, x_t). The difficulty we are facing is, due to noise in the translations, the conditional probabilities $p(y|x_s)$ and the one in the translated texts $p(y|x_{s'})$ may be quite different. Consider the following two straightforward strategies of using automatic machine translations: one can translate the original English labeled data (y, x_s) into $(y, x_{t'})$ in Chinese and train a classifier, or one can train a classifier on (y, x_s) and translate x_t in Chinese into $x_{s'}$ in English so as to use the classifier. But as the conditional distribution can be quite different for the original language and the pseudo language produced by the machine translators, these two strategies give poor performance as reported in (Wan, 2009).

Our solution to this problem is simple: instead of using all the features as $(x_s, x_{t'})$ and $(x_{s'}, x_t)$, we only preserves the pivot features in the translated texts $x_{s'}$ and $x_{t'}$ respectively and discard the other features produced by the translator. So, now we will have (x_s, x_{tp}) and (x_{sp}, x_t) where $x_{(s|t)p}$ are pivot features in the source and the target languages. In other words, when we use the SCL on our problem, the translations are only used to decide if a certain pivot feature occurs or not in the training of the linear predictors. All the other non-pivot features in the translators are blocked to reduce the noise.

In the original SCL as we mentioned earlier, the final classifier is trained on the extended features $(x, w * x)$. However, as mentioned above we will only use the pivot features. To represent this constraint, we can modify the vector to be $(w_p * x, w * x)$ where w_p is a constant matrix that only selects the pivot features. This modification will not affect the deduction procedure and results in (Ando and Zhang, 2005). Experiments show that using only pivot features actually outperforms the full feature setting.

For the selection of the pivot features, we follow the automatic selection method proposed in (Blitzer et al., 2007a). We first select some candidates that occur at least some constant number of times in reviews of the two languages. Then, we rank these features according to their conditional entropy to the labels on the training set. In table 1, we give some of the pivot features with English translations associated with the Chinese pivot features. As we can see from the table, although we only have text-level translations we still get some features with similar meaning from different languages, just like performing an alignment of words.

English Pivot Features
"poor quality", "not buy", "easy use", "very easy" "excellent", "perfect", "still very", "garbage", "poor", "not work", "not to", "very comfortable"
Chinese Pivot Features
wanmei(perfect), xiaoguo hen(effect is very...) tisheng(improve),feichang hao(very good), cha(poor), shushi(comfortable), chuse(excellent)

Table 1: Some pivot features.

2.3 Utilizing the Search Engine

Data sparseness is a common problem in NLP tasks. On the other hand, search engines nowadays usually index a huge amount of web pages. We now show how they can also be used as a valuable data source in a less obvious way. Previous studies like (Bollegala, 2007) have shown that search engine results can be comparable to language statistics from a large scale corpus for some NLP tasks like word sense disambiguation. For our problem, we use the query counts returned by a search engine to compute the correlations between a normal feature and the pivot features.

Consider the word "PIANYI" which is mostly used in positive comments, the query "CHANPIN(product) PING(comment) CHA(bad) PIANYI" has 2,900,000 results, while "CHANPIN(product) PING(comment) HAO(good) PIANYI" returns 57,400,000 pages. The results imply the word "PIANYI" is closer to the pivot feature "good" and it behaves less similar with the pivot feature "bad".

To add the query counts into the SCL scheme, we create pseudo examples when training linear predictors for pivot features. To construct a pseudo-positive example between a certain feature x_i and a certain pivot feature x_p, we simply query the term $x_i x_p$ and get a count c_1. We also query x_p alone and get another count c_2.Then we can create an example $(1, \{0, ..., 0, x_i = \frac{c_1}{c_2}, 0, ..., 0\})$. The pseudo-negative examples are created similarly. These pseudo examples are equivalent to texts with a single word and the count is used to

approximate the empirical expectation. As an initial experiment, we select 10,000 Chinese features that occur more than once in the Chinese unlabeled data set but not frequent enough to be captured by the original SCL. And we also select the top 20 most informative Chinese pivot features to perform the queries.

3 Experiment

3.1 Data Set

For comparsion, we use the same data set in (Wan, 2009):

Test Set(Labeled Chinese Reviews): The data set contains a total of 886 labeled product reviews in Chinese (451 positive reviews and 435 negative ones). These reviews are extracted from a popular Chinese IT product website IT168 [1]. The reviews are mainly about electronic devices like mp3 players, mobile phones, digital cameras and computers.

Training Set(Labeled English Reviews): This is the data set used in the domain adaption experiment of (Blitzer et al., 2007b). It contains four major categories: books, DVDs, electronics and kitchen appliances. The data set consists of 8000 reviews with 4000 positive and 4000 negative, It is a public data set available on the web [2].

Unlabeled Set (Unlabeled Chinese Reviews): 1000 Chinese reviews downloaded from the same website as the Chinese training set. They are of the same domain as the test set.

We translate each English review into Chinese and vice versus through the public Google Translation service. Also following the setting in (Wan, 2009), we only use the Chinese unlabeled data and English training sets for our SCL training procedures. The test set is blind to the training stage.

The features we used are bigrams and unigrams in the two languages as in (Wan, 2009). In Chinese, we first apply the stanford Chinese word segmenter [3] to segment the reviews. Bigrams refers to a single Chinese word and a bigram refers to two adjacent Chinese words. The features are also pre-processed and normalized as in (Blitzer et al., 2007b).

[1] http://www.it168.com

[2] http://www.cis.upenn.edu/ mdredze/datasets/sentiment/

[3] http://nlp.stanford.edu/software/segmenter.shtml

Models	Precision	Recall	F-Score
CoTrain	0.768	0.905	0.831
SCL-B	0.772	0.914	0.837
SCL-C	0.764	0.896	0.825
SCL-O	0.760	0.909	0.828
SCL-E	0.801	0.909	0.851

Table 2: Results on the Positive Reviews

Models	Precision	Recall	F-Score
CoTrain	0.879	0.717	0.790
SCL-B	0.931	0.752	0.833
SCL-C	0.908	0.743	0.817
SCL-O	0.928	0.739	0.823
SCL-E	0.928	0.796	0.857

Table 3: Results on the Negative Reviews

3.2 Comparisons

We compare our procedure with the co-training scheme reported in (Wan, 2009):

CoTrain: The method with the best performance in (Wan, 2009). Two standard SVMs are trained using the co-training scheme for the Chinese views and the English views. And the results of the two SVMs are combined to give the final output.

SCL-B: The basic SCL procedure as explained.

SCL-O: The basic SCL except that we use all features from the translated texts instead of only the pivot features.

SCL-C: The training procedure is still the same as **SCL-B** except in the test time we only use the Chinese pivot features and neglect the English pivot features from translations.

SCL-E: The same as **SCL-B** except that in the training of linear pivot predictors, we also use the pseudo examples constructed from queries of the search engine.

Table 2 and 3 give results measured on the positive labeled reviews and negative reviews separately. Table 4 gives the overall accuracy on the whole 886 reviews. Our basic SCL approach **SCL-B** outperforms the original **Co-Training** approach by 2.2% in the overall accuracy. We can

CoTrain	SCL-B	SCL-O	SCL-C	SCL-E
0.813	0.835	0.826	0.822	0.854

Table 4: Overall Accuracy of Different Methods

also notice that using all the features including the ones from translations actually deteriorate the performance from 0.835 to 0.826.

The model incorporating the co-occurrence count information from the search engine has the best overall performance of 0.857. It is interesting to note that the simple scheme we have adopted increased the recall performance on the negative reviews significantly. After examining the reviews, we find the negative part contains some idioms and words mainly used on the internet and the query count seems to be able to capture their usage.

Finally, as our final goal is to train a Chinese sentiment classifier, it will be best if our model can only rely on the Chinese features. The **SCL-C** model improves the performance from the **Co-Training** method a little but not as much as the $SCL-B$ and the $SCL-O$ approaches. This observation suggests that the translations are still helpful for the cross-lingual adaptation problem as the translators perform some implicit semantic mapping.

4 Conclusion

In this paper, we are interested in adapting existing knowledge to a new language. We show that instead of fully relying on automatic translation, which may be misleading for a highly semantic task like the sentiment analysis, using techniques like SCL to connect the two languages through feature-level mapping seems a more suitable choice. We also perform an initial experiment using the co-occurrence statistics from a search engine to handle the data sparseness problem in the adaptation process, and the result is encouraging.

As future research we believe a promising avenue of exploration is to construct a probabilistic version of the SCL approach which could offer a more explicit model of the relations between the two domains and the relations between the search engine results and the model parameters. Also, in the current work, we select the pivot features by simple ranking with mutual information, which only considers the distribution information. Incorporating the confidence from the translator may further improve the performance.

References

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*.

Carmen Banea, Rada Mihalcea, Janyce Wiebe, and Samer Hassan. 2008. Multilingual subjectivity analysis using machine translation. In *Proceedings of EMNLP*.

Nuria Bel, Cornelis H. A. Koster, and Marta Villegas. 2003. Cross-lingual text categorization. In *Research and AdvancedTechnology for Digital Libraries*.

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of EMNLP*.

John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman. 2007a. Learning bounds for domain adaptation. In *Proceedings of NIPS*.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007b. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of ACL*.

Danushka Bollegala. 2007. Measuring semantic similarity between words using web search engines. In *Proceedings of WWW 07*.

Wenyuan Dai, Gui-Rong Xue, Qiang Yang, and Yong Yu. 2007. Co-clustering based classification for out-of-domain documents. In *Proceedings of KDD*.

Jing Jiang and ChengXiang Zhai. 2007. A two-stage approach to domain adaptation for statistical classifiers. In *Proceedings of CIKM*.

Ryan T. McDonald, Kerry Hannan, Tyler Neylon, Mike Wells, and Jeffrey C. Reynar. 2007. Structured models for fine-to-coarse sentiment analysis. In *Proceedings of ACL*.

Rada Mihalcea, Carmen Banea, and Janyce Wiebe. 2007. Learning multilingual subjective language via cross-lingual projections. In *Proceedings of ACL*.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of EMNLP*.

Peter D. Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of ACL*.

Xiaojun Wan. 2008. Using bilingual knowledge and ensemble techniques for unsupervised chinese sentiment analysis. In *Proceedings of EMNLP*.

Xiaojun Wan. 2009. Co-training for cross-lingual sentiment classification. In *Proceedings of ACL*.