Opinion Target Extraction Using Word-Based Translation Model

Kang Liu, Liheng Xu, Jun Zhao

National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China
{kliu, lhxu, jzhao}@nlpr.ia.ac.cn

Abstract

This paper proposes a novel approach to extract opinion targets based on word-based translation model (WTM). At first, we apply WTM in a monolingual scenario to mine the associations between opinion targets and opinion words. Then, a graph-based algorithm is exploited to extract opinion targets, where candidate opinion relevance estimated from the mined associations, is incorporated with candidate importance to generate a global measure. By using WTM, our method can capture opinion relations more precisely, especially for long-span relations. In particular, compared with previous syntax-based methods, our method can effectively avoid noises from parsing errors when dealing with informal texts in large Web corpora. By using graph-based algorithm, opinion targets are extracted in a global process, which can effectively alleviate the problem of error propagation in traditional bootstrap-based methods, such as *Double Propagation*. The experimental results on three real world datasets in different sizes and languages show that our approach is more effective and robust than state-of-art methods.

1 Introduction

With the rapid development of e-commerce, most customers express their opinions on various kinds of entities, such as products and services. These reviews not only provide customers with useful information for reference, but also are valuable for merchants to get the feedback from customers and enhance the qualities of their products or services. Therefore, mining opinions from these vast amounts of reviews becomes urgent, and has attracted a lot of attentions from many researchers.

In opinion mining, one fundamental problem is opinion target extraction. This task is to extract items which opinions are expressed on. In reviews, opinion targets are usually nouns/noun phrases. For example, in the sentence of "*The phone has a colorful and even amazing screen*", "*screen*" is an opinion target. In online product reviews, opinion targets often are products or product features, so this task is also named as product feature extraction in previous work (Hu et al., 2004; Ding et al., 2008; Liu et al., 2005; Popescu et al., 2005; Wu et al., 2005; Su et al., 2008).

To extract opinion targets, many studies regarded opinion words as strong indicators (Hu et al., 2004; Popescu et al., 2005; Liu et al., 2005; Qiu et al., 2011; Zhang et al., 2010), which is based on the observation that opinion words are usually located around opinion targets, and there are associations between them. Therefore, most pervious methods iteratively extracted opinion targets depending upon the associations between opinion words and opinion targets (Qiu et al., 2011; Zhang et al., 2010). For example, "*colorful*" and "*amazing*" is usually used to modify "*screen*" in reviews about cell phone, so there are strong associations between them. If "*colorful*" and "*amazing*" had been known to be opinion words, "*screen*" is likely to be an opinion target in this domain. In addition, the extracted opinion targets can be used to expand more opinion words according to their associations. It's a mutual reinforcement procedure.

Therefore, mining associations between opinion targets and opinion words is a key for opinion target extraction (Wu et al., 2009). To this end, most previous methods (Hu et al., 2004; Ding et al., 2004; Wang et al., 2008), named as *adjacent methods*, employed the *adjacent rule*, where an opinion target was regarded to have opinion relations with the surrounding opinion words in a given window. However, because of the limitation of window size, opinion relations cannot be captured precisely, especially for long-span relations, which would hurt estimating associations between opinion targets and opinion words. To resolve this problem, several studies exploited syntactic information such as dependency trees (Popescu et al., 2005; Qiu et al., 2009; Qiu et al., 2011; Wu et al., 2009; Zhang et al., 2010). If the syntactic relation between an opinion word and an opinion target satisfied a designed pattern, then there was an opinion relation between them. Experiments consistently reported that *syntax-based methods* could yield better performance than *adjacent methods* for small or medium corpora (Zhang et al., 2010). The performance of *syntax-based methods* heavily depends on the parsing performance. However, online reviews are often informal texts (including grammar mistakes, typos, improper punctuations etc.). As a result, parsing may generate many mistakes. Thus, for large corpora from Web including a great deal of informal texts, these *syntax-based methods* may suffer from parsing errors and introduce many noises. Furthermore, this problem maybe more serious on non-English language reviews, such as Chinese reviews, because that the performances of parsing on these languages are often worse than that on English.

To overcome the weakness of the two kinds of methods mentioned above, we propose a novel unsupervised approach to extract opinion targets by using word-based translation model (WTM). We formulate identifying opinion relations between opinion targets and opinion words as a word alignment task. We argue that an opinion target can find its corresponding modifier through monolingual word alignment. For example in Figure 1, the opinion words "*colorful*" and "*amazing*" are aligned with the target "*screen*" through word alignment. To this end, we use WTM to perform monolingual word alignment for mining associations between opinion targets and opinion words. In this process, several factors, such as word co-occurrence frequencies, word positions etc., can be considered globally. Compared with *adjacent methods*, WTM doesn't identify opinion relations between words in a given window, so long-span relations can be effectively captured (Liu et al., 2009). Compared with *syntax-based methods*, without using parsing, WTM can effectively avoid errors from parsing informal texts. So it will be more robust. In addition, by using WTM, our method can capture the "one-to-many" or "many-to-one" relations ("one-to-many" means that, in a sentence one opinion word modifies several opinion targets, and "many-to-one" means several opinion words modify one opinion target). Thus, it's reasonable to expect that WTM is likely to yield better performance than traditional methods for mining associations between opinion targets and opinion words.

Based on the mined associations, we extract opinion targets in a ranking framework. All nouns/noun phrases are regarded as opinion target candidates. Then a graph-based algorithm is exploited to assign confidences to each candidate, in which candidate opinion relevance and importance are incorporated to generate a global measure. At last, the candidates with higher ranks are extracted as opinion targets. Compared with most traditional methods (Hu et al. 2004; Liu et al., 2005; Qiu et al., 2011), we don't extract opinion targets iteratively based on the bootstrapping strategy, such as *Double Propagation* (Qiu et al., 2011), instead all candidates are dynamically ranked in a global process. Therefore, error propagation can be effectively avoided and the performance can be improved.

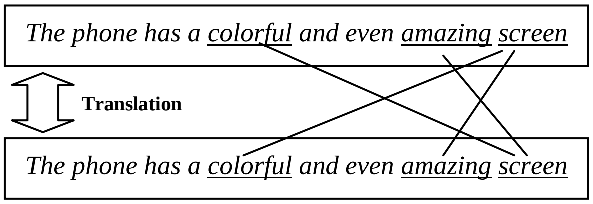
The phone has a colorful and even amazing screen
Translation
The phone has a colorful and even amazing screen

Figure 1: Word-based translation model for opinion relation identification

The main contributions of this paper are as follows.

1) We formulate the opinion relation identification between opinion targets and opinion words as a word alignment task. To our best knowledge, none of previous methods deal with this task using monolingual word alignment model (in Section 3.1).

2) We propose a graph-based algorithm for opinion target extraction in which candidate opinion relevance and importance are incorporated into a unified graph to estimate candidate confidence. Then the candidates with higher confidence scores are extracted as opinion targets (in Section 3.2).
3) We have performed experiments on three datasets in different sizes and languages. The experimental results show that our approach can achieve performance improvement over the traditional methods. (in Section 4).

The rest of the paper is organized as follows. In the next section, we will review related work in brief. Section 3 describes our approach in detail. Then experimental results will be given in Section 4. At the same time, we will give some analysis about the results. Finally, we give the conclusion and the future work.

2 Related Work

Many studies have focused on the task of opinion target extraction, such as (Hu et al., 2004; Ding et al., 2008; Liu et al., 2006; Popescu et al., 2005; Wu et al., 2005; Wang et al., 2008; Li et al., 2010; Su et al., 2008; Li et al., 2006). In general, the existing approaches can be divided into two main categories: supervised and unsupervised methods.

In supervised approaches, the opinion target extraction task was usually regarded as a sequence labeling task (Jin et al. 2009; Li et al. 2010; Wu et al., 2009; Ma et al. 2010; Zhang et al., 2009). Jin et al. (2009) proposed a lexicalized HMM model to perform opinion mining. Li et al. (2010) proposed a Skip-Tree CRF model for opinion target extraction. Their methods exploited three structures including linear-chain structure, syntactic structure, and conjunction structure. In addition, Wu et al. (2009) utilized a SVM classifier to identify relations between opinion targets and opinion expressions by leveraging phrase dependency parsing. The main limitation of these supervised methods is that labeling training data for each domain is impracticable because of the diversity of the review domains.

In unsupervised methods, most approaches regarded opinion words as the important indicators for opinion targets (Hu et al., 2004; Popsecu et al., 2005; Wang et al., 2008; Qiu et al., 2011; Zhang et al., 2010). The basic idea was that reviewers often use the same opinion words when they comment on the similar opinion targets. The extraction procedure was often a bootstrapping process which extracted opinion words and opinion targets iteratively, depending upon their associations. Popsecu et al. (2005) used syntactic patterns to extract opinion target candidates. After that they computed the point-wise mutual information (PMI) score between a candidate and a product category to refine the extracted results. Hu et al. (2004) exploited an association rule mining algorithm and frequency information to extract frequent explicit product features. The adjective nearest to the frequent explicit feature was extracted as an opinion word. Then the extracted opinion words were used to extract infrequent opinion targets. Wang et al. (2008) adopted the similar idea, but their method needed a few seeds to weakly supervise the extraction process. Qiu et al. (2009, 2011) proposed a *Double Propagation* method to expand a domain sentiment lexicon and an opinion target set iteratively. They exploited direct dependency relations between words to extract opinion targets and opinion words iteratively. The main limitation of Qiu's method is that the patterns based on dependency parsing tree may introduce many noises for the large corpora (Zhang et al., 2010). Meanwhile, *Double Propagation* is a bootstrapping strategy which is a greedy process and has the problem of error propagation. Zhang et al. (2010) extended Qiu's method. Besides the patterns used in Qiu's method, they adopted some other patterns, such as phrase patterns, sentence patterns and "no" pattern, to increase recall. In addition they used the HITS (Klernberg et al., 1999) algorithm to compute the feature relevance scores, which were simply multiplied by the log of feature frequencies to rank the extracted opinion targets. In this way, the precision of result can be improved.

3 Opinion Target Extraction Using Word-Based Translation Model

3.1 Method Framework

As mentioned in the first section, our approach for opinion target extraction is composed of the following two main components:

1) ***Mining associations between opinion targets and opinion words***: Given a collection of reviews, we adopt a word-based translation

model to identify potential opinion relations in all sentences, and then the associations between opinion targets and opinion words are estimated.

2) ***Candidate confidence estimation***: Based on these associations, we exploit a graph-based algorithm to compute the confidence of each opinion target candidate. Then the candidates with higher confidence scores are extracted as opinion targets.

3.2 Mining associations between opinion targets and opinion words using Word-based Translation Model

This component is to identify potential opinion relations in sentences and estimate associations between opinion targets and opinion words. We assume opinion targets and opinion words respectively to be nouns/noun phrases and adjectives, which have been widely adopted in previous work (Hu et al., 2004; Ding et al., 2008; Wang et al., 2008; Qiu et al., 2011). Thus, our aim is to find potential opinion relations between nouns/noun phrases and adjectives in sentences, and calculate the associations between them. As mentioned in the first section, we formulate opinion relation identification as a word alignment task. We employ the word-based translation model (Brown et al. 1993) to perform monolingual word alignment, which has been widely used in many tasks, such as collocation extraction (Liu et al., 2009), question retrieval (Zhou et al., 2011) and so on. In our method, every sentence is replicated to generate a parallel corpus, and we apply the bilingual word alignment algorithm to the monolingual scenario to align a noun/noun phase with its modifier.

Given a sentence with n words $S=\{w_1, w_2, ..., w_n\}$, the word alignment $A=\{(i,a_i)\,|\,i\in[1,n]\}$ can be obtained by maximizing the word alignment probability of the sentence as follows.

$$\hat{A}=\arg\max_{A} P(A\,|\,S) \qquad (1)$$

where (i,a_i) means that a noun/noun phrase at position i is aligned with an adjective at position a_i. If we directly use this alignment model to our task, a noun/noun phrase may align with the irrelevant words other than adjectives, like prepositions or conjunctions and so on. Thus, in the alignment procedure, we introduce some constrains: 1) nouns/noun phrases (adjectives) must be aligned with adjectives (nouns/noun phrases) or null words; 2) other words can only align with themselves. Totally, we employ the following 3 WTMs (IBM 1~3) to identify opinion relations.

$$P_{IBM-1}(A\,|\,S)\propto\prod_{j=1}^{n} t(w_j\,|\,w_{a_j})$$

$$P_{IBM-2}(A\,|\,S)\propto\prod_{j=1}^{n} t(w_j\,|\,w_{a_j})d(j\,|\,a_j,n)$$

$$P_{IBM-3}(A\,|\,S)\propto\prod_{i=1}^{n} n(\phi_i\,|\,w_i)\prod_{j=1}^{n} t(w_j\,|\,w_{a_j})d(j\,|\,a_j,n) \qquad (2)$$

There are three main factors: $t(w_j\,|\,w_{a_j})$, $d(j\,|\,a_j,n)$ and $n(\phi_i\,|\,w_i)$, which respectively models different information.

1) $t(w_j\,|\,w_{a_j})$ models the co-occurrence information of two words in corpora. If an adjective co-occurs with a noun/noun phrase frequently in the reviews, this adjective has high association with this noun/noun phrase. For example, in reviews of cell phone, "*big*" often co-occurs with "*phone's size*", so "*big*" has high association with "*phone's size*".

2) $d(j\,|\,a_j,l)$ models word position information, which describes the probability of a word in position a_j aligned with a word in position j.

3) $n(\phi_i\,|\,w_i)$ models the fertility of words, which describe the ability of a word for "one-to-many" alignment. ϕ_i denotes the number of words that are aligned with w_i. For example, "*Iphone4 has amazing screen and software*". In this sentence, "*amazing*" is used to modify two words: "*screen*" and "*software*". So ϕ equals to 2 for "*amazing*".

Therefore, in Eq. (2), $P_{IBM-1}(A\,|\,S)$ only models word co-occurrence information. $P_{IBM-2}(A\,|\,S)$ additionally employs word position information. Besides these two information, $P_{IBM-3}(A\,|\,S)$ considers the ability of a word for "one-to-many" alignment. In the following experiments section, we will discuss the performance difference among these models in detail. Moreover, these models may capture “one-to-many” or “many-to-one” opinion relations (mentioned in the first section). In our knowledge, it isn’t specifically considered by previous methods including *adjacent methods* and *syntax-based methods*. Meanwhile， the alignment results may contain empty-word alignments, which means a noun/noun phrase has no modifier or an adjective modify nothing in the sentence.

After gathering all word pairs from the review sentences, we can estimate the translation probabilities between nouns/noun phrases and adjectives as follows.

$$p(w_N \mid w_A) = \frac{Count(w_N, w_A)}{Count(w_A)} \qquad (3)$$

where $p(w_N \mid w_A)$ means the translation probabilities from adjectives to nouns/noun phrases. Similarly, we can obtain translation probability $p(w_A \mid w_N)$. Therefore, similar to (Liu et al. 2009), the association between a noun/noun phrase and an adjective is estimated as follows.

$$\begin{aligned}&Association(w_N, w_A)\\&= (t/p(w_N \mid w_A) + (1-t)/p(w_A \mid w_N))^{-1}\end{aligned} \qquad (4)$$

where t is the harmonic factor to combine these two translation probabilities. In this paper, we set $t = 0.5$. For demonstration, we give some examples in Table 1. We can see that our method using WTM can successfully capture associations between opinion targets and opinion words.

	battery life	sound	software
wonderful	0.000	0.042	0.000
poor	0.032	0.000	0.026
long	0.025	0.000	0.000

Table 1: Examples of associations between opinion targets and opinion words.

3.3 Candidate Confidence Estimation

In this component, we compute the confidence of each opinion target candidate and rank them. The candidates with higher confidence are regarded as the opinion targets. We argue that the confidence of a candidate is determined by two factors: 1) *Opinion Relevance*; 2) *Candidate Importance*.

Opinion Relevance reflects the degree that a candidate is associated to opinion words. If an adjective has higher confidence to be an opinion word, the noun/noun phrase it modifies will have higher confidence to be an opinion target. Similarly, if a noun/noun phrase has higher confidence to be an opinion target, the adjective which modifies it will be highly possible to be an opinion word. It’s an iterative reinforcement process, which indicates that existing graph-based algorithms are applicable.

Candidate Importance reflects the salience of a candidate in the corpus. We assign an importance score to an opinion target candidate f according to its $tf\text{-}idf$ score, which is further normalized by the sum of $tf\text{-}idf$ scores of all candidates.

$$Importance(c) = \frac{tf\text{-}idf\,(c)}{\sum_c tf\text{-}idf\,(c)} \qquad (5)$$

where c represents a candidate, tf is the term frequency in the dataset, and df is computed by using the Google n-gram corpus[1].

To model these two factors, a bipartite graph is constructed, the vertices of which include all nouns/noun phrases and adjectives. As shown in Figure 2, the white vertices represent nouns/noun phrases and the gray vertices represent adjectives. An edge between a noun/noun phrase and an adjective represents that there is an opinion relation between them. The weight on the edges represents the association between them, which are estimated by using WTM, as shown in Eq. (4).

To estimate the confidence of each candidate on this bipartite graph, we exploit a graph-based algorithm, where we use C to represent candidate confidence vector, a $n \times 1$ vector. We set the candidate initial confidence with candidate importance score, i.e. $C^0 = S$, where S is the candidate initial confidence vector and each item in S is computed using Eq. (5).

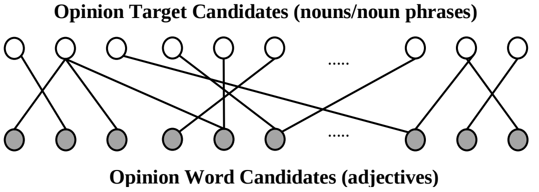
Opinion Target Candidates (nouns/noun phrases)
.....
.....
Opinion Word Candidates (adjectives)

Figure 2: Bipartite graph for modeling relations between opinion targets and opinion words

[1] http://books.google.com/ngrams/datasets

Then we compute the candidate confidence by using the following iterative formula.

$$C^{t+1} = M^T \times M \times C^t \qquad (6)$$

where C^t is the candidate confidence vector at time t, and C^{t+1} is the candidate confidence vector at time $t+1$. M is an opinion relevance matrix, a $m \times n$ matrix, where $M_{i,j}$ is the associated weight between a noun/noun phrase i and an adjective j.

To consider the candidate importance scores, we introduce a reallocate condition: combining the candidate opinion relevance with the candidate importance at each step. Thus we can get the final recursive form of the candidate confidence as follows.

$$C^{t+1} = (1-\lambda) \times M^T \times M \times C^t + \lambda \times S \qquad (7)$$

where $\lambda \in [0,1]$ is the proportion of candidate importance in the candidate confidence. When $\lambda = 1$, the candidate confidence is completely determined by the candidate importance; and when $\lambda = 0$, the candidate confidence is determined by the candidate opinion relevance. We will discuss its effect in the section of experiments.

To solve Eq. (7), we rewrite it as the following form.

$$C = \lambda \times (I - (1-\lambda) \times M^T \times M)^{-1} \times S \qquad (8)$$

where I is an identity matrix. To handle the inverse of the matrix, we expand the Eq. (8) as a power series as following.

$$C = \lambda \times [I + B + \ldots + B^k] \times S \qquad (9)$$

where $B = (1-\lambda) \times M^T \times M$ and $k \in [0, \infty)$ is an approximate factor. In experiments, we set $k = 100$. Using this equation, we estimate confidences for opinion target candidates. The candidates with higher confidence scores than the threshold will be extracted as the opinion targets.

4 Experiments

4.1 Datasets and Evaluation Metrics

In our experiments, we select three real world datasets to evaluate our approach. The first dataset is *COAE2008 dataset2*[2], which contains Chinese reviews of four different products. The detailed information can be seen in Table 2. Moreover, to evaluate our method comprehensively, we collect a larger collection named by *Large*, which includes three corpora from three different domains and different languages. The detailed statistical information of this dataset is also shown in Table 2. Restaurant is crawled from the Chinese Web site: www.dianping.com. The Hotel and MP3[3] were used in (Wang et al., 2011), which are respectively clawed from www.tripadvisor.com and www.amazon.com. For each collection, we perform random sampling to generate testing dataset, which include 6,000 sentences for each domain. Then the opinion targets in *Large* were manually annotated as the gold standard for evaluations. Three annotators are involved in the annotation process as follows. First, every noun/noun phrase and its contexts in review sentences are extracted. Then two annotators were required to judge whether every noun/noun phrase is opinion target or not. If a conflict happens, a third annotator will make judgment for finial results. The inter-agreement was 0.72. In total, we respectively obtain 1,112, 1,241 and 1,850 opinion targets in Hotel, MP3 and Restaurant. The third dataset is Customer Review Datasets[4] (English reviews of five products), which was also used in (Hu et al., 2004; Qiu et al., 2011). They have labeled opinion targets. The detailed information can be found in (Hu et al., 2004).

Domain	Language	#Sentence	#Reviews
Camera	Chinese	2075	137
Car	Chinese	4783	157
Laptop	Chinese	1034	56
Phone	Chinese	2644	123

(a) *COAE2008 dataset2*

Domain	Language	#Sentence	#Reviews
Hotel	English	1,855,351	185,829
MP3	English	289,931	30,837
Restaurant	Chinese	1,683,129	395,124

(b) *Large*

Table 2: Experimental Data Sets, # denotes the size of the reviews/sentences

In experiments, each review is segmented into sentences according to punctuations. Then sentences are tokenized and the part-of-speech of

[2] http://ir-china.org.cn/coae2008.html

[3] http://sifaka.cs.uiuc.edu/~wang296/Data/index.html

[4] http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

Methods	Camera			Car			Laptop			Phone		
	P	R	F	P	R	F	P	R	F	P	R	F
Hu	0.63	0.65	0.64	0.62	0.58	0.60	0.51	0.67	0.58	0.69	0.60	0.64
DP	0.71	0.70	0.70	0.72	0.65	0.68	0.58	0.69	0.63	0.78	0.66	0.72
Zhang	0.71	0.78	0.74	0.69	0.68	0.68	0.57	0.80	0.67	0.80	0.71	0.75
Ours	0.75	0.81	0.78	0.71	0.71	0.71	0.61	0.85	0.71	0.83	0.74	0.78

Table 3: Experiments on *COAE2008 dataset2*

Methods	Hotel			MP3			Restaurant		
	P	R	F	P	R	F	P	R	F
Hu	0.60	0.65	0.62	0.61	0.68	0.64	0.64	0.69	0.66
DP	0.67	0.69	0.68	0.69	0.70	0.69	0.74	0.72	0.73
Zhang	0.67	0.76	0.71	0.67	0.77	0.72	0.75	0.79	0.77
Ours	0.71	0.80	0.75	0.70	0.82	0.76	0.80	0.84	0.82

Table 4: Experiments on *Large*

Methods	D1			D2			D3			D4			D5		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
Hu	0.75	0.82	0.78	0.71	0.79	0.75	0.72	0.76	0.74	0.69	0.82	0.75	0.74	0.80	0.77
DP	0.87	0.81	0.84	0.90	0.81	0.85	0.90	0.86	0.88	0.81	0.84	0.82	0.92	0.86	0.89
Zhang	0.83	0.84	0.83	0.86	0.85	0.85	0.86	0.88	0.87	0.80	0.85	0.83	0.86	0.86	0.86
Ours	0.84	0.85	0.84	0.87	0.85	0.86	0.88	0.89	0.88	0.81	0.85	0.83	0.89	0.87	0.88

Table 5: Experiments on *Customer Review Dataset*

each word is assigned. Stanford NLP tool[5] is used to perform POS-tagging and dependency parsing. The method in (Zhu et al., 2009) is used to identify noun phrases. We select precision, recall and F-measure as the evaluation metrics. We also perform a significant test, i.e., a t-test with a default significant level of 0.05.

4.2 Our Methods vs. State-of-art Methods

To prove the effectiveness of our method, we select the following state-of-art unsupervised methods as baselines for comparison.

1) ***Hu*** is the method described in (Hu et al., 2004), which extracted opinion targets by using adjacent rule.

2) ***DP*** is the method described in (Qiu et al., 2011), which used *Double Propagation* algorithm to extract opinion targets depending on syntactic relations between words.

3) ***Zhang*** is the method described in (Zhang et al., 2010), which is an extension of *DP*. They extracted opinion targets candidates using syntactic patterns and other specific patterns. Then HITS (Kleinberg 1999) algorithm combined with candidate frequency is employed to rank the results for opinion target extraction.

Hu is selected to represent *adjacent* methods for opinion target extraction. And ***DP*** and ***Zhang*** are selected to represent *syntax-based* methods. The parameter settings in these three baselines are the same as the original papers. In special, for *DP* and *Zhang*, we used the same patterns for different language reviews. The overall performance results are shown in Table 3, 4 and 5, respectively, where "P" denotes precision, "R" denotes recall and "F" denotes F-measure. *Ours* denotes full model of our method, in which we use IBM-3 model for identifying opinion relations between words. Moreover, we set $\phi_{\max} = 2$ in Eq. (2) and $\lambda = 0.3$ in Eq. (7). From results, we can make the following observations.

1) *Ours* achieves performance improvement over other methods. This indicates that our method based on word-based translation model is effective for opinion target extraction.
2) The graph-based methods (*Ours* and *Zhang*) outperform the methods using *Double Propagation* (*DP*). Similar observations have been made by Zhang et al. (2010). The reason is that graph-based methods extract opinion targets in a global framework and they can effectively avoid the error propagation made by traditional methods based on *Double Propagation*. Moreover, *Ours* outperforms *Zhang*. We believe the reason is that *Ours* consider the opinion relevance and the candidate importance in a unified graph-based framework. By contrast, Zhang only simply

[5] http://nlp.stanford.edu/software/tagger.shtml

plus opinion relevance with frequency to determine the candidate confidence.

3) In Table 4, the improvement made by *Ours* on Restaurant (Chinese reviews) is larger than that on Hotel and MP3 (English reviews). The same phenomenon can be found when we compare the improvement made by *Ours* in Table 3 (Chinese reviews) with that in Table 5 (English reviews). We believe that reason is that syntactic patterns used in *DP* and *Zhang* were exploited based on English grammar, which may not be suitable to Chinese language. Moreover, another reason is that the performance of parsing on Chinese texts is not better than that on English texts, which will hurt the performance of *syntax-based methods* (*DP* and *Zhang*).
4) Compared the results in Table 3 with the results in Table 4, we can observe that *Ours* obtains larger improvements with the increase of the data size. This indicates that our method is more effective for opinion target extraction than state-of-art methods, especially for large corpora. When the data size increase, the methods based on syntactic patterns will introduce more noises due to the parsing errors on informal texts. On the other side, *Ours* uses WTM other than parsing to identify opinion relations between words, and the noises made by inaccurate parsing can be avoided. Thus, *Ours* can outperform baselines.
5) In Table 5, *Ours* makes comparable results with baselines in *Customer Review Datasets*, although there is a little loss in precision in some domains. We believe the reason is that the size of *Customer Review Datasets* is too small. As a result, WTM may suffer from data sparseness for association estimation. Nevertheless, the average recall is improved.

An Example In Table 6, we show top 10 opinion targets extracted by *Hu*, *DP*, *Zhang* and *Ours* in MP3 of *Large*. In *Hu* and *DP*, since they didn't rank the results, their results are ranked according to frequency in this experiment. The errors are marked in bold face. From these examples, we can see *Ours* extracts more correct opinion targets than others. In special, *Ours* outperforms *Zhang*. It indicates the effectiveness of our graph-based method for candidate confidence estimation. Moreover, *Ours* considers candidate importance besides opinion relevance, so some specific opinion targets are ranked to the fore, such as "voice recorder", "fm radio" and "lcd screen".

4.3 Effect of Word-based Translation Model

In this subsection, we aim to prove the effectiveness of our WTM for estimating associations between opinion targets and opinion words. For comparison, we select two baselines for comparison, named as *Adjacent* and *Syntax*. These baselines respectively use adjacent rule (Hu et al. 2004; Wang et al., 2008) and syntactic patterns (Qiu et al., 2009) to identify opinion relations in sentences. Then the same method (Eq.3 and Eq.4) is used to estimate associations between opinion targets and opinion words. At last the same graph-based method (in Section 3.3) is used to extract opinion targets. Due to the limitation of the space, the experimental results only on *COAE2008 dataset2* and *Large* are shown in Figure 3.

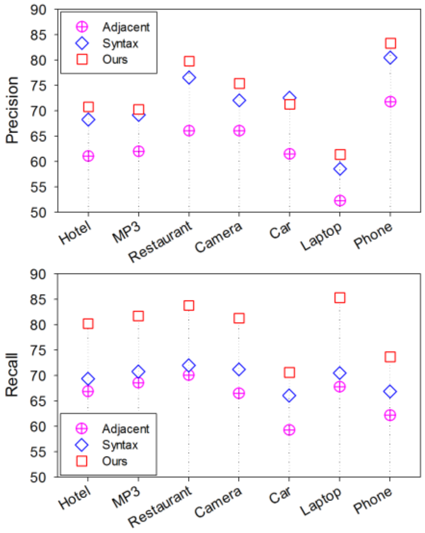

Figure 3: Experimental comparison among different relation identification methods

Hu	quality, **thing**, drive, feature, battery, sound, **time**, music, price
DP	quality, battery, software, device, screen, file, **thing**, feature, battery life
Zhang	quality, size, battery life, **hour**, version, function, upgrade, **number**, music
Ours	quality, battery life, voice recorder, video, fm radio, battery, file system, screen, lcd screen

Table 6: Top 10 opinion targets extracted by different methods.

In Figure 3, we observe that *Ours* using WTM makes significant improvements compared with two baselines, both on precision and recall. It indicates that WTM is effective for identifying opinion relations, which makes the estimation of the associations be more precise.

4.4 Effect of Our Graph-based Method

In this subsection, we aim to prove the effectiveness of our graph-based method for opinion target extraction. We design two baselines, named as *WTM_DP* and *WTM_HITS*. Both *WTM_DP* and *WTM_HITS* use WTM to mine associations between opinion targets and opinion words. Then, *WTM_DP* uses *Double Propagation* adapted in (Wang et al. 2008; Qiu et al. 2009) to extract opinion targets, which only consider the candidate opinion relevance. *WTM_HITS* uses a graph-based method of Zhang et al. (2010) to extract opinion targets, which consider both candidate opinion relevance and frequency. Figure 4 gives the experimental results on *COAE2008 dataset2* and *Large*. In Figure 4, we can observe that our graph-based algorithm outperforms not only the method based on *Double Propagation*, but also the previous graph-based approach.

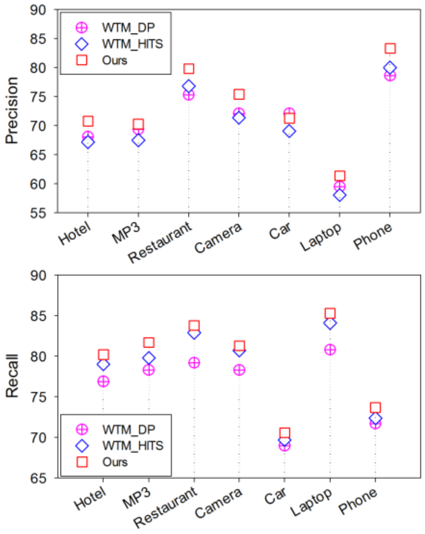

Figure 4: Experimental Comparison between different ranking algorithms

4.5 Parameter Influences

4.5.1 Effect of Different WTMs

In section 3, we use three different WTMs in Eq. (2) to identify opinion relations. In this subsection, we make comparison among them. Experimental results on *COAE2008 dataset2* and *Large* are shown in Figure 5. *Ours_1*, *Ours_2* and *Ours_3* respectively denote our method using different WTMs (IBM 1~3). From the results in Figure 5, we observe that *Ours_2* outperforms *Ours_1*, which indicates that word position is useful for identifying opinion relations. Furthermore, *Ours_3* outperforms other models, which indicates that considering the fertility of a word can produce better performance.

4.5.2 Effect of λ

In our method, when we employ Eq. (7) to assign confidence score to each candidate, $\lambda \in [0,1]$ decides the proportion of candidate importance in our method. Due to the limitation of space, we only show the F-measure of *Ours* on COAE2008 dataset2 and *Large* when varying λ in Figure 6.

In Figure 6, curves increase firstly, and decrease with the increase of λ. The best performance is obtained when λ is around 0.3. It indicates that candidate importance and candidate opinion relevance are both important for candidate confidence estimation. The performance of opinion target extraction benefits from their combination.

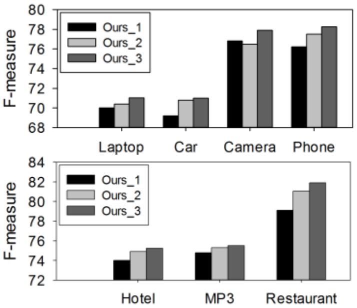

Figure 5. Experimental results by using different word-based translation model.

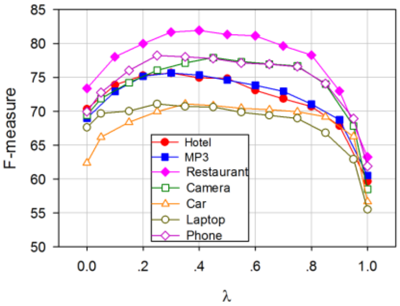

Figure 6. Experimental results when varying λ

5 Conclusions and Future Work

This paper proposes a novel graph-based approach to extract opinion targets using WTM. Compared with previous *adjacent methods* and *syntax-based methods*, by using WTM, our method can capture opinion relations more precisely and therefore be more effective for opinion target extraction, especially for large informal Web corpora.

In future work, we plan to use other word alignment methods, such as discriminative model (Liu et al., 2010) for this task. Meanwhile, we will add some syntactic information into WTM to constrain the word alignment process, in order to identify opinion relations between words more precisely. Moreover, we believe that there are some verbs or nouns can be opinion words and they may be helpful for opinion target extraction. And we think that it's useful to add some prior knowledge of opinion words (sentiment lexicon) in our model for estimating candidate opinion relevance.

Acknowledgements

The work is supported by the National Natural Science Foundation of China (Grant No. 61070106), the National Basic Research Program of China (Grant No. 2012CB316300), Tsinghua National Laboratory for Information Science and Technology (TNList) Cross-discipline Foundation and the Opening Project of Beijing Key Laboratory of Internet Culture and Digital Dissemination Research (Grant No. 5026035403). We thank the anonymous reviewers for their insightful comments.

References

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. *The Mathematics of Statistical Machine Translation: Parameter Estimation. Computational Linguistics*, 19(2): 263-311.

Xiaowen Ding, Bing Liu and Philip S. Yu. 2008. A Holistic Lexicon-Based Approach to Opinion Mining. In *Proceedings of WSDM 2008*.

Xiaowen Ding and Bing Liu. 2010. Resolving Object and Attribute Reference in Opinion Mining. In *Proceedings of COLING 2010*.

Mingqin Hu and Bing Liu. 2004. Mining and Summarizing Customer Reviews. In *Proceedings of KDD 2004*

Minqing Hu and Bing Liu. 2004. Mining Opinion Features in Customer Reviews. In *Proceedings of AAAI-2004*, San Jose, USA, July 2004.

Wei Jin and Huang Hay Ho. A Novel Lexicalized HMM-based Learning Framework for Web Opinion Mining. In *Proceedings of ICML 2009*.

Jon Klernberg. 1999. Authoritative Sources in Hyperlinked Environment. *Journal of the ACM* 46(5): 604-632

Zhuang Li, Feng Jing, Xiao-yan Zhu. 2006. Movie Review Mining and Summarization. In *Proceedings of CIKM 2006*

Fangtao Li, Chao Han, Minlie Huang and Xiaoyan Zhu. 2010. Structure-Aware Review Mining and Summarization. In *Proceedings of COLING 2010*.

Zhichao Li, Min Zhang, Shaoping Ma, Bo Zhou, Yu Sun. Automatic Extraction for Product Feature Words from Comments on the Web. In *Proceedings of AIRS 2009*.

Bing Liu, Hu Mingqing and Cheng Junsheng. 2005. Opinion Observer: Analyzing and Comparing Opinions on the Web. In *Proceedings of WWW 2005*

Bing Liu. 2006. Web Data Mining: Exploring Hyperlinks, contents and usage data. *Springer*, 2006

Bing Liu. 2010. Sentiment analysis and subjectivity. *Handbook of Natural Language Processing, second edition*, 2010.

Yang Liu, Qun Liu, and Shouxun Lin. 2010. Discriminative word alignment by linear modeling. *Computational Linguistics*, 36(3):303–339.

Zhanyi Liu, Haifeng Wang, Hua Wu and Sheng Li. 2009. Collocation Extraction Using Monolingual Word Alignment Model. In *Proceedings of EMNLP 2009*.

Tengfei Ma and Xiaojun Wan. 2010. Opinion Target Extraction in Chinese News Comments. In *Proceedings of COLING 2010*.

Popescu, Ana-Maria and Oren, Etzioni. 2005. Extracting produt fedatures and opinions from reviews. In *Proceedings of EMNLP 2005*

Guang Qiu, Bing Liu., Jiajun Bu and Chun Che. 2009. Expanding Domain Sentiment Lexicon through Double Popagation. In *Proceedings of IJCAI 2009*

Guang Qiu, Bing Liu, Jiajun Bu and Chun Chen. 2011. Opinion Word Expansion and Target Extraction through Double Propagation. *Computational Linguistics*, March 2011, Vol. 37, No. 1: 9.27

Qi Su, Xinying Xu., Honglei Guo, Zhili Guo, Xian Wu, Xiaoxun Zhang, Bin Swen and Zhong Su. 2008. Hidden Sentiment Association in Chinese Web Opinion Mining. In *Proceedings of WWW 2008*

Bo Wang, Houfeng Wang. Bootstrapping both Product Features and Opinion Words from Chinese Customer Reviews with Cross-Inducing. In *Proceedings of IJCNLP 2008*.

Hongning Wang, Yue Lu and Chengxiang Zhai. 2011. Latent Aspect Rating Analysis without Aspect Keyword Supervision. In *Proceedings of KDD 2011*.

Yuanbin Wu, Qi Zhang, Xuangjing Huang and Lide Wu, 2009, Phrase Dependency Parsing For Opinion Mining, In *Proceedings of EMNLP 2009*

Lei Zhang, Bing Liu, Suk Hwan Lim and Eamonn O'Brien-Strain. 2010. Extracting and Ranking Product Features in Opinion Documents. *In Proceedings of COLING 2010.*

Qi Zhang, Yuanbin Wu, Tao Li, Mitsunori Ogihara, Joseph Johnson, Xuanjing Huang. 2009. Mining Product Reviews Based on Shallow Dependency Parsing, In Proceedings of SIGIR 2009.

Guangyou Zhou, Li Cai, Jun Zhao and Kang Liu. 2011. Phrase-based Translation Model for Question Retrieval in Community Question Answer Archives. In Proceedings of ACL 2011.

Jingbo Zhu, Huizhen Wang, Benjamin K. Tsou and Muhua Zhu. 2009. Multi-aspect Opinion Polling from Textual Reviews. *In Proceedings of CIKM 2009*.