2012 IEEE 12th International Conference on Data Mining

# Cross-Language Opinion Target Extraction in Review Texts

Xinjie Zhou, Xiaojun Wan* and Jianguo Xiao
Institute of Computer Science and Technology
The MOE Key Laboratory of Computational Linguistics
Peking University
Beijing, China
{zhouxinjie, wanxiaojun, xiaojianguo}@pku.edu.cn

***Abstract*—Opinion target extraction is a subtask of opinion mining which is very useful in many applications. In this study, we investigate the problem in a cross-language scenario which leverages the rich labeled data in a source language for opinion target extraction in a different target language. The English labeled corpus is used as training set. We generate two Chinese training datasets with different features. Two labeling models for Chinese opinion target extraction are learned based on Conditional Random Fields (CRF). After that, we use a monolingual co-training algorithm to improve the performance of both models by leveraging the enormous unlabeled Chinese review texts on the web. Experimental results show the effectiveness of our proposed approach.**

***Keywords-opinion mining; opinion target extraction; cross-language information extraction;***

## I. INTRODUCTION

The rapid development of e-commerce has boosted the research on product review analysis. Reviews from different users may focus on different product features. The task of opinion target extraction aims to extract the opinionated product features automatically. Two product reviews in English and Chinese and their product features are shown as below. The opinion targets which we aim to extract are underlined in the sentences.

(1) The iPod's sound quality is pretty good.
(2) iPod 的音质非常好.

Conditional Random Fields (CRF) [1] has been successfully applied to the opinion target extraction task in the past [2]. However, supervised learning techniques such as CRF exploit large amounts of annotated data to learn models that can label unseen data. Acquiring such annotated data in a language is important for opinion target extraction and it usually involves significant human efforts. Such corpuses in different languages are very imbalanced. To overcome this difficulty, we propose a novel cross-language method which leverages the English annotated opinion data for Chinese opinion target extraction. Though we focus on English-to-Chinese cross-language opinion target extraction in this study, the proposed method can be easily adapted for other languages.

*Xiaojun Wan is the corresponding author.

In our study, an English annotated corpus is translated into Chinese with the help of machine translation service. We then use natural language processing tools to parse both the original English corpus and the translated Chinese corpus. We can directly generate features from the Chinese corpus, and we can also project the features of the English corpus into Chinese using word alignment information. Thus, we get two Chinese training datasets with different features, one of which is generated from the Chinese corpus, and the other is projected from the English corpus. After training two labeling models with CRF on the two training sets, we use the co-training algorithm to improve the performance of both models by exploiting unlabeled Chinese data. Experimental results show the effectiveness of our proposed approach.

Our contributions in this study are summarized as follows: 1) we investigate a cross-language scenario for opinion target extraction in review texts, which can solve the resource-poor problem in a particular language. 2) We propose a monolingual co-training approach to improve the performance of cross-language opinion target extraction. 3) We empirically compare the proposed co-training approach and several baselines. The experimental results show the effectiveness of our proposed approach.

The rest of this paper is organized as follows: Section II introduces related work. The detailed method is revealed in Section III. Section IV shows the experimental results. Lastly we conclude this paper in Section V.

## II. RELATED WORK

### A. Opinion Mining

Opinion mining or sentiment analysis has attracted great attention with the explosive growth of online reviews and comments. Opinion mining can be performed on word, sentence and document levels. Opinion target extraction can be classified as a word level task.

Opinion target extraction is mostly performed on the product reviews in which opinion targets are always described as product features [4]. Hu and Liu [8] propose a method which extracts frequent nouns and noun phrases as the opinion targets, relying on a statistical analysis of the review terms based on

1550-4786/12 $26.00 © 2012 IEEE
DOI 10.1109/ICDM.2012.32

association mining. Zhuang et al. [6] present a supervised algorithm for the extraction of opinion word and opinion target pairs. Their algorithm learns the opinion target candidates and a combination of dependency and part-of-speech paths connecting such pairs from an annotated dataset. Jacob and Gurevych [2] model the problem as an information extraction task based on CRF. They compare the extraction performance in two different settings: single-domain and cross-domain.

### B. Cross-language Information Extraction

Opinion target extraction is considered as a special information extraction task [8]. Information extraction (IE) systems are costly to build because they require large training corpus and tool development. It is always a time-consuming and expensive work to annotate data for a particular language. Cross-language information extraction has become a popular solution to this problem which uses the corpus in one language to perform a certain task in another language.

Cross language information extraction has been investigated on several common subtasks. Yarowsky et al. [9] describe a system and a set of algorithms for automatically inducing stand-alone monolingual part-of-speech taggers, base noun-phrase bracketers, named-entity taggers and morphological analyzers for an arbitrary foreign language. Oh et al. [10] propose a bilingual co-training algorithm for hyponymy-relation acquisition from Wikipedia, which is, however, not a sequence labeling task.

To the best of our knowledge, cross-language opinion target extraction has not yet been investigated yet. Moreover, our approach trains and combines two models in a single language while previous cross-language methods for sequence labeling are based on a single model.

## III. Our Proposed Approach

### A. Motivation

Cross-language information extraction systems are usually built with cross-language projection which attempts to make training corpora available for new languages. If a parallel bilingual corpus is available, all that required is a tagged training corpus for an already developed language. Using the tagged training corpus, we can train a model and tag the parallel corpus in this language. We then project the tags across the parallel corpus based on text alignment. However, there are only a few bilingual parallel text corpora available, restricting the number of occasions when this architecture can be of use, for example, in the occasion of opinion target extraction. In such cases, machine translation is widely used for creating bilingual text corpus. Figure 1 shows the basic framework for this scenario. In the figure, $L_1$ and $L_2$ represent the developed language and the new language, respectively.

Besides translating the training corpus, we can also choose to translate the test corpus. In this circumstance, the model is directly trained using the dataset in the source language. However, it is not suitable for word level task. If it is applied to extract opinion target, we need to translate the test data in $L_2$ into $L_1$ for the labeler. After labeling the translated test data, the tagged opinion target must be projected back to $L_2$ again based on word alignment. Such approach will be very sensitive to the alignment error because it will directly cause a wrong target label. Therefore, we originally present a framework which builds two different models both in the new language and adopts the monolingual co-training algorithm to improve the performance.

### B. Framework

Our approach aims to leverage the English annotated corpus to train labeling models for Chinese opinion target extraction.

An overall framework of our approach is shown in Figure 2. We first translate the original English dataset into Chinese. Features are generated for both the two datasets. The feature projection stage helps to get two different Chinese datasets. Based on these two datasets and unlabeled Chinese review texts, the labeling model is trained using CRF and monolingual co-training algorithm.

The original English annotated dataset are first translated into Chinese using the online machine translation service - Bing Translate. We also use the word alignment results and the Chinese word segmentation results provided by Bing Translate. The word alignments are used to project labels of opinion targets in English annotated data into translated Chinese data. We directly label a Chinese word as Chinese opinion target if the word is aligned to an English opinion target word. If an English opinion target is aligned to separate Chinese words, we label all these Chinese words as different targets. If an English target is aligned to one or more continuous Chinese words, we label the sequence of these words as a single target.

After that, NLP tools are used to parse both the translated Chinese corpus and the original English corpus to generate the part-of-speech tag based features and typed dependency based features.

In the English feature projection stage, we project

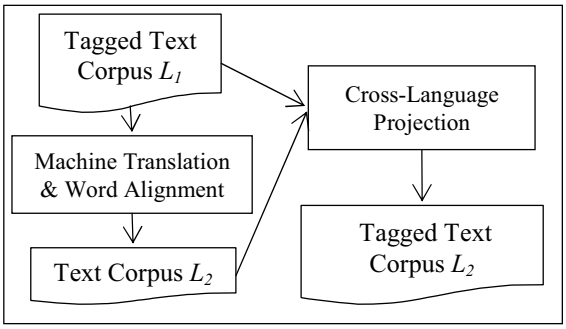


Figure 1. Cross-language projection with machine translation

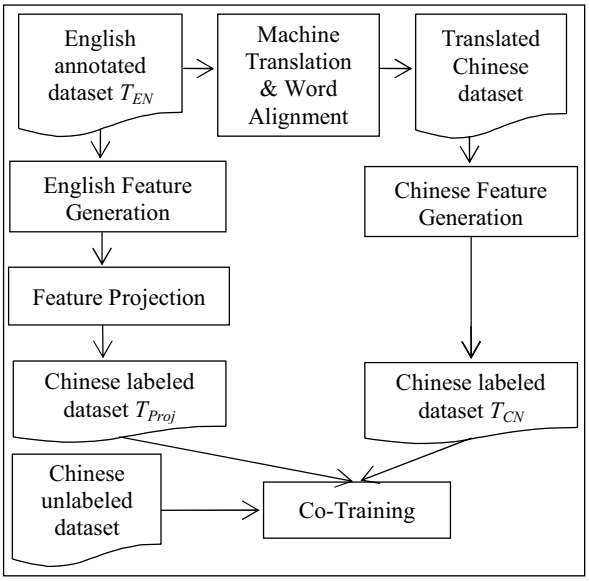


Figure 2. Framework of our approach

the English-side features to Chinese side based on word alignment results. For example, the POS tag of an English word is projected to the Chinese word which is aligned to this word. The other features can be similarly projected to the aligned Chinese words. Thus, we get two different views of features both on the translated Chinese data, one of which is directly obtained from the Chinese side, and the other of which is obtained by projecting the features from the English side. We consider the two views of the Chinese dataset as two labeled training datasets.

The linear-chain Conditional Random Felds model is used as the basic model in the monolingual co-training algorithm which learns opinion target labelers based on the two labeled datasets and an unlabeled dataset. We choose CRF++[1] for all the experiments.

### C. Feature Generation

#### 1) Feature Set

In our approach, we use four kinds of features. Word-based features are obtained from the Bing Translate service because it directly returns segmented Chinese words after translation. Part-of-speech tag based features and typed dependency based features are generated for the English and translated Chinese data using the Stanford Parser[2]. Opinion word type features are generated based on opinion lexicons in the two languages. The detailed feature types used in our model are introduced as below.

*a) Word-based Features*

The translated Chinese texts are segmented by the Bing Translate tool. Each Chinese word and English word is regarded as a feature. We also regard the combination of two continuous word pairs as features.

*b) POS-based Features*

The part-of-speech tag of a word is used as a feature. We also regard the combination of two continuous part-of-speech tag pairs as features. However, the English-side POS feature is different from the Chinese-side POS feature because they are based on the Penn English Treebank tag set and the Penn Chinese Treebank tag set, respectively. We will introduce the mapping strategy which makes them equivalent in the next subsection.

*c) Dependency Path-based Features*

Previous research [7] has shown the effectiveness of dependency path in opinion target extraction. Dependency path is formed by one or more dependency relations which connect two words in the dependency tree. The dependency path between the target and an opinion word is more likely to collapse into several types, such as “amod” (adjectival modifier), “nsubj” (nominal subject).

However, the accurate recognition of opinion word is also another difficult task, which will not be discussed in this study. We simply use a Chinese opinion lexicon and an English opinion lexicon to identify the opinion words. The Chinese opinion lexicon used here is the only Chinese sentiment resource in our approach. Compared to the domain-specific annotated corpus, the opinion lexicon is much easier to obtain. Alternatively, we can also translate the English opinion lexicon into other languages when the opinion lexicon does not exist. The Chinese NTU Sentiment Dictionary (NTUSD) and the English MPQA Subjectivity Lexicon are used in our experiments. They contain 10542 Chinese opinion words and 8221 English opinion words respectively. We only regard adjectives and verbs in the lexicon as opinion words. After that the dependency path-based feature of each word is defined as the shortest dependency path between the word and every opinion word in the sentence. If there is no opinion word, we use the path between the current word and the root of the dependency tree. We use the Stanford Parser to generate the dependency path for both English and Chinese.

The dependency path-based features have similar problems as the POS-based features because different languages have different relation sets. We will discuss the problem in the next subsection.

*d) Opinion Word Type Feature*

We use three different numbers to label each word in a sentence according to which type of opinion word the sentence has: verb opinion word, adjective opinion word or no opinion word. If a sentence contains several opinion words with different part-of-speech tags, we use the type of the nearest opinion word in the dependency tree to label each word. It is reasonable to induce this feature because different part-of-speech tags of the opinion word may indicate the different dependency path-based features.

#### 2) Feature Mapping

[1] http://crfpp.sourceforge.net/

[2] http://nlp.stanford.edu/software/lex-parser.shtml. The parser supports both Chinese and English languages.

As mentioned above, the English POS tag set is based on Penn English Treebank while the Chinese POS tag set is based on Penn Chinese Treebank. These two tag sets differ from each other. For example, the English language has morphological variation while Chinese does not. So English has different verb tags such as VB, VBD, VBP, VBN, VBZ and VBG to indicate different inflection categories. To address this problem, we map the two tag sets into coarse-grained POS categories. We rely on the twelve universal part-of-speech tags of [11]. Because there might be some controversy about the exact definitions of such universal tags, this set of coarse-grained POS categories is defined operationally, by collapsing language (or treebank) specific distinctions to a set of categories that exists across both languages.

The dependency path-based features have similar problems. Dependency path is formed by one or more dependency relations which are designed to provide a simple description of the grammatical relationships in a sentence. Chinese dependency relations are different from the English ones because of the different grammatical structure of the two languages. Chang et al. [12] find 45 distinct dependency types in Chinese, and 50 in English. They only share a subset of 18 types. However, it is very difficult to map the dependency relations into coarse-grained categories. Fortunately, the frequent dependency relations between opinion targets and opinion words are included in the shared subset. So, we keep all the 50 English relations for English-side feature and 45 Chinese relations for Chinese-side feature.

### D. Featue Projection

In the feature projection stage, we project the features in $T_{EN}$ to the translated Chinese corpus to get another training dataset $T_{Proj}$ , which means the aligned Chinese-English word pair in $T_{Proj}$ and $T_{EN}$ share the same features except the word-based features.

The two datasets $T_{Proj}$ and $T_{CN}$ have the same word-based feature and the same target label for each word. However, the features in $T_{CN}$ are directly generated from the translated Chinese text while the features in $T_{Proj}$ are projected from the English corpus $T_{EN}$. Thus, we get two different Chinese datasets.

### E. Monolingual Co-Training

The co-training algorithm [13] is a semi-supervised learning technique that requires two views of the data. It uses an unlabeled dataset to increase the amount of annotated data in an incremental way. Co-training has been successfully used for a few NLP tasks, including relation extraction [10], text classification [3] and so on. We use the co-training algorithm for the cross-language target extraction task due to the following two reasons: 1) we can train two different models based on two different training datasets. 2) Although we are lack of annotated Chinese corpus, the unlabeled Chinese product reviews can be easily obtained from the web.

Given:

$T_{CN}$ = Chinese training data with features generated from translated Chinese corpus.

$T_{Proj}$ = Chinese training data with features projected from English corpus.

$UD_1 = UD_2$ =Unlabeled Chinese data.

Algorithm:

1. Train model $M_1$ using $T_{CN}$.
2. Train model $M_2$ using $T_{Proj}$.
3. Loop for $I$ iterations:
   1) Get the labeled data $LUD_1$ by labeling $UD_1$ with $M_1$.
   2) Get the labeled data $LUD_2$ by labeling $UD_2$ with $M_2$.
   3) Select a subset $SUD_1$ from $LUD_1$ that contains $N$ most confidently labeled examples.
   4) Add $SUD_1$ to $T_{Proj}$ and remove $SUD_1$ from $UD_1$
   5) Select a subset $SUD_2$ from $LUD_2$ that contains $N$ most confidently labeled examples.
   6) Add $SUD_2$ to $T_{CN}$ and remove $SUD_2$ from $LUD_2$
   7) Re-train model $M_1$ using $T_{CN}$.
   8) Re-train model $M_2$ using $T_{Proj}$.

   End of loop.
4. Combine $M_1$ and $M_2$ using OR merger.

Figure 3. The monolingual co-training algorithm

As shown in Figure 3, we start with two different labeled datasets ($T_{CN}$ and $T_{Proj}$). Two models $M_1$ and $M_2$ are trained on these datasets using CRF. In each iteration we use $M_1$ and $M_2$ to label the unlabeled data $UD_1$ and $UD_2$, respectively. Note that $UD_1$ and $UD_2$ are the same before the co-training starts. We select $N$ most confidently labeled examples by $M_1$ and add them to $T_{Proj}$. Similarly, $N$ most confidently labeled examples by $M_2$ are added to $T_{CN}$. These examples with high confidence are removed from $UD_1$ and $UD_2$. Then $M_1$ and $M_2$ are re-trained with the enlarged datasets $T_{CN}$ and $T_{Proj}$, respectively. This process is repeated for $I$ iterations. At last, we use the OR merger which is used in [14] to combine the labeling results of the two component models together. It means that a word will be regarded as a target if it is labeled as a target by one or more models. The two parameters $N$ and $I$ will be referred as growth size and iteration in the later discussion.

## IV. EXPERIMENTS

### A. Dataset

The following three datasets are collected and used in the experiments:

TABLE I. COMPARISON RESULTS OF DIFFERENT MODELS ($N$=1000, $I$=20 FOR CO-TRAINING)

| Method | Strict | | | Lenient | | |
|---|---|---|---|---|---|---|
| | Precision | Recall | F-measure | Precision | Recall | F-measure |
| **UN** | 0.120 | **0.360** | 0.180 | 0.292 | **0.875** | 0.438 |
| **M1** | **0.419** | 0.128 | 0.196 | **0.832** | 0.254 | 0.389 |
| **M2** | 0.283 | 0.157 | 0.202 | 0.728 | 0.405 | 0.520 |
| **CT(M1)** | 0.336 | 0.299 | 0.316 | 0.748 | 0.686 | 0.715 |
| **CT(M2)** | 0.317 | 0.281 | 0.298 | 0.747 | 0.662 | 0.702 |
| **CT** | 0.313 | 0.327 | **0.320** | 0.721 | 0.754 | **0.737** |

*1) Chinese Test Set:* We use the dataset of Chinese Opinion Analysis Evaluation (COAE) 2008 [3] which includes the task of opinion target extraction. This test set contains reviews on four domains including camera, car, notebook and phone. The dataset contains 473 files and a total of 8177 targets.

*2) English Training Set:* We use the customer review collection from [5] as the training set. The collection contains five English review datasets: two on two different digital cameras, one on a DVD player, one on an mp3 player, and one on a cell phone. It contains 3945 sentences and 2962 targets totally.

*3) Chinese Unlabeled Set:* We download product reviews on the four testing domains. The unlabeled reviews on camera, phone and notebook are downloaded from the popular Chinese IT product website ZOL [4]. The unlabeled car reviews are downloaded from the Chinese car website Bitauto[5]. The final unlabeled dataset is formed by mixing the four domain datasets with equal amount. It totally contains 20,000 reviews and about 100,000 sentences.

For Chinese test set and unlabeled set, the reviews are firstly segmented with a popular Chinese word segmenter - ICTCLAS [6] to generate the word-based feature. The other features are derived in the same way as the dataset $T_{CN}$ which has been discussed in Section III.

## B. Evaluation Metrics

We use the same evaluation metrics as COAE. Precision, recall and F-measure are used to measure the performance. Results are reported in two different ways: strict and lenient. Strict evaluation means that a proposed target is right if it covers exactly the same span with the correct answer target. Lenient evaluation means that a proposed target is right if the spans of the proposed target and the correct answer target overlap. Following the COAE 2008 guidelines, only the opinionated sentences are used in the evaluation.

[3] http://ir-china.org.cn/coae2008.html. COAE is the most authoritative evaluation for opinion mining in China.
[4] http://www.zol.com.cn/
[5] http://www.bitauto.com/
[6] http://ictclas.org/

## C. Baselines

In the experiments, we compare our proposed CT model with five baseline models, and they are described as follows:

**UN:** We implement an unsupervised method based on [5], which relies on association mining and a sentiment dictionary to extract frequent and infrequent product features.

**M1:** The model is trained using $T_{CN}$ without co-training.

**M2:** The model is trained using $T_{Proj}$ without co-training.

**CT(M1):** After training the model M1 using $T_{CN}$, the monolingual co-training algorithm is used to improve the performance of M1.

**CT(M2):** After training the model M2 using $T_{Proj}$, the monolingual co-training algorithm is used to improve the performance of M2.

**CT:** After obtaining two component models CT(M1) and CT(M2) with co-training, we use the OR merger to combine the components together.

We simply set the co-training parameters as $N$=1000 and $I$=20 in our experiment, the influence of the parameters will be discussed later.

## D. Results

Among the six models in Table 1, the unsupervised method achieves poor results on both strict and lenient evaluation. It gets high recall but very low precision, which means that many frequent nouns or nouns groups in the dataset are not opinion target. The co-training model CT achieves the best results. Each component model of co-training gets significant improvement over the original model. For example, the F-measure of model CT(M1) increases by 0.12 and 0.326 on strict and lenient evaluation compared to the original model M1. The OR merger helps the model CT to increase slightly over F-measure compared to the two component models CT(M1) and CT(M2). We can see a decline in precision but an increase in recall on both strict and lenient evaluation, which is reasonable because the OR operator is used. The use of unlabeled data improves the performance for the co-training algorithm. The good performance of our monolingual co-training algorithm proves that the two different views of the dataset can make up for the shortage of each other.

### E. Discussion

*Influence of Growth Size N in Co-Training*

Figures 4 and 5 show how the growth size influences the F-measure score of the proposed co-training approach. We plot the F-measure scores of the co-training model on both lenient evaluation and strict evaluation in the two figures respectively. In Figure 4 we can see that the F-measure score increases faster during the initial few iterations with the increase of *N*. However, the curve with a growth size of 1500 becomes steady after 30 iterations while the curve with a growth size of 1000 keeps increasing and gets the highest F-measure score of 0.758. The F-measure score on strict evaluation in Figure 5 shows the similar trend that a larger *N* helps the performance increase faster. A significant difference between the lenient evaluation in Figure 4 and the strict evaluation in Figure 5 is that the curves in Figure 4 become steady after several iterations while the curves in Figure 5 decline. Furthermore, the curves in Figure 5 decline faster for larger growth size. This is because the strict evaluation is more sensitive to wrong target labels than the lenient evaluation.

## V. Conclusion and Future Work

In this paper, we propose a cross-language opinion target extraction approach using the monolingual co-training algorithm. We do not use any labeled Chinese dataset except for an annotated English product review dataset. The online unlabeled Chinese review data are downloaded to improve the performance in the co-training approach. Both of our two component models are trained with the translated Chinese dataset containing much noise. We successfully overcome this difficulty with the co-training algorithm. Evaluation results show the effectiveness of our approach.

In future work, we will try to exploit more useful features for further improving the opinion target extraction performance, including Semantic role labeling (SRL), etc.

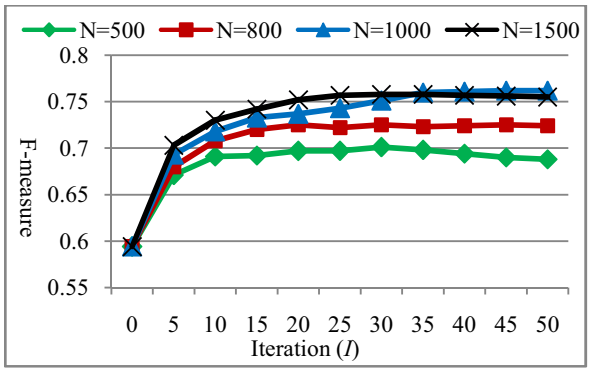


Figure 4. F-measure on lenient evaluation different growth size.

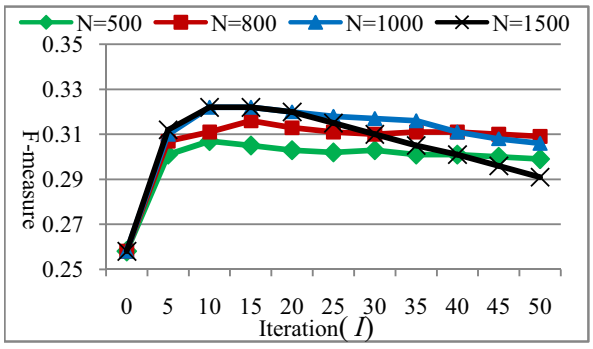


Figure 5. F-measure on strict evaluation with different growth size.


## Acknowledgment

The work was supported by NSFC (61170166), Beijing Nova Program (2008B03), NCET (NCET-08-0006) and National High-Tech R&D Program (2012AA011101).


## References


[1] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira, "Conditional Random Felds: Probabilistic Models for Segmenting and Labeling Sequence Data," In Proceedings of the 18th International Conference on Machine Learning, 2001, pp. 282–289.

[2] Niklas Jakob and Iryna Gurevych, "Extracting Opinion Targets in a Single- and Cross-Domain Setting with Conditional Random Fields," In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, 2010, pp. 1035–1045.

[3] Xiaojun Wan, "Co-Training for Cross-Lingual Sentiment Classification," In Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, 2009, pp. 235–243.

[4] Bing Liu, "Web Data Mining: Exploring Hyperlinks, Contents and Usage Data," Springer, Berlin, 2006.

[5] Mingqing Hu and Bing Liu, "Mining and Summarizing Customer Reviews," In Proceedings of SIGKDD'04, 2004, pp. 168–177.

[6] Li Zhuang, Feng Jing, and Xiao-Yan Zhu, "Movie Review Mining and Summarization," In Proceedings of the ACM 15th Conference on Information and Knowledge Management, 2006, pp. 43–50.

[7] Guang Qiu, Bing Liu, Jiajun Bu and Chun Chen, "Opinion Word Expansion and Target Extraction through Double Propagation," Computational Linguistics, 37(1), 2011, pp. 9–27.

[8] James R. Cowie and Wendy G. Lehnert, "Information Extraction," Communications of the ACM, 39(1), 1996, pp. 80–91.

[9] D. Yarowsky, G. Ngai, and R. Wicentowski, "Inducing Multilingual Text Analysis Tools via Robust Projection across Aligned Corpora," In Proceedings of HLT 2001, 2011 pp. 1–8.

[10] Jong-Hoon Oh, Kiyotaka Uchimoto and Kentaro Torisawa, "Bilingual Co-Training for Monolingual Hyponymy-Relation Acquisition," In Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, 2009, pp. 432–440.

[11] Slav Petrov, Dipanjan Das, and Ryan McDonald, "A Universal Part-of-Speech Tagset," 2011, ArXiv:1104.2086.

[12] Pi-Chuan Chang, Huihsin Tseng, Dan Jurafsky, and Christopher D. Manning, "Discriminative Reordering with Chinese Grammatical Relations Features," In Proceedings of SSST-3, Third Workshop on Syntax and Structure in Statistical Translation, 2009, pp. 51–59.

[13] Avrim Blum and Tom Mitchell, "Combining Labeled and Unlabeled Data with Cotraining," In Proceedings of COLT-98. 1998, pp. 92–100.

[14] Ferenc P. Szidarovszky, Ill´es Solt, Domonkos Tikk, "A Simple Ensemble Method for Hedge Identification," In Proceedings of the Fourteenth Conference on Computational Natural Language Learning: Shared Task, 2010, pp. 144–147.