CLOpinionMiner: Opinion Target Extraction in a Cross-Language Scenario

Xinjie Zhou, Xiaojun Wan* and Jianguo Xiao

***Abstract*—Opinion target extraction is a subtask of opinion mining which is very useful in many applications. The problem has usually been solved by training a sequence labeler on manually labeled data. However, the labeled training datasets are imbalanced in different languages, and the lack of labeled corpus in a language limits the research progress on opinion target extraction in this language. In order to address the above problem, we propose a novel system called CLOpinionMiner which investigates leveraging the rich labeled data in a source language for opinion target extraction in a different target language. In this study, we focus on English-to-Chinese cross-language opinion target extraction. Based on the English dataset, our method produces two Chinese training datasets with different features. Two labeling models for Chinese opinion target extraction are trained 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.**

***Index Terms*—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. The feedback in the reviews can help the customers choose among different products and help the manufacturers improve the product quality. The task of opinion target extraction aims to automatically extract the entity to which the opinion is expressed. Two product reviews in English and Chinese and their opinion targets 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.

Opinion target extraction usually relies on supervised learning algorithms such as Conditional Random Fields (CRF) [2]. However, these techniques exploit large amounts of annotated data to train 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. Besides, such corpora in different languages are very imbalanced. The amount of labeled sentiment data in English is much larger than those in other languages such as Chinese. To overcome this difficulty, we propose a new system called CLOpinionMiner 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.

Most of existing cross-language opinion mining work focuses on the task of sentiment classification. It aims to classify the sentiment polarity of texts into positive or negative. In most approaches, machine translation engines are directly used to adapt labeled data from the source language to the target language. To overcome the defection of machine translation, Wan [3] tried to translate both the training data (English to Chinese) and the test data (Chinese to English). Two models for sentiment classification are trained in both the source and target languages. A co-training algorithm is used to combine the bilingual models and improve the performance. Inspired by [3], an intuitive approach is to directly use this method to solve our opinion target extraction problem. However, the approach of [3] is not suitable for word level task. If it is applied to extract opinion target, we need to translate the test data for the labeler. After labeling the translated test data, the tagged opinion target must be projected back to the source language again based on word alignment. Such approach will be very sensitive to the alignment error because each alignment error will directly cause a wrong target label. Therefore, we originally present a framework which builds two different models both in the source language and adopts the monolingual co-training algorithm to improve the performance.

In our approach, an English annotated corpus is translated into Chinese with the help of machine translation service. We use natural language processing tools to parse both the original English corpus and the translated Chinese corpus. We can directly use features generated from the Chinese corpus, and we can also project the features of the English corpus into Chinese using word alignment information. For example, to get the part-of-speech tag feature of a Chinese word “相机” (“camera”), we can directly use a Chinese POS tagger to tag the Chinese word “相机” or use an English part-of-speech tagger to tag the English word “camera” and project the result to the Chinese word “相机” based on the alignment information between them.

This work was supported by National Hi-Tech Research and Development Program (863 Program) of China (2014AA015102, 2012AA011101) and National Natural Science Foundation of China (61170166, 61331011). *(Corresponding author: X. Wan.)*

Xinjie Zhou is with Institute of Computer Science and Technology, the MOE Key Laboratory of Computational Linguistics, Peking University, Beijing 100871, China. (e-mail: zhouxinjie@pku.edu.cn).

Xiaojun Wan is with Institute of Computer Science and Technology, the MOE Key Laboratory of Computational Linguistics, Peking University, Beijing 100871, China. (e-mail: wanxiaojun@pku.edu.cn).

Jianguo Xiao is with Institute of Computer Science and Technology, the MOE Key Laboratory of Computational Linguistics, Peking University, Beijing 100871, China. (e-mail: xiaojianguo@pku.edu.cn).

Thus, we get two Chinese training datasets with different features, one of which is generated from the translated Chinese corpus, and the other is projected from the original English corpus. We map the features in both datasets into a unified feature space, which means the two training datasets can adapt to the same Chinese test dataset. After training two labeling models with CRF based on the two training sets, we use the co-training algorithm to improve the performance of both models by exploiting unlabeled Chinese data.

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 and has not been investigated yet. 2) We propose a monolingual co-training approach to improve the performance of cross-language opinion target extraction. The proposed approach can also be used for other cross-language information extraction tasks. 3) We empirically compare the proposed co-training approach and several baselines. The experimental results show the effectiveness of our approach.

This journal article is substantially extended from our previous work [1]. First, we give a more detailed and thorough description of our approach in this paper. Several examples are added to help the readers understand our strategy. Second, we explicitly discuss the differences of our two components and explain why the co-training algorithm will be effective in our scenario. Third, new experimental results are added using the self-training algorithm. Performances of self-training and co-training are compared and discussed. A new rule-based baseline is added. Four, we compare our results with those from the Chinese Opinion Analysis Evaluation (COAE) 2008 and discuss the differences. Five, we analyze the result of our system and discuss some drawbacks.

The rest of this paper is organized as follows: Section II briefly presents some preliminaries. Section III introduces related work. We introduce our motivation in Section IV. The detailed approach is revealed in Section V. Section VI shows the experimental results. Lastly we conclude this paper in Section VII.

II. Preliminaries

A. Definition of opinion target

In this study, we aim to extract opinion targets from review texts which are very common on the e-commerce websites. Figure 1 shows an example of a review.

> Posted by: John Smith Date: September 10, 2011
>
> (1) I bought a Canon G12 camera six months ago. (2) I simply love it. (3) The picture quality is amazing. (4) The battery life is also long. (5) It has a 2.8-inch PureColor System LCD screen. (6) I just returned from a wonderful trip to Jamaica where I took many pics with it.

Figure 1. Example of a review

Opinion target is defined as the entity to which the opinion is expressed. For example, sentence (2) in the above figure expresses a positive opinion about "it" (i.e. Canon G12). Sentence 3 expresses a positive opinion about the "picture quality". In this study, we aim to locate the opinion target of a sentence automatically.

B. Opinion target extraction

In review texts, the opinion targets are always aspects of a product, such as battery life, screen, signal etc. Therefore, opinion target extraction is often referred to as aspect extraction. However, the two tasks have a major difference. Aspect extraction aims to find a lexicon of aspects for a given product while opinion target extraction aims to find the opinion target of each review sentence. In sentence (5) of Figure 1, "LCD screen" should not be regarded as an opinion target, but aspect extraction should identify it as an aspect from a large corpus of review. In sentence (6), "trip" should be extracted in an opinion target task but should be ignored in an aspect extraction task. Despite of the differences, some literatures do not distinguish between the two tasks and always describe an aspect extraction task as an opinion target extraction task. In this study, we mainly focus on opinion target extraction instead of aspect extraction.

Opinion target extraction is difficult due to the following two reasons: 1) Opinion target extraction is a fine-grained task. While coarse-grained tasks like document or sentence-level sentiment classification (Pang et al., 2002) only need to employ simple features such as word tokens or part-of-speech tags, opinion target extraction relies on deeper knowledge such as syntax structure. 2) Opinion target extraction can be regarded as a specific information extraction task, but it is more complicated. Opinion target is always bounded to an opinion expression. However, it is still a difficult issue to model the close relationship between an opinion expression and its targets in supervised learning approaches [6].

We model problem as a sequence labeling task. Denote $T = \{(w^i, y^i)_{i=1}^{N}\}$ as the training dataset, where w^i represents each word in the dataset, y^i represents the corresponding label of w^i and N is the number of words. We adopt the IOB scheme, which means

$$y^i = \begin{cases} B & w^i\ is\ the\ beginning\ of\ a\ target \\ I & w^i\ is\ inside\ a\ target \\ O & w^i\ is\ outside\ a\ target \end{cases} \tag{2}$$

We use linear-chain Conditional Random Fields to train the model. A linear-chain CRF is a distribution p(y|x) that takes the form

$$p(y\mid x) = \frac{1}{Z(x)}\prod_{t=1}^{T}\exp\left(\sum_{k=1}^{K} I_k f_k\left(y_{t-1}, y_t, x_t\right)\right) \tag{3}$$

where $I = \{I_k\} \in R^K$ is the parameter vector, $\{f_k(y_{t-1}, y_t, x_t)\}_{k=1}^{K}$ is a set of real-valued feature functions. The parameters can be estimated by maximum likelihood, that is, the parameters are chosen such that the training data have highest probability under the model. Z(x) is an instance-specific normalization function.

$$Z(x) = \sum_{y\in Y}\prod_{t=1}^{T}\exp\left(\sum_{k} I_k f_k\left(y_{t-1}, y_t, x_t\right)\right) \tag{4}$$

C. Cross-language opinion target extraction

The task of cross-language opinion task extraction is developed to address the difficulties when we only have an annotated corpus in a source language and we need to build an opinion target extractor in a different target language. Particularly, we intend to develop a Chinese opinion target extraction system by leveraging English training data. Denote the English training dataset as $T_E = \{(w_E^i, y_E^i)_{i=1}^{N_E}\}$ and the Chinese test dataset as $T_C = \{(w_C^i)_{i=1}^{N_C}\}$. Our task is to label each Chinese word w_C^i as B, I or O using the CRF model trained by T_E.

In our study, the original English annotated corpus is first translated into Chinese with the help of a machine translation service. We generate two Chinese training datasets with different features, one of which is generated from the translated Chinese corpus, and the other is projected from the English corpus. In addition, we propose a monolingual co-training algorithm to improve the performance.

III. Related Work

A. Opinion Mining and Opinion Target Extraction

Opinion mining is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions [7].

Most of the previous sentiment analysis researches focus on customer reviews [9][11] and some of them focus on news [13] and blogs [14]. Classification of opinion polarity is the most common task studied in review texts. Pang et al. [9] regarded it as a text classification problem. They used existing supervised learning methods such as naive Bayes classification, support vector machines (SVM) and achieved promising results. In subsequent research, more features and learning algorithms were tried for sentiment classification by a large number of researchers, such as the syntactic relation feature [15], Delta TF-IDF weighting scheme [16], minimum cut algorithm [17], non-negative matrix factorization method [18] and so on.

Compared to sentiment classification, opinion target extraction is a finer-grained and more complicated task. It requires deeper natural language processing capabilities and produces a richer set of results. Hu and Liu [8] proposed a method which extracted frequent nouns and noun phrases as the opinion targets, relying on a statistical analysis of the review terms based on association mining. The same dataset of product reviews was used in the work of [12]. They presented and evaluated a complete system for opinion extraction, and it used the Likelihood Ratio Test for opinion target extraction. Besides the product reviews, Kim and Hovy [13] aimed at extracting opinion holders and opinion targets in newswire. Their method relied on semantic role labeling. It defined a mapping of the semantic roles identified with FrameNet to the respective opinion elements. Liu et al. [19] used the word translation model in a monolingual scenario to mine the associations between opinion targets and opinion words.

Besides the above unsupervised methods, Zhuang et al. [20] presented a supervised algorithm for the extraction of opinion word - opinion target pairs. Their algorithm learned the opinion target candidates and a combination of dependency and part-of-speech paths connecting such pairs from an annotated dataset. Jacob and Gurevych [5] modeled the problem as an information extraction task based on CRF. They compared the extraction performance in two different settings: single-domain and cross-domain. Qiu et al. [21] proposed a double propagation method to extract opinion word and opinion target simultaneously. Li et al. [6] explored supervised opinion target extraction from a parse tree structure perspective and formulated it as a shallow semantic parsing problem. Yang and Cardie [22] jointly extracted the opinion expressions, the opinion holders, and the targets of the opinions, and the relations. Their approach is evaluated based on a standard corpus for fine-grained opinion analysis - the MPQA corpus and the results outperform traditional baselines by a significant margin.

B. Cross-language Opinion Mining and Cross-language Information Extraction

Cross-language opinion mining has been extensively studied in the very recent years. However, almost all of the existing works focus on the task of cross-language sentiment classification. Mihalcea et al. [23] experimented with translating English subjectivity words and phrases into the target language to build a lexicon-based classifier in the target language. Wan [3] translated both the training data (English to Chinese) and the test data (Chinese to English) to train different models in both the source and target languages. The co-training algorithm [35] was used to combine the bilingual models together and improve the performance. Lu et al., [8]attempted to jointly classify the sentiment for both source language and target language, assuming that there was a certain amount of sentiment labeled data available for both the source and target languages, and there was also an unlabeled parallel corpus. Meng et al. [24] proposed a generative cross-lingual mixture model to leverage unlabeled bilingual parallel data for sentiment classification. Wan [25] conducted a comparative study to explore the challenges of cross-lingual sentiment classification and proposed an ensemble system which combines different individual schemes.

Opinion target extraction is also considered as a special information extraction task [26]. Information extraction (IE) systems are costly to build because they require large training corpus and tool development. Cross language information extraction has been investigated on several common subtasks. Yarowsky et al. [27] described 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. Case studies included French, Chinese, Czech and Spanish. Kim et al. [28] developed a cross-lingual annotation projection method that leverages parallel corpora to bootstrap a relation detector without significant annotation efforts for a resource-poor language. Zitouni and Florian [29] presented and investigated a method of propagating mention detection from a resource-rich language

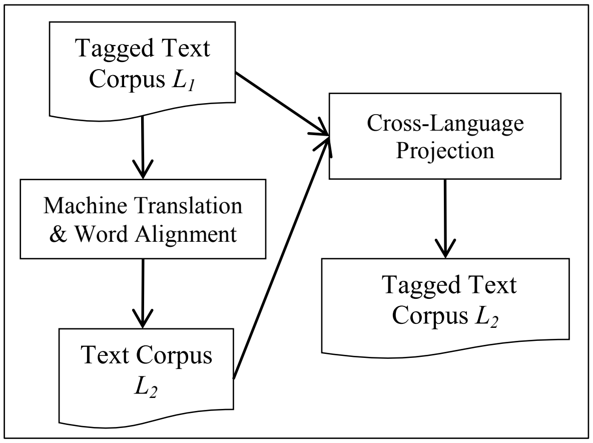


Figure 2. Cross-language projection with machine translation

into a relatively resource-poor language such as Arabic.

In addition to cross-language opinion mining and information extraction, there are many other tasks studied in the cross-language scenario, such as cross-language hyponymy-relation acquisition [30], cross-language information retrieval [31], cross-language summarization [32] etc. However, these works are not quite related to the presented study.

To the best of our knowledge, cross-language opinion target extraction has not yet been well investigated yet. Our CLOpinionMiner system trains and combines two models in a single language which is quite different from traditional cross-language opinion mining methods or cross-language information extraction methods. We believe that our method can be applied not only to opinion target extraction but also to other information extraction tasks.

IV. 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 is 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. Based on the translated corpus, we can train a model in the new language. Figure 2 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.

Wan [3] adopted both of the above two strategies and used the co-training algorithm to combine the bilingual models together. However, the approach of [2] is not suitable for our 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. In our method, one of the dataset projects the features in the source language into the target language, which is the biggest difference with traditional methods.

V. Our Proposed Approach

A. Framework

Our approach aims to leverage the English annotated corpus to train labeling models for Chinese opinion target extraction. An overall framework is shown in Figure 3. 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 reviews, the labeling model is trained using CRF and monolingual co-training algorithm. 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. Actually, it would be easier to implement for similar language pairs such as English and German which are both of the Germanic languages. Chinese and English are quite different which makes the problem even harder.

The original English annotated dataset is first translated into Chinese using the online machine translation service - Bing Translator, which was developed by Microsoft Research and achieved the top Chinese-English MT performance in the 2008

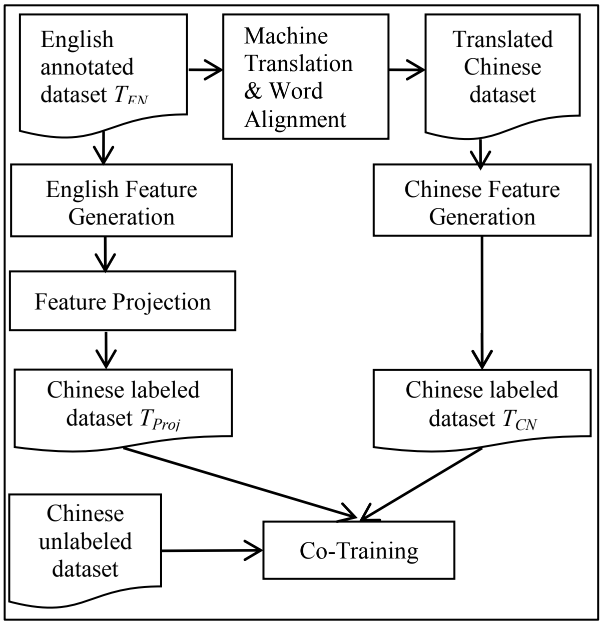

English annotated dataset T_{EN}
Machine Translation & Word Alignment
Translated Chinese dataset
English Feature Generation
Chinese Feature Generation
Feature Projection
Chinese labeled dataset T_{Proj}
Chinese labeled dataset T_{CN}
Chinese unlabeled dataset
Co-Training

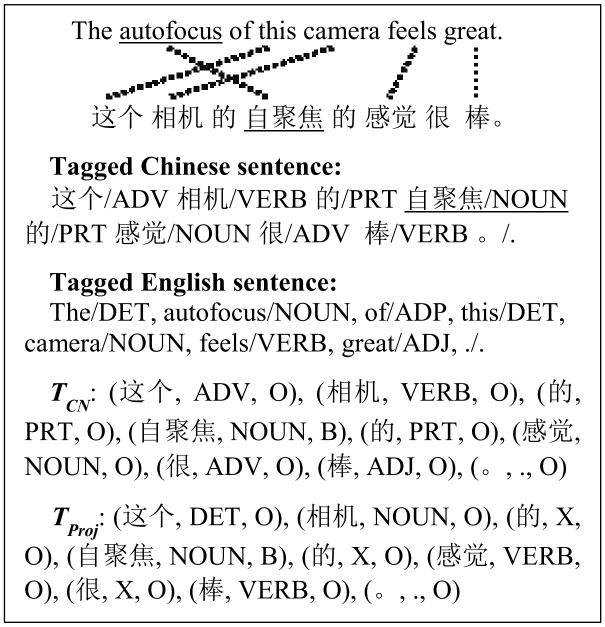
Figure 3. Framework of our approach

National Institute of Standards and Technology's (NIST) Open MT evaluation series. 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. In addition to word-based features, two kinds of features can be generated in English and Chinese languages based on the parsing results: part-of-speech tag based features and typed dependency based features. Since the above features on the English and Chinese sides are based on different POS tag sets and dependency relation type sets, we need to map the tag sets and relation type sets of the two sides into a unified feature space to make them equivalent. Detailed strategy will be discussed in next subsections.

In the English feature projection stage, we project the English-side features to Chinese side based on word alignment results. For example, the POS tag *NNS* of an English word "camera" is projected to the Chinese word "相机" which is aligned to "camera". 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 T_{CN} and T_{Proj}.

The two datasets have the same word-based features but are different for all other kinds of features. Both T_{CN} and T_{Proj} suffer from the imperfection of the machine translation and word alignment tools. The noise induced by machine translation will cause a large error rate during the parsing stage in T_{CN}. Compared to the Chinese side, the English-side parsing results will be more reliable. However, T_{Proj} will be influenced by the alignment error because they are projected from the English side. Besides, some Chinese words may get all the features except the word-based feature as null value if they are not aligned to any English word.


The autofocus of this camera feels great.
这个 相机 的 自聚焦 的 感觉 很 棒。

Tagged Chinese sentence:
这个/ADV 相机/VERB 的/PRT 自聚焦/NOUN 的/PRT 感觉/NOUN 很/ADV 棒/VERB 。/.

Tagged English sentence:
The/DET, autofocus/NOUN, of/ADP, this/DET, camera/NOUN, feels/VERB, great/ADJ, ./.

$\boldsymbol{T_{CN}}$: (这个, ADV, O), (相机, VERB, O), (的, PRT, O), (自聚焦, NOUN, B), (的, PRT, O), (感觉, NOUN, O), (很, ADV, O), (棒, ADJ, O), (。, ., O)

$\boldsymbol{T_{Proj}}$: (这个, DET, O), (相机, NOUN, O), (的, X, O), (自聚焦, NOUN, B), (的, X, O), (感觉, VERB, O), (很, X, O), (棒, VERB, O), (。, ., O)

Figure 4. An example of T_{CN} and T_{Proj}

A simple example is shown in Figure 4. T_{CN} and T$_{Proj}$ are the two training datasets. For simplicity, we only list the word-based feature, pos-based feature and the label. Each instance is represented as (*word*, *part-of-speech tag*, *label*). The original English sentence is "*The autofocus of this camera feels great.*" which contains the target "*autofocus*". The Chinese sentence is translated from English with machine translation tools. The dotted lines represent the word alignments. At the beginning, only the English sentence is labeled. Since the word "autofocus" is the opinion target, the sentence is labeled as "The/O autofocus/B of/O this/O camera/O feels/O great/O." Firstly, we label the Chinese opinion target "*自聚焦*" (*autofocus*) as "*B*" based on the word alignment. The other words are labeled as "O" since they are not aligned to any English opinion targets. The part-of-speech tags in T_{CN} are generated by the Chinese-side parser. We use an English-side parser to generate the POS tags for the English sentence. These English POS tags and labels are projected to T_{Proj} based on word alignment. For example, the Chinese word "这个" in T_{Proj} gets the POS tag "DET" and the label "O" because it is aligned to the English word "this". Both the English and Chinese POS tags have been mapped to the universal part-of-speech tags which will be discussed later. We can find that the part-of-speech tags in the two Chinese sentences are a bit different. In T_{CN} "*相机*" (*camera*) is wrongly recognized as verb. T_{Proj} correctly recognizes it as noun but it tags the word "的" and "很" as "X" because these two words are not aligned to any English word. T_{CN} and T_{Proj} have different properties and can work complementarily to make up the shortage of each other. It is an important reason that we use them together. In conclusion, the features in T_{Proj} are more accurate because the machine translation process introduces much noise for parsing T_{CN}. The other training dataset T_{CN} is more adaptable for the test dataset because the features are directly parsed on Chinese texts.

The linear-chain Conditional Random Fields 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.

B. 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.

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

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. All the word-based features are referred as w_{CN} for Chinese and w_{EN} for English.

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. All the POS-based features are referred as pos_{CN} for Chinese and pos_{EN} for English. 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 [21] 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 CLOpinionMiner. 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 [23]. 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. We refer to this feature dep_{CN} for Chinese and dep_{EN} for English. Figure 5.a shows two examples for dep_{CN} and dep_{EN}, respectively.

The relations between opinion target and opinion word are also illustrated in Figure 5.b. The opinion targets are underlined and the opinion words or roots are shown in italics. We only display the dependency path feature of the target word in Figure 5.b. The first one contains a verb opinion word "*love*" and the target "*shape*" is the direct object of "*love*". The second one contains an adjective opinion word and has the dependency path "nsubj" for the target "*sound*". The third sentence contains none opinion word. Chinese sentences also have similar dependency relations between opinion targets and opinion words.

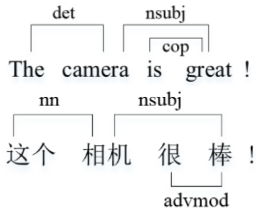
The dependency path-based features have similar problems as the POS-based features because different languages have different relation sets. We will discuss the problem in the next subsection.

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. For example, all the words in the first sentence in Figure 5.b will get the feature value of 0 because the sentence contains a verb opinion word. The words in the second sentence will be labeled as 1 and the words in the third sentence will be labeled as 2. 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. We refer to this feature as owt_{CN} for Chinese and owt_{EN} for English.

To sum up the above, the original English training dataset can be represented with all the features as

$$T_{EN} = \{(w^i_{EN}, pos^i_{EN}, dep^i_{EN}, owt^i_{EN}, y^i_{EN})_{i=1}^{N_{EN}}\} \qquad (5)$$



(a). Dependency path for Chinese and English datasets. The important relation "nsubj" remains the same for the two languages.

1) I *love* the shape of the MP3 player.
dobj – 'shape' is the *direct object* of 'love'.

2) The sound of iPod is *amazing*.
nsubj – 'sound' is the *nominal subject* of 'amazing'

3) Capacity is not a *problem* here.
nsubj – 'capacity' is the *nominal subject* of 'problem' while 'problem' is parsed as the root of the sentence in dependency tree.

(b). Relations between opinion word and opinion target

Figure 5. Examples of dependency path

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

The translated Chinese dataset can be denoted as

$$T_{CN} = \{(w_{CN}^{i}, pos_{CN}^{i}, dep_{CN}^{i}, owt_{CN}^{i}, y_{CN}^{i})_{i=1}^{N_{CN}}\} \tag{6}$$

Where y_{CN}^i and y_{EN}^i are the target labels and

$$y_{CN}^{i} = \begin{cases} y_{EN}^{j} & if\ w_{CN}^{i}\ is\ aligned\ to\ w_{EN}^{j} \\ 0 & other \end{cases} \tag{7}$$

All the features in T_{CN} are generated directly from the translated Chinese text.

2) Feature Mapping

In the feature mapping stage, we try to map the features in the source and target languages into a unified feature space. 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 [33]. As 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. [34] 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.

Thus, the two dataset T_{EN} and T_{CN} are represented with all the features as

$$T_{EN} = \{(w_{EN}^{i}, upos_{EN}^{i}, dep_{EN}^{i}, owt_{EN}^{i}, y_{EN}^{i})_{i=1}^{N_{EN}}\} \tag{8}$$

$$T_{CN} = \{(w_{CN}^{i}, upos_{CN}^{i}, dep_{CN}^{i}, owt_{CN}^{i}, y_{CN}^{i})_{i=1}^{N_{CN}}\} \tag{9}$$

where *upos*^{i} represents the universal part-of-speech tag corresponding to *pos*^{i}. After the feature mapping stage, all the part-of-speech tags in T_{EN} and T_{CN} are transformed into the universal part-of-speech tag set.

C. Feature 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} .

$$T_{Proj} = \{(w_{CN}^{j}, upos_{Proj}^{i}, dep_{Proj}^{i}, owt_{Proj}^{i}, y_{CN}^{i})_{i=1}^{N_{CN}}\} \tag{10}$$

where w_{CN}^i is directly derived from the translated Chinese words and

$$upos_{Proj}^{i} = \begin{cases} upos_{EN}^{j} & if\ w_{CN}^{i}\ is\ aligned\ to\ w_{EN}^{j} \\ "X" & other \end{cases} \tag{11}$$

$$dep_{Proj}^{i} = \begin{cases} dep_{EN}^{j} & if\ w_{CN}^{i}\ is\ aligned\ to\ w_{EN}^{j} \\ "X" & other \end{cases} \tag{12}$$

$$owt_{Proj}^{i} = \begin{cases} owt_{EN}^{j} & if\ w_{CN}^{i}\ is\ aligned\ to\ w_{EN}^{j} \\ "X" & other \end{cases} \tag{13}$$

"*X*" represents the null value when the Chinese word is not aligned to any English word.

The two datasets T_{Proj} and T_{CN} share the same word-based feature and the same 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. Thus, we get two different Chinese datasets.

D. Monolingual Co-Training

The co-training algorithm [35] 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 [30], text classification [3][36], word sense disambiguation [37], and so on. We use the co-training algorithm for the cross-language target extraction task due to the following three reasons: 1) we can train two different models based on two different training datasets. 2) Although we lack an annotated Chinese corpus, the unlabeled Chinese product reviews can be easily obtained from the web. 3) The co-training algorithm helps to narrow the domain barrier between training and test dataset which is analyzed in Section VI.

As shown in Figure 6, 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 [38] to combine the labeling results of the two component models together. It means that a word will

```
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 an OR merger.
```

Figure 6. The monolingual co-training algorithm

be regarded as a target if it is labeled as "B" or "I" by one or more models. For example, if a word "相机" ("camera") is labeled as "B" by one model and "O" by the other model, we adopt "B" as the final answer. The two parameters N and I will be referred as growth size and iteration in the later discussion.

We will also compare our algorithm with self-training. Different from co-training, the self-training progress trains the two models separately. Taking M_1 for example, N most confidently labeled examples by M_1 is added to T_{CN} and removed from UD_1. Then M_1 is re-trained with the enlarged datasets T_{CN}. We loop the progress for I iterations. The other self-training model M_2 is trained in a similar way.

VI. Experiments

A. Dataset

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

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 detailed information is shown in Table I(a).

2) English Training Set: We use the customer review collection from [11] as the training dataset. 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. The detailed information is shown in Table I(b).

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 segmentation tool - 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 V.

TABLE I. Detailed description of datasets

Domain	#Review	#Sentence	#Opinion target
Camera	137	2075	1979
Car	157	4783	2687
Laptop	56	1034	1035
Phone	123	2644	2416
Total	473	10536	8117

(a) COAE dataset

Domain	#Review	#Sentence	#Opinion target
Camera-1	45	597	286
Camera-2	34	346	203
DVD Player	99	739	431
MP3 Player	95	1716	848
Phone	41	546	340
Total	314	3944	2108

(b) Customer review dataset

B. Evaluation Metrics

We use the same evaluation metrics as COAE. Precision, recall and F-measure are used to measure the performance. Precision and recall are calculated as follows, F-measure is the harmonic mean of them.

$$Precision = \frac{\#system\_correct}{\#system\_proposed}$$

$$Recall = \frac{\#system\_correct}{\#gold}$$

where *#system_proposed* is the number of proposed opinion targets of our system, *#gold* is the number of human labeled opinion targets. *#system_correct* is the number of correct opinion targets proposed by our system. COAE adopts two different criteria to judge whether a proposed opinion target is correct: strict and lenient. In strict evaluation, a proposed target

[3] http://ir-china.org.cn/coae2008.html. COAE is one of the most authoritative evaluation for Chinese opinion mining.

[4] http://www.zol.com.cn/

[5] http://www.bitauto.com/

[6] http://ictclas.org/

must cover exactly the same span with the correct answer target. In lenient evaluation, a proposed target is correct if the spans of the proposed target and the human labeled target overlap. For example, if the human labeled target is "Canon G12" and the proposed target is "G12", it will be regarded as a wrong answer in strict evaluation but a correct answer in lenient evaluation.

C. Baselines

In the experiments, we compare our proposed CLOpinionMiner system with eight baseline models, and they are described as follows:

UN: We implement an unsupervised method based on [11], which relies on association mining and a sentiment dictionary to extract frequent and infrequent product aspects.

Rule: We implement a rule-based method which directly uses dependency relation patterns to extract opinion targets. We adopt the two rules similar to those in [13] which extract opinion target based on opinion word. As illustrated bellow, *O* represents the opinion word identified by the opinion lexicon and *T* represents the output (opinion target). MR = {amod, nsubj, dobj} is the possible dependency path set between opinion word and opinion target. The part-of-speech tag of opinion target is restricted as noun.

Rule	Example
$O \rightarrow \text{Dep} \rightarrow T$ *s.t.* $\text{Dep} \in \{\text{MR}\}, \text{POS}(T) \in \{\text{N}\}$	The phone has a good screen. (*good*→amod→*screen*)
$O \rightarrow O\text{-Dep} \rightarrow H \leftarrow T\text{-Dep} \leftarrow T$ *s.t.* $O/T\text{-Dep} \in \{\text{MR}\}, \text{POS}(T) \in \{\text{N}\}$	iPod is the best mp3 player. (*best*→amod→player←nsubj← *iPod*)

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

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

ST(M1): After training the model M1 using T_{CN}, the self-training algorithm is used to improve the performance.

ST(M2): After training the model M2 using T_{Proj}, the self-training algorithm is used to improve the performance.

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

TABLE II. COMPARISON RESULTS OF DIFFERENT MODELS (N=1000, I=20 FOR CO-TRAINING AND SELF-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
Rule	0.260	0.198	0.225	0.683	0.521	0.591
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
ST(M1)	0.337	0.262	0.295	0.767	0.596	0.670
ST(M2)	0.306	0.220	0.256	0.751	0.542	0.630
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
CLOpinionMiner	0.313	0.327	**0.320**	0.721	0.754	**0.737**

TABLE III. COMPARISON RESULTS OF THE CLOPINIONMINER MODEL AND THE COAE RESULTS.

Method	Strict		
	Precision	Recall	F-measure
COAE-1	0.3798	0.4172	0.3976
COAE-2	0.3275	0.4058	0.3625
CLOpinionMiner	0.313	0.327	0.32
COAE-3	0.5641	0.1942	0.2889
COAE-4	0.2354	0.3411	0.2786
COAE-5	0.2796	0.2009	0.2338
COAE-6	0.1984	0.2805	0.2324
COAE-Ave	0.2526	0.2048	0.2262
COAE-7	0.2225	0.1994	0.2103
COAE-8	0.1356	0.2571	0.1776
COAE-9	0.1946	0.1285	0.1548
COAE-10	0.2751	0.09053	0.1362
COAE-11	0.243	0.06551	0.1032
COAE-12	0.1789	0.04223	0.06833
COAE-13	0.04968	0.03955	0.04404

Method	Lenient		
	Precision	Recall	F-measure
CLOpinionMiner	0.721	0.754	0.737
COAE-2	0.467	0.5788	0.5169
COAE-1	0.4934	0.5421	0.5166
COAE-4	0.3829	0.5547	0.4531
COAE-5	0.4932	0.3544	0.4124
COAE-6	0.3239	0.4581	0.3795
COAE-3	0.7206	0.2481	0.3691
COAE-Ave	0.4128	0.3301	0.3668
COAE-9	0.4283	0.2828	0.3407
COAE-7	0.3488	0.3125	0.3296
COAE-8	0.2298	0.4357	0.3009
COAE-10	0.4723	0.1554	0.2339
COAE-11	0.4516	0.1218	0.1918
COAE-13	0.2076	0.1653	0.184
COAE-12	0.3471	0.08193	0.1326

performance of M2.

CLOpinionMiner: 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 and self-training parameters as *N*=1000 and *I*=20 in our experiment, the influence of the parameters will be discussed later.

In addition, we compare the performance of our system with the results reported by COAE.

COAE-*i*: COAE 2008 reported a total of 16 results from 13 teams. We list the best run for each team *i*.

COAE-Ave: The average result is used as a baseline here. We first calculate the average precision and recall scores and then derive the F-measure score.

Zhang et al. [39] achieve the best results out of all the 13 teams. The CRF model is adopted in their system, but they rely on manually annotated Chinese training datasets on the four testing domains to train the model. Besides, they manually build a Chinese product aspect dictionary to extract opinion target. The most significant difference between our method and the COAE teams' is that we only use English annotated corpus and solve the problem from a cross-language view, while the COAE teams develop their systems based on Chinese datasets.

D. Results

Among the nine models in Table II, the unsupervised method achieves poor results on both strict and lenient evaluation. It gets high recall but very low precision, which means many frequent nouns or nouns groups in the dataset are not opinion targets. The rule-based method gets much better result than UN which shows the effectiveness of dependency relations in opinion target extraction. The final co-training model CLOpinionMiner 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 CLOpinionMiner 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 both co-training and self-training algorithms. Furthermore, the co-training component models CT(M1) and CT(M2) outperform the self-training models ST(M1) and ST(M2), respectively, which means the co-training algorithm is more effective than self-training. 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.

Compared to the COAE results in Table III, our proposed CLOpinionMiner system outperforms most COAE systems on both strict and lenient evaluations. Our system ranks first in lenient evaluation and third in strict evaluation. Considering that our method does not use any Chinese training corpus except the Chinese opinion lexicon, the overall results are very promising. Compared to the COAE best result on both evaluation metrics, we get a 22 percent higher F-measure on lenient evaluation but an 8 percent lower F-measure on strict evaluation. The high performance of COAE-1 on strict evaluation is mainly caused by two reasons: 1) Zhang et al. [39] manually annotated the Chinese corpus for the four testing domains, while our original English datasets do not contain reviews on car and notebook. 2) Zhang et al. [39] relied on a hand-crafted Chinese aspect dictionary to identify the opinion targets.

E. Result Analysis

Our performance in exact evaluation is relatively low compared to that in lenient evaluation. To make better sense of this point, we list the average length of target spans for the human annotation results of test dataset, the proposed results of our system on test dataset and the gold-standard English dataset in Table IV. From the table, we can observe that the average target length of our gold-standard test data is close to 2 which means many opinion targets in the test data have two or more words. However, the average target length in our proposed result on the test dataset is close to 1 which means most of the proposed opinion targets contain only one word. Such difference in span length distribution makes our system perform not well in strict evaluation which requires the proposed targets have exactly the same spans with human annotated results. Actually, this problem is caused by our original English dataset. In the original English dataset, most of opinion targets have only one word. Besides, some multi-words English opinion targets become single-word Chinese opinion target after machine translation, such as "picture quality" to "画质". The above factors make both our two Chinese training datasets T_{CN} and T_{Proj} contain much more single-word opinion targets than multi-words opinion targets. Thus, both the models tend to identify single word opinion target and cannot capture enough information for long opinion target phrases. For example, in the opinion target "防抖技术" ("image-stabilization technology") only "技术"("technology") will be tagged as a target. Sometimes, the wrong Chinese word segmentation results also affect the performance. For example, the word "便携性"("portability") is incorrectly segmented into three words "便/携/性" in some sentences and all the other features such as part-of-speech and dependency relation related to them become wrong. These errors make our model fail to identify the target.

TABLE IV. AVERAGE SPAN LENGTH OF OPINION TARGET

Gold-Standard Test Dataset	1.91
Results of CLOpinionMiner	1.04
Gold-Standard English Dataset	1.22

The monolingual co-training algorithm helps our system gets large gain in both of the two evaluation metrics. One reason for the improvement is the different properties of the two Chinese datasets. They can make up for the shortage of each other as explained in Section V-D. The other reason is that the co-training algorithm helps to narrow the domain barrier

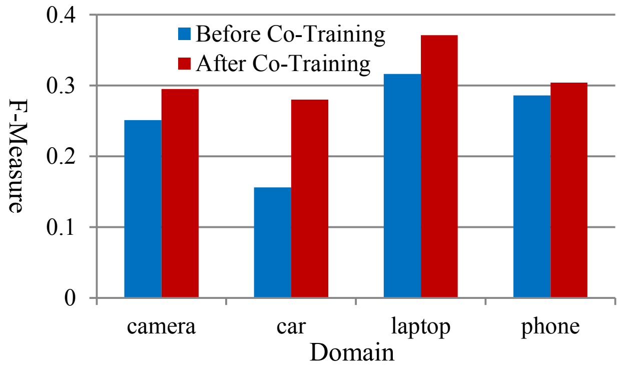


Figure 7. Performance before and after co-training in strict evaluation

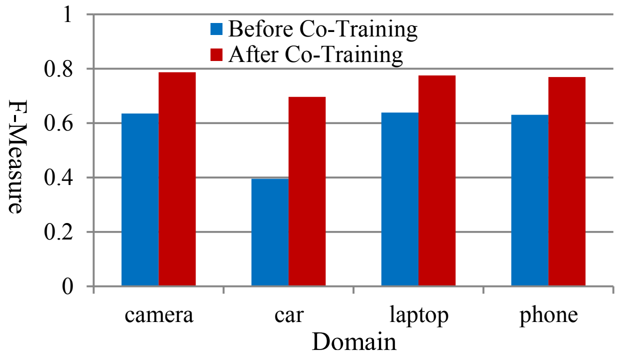


Figure 8. Performance before and after co-training in lenient evaluation

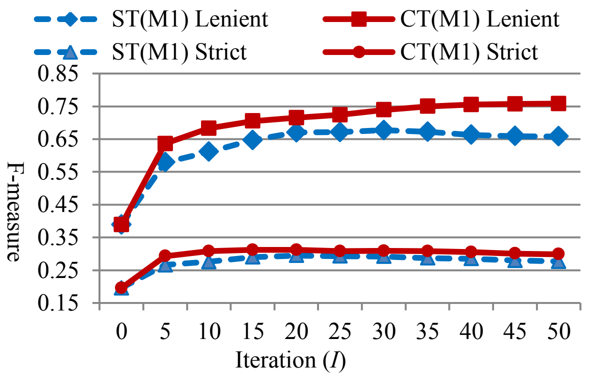


Figure 9. Comparison between CT(M1) and ST(M1)

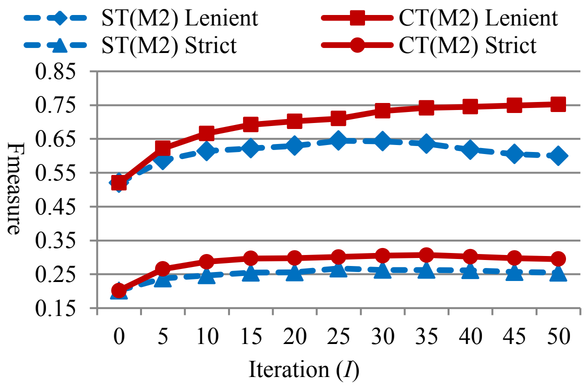


Figure 10. Comparison between CT(M2) and ST(M2)

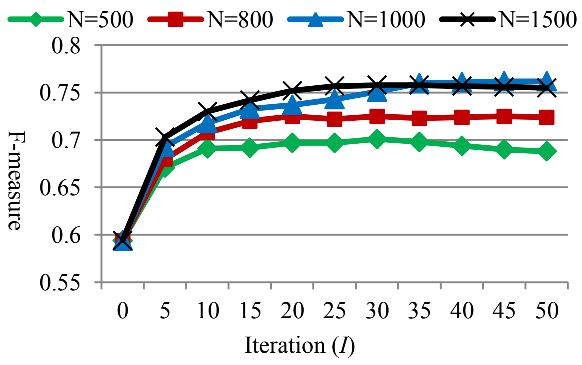


Figure 11. F-measure on lenient evaluation different growth size.

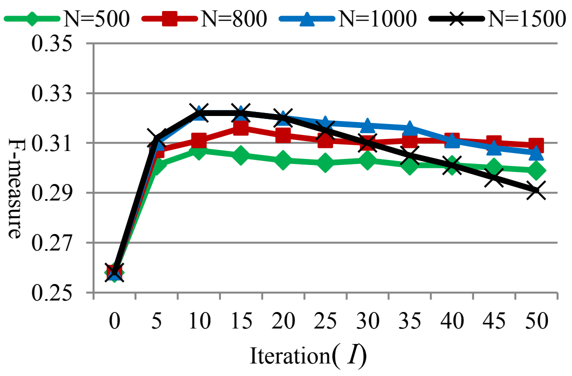


Figure 12. F-measure on strict evaluation with different growth size.

between the test dataset and the original English dataset [36]. The four domains of our English dataset listed in Table I are all about electronic products. The domains of the Chinese test data are not the same. Among the four different domains in the test dataset, thr ee of them (camera, phone and laptop) are also electronic products, but the remaining one (car) is quite different. The domain barrier will make our model perform poorly especially on the car domain. For example, most of the opinion target in the car domain such as “油耗”(“oil consumption”) and “加速”(“acceleration”) never appears in the training datasets. The word-based feature which is quite important becomes useless in the cross-domain scenario. However, the co-training algorithm can help to narrow the barrier between different domains. By leveraging the unlabeled data in the car domain, we can firstly identify some high-confidence opinion target in the unlabeled data such as “acceleration”. Then, we add them into the training data which helps to improve the result on the test set. The detailed results are shown in Figures 7 and 8. The results before co-training are derived by combining M1 and M2 using OR merger and the results after co-training are the final outputs of our system. As can be observed from the two figures, the car domain gets much worse result than the other domains before co-training. F-measure on all the four domains is increased after using the co-training algorithm. However, the car domain achieves much larger improvement than the other domains. Although it still gets the worst result after co-training, the performance gaps with other domains are significantly reduced. The result shows that the co-training algorithm not only makes good use of the two different Chinese training datasets but also helps to narrow the domain barrier especially for the car domain in our test dataset.

F. Discussion

1) Co-Training vs. Self-Training

From the experimental results in Table II, we can conclude that the co-training algorithm outperforms the self-training algorithm. Using ST(M1) and CT(M1) in Table II for comparison, these two models have similar precision score. However, CT(M1) is superior to ST(M1) in terms of recall. The better recall performance can be attributed to the fact that the self-training algorithm can label only those instances with high confidence which it has already seen in the training data. Hence, in successive iterations, very little new information becomes available to the algorithm. While in co-training the model learns more information from the other model's labeled instances. A visual comparison of co-training and self-training over the F-measure is shown in Figures 9 and 10. Figure 9 compares the F-measure scores of the co-training component model CT(M1) and the self-training model ST(M1) on both strict and lenient evaluations. Figure 10 shows the comparison between CT(M2) and ST(M2). We can see that the two component models in co-training can always outperform the two self-training models with respect to different iteration numbers on both lenient and strict evaluations.

2) Influence of Growth Size N in Co-Training

Figures 11 and 12 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 11 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 12 shows the similar trend that a larger *N* helps the performance increase faster. The growth sizes of 1000 and 1500 get the same highest F-measure scores. A significant difference between the lenient evaluation in Figu re 11 and the strict evaluation in Figure 12 is that the curves in Figure 11 become steady after several iterations while the curves in Figure 12 decline. Furthermore, the curves in Figure 12 decline faster for larger growth size. This is because the strict evaluation is more sensitive to wrong target labels than the lenient evaluation.

VII. Conclusion and Future Work

In this paper, we propose a cross-language opinion target extraction system CLOpinionMiner using the monolingual co-training algorithm, which can be easily adapted to other cross-language information extraction tasks. 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) [40], etc. We will also use our approach to build opinion target extraction models for other languages to test the robustness of our method.

In our experiments, the training dataset and test dataset covers different domains, which also infulences the result. We will try to select English training dataset and Chinese test dataset from a single domain to have a further analysis of the system.

References

[1] X. Zhou, X. Wan, J. Xiao, "Cross-language opinion target extraction in review texts," In *Proceedings of the 2012 IEEE 12th International Conference on Data Mining*. IEEE Computer Society, 2012: 1200-1205.

[2] J. D. Lafferty, A. McCallum, and F. 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.

[3] X. 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] E. Breck, Y. Choi, and C. Cardie, "Identifying expressions of opinion in context," In *Proceedings of IJCAI'07*, 2007, pp 2683–2688.

[5] N. Jakob and I. 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.

[6] S. Li, R. Wang, and G. Zhou, "Opinion target extraction using a shallow semantic parsing framework". In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[7] B. Liu, "Sentiment analysis and subjectivity," In *Handbook of Natural Language Processing*, Second Edition, N. Indurkhya and F.J. Damerau, Editors. 2010.

[8] B. Lu, C. Tan, C. Cardie, and B. K. Tsou, "Joint bilingual sentiment classification with unlabeled parallel corpora," In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 2011, pp. 320–330.

[9] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? sentiment classification using machine learning techniques," In *Proceedings of EMNLP'02*, 2002, pp. 79–86.

[10] B. Liu, *Web Data Mining: Exploring Hyperlinks, Contents and Usage Data*. Springer, Berlin, 2006.

[11] M. Hu and B. Liu, "Mining and summarizing customer reviews," In *Proceedings of SIGKDD*'04, 2004, pp. 168–177.

[12] J. Yi, T. Nasukawa, R. Bunescu, and W. Niblack, "Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques," In *Proceedings of the 3rd IEEE International Conference on Data Mining*, 2003, pp. 427–434.

[13] S.-M. Kim and E. Hovy, "Extracting opinions, Opinion Holders, and Topics Expressed in Online News Media Text," In *Proceedings of the ACL Workshop on Sentiment and Subjectivity in Text*, 2006, pp. 1–8.

[14] G. Draya, M. Plantié, A. Harb, P. Poncelet, M. Roche, and F. Trousset, "Opinion mining from blogs," *International Journal of Computer Information Systems and Industrial Management Applications (IJCISIM)*.

[15] T. Mullen, and N. Collier, "Sentiment analysis using support vector machines with diverse information sources," In *Proceedings of EMNLP-2004*.

[16] J. Martineau, and T. Finin, "Delta tfidf: An improved feature space for sentiment analysis," In *Proceedings of the Third International AAAI Conference on Weblogs and Social Media*. 2009.

[17] B. Pang and L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," In *Proceedings of Meeting of the Association for Computational Linguistics*, 2004.

[18] T. Li, Y. Zhang, and V. Sindhwani, "A non-negative matrix tri-factorization approach to sentiment classification with lexical prior knowledge," In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. 2009.

[19] K. Liu, L. Xu, and J. Zhao. "Opinion target extraction using word-based translation model," In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, 2012.

[20] L.Zhuang, F. Jing, and X. Zhu, "Movie review mining and summarization," In *Proceedings of the ACM 15th Conference on Information and Knowledge Management*, 2006, pp. 43–50.

[21] G. Qiu, B. Liu, J. Bu and C. Chen, "Opinion word expansion and target extraction through double propagation," *Computational Linguistics*, 37(1), 2011, pp. 9–27.

[22] B. Yang, and C. Cardie. "Joint inference for fine-grained opinion extrac-tion," In *Proceedings of ACL 2013*.

[23] R. Mihalcea, C. Banea, and J. Wiebe. "Learning multilingual subjective language via cross-lingual projections," In *Proceedings of the Annual Meeting of the Association for Computational Linguistics,* 2007.

[24] X. Meng, F. Wei, X. Liu, M. Zhou, G. Xu and H. Wang. "Cross-lingual mixture model for sentiment classification," In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers*, Volume 1. Association for Computational Linguistics, 2012: 572-581.

[25] X. Wan. "A comparative study of cross-Lingual sentiment classification," In *Proceedings of 2012 IEEE/WIC/ACM International Conference on Web Intelligence*.

[26] J. R. Cowie and W. G. Lehnert, "Information extraction," *Communications of the ACM*, 39(1), 1996, pp. 80–91.

[27] 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.

[28] S. Kim, M. Jeong, J. Lee, and G. G. Lee, "A cross-lingual annotation projection approach for relation detection," In *Proceedings of the 23rd International Conference on Computational Linguistics* (Coling 2010), 2010, pp. 564–571.

[29] I. Zitouni and R. Florian, "Cross-language information propagation for arabic mention detection," *ACM Transactions on Asian Language Information Processing*, Vol. 8, No. 4, Article 17.

[30] J.-H. Oh, K. Uchimoto, and K. 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.

[31] G. Grefenstette, "The problem of cross-language information retrieval," *Springer* US, 1998.

[32] W. Ogden, J. Cowie, M. Davis, E . Ludovik, H. Molina-Salgado, and H. Shin. "Getting information from documents you cannot read: An interactive cross-language text retrieval and summarization system," *Joint ACM DL/SIGIR workshop on multilingual information discovery and access*. 1999.

[33] S. Petrov, D. Das, and R. McDonald, "A universal part-of-speech tagset," 2011, *ArXiv*:1104.2086.

[34] P.-C. Chang, H. Tseng, D. Jurafsky, and C. 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.

[35] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," In *Proceedings of COLT-98*, 1998, pp. 92–100.

[36] M. Chen, K. Q. Weinberger, J. C. Blitzer, "Co-Training for domain adaptation," In *Proceedings of NIPS-2011*, 2011.

[37] R. Mihalcea, "Co-training and self-training for word sense disambiguation," In *Proceedings of CoNLL-04*, 2004, pp. 33–49.

[38] F. P. Szidarovszky, I. Solt, D. 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.

[39] S. Zhang, W. Jia, Y. Xia, Y. Meng, and H. Yu, "Research on CRF-based evaluated object extraction," In *Proceedings of the COAE 2008 Workshop*, 2008, pp.70–76.

[40] L. Zhou, Y. Xia, B. Li, and K.-F. Wong, "WIA-Opinmine system in NTCIR-8 MOAT evaluation," In *Proceedings of NTCIR-8 Workshop Meeting*, 2010, 286–292.