Co-Regression for Cross-Language Review Rating Prediction

Xiaojun Wan

Institute of Computer Science and Technology, The MOE Key Laboratory of
Computational Linguistics, Peking University, Beijing 100871, China
`wanxiaojun@pku.edu.cn`

Abstract

The task of review rating prediction can be well addressed by using regression algorithms if there is a reliable training set of reviews with human ratings. In this paper, we aim to investigate a more challenging task of cross-language review rating prediction, which makes use of only rated reviews in a source language (e.g. English) to predict the rating scores of unrated reviews in a target language (e.g. German). We propose a new co-regression algorithm to address this task by leveraging unlabeled reviews. Evaluation results on several datasets show that our proposed co-regression algorithm can consistently improve the prediction results.

1 Introduction

With the development of e-commerce, more and more people like to buy products on the web and express their opinions about the products by writing reviews. These reviews usually contain valuable information for other people's reference when they buy the same or similar products. In some applications, it is useful to categorize a review into either positive or negative, but in many real-world scenarios, it is important to provide numerical ratings rather than binary decisions.

The task of review rating prediction aims to automatically predict the rating scores of unrated product reviews. It is considered as a finer-grained task than the binary sentiment classification task. Review rating prediction has been modeled as a multi-class classification or regression task, and the regression based methods have shown better performance than the multi-class classification based methods in recent studies (Li et al. 2011). Therefore, we focus on investigating regression-based methods in this study.

Traditionally, the review rating prediction task has been investigated in a monolingual setting, which means that the training reviews with human ratings and the test reviews are in the same language. However, a more challenging task is to predict the rating scores of the reviews in a target language (e.g. German) by making use of the rated reviews in a different source language (e.g. English), which is called Cross-Language Review Rating Prediction. Considering that the resources (i.e. the rated reviews) for review rating prediction in different languages are imbalanced, it would be very useful to make use of the resources in resource-rich languages to help address the review rating prediction task in resource-poor languages.

The task of cross-language review rating prediction can be typically addressed by using machine translation services for review translation, and then applying regression methods based on the monolingual training and test sets. However, due to the poor quality of machine translation, the reviews translated from one language A to another language B are usually very different from the original reviews in language B, because the words or syntax of the translated reviews may be erroneous or non-native. This phenomenon brings great challenges for existing regression algorithms.

In this study, we propose a new co-regression algorithm to address the above problem by leveraging unlabeled reviews in the target language. Our algorithm can leverage both views of the reviews in the source language and the target language to collaboratively determine the confidently predicted ones out of the unlabeled reviews, and then use the selected examples to enlarge the training set. Evaluation results on several datasets show that our proposed co-regression algorithm can consistently improve the prediction results.

2 Related Work

Most previous works on review rating prediction model this problem as a multi-class classification task or a regression task. Various features have been exploited from the review text, including words, patterns, syntactic structure, and semantic topic (Qu et al. 2010; Pang and Lee, 2005; Leung et al. 2006; Ganu et al. 2009). Traditional learn-

ing models, such as SVM, are adopted for rating prediction. Most recently, Li et al. (2011) propose a novel tensor-based learning framework to incorporate reviewer and product information into the text based learner for rating prediction. Saggion et al. (2012) study the use of automatic text summaries instead of the full reviews for movie review rating prediction. In addition to predicting the overall rating of a full review, multi-aspect rating prediction has also been investigated (Lu et al. 2011b; Snyder and Barzilay, 2007; Zhu et al. 2009; Wang et al. 2010; Lu et al. 2009; Titov and McDonald, 2008). All the above previous works are working under a monolingual setting, and to the best of our knowledge, there exists no previous work on cross-language review rating prediction.

It is noteworthy that a few studies have been conducted for the task of cross-lingual sentiment classification or text classification, which aims to make use of labeled data in a language for the binary classification task in a different language (Mihalcea et al., 2007; Banea et al., 2008; Wan 2009; Lu et al. 2011a; Meng et al. 2012; Shi et al., 2010; Prettenhofer and Stein 2010). However, the binary classification task is very different from the regression task studied in this paper, and the proposed methods in the above previous works cannot be directly applied.

3   Problem Definition and Baseline Approaches

Let $L=\{(x_1, y_1), \ldots, (x_i, y_i), \ldots, (x_n, y_n)\}$ denote the labeled training set of reviews in a source language (e.g. English), where x_i is the i-th review and y_i is its real-valued label, and n is the number of labeled examples; Let T denote the test review set in a different target language (e.g. German); Then the task of cross-language review rating prediction aims at automatically predicting the rating scores of the reviews in T by leveraging the labeled reviews in L. No labeled reviews in the target language are allowed to be used.

The task is a regression problem and it is challenging due to the language gap between the labeled training dataset and the test dataset. Fortunately, due to the development of machine translation techniques, a few online machine translation services can be used for review translation. We adopt *Google Translate*[1] for review translation. After review translation, the training reviews and the test reviews are now in the same

language, and any regression algorithm (e.g. logistic regression, least squares regression, KNN regressor) can be applied for learning and prediction. In this study, without loss of generality, we adopt the widely used regression SVM (Vapnik 1995; Joachims 1999) implemented in the SVMLight toolkit[2] as the basic regressor. For comparative analysis, we simply use the default parameter values in SVMLight with linear kernel. The features include all unigrams and bigrams in the review texts, and the value of each feature is simply set to its frequency (TF) in a review.

Using features in different languages, we have the following baseline approaches for addressing the cross-language regression problem.

REG_S: It conducts regression learning and prediction in the source language.

REG_T: It conducts regression learning and prediction in the target language.

REG_ST: It conducts regression learning and prediction with all the features in both languages.

REG_STC: It combines **REG_S** and **REG_T** by averaging their prediction values.

However, the above regression methods do not perform very well due to the unsatisfactory machine translation quality and the various language expressions. Therefore, we need to find new approaches to improve the above methods.

4   Our Proposed Approach

4.1   Overview

Our basic idea is to make use of some amounts of unlabeled reviews in the target language to improve the regression performance. Considering that the reviews have two views in two languages and inspired by the co-training style algorithms (Blum and Mitchell, 1998; Zhou and Li, 2005), we propose a new co-training style algorithm called co-regression to leverage the unlabeled data in a collaborative way. The proposed co-regression algorithm can make full use of both the features in the source language and the features in the target language in a unified framework similar to (Wan 2009). Each review has two versions in the two languages. The source-language features and the target-language features for each review are considered two redundant views of the review. In the training phase, the co-regression algorithm is applied to learn two regressors in the two languages. In the prediction phase, the two regressors are applied to predict two rating scores of the review. The

---

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

[2] http://svmlight.joachims.org

final rating score of the review is the average of the two rating scores.

4.2 Our Proposed Co-Regression Algorithm

In co-training for classification, some confidently classified examples by one classifier are provided for the other classifier, and vice versa. Each of the two classifiers can improve by learning from the newly labeled examples provided by the other classifier. The intuition is the same for co-regression. However, in the classification scenario, the confidence value of each prediction can be easily obtained through consulting the classifier. For example, the SVM classifier provides a confidence value or probability for each prediction. However, in the regression scenario, the confidence value of each prediction is not provided by the regressor. So the key question is how to get the confidence value of each labeled example. In (Zhou and Li, 2005), the assumption is that the most confidently labeled example of a regressor should be with such a property, i.e. the error of the regressor on the labeled example set (i.e. the training set) should decrease the most if the most confidently labeled example is utilized. In other words, the confidence value of each labeled example is measured by the decrease of the error (e.g. mean square error) on the labeled set of the regressor utilizing the information provided by the example. Thus, each example in the unlabeled set is required to be checked by training a new regression model utilizing the example. However, the model training process is usually very time-consuming for many regression algorithms, which significantly limits the use of the work in (Zhou and Li, 2005). Actually, in (Zhou and Li, 2005), only the lazy learning based KNN regressor is adopted. Moreover, the confidence of the labeled examples is assessed based only on the labeled example set (i.e. the training set), which makes the generalization ability of the regressor not good.

In order to address the above problem, we propose a new confidence evaluation strategy based on the consensus of the two regressors. Our intuition is that if the two regressors agree on the prediction scores of an example very well, then the example is very confidently labeled. On the contrary, if the prediction scores of an example by the two regressors are very different, we can hardly make a decision whether the example is confidently labeled or not. Therefore, we use the absolute difference value between the prediction scores of the two regressors as the confidence value of a labeled example, and if the ex-

ample is chosen, its final prediction score is the average of the two prediction scores. Based on this strategy, the confidently labeled examples can be easily and efficiently chosen from the unlabeled set as in the co-training algorithm, and these examples are then added into the labeled set for re-training the two regressors.

Given:
- F_{source} and F_{target} are redundantly sufficient sets of features, where F_{source} represents the source language features, F_{target} represents the target language features;
- L is a set of labeled training reviews;
- U is a set of unlabeled reviews;

Loop for I iterations:
1. Learn the first regressor R_{source} from L based on F_{source};
2. Use R_{source} to label reviews from U based on F_{source}; Let $\hat{y}_i^{source}$ denote the prediction score of review x_i;
3. Learn the second classifier R_{target} from L based on F_{target};
4. Use R_{target} to label reviews from U based on F_{target}; Let $\hat{y}_i^{target}$ denote the prediction score of review x_i;
5. Choose m most confidently predicted reviews $E=\{$ top m reviews with the smallest value of $\left|\hat{y}_i^{target} - \hat{y}_i^{source}\right|\}$ from U, where the final prediction score of each review in E is $\left|\hat{y}_i^{target} + \hat{y}_i^{source}\right|/2$;
6. Removes reviews E from U and add reviews E with the corresponding prediction scores to L;
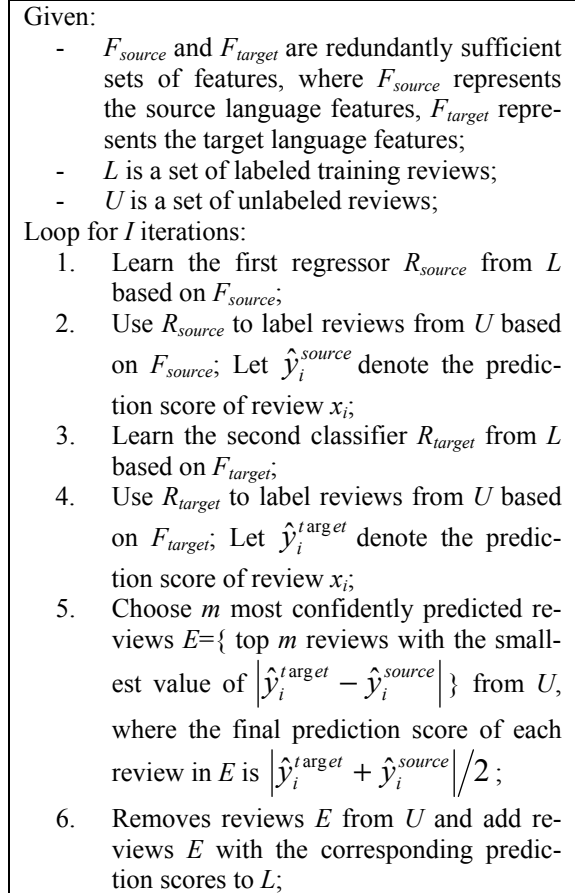
Figure 1. Our proposed co-regression algorithm

Our proposed co-regression algorithm is illustrated in Figure 1. In the proposed co-regression algorithm, any regression algorithm can be used as the basic regressor to construct R_{source} and R_{target}, and in this study, we adopt the same regression SVM implemented in the SVMLight toolkit with default parameter values. Similarly, the features include both unigrams and bigrams and the feature weight is simply set to term frequency. There are two parameters in the algorithm: I is the iteration number and m is the growth size in each iteration. I and m can be empirically set according to the total size of the unlabeled set U, and we have $I\times m\le |U|$.

Our proposed co-regression algorithm is much more efficient than the COREG algorithm (Zhou and Li, 2005). If we consider the time-consuming regression learning process as one
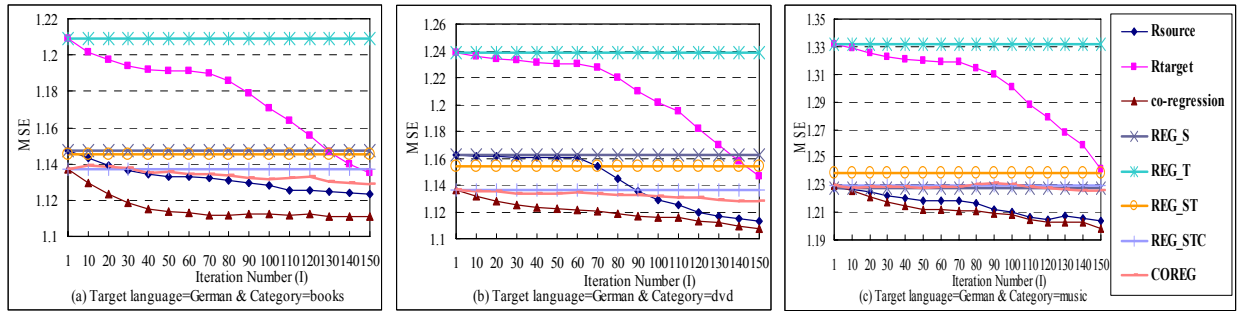
Figure 2. Comparison results vs. Iteration Number (I) (R_{source} and R_{target} are the two component regressors)

basic operation and make use of all unlabeled examples in U, the computational complexity of COREG is $O(|U|+I)$. By contrast, the computational complexity of our proposed co-regression algorithm is just $O(I)$. Since $|U|$ is much larger than I, our proposed co-regression algorithm is much more efficient than COREG, and thus our proposed co-regression algorithm is more suitable to be used in applications with a variety of regression algorithms.

Moreover, in our proposed co-regression algorithm, the confidence of each prediction is determined collaboratively by two regressors. The selection is not restricted by the training set, and it is very likely that a portion of good examples can be chosen for generalize the regressor towards the test set.

5   Empirical Evaluation

We used the WEBIS-CLS-10 corpus[3] provided by (Prettenhofer and Stein, 2010) for evaluation. It consists of Amazon product reviews for three product categories (i.e. books, dvds and music) written in different languages including English, German, etc. For each language-category pair there exist three sets of training documents, test documents, and unlabeled documents. The training and test sets comprise 2000 documents each, whereas the number of unlabeled documents varies from 9000 – 170000. The dataset is provided with the rating score between 1 to 5 assigned by users, which can be used for the review rating prediction task. We extracted texts from both the summary field and the text field to represent a review text. We then extracted the rating score as a review's corresponding real-valued label. In the cross-language scenario, we regarded English as the source language, and regarded German as the target language. The experiments were conducted on each product category separately. Without loss of generality, we sampled and used

only 8000 unlabeled documents for each product category. We use Mean Square Error (MSE) as the evaluation metric, which penalizes more severe errors more heavily.

In the experiments, our proposed co-regression algorithm (i.e. "co-regression") is compared with the COREG algorithm in (Zhou and Li, 2005) and a few other baselines. For our proposed co-regression algorithm, the growth size m is simply set to 50. We implemented the COREG algorithm by replacing the KNN regressor with the regression SVM and the pool size is also set to 50. The iteration number I varies from 1 to 150. The comparison results are shown in Figure 2.

We can see that on all product categories, the MSE values of our co-regression algorithm and the two component regressors tend to decline over a wide range of I, which means that the selected confidently labeled examples at each iteration are indeed helpful to improve the regressors. Our proposed co-regression algorithm outperforms all the baselines (including COREG) over different iteration members, which verifies the effectiveness of our proposed algorithm. We can also see that the COREG algorithm does not perform well for this cross-language regression task. Overall, our proposed co-regression algorithm can consistently improve the prediction results.

6   Conclusion and Future Work

In this paper, we study a new task of cross-language review rating prediction and propose a new co-regression algorithm to address this task. In future work, we will apply the proposed co-regression algorithm to other cross-language or cross-domain regression problems in order to verify its robustness.

---

[3] http://www.uni-weimar.de/medien/webis/research/corpora/corpus-webis-cls-10.html

References

Carmen Banea, Rada Mihalcea, Janyce Wiebe, and Samer Hassan. 2008. Multilingual subjectivity analysis using machine translation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 127-135.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In Annual Meeting-Association For Computational Linguistics.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In Proceedings of the eleventh annual conference on Computational learning theory, pp. 92-100.

Hang Cui, Vibhu Mittal, and Mayur Datar. 2006. Comparative experiments on sentiment classification for online product reviews. In Proceedings of the National Conference on Artificial Intelligence.

Gayatree Ganu, Noemie Elhadad, and Amélie Marian. 2009. Beyond the stars: Improving rating predictions using review text content. In WebDB.

Thorsten Joachims, 1999. Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning, MIT-Press.

CaneWing Leung, Stephen Chi Chan, and Fu Chung. 2006. Integrating collaborative filtering and sentiment analysis: A rating inference approach. In ECAI Workshop, pages 300–307.

Fangtao Li, Nathan Liu, Hongwei Jin, Kai Zhao, Qiang Yang and Xiaoyan Zhu. 2011. Incorporating reviewer and product information for review rating prediction. In Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence (IJCAI2011).

Yue Lu, ChengXiang Zhai, Neel Sundaresan. 2009. Rated Aspect Summarization of Short Comments. Proceedings of the World Wide Conference 2009 (WWW'09), pages 131-140.

Bin Lu, Chenhao Tan, Claire Cardie, Ka Yin Benjamin TSOU. 2011a. Joint bilingual sentiment classification with unlabeled parallel corpora. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 320-330.

Bin Lu, Myle Ott, Claire Cardie and Benjamin K. Tsou. 2011b. Multi-aspect sentiment analysis with topic models. In Proceedings of Data Minig Workshps (ICDMW), 2011 IEEE 11th International Conference on, pp. 81-88, IEEE.

Xinfan Meng, Furu Wei, Xiaohua Liu, Ming Zhou, Ge Xu, and Houfeng Wang. 2012. Cross-Lingual Mixture Model for Sentiment Classification. In Proceedings of ACL-2012.

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

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, pp. 79-86, 2002.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In Proceedings of the ACL, pages 115–124.

Peter Prettenhofer and Benno Stein. 2010. Cross-Language Text Classification using Structural Correspondence Learning. In 48th Annual Meeting of the Association of Computational Linguistics (ACL 10), 1118-1127.

Lizhen Qu, Georgiana Ifrim, and Gerhard Weikum. 2010. The bag-of-opinions method for review rating prediction from sparse text patterns. In COLING, pages 913–921, Stroudsburg, PA, USA, 2010. ACL.

Horacio Saggion, Elena Lloret, and Manuel Palomar. 2012. Can text summaries help predict ratings? a case study of movie reviews. Natural Language Processing and Information Systems (2012): 271-276.

Lei Shi, Rada Mihalcea, and Mingjun Tian. 2010. Cross language text classification by model translation and semi-supervised learning. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp. 1057-1067, 2010.

Benjamin Snyder and Regina Barzilay. 2007. Multiple aspect ranking using the good grief algorithm. Proceedings of the Joint Human Language Technology/North American Chapter of the ACL Conference (HLT-NAACL).

Ivan Titov and Ryan McDonald. 2008. A joint model of text and aspect ratings for sentiment summarization. In Proceedings of ACL-08:HLT, pages 308-316.

Peter D. Turney. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 417-424.

Vladimir N. Vapnik, 1995. The Nature of Statistical Learning Theory. Springer.

Xiaojun Wan. 2009. 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, pp. 235-243.

Hongning Wang, Yue Lu, ChengXiang Zhai. 2010. Latent Aspect Rating Analysis on Review Text Data: A Rating Regression Approach. Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'10), pages 115-124.

Jingbo Zhu, Huizhen Wang, Benjamin K. Tsou, and Muhua Zhu. 2009. Multi-aspect opinion polling from textual reviews. In Proceedings of the 18th ACM conference on Information and knowledge management, pp. 1799-1802. ACM.

Zhi-Hua Zhou and Ming Li. 2005. Semi-supervised regression with co-training. In Proceedings of the 19th international joint conference on Artificial intelligence, pp. 908-913. Morgan Kaufmann Publishers Inc.