IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 11, November 2017

Word Alignment Model for Co-Extracting Opinion Targets and Opinion Words from Online Reviews

Padmaja Katta [1], **Nagaratna P. Hegde** [2]

Assistant Professor, Computer Science and Engineering, University College of Engineering Kakatiya University, Kothagudem, India[1]

Professor, Computer Science and Engineering, Vasavi College of Engineering, Hyderabad, India[2]

Abstract: Mining opinion targets and opinion words from online reviews are important tasks for fine-grained opinion mining, the key component of which involves detecting opinion relations among words. It proposes a novel approach based on the alignment model, which regards identifying opinion relations as an alignment process. Then, a graph-based co-ranking algorithm is exploited to estimate the confidence of each candidate. Then the candidates with higher confidence are extracted as opinion targets or opinion words. Compared to previous methods based on the nearest-neighbor rules, the model captures opinion relations more precisely, especially for long-span relations. Compared to syntax-based methods, the word alignment model effectively alleviates the negative effects of parsing errors when dealing with informal online texts. In particular, compared to the traditional unsupervised alignment model, the proposed model obtains better precision. In addition, when estimating candidate confidence, we penalize higher-degree vertices in the graph-based co-ranking algorithm to decrease the probability of error generation.

Keywords: Opinion, Opinion targets, Co-ranking, alignment.

I. INTRODUCTION

This system aims to develop a novel approach based on the alignment model, which regards identifying opinion relations as an alignment process. Then, a graph-based co-ranking algorithm is exploited to estimate the confidence of each candidate. Finally, candidates with higher confidence are extracted as opinion targets or opinion words. Compared to previous methods based on the nearest-neighbor rules, the model captures opinion relations more precisely, especially for long-span relations [1]. Compared to syntax-based methods, the word alignment model effectively alleviates the negative effects of parsing errors when dealing with informal online texts. In particular, compared to the traditional unsupervised alignment model, the proposed model obtains better precision because of the usage of partial supervision [6].

II. RELATED WORK

Mining and Summarizing Customer Reviews [8]

With the rapid expansion of e-commerce, more and more products are sold on the Web, and more and more people are also buying products online. In order to enhance customer satisfaction and shopping experience, it has become a common practice for online merchants to enable their customers to review or to express opinions on the products that they have purchased [8]. With more and more common users becoming comfortable with the Web, an increasing number of people are writing reviews. As a result, the number of reviews that a product receives grows rapidly. Some popular products can get hundreds of reviews at some large merchant sites. Furthermore, many reviews are long and have only a few sentences containing opinions on the product [8]. This makes it hard for a potential customer to read them to make an informed decision on whether to purchase the product. If he/she only reads a few reviews, he/she may get a biased view.

The large number of reviews also makes it hard for product manufacturers to keep track of customer opinions of their products. For a product manufacturer, there are additional difficulties because many merchant sites may sell its products, and the manufacturer may (almost always) produce many kinds of products [8]. Here, features broadly mean product features (or attributes) and functions. Given a set of customer reviews of a particular product, the task involves three subtasks: (1) identifying features of the product that customers have expressed their opinions on product features; (2) for each feature, identifying review sentences that give positive or negative opinions; and (3) producing a summary using the discovered information. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Picture quality and (camera) size are the product

DOI 10.17148/IJARCCE.2017.61136

features [8]. There are 253 customer reviews that express positive opinions about the picture quality, and only 6 that express negative opinions.

The link points to the specific sentences and/or the whole reviews that give positive or negative comments about the feature. With such a feature-based summary, a potential customer can easily see how the existing customers feel about the digital camera. If he/she is very interested in a particular feature, he/she can drill down by following the <individual review sentences> link to see why existing customers like it and/or what they complain about. For a manufacturer, it is possible to combine summaries from multiple merchant sites to produce a single report for each of its products [8].

The task is different from traditional text summarization in a number of ways. First of all, a summary in the case is structured rather than another (but shorter) free text document as produced by most text summarization systems. Second, customers are only interested in features of the product that customers have opinions on and also whether the opinions are positive or negative. Customers do not summarize the reviews by selecting or rewriting a subset of the original sentences from the reviews to capture their main points as in tra1ditional text summarization [8]. Mining product features that have been commented on by customers. The use of both data mining and natural language processing techniques to perform this task. This part of the study has been reported.

Identifying opinion sentences in each review and deciding whether each opinion sentence is positive or negative. These opinion sentences must contain one or more product features identified above. To decide the opinion orientation of each sentence (whether the opinion expressed in the sentence is positive or negative), we perform three subtasks. First, a set of adjective words (which are normally used to express opinions) is identified using a natural language processing method. These words are also called opinion words in this project [8]. Second, for each opinion word, we determine its semantic orientation, e.g., positive or negative.

A bootstrapping technique is proposed to perform this task using Word Net. Then, decide the opinion orientation of each sentence. An effective algorithm is also given for this purpose.

Cross-Domain Co-Extraction of Sentiment and Topic Lexicon [1]

In the past few years, opinion mining and sentiment analysis have attracted much attention in Natural Language Processing (NLP) and Information Retrieval (IR) (Pang and Lee, Liu,). Sentiment lexicon construction and topic lexicon extraction are two fundamental subtasks for opinion mining [1]. A sentiment lexicon is a list of sentiment expressions, which are used to indicate sentiment polarity (e.g., positive or negative). The sentiment lexicon is domain dependent as users may use different sentiment words to express their opinion in different domains (e.g., different products). A topic lexicon is a list of topic expressions, on which the sentiment words are expressed.

Extracting the topic lexicon from a specific domain is important because users not only care about the overall sentiment polarity of a review but also care about which aspects are mentioned in review. The, similar to sentiment lexicons, different domains may have very different topic lexicons. Recently, Jin and Ho and Li et al. showed that supervised learning methods can achieve state-of-the-art results for lexicon extraction. However, the performance of these methods highly relies on manually annotated training data. In most cases, the labeling work may be time consuming and expensive. It is impossible to annotate each domain of interest to build precise domain dependent lexicons [1]. It is more desirable to automatically construct precise lexicons in domains of interest by transferring knowledge from other domains. The co-extraction task of sentiment and topic lexicons in a target domain does not have any labeled data, but have plenty of labeled data in a source domain.

The goal is to leverage the knowledge extracted from the source domain to help lexicon co-extraction in the target domain. To address this problem, a two-stage domain adaptation method is proposed. In the first step, build a bridge between the source and target domains by identifying some common sentiment words as sentiment seeds in the target domain, such as “good”, “bad”, “nice”, etc. After that, generate topic seeds in the target domain by mining some general syntactic relation patterns between the sentiment and topic words from the source domain [1]. In the second step, a Relational Adaptive bootstrapping (RAP) algorithm is proposed to expand the seeds in the target domain. The useful labeled data from the source domain as well as exploit the relationships between the topic and sentiment words to propagate information for lexicon construction in the target domain.

Experimental results show that the proposed method is effective for cross-domain lexicon co-extraction In summary, we have three main contributions: 1) A systematic study on cross-domain sentiment analysis in word level [1]. While, most of previous work focused on document level; 2) A new two-step domain adaptation framework, with a novel RAP algorithm for seed expansion, is proposed. 3) It conducts extensive evaluation, and the experimental results demonstrate the effectiveness of our methods. Sentiment or topic lexicon extraction is to identify the sentiment or topic words from text. In the past, many machine learning techniques have been proposed for this task. Hu and Liu et al. (2004) proposed an association-rule-based method to extract topic words and a dictionary-based method to identify sentiment words, independently. Wiebe et al. (2004) and Rioff et al. (2003) proposed to identify subjective adjectives and nouns using word clustering based on their distributional similarity. Popescu and Etzioni (2005) proposed a relaxed labelling approach to utilize linguistic rules for opinion polarity detection.

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

DOI 10.17148/IJARCCE.2017.61136

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) [5]. It formulates identifying opinion relations between opinion targets and opinion words as a word alignment task.

They argue that an opinion target can find its corresponding modifier through monolingual word alignment. For example the opinion words “colorful” and “amazing” are aligned with the target “screen” through word alignment. In this WTM is used to perform monolingual word alignment for mining associations between opinion targets and opinion words.

Expanding Domain Sentiment Lexicon through Double Propagation [3]

Sentiment analysis is an important problem in opinion mining and has attracted a great deal of attention. The task is to predict the sentiment polarities (also known as semantic orientations) of opinions by analyzing sentiment words and expressions in sentences and documents [3]. Sentiment words are words that convey positive or negative sentiment polarities. A comprehensive sentiment lexicon is essential for sentiment analysis.

This is closely related to which extracts domain specific sentiment words in Japanese text. In their work, they exploit sentiment coherency within sentence and among sentences to extract sentiment candidates and then use a statistical method to determine whether a candidate is correct [3]. However, their idea of selecting candidates restricts the extracted sentiment words only to contexts with known sentiment words (seeds), and the statistical estimation can be unreliable when the occurrences of candidates are infrequent with small corpora. The key difference between our work and theirs is that we exploit the relationships between sentiment words and product features (or topics) in extraction [3]. This important information is not used in this work. Here product features (or features) mean product components and attributes [Liu, 2006].

Experimental results show that our approach, even without propagation, outperforms their method by 18% and 11% in precision and recall respectively. With propagation, our method improves even further. The proposed method identifies domain specific sentiment words from relevant reviews using only some seed sentiment words (we currently focus on product domains). The key idea is that in reviews sentiment words are almost always associated with features. Thus, sentiment words can be recognized by identified features. Since feature extraction itself is also a challenging problem, to extract features using the same seed sentiment words in a similar way (no seed feature is needed from the user) [3].

The newly extracted sentiment words and features are utilized to extract new sentiment words and new features which are used again to extract more sentiment words and features. The propagation ends until no more sentiment words or features can be identified. As the process involves propagation through both sentiment words and features, we call the method double propagation [3]. To the knowledge, no previous work on sentiment word extraction employed this approach. The extraction of sentiment words and features is performed Hatzivassiloglou and McKeown [1997] did the first work on tackling the problem of determining the semantic orientation (or polarity) of words.

Their method predicts the orientation of adjectives by analyzing pairs of adjectives extracted from a large document set. These pairs of adjectives are conjoined by and, or, but, either-or, or neither-nor. The underlying intuition is that the conjoining adjectives subject to linguistic constraints on the orientation of the adjectives involved [3]. For example, and usually conjoins two adjectives of the same orientation while but conjoins two adjectives of opposite orientations. It differs from theirs in that they are unable to extract unpaired adjectives while we could extract through features. Wiebe [2000] focused on the problem of subjectivity tagging and proposed an approach to finding subjective adjectives using the results of a method for clustering words according to their distributional similarity, seeded by a small number of simple adjectives extracted from a manually annotated corpus.

The basic idea is that subjective words are similar in distribution as they share pragmatic usages. However, the approach is unable to predict sentiment orientations of the found subjective adjectives. Turney and Littman adopt a different methodology which requires little linguistic knowledge. They first define two minimal sets of seed terms for positive and negative categories [3]. Then they compute the point wise mutual information (PMI) of the target term with each seed term as a measure of their semantic association. Positive value means positive orientation and higher absolute value means stronger orientation. Their work requires additional Web access. In [Kaji and Kitsuregawa, the authors propose to use language and layout structural clues of Web pages to extract sentiment sentences from Japanese HTML documents. The structural clues are set in advance. Adjectives/Adjective phrases in these sentences are treated as candidate sentiment phrases.

DOI 10.17148/IJARCCE.2017.61136

III. PROBLEM STATEMENT AND OBJECTIVE

a. Purpose

Frequency statistics and phrase detection, to detect the proper opinion targets/words.

b. Scope

An opinion target can find its corresponding modifier through word alignment. “Colorful” and “big” are aligned with the target word “screen” [1].

c. Problem Statement

In existing system unsupervised alignment and syntactic patterns are used to find relation between OT and OW. In figure1 due to unsupervised alignment “courteous” OW wrongly get align with OT “food” [1].

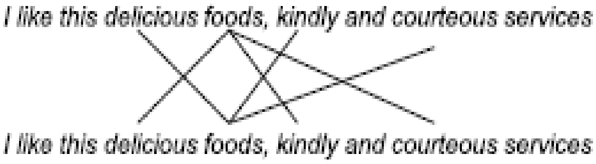
I like this delicious foods, kindly and courteous services
I like this delicious foods, kindly and courteous services

Figure 1.3 Unsupervised Alignment

d. Motivation

Consumers are often forced to wade through many on-line reviews in order to make an informed product choice, an unsupervised information extraction system which mines reviews in order to build a model of important product features, their evaluation by reviewers, and their relative quality across products [1].

e. Objectives

Step 1**:** Select a dataset of reviews.

Step2: Apply Parts Of Speech Tagging.

Step3: Extraction of candidates is done that is extracting opinion targets and opinion words.

Step4: Then calculate opinion associations among words. It generates Opinion Target Candidates, Opinion Word Candidates.

Step 5: Calculate the candidate confidence; it generates Opinion Target Confidence, Opinion Word Confidence.

Step 6: Finding the opinion targets and opinion words by giving a threshold value. It generates a graph of Opinion Targets and Opinion Words before and after Co-extraction

IV. PROPOSED SYSTEM

- To precisely mine the opinion relations among words ,a method based on a monolingual word alignment model (WAM) is proposed. An opinion target can find its corresponding modifier through word alignment.
- The standard word alignment models are often trained in a completely unsupervised manner, which results in alignment quality that may be unsatisfactory. It certainly can improve alignment quality by using supervision. However, it is both time consuming and impractical to manually label full alignments in sentences. Thus, employ a partially-supervised word alignment model (PSWAM).
- It can easily obtain a portion of the links of the full alignment in a sentence. These can be used to constrain the alignment model and obtain better alignment results. To obtain partial alignments, to resort syntactic parsing.

a. Advantages

- Compared to previous nearest-neighbor rules, the WAM does not constrain identifying modified relations to a limited window; therefore, it can capture more complex relations, such as long-span modified relations.
- Compared to syntactic patterns, the WAM is more robust because it does not need to parse informal texts. In addition, the WAM can integrate several intuitive factors, such as word co-occurrence frequencies and word positions, into a unified model for indicating the opinion relations among words. Thus, it expects to obtain more precise results on opinion relation identification.

The alignment model used has proved to be effective for opinion target extraction

V. METHODOLOGY

a. Parts of speech Tagging

- Part-of-speech tagging also called as grammatical tagging is the process of making up a word in a text as corresponding to particular part of speech.
- It is useful in Information Retrieval.

DOI 10.17148/IJARCCE.2017.61136

- Part-of-Speech tagger (POS tagger) was used to define boundaries and to produce for each word. The reason why opinions are split into sentences is basically to achieve a finer granularity as many discussed aspects may reside in different sentences that compose the whole text. Later on it will be discussed the optimal granularity level to analyze opinions. The tagged sentences produced by the NL Processor in this step, will play a very important role for the rest of the system. In feature identification, a data mining system will depend on the noun or noun phrases (two up to three neighbor nouns in a sentence) generated on this step to produce a number of frequent features. Also, the classification of sentiment will depend on the words classified both as adjectives and adverbs in this step to produce a set of possible opinion words [1].

b. Word Alignment Model

WAM method is based on the monolingual model, which precisely mine the opinion relations among the words.

Example: This phone has an amazing and colourful screen‖

Based on WAM, the opinion word and opinion target was extracted. In the above example, amazing and colourful is the opinion target and the screen is an opinion word. When compare to previous method syntactic patterns, the WAM precisely mine the words and target. The previous nearest-neighbour method precisely mines the relation for short span sentences. But WAM method precisely mine relation for both short span and long span relations [6]. The WAM method has some following constrains:

- Nouns/noun phrases should be aligned with adjectives/verbs/a null word.
- Other unrelated words, such as prepositions conjunctions and adverbs should be aligned only with themselves.

Given a sentence S with n words,

S = (w1, w2, . . . , wn), the word alignment is,

A^ = {(i, ai)| i ϵ [1, n], ai ϵ [1, n]}

Where (i, ai) means that a noun/noun phrase (opinion target) at position *i* is aligned with its modifier (opinion word) at position ai.

c. Partially-Supervised Word Alignment Model

Word Alignment Model (WAM) is trained in a completely unsupervised manner. This may not obtain accuracy of the alignment result. Thus to improve alignment performance we perform a partially-supervision on the statistic model and to produce a partially supervised alignment model [6]. Here partial alignments are regarded as constraints.

Word Alignment approach has shown significant performance improvement. This unsupervised approach could lead to wrong associations. So, in this model, a portion of links could be used to supervise this model [6]. Initially, syntactic methods with high precision and low recall are used, which will give at least a short dependency relation.

Partially-supervised word alignment model (PSWAM). At first, we apply PSWAM in a monolingual scenario to mine opinion relations in sentences and estimate the associations between words. Then, a graph-based algorithm is exploited to estimate the confidence of each candidate, and the candidates with higher confidence will be extracted as the opinion targets. Compared with existing syntax-based methods, PSWAM can effectively avoid parsing errors when dealing with informal sentences in online reviews [6].

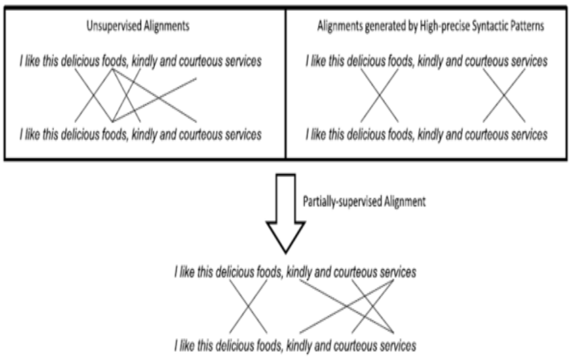
Unsupervised Alignments
I like this delicious foods, kindly and courteous services
I like this delicious foods, kindly and courteous services
Alignments generated by High-precise Syntactic Patterns
I like this delicious foods, kindly and courteous services
I like this delicious foods, kindly and courteous services
Partially-supervised Alignment
I like this delicious foods, kindly and courteous services
I like this delicious foods, kindly and courteous services

Figure (5.3) Mining opinion relations between words using partially supervised alignment model.

d. Partially Supervised Candidate Extraction & word alignment

These methods usually adopted coarser techniques, such as frequency statistics and phrase detection, to detect the proper opinion targets/words. They put more emphasis on how to cluster these words into their corresponding topics (or) aspects.

It obtains a set of word pairs, each of which is composed of a noun/noun phrase (opinion target candidate) and its corresponding modified word (opinion word candidate). Next, the alignment probabilities between a potential opinion target wt and a potential opinion word wt are estimated [6].

DOI 10.17148/IJARCCE.2017.61136

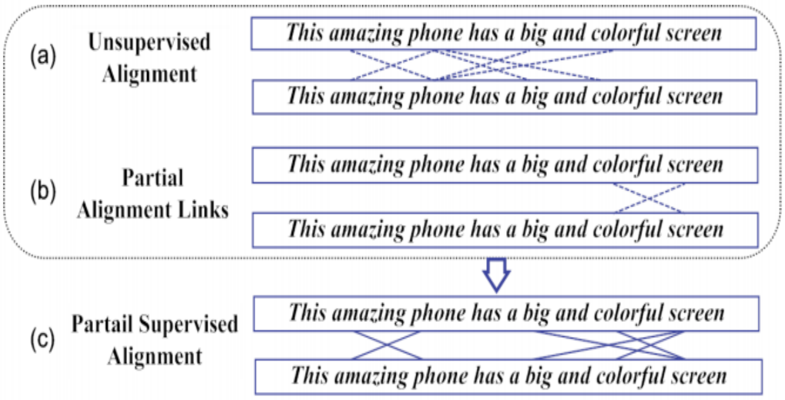
(a)
Unsupervised Alignment
This amazing phone has a big and colorful screen
This amazing phone has a big and colorful screen
(b)
Partial Alignment Links
This amazing phone has a big and colorful screen
This amazing phone has a big and colorful screen
(c)
Partait Supervised Alignment
This amazing phone has a big and colorful screen
This amazing phone has a big and colorful screen

Figure (5.3.1) Partially Supervised Alignment

e. Graph Co-Ranking Algorithm

After extracting the opinion word and the opinion target, the relations have been constructed by the opinion relation graph. Graph co-ranking method is estimated by candidate confidence of each opinion word and opinion target and this can be constructed on the graph. The word which has higher problem will be extracted as opinion word or opinion target [2]. The candidate confidence can be estimated by random walking method. Here the confidence of an opinion target candidates and opinion word candidates in the iterations, then the higher confidence than the threshold are obtained as an opinion word or opinion target [6]. The previous bootstrapping method has the error propagation problem. The graph based co-ranking algorithm effectively decreases the error problem.

The following features are used to represent the candidates:

- Salience feature: This feature indicates the salience degree of the candidates.
- Domain relevance feature: The opinion targets are domain specific and the difference between them has different domains.

In this process, we penalize high-degree vertices to weaken their impacts and decrease the probability of a random walk running into unrelated regions on the graph. Meanwhile, we calculate the prior knowledge of candidates for indicating some noises and incorporating them into our ranking algorithm to make collaborated operations on candidate confidence estimations [6]. Finally, candidates with higher confidence than a threshold are extracted.

Instead, the confidence of each candidate is estimated in a global process with graph co-ranking. Intuitively, the error propagation is effectively alleviated. The WAM does not constrain identifying modified relations to a limited window; therefore, it can capture more complex relations, such as long-span modified relations. Compared to syntactic patterns, the WAM is more robust because it does not need to parse informal texts. An opinion target can find its corresponding modifier through word alignment [9]. The confidence of each candidate is estimated in a global process with graph co-ranking. Intuitively, the error propagation is effectively alleviated. To help your readers, avoid using footnotes altogether and include necessary peripheral observations in the text (within parentheses, if you prefer, as in this sentence). Number footnotes separately from reference numbers, and in superscripts. Do not put footnotes in the reference list. Use letters for table footnotes.

f. Opinion Associations among Words

After alignment results [1], To get a set of opinion word and opinion target (candidates)pairs, each pair composed of a noun/noun phrase (opinion target candidates) and its corresponding modified word (opinion word candidate). Next, estimate the alignment probabilities between an opinion target Wt and a opinion word Wo,

$$P(Wt|Wo)=Count(Wt,Wo)/Count(Wo) \qquad (5.6.1)$$

Where P (Wt|Wo) means the alignment probability between opinion target and opinion words. Similarly, the alignment probability P (Wo|Wt) by changing the alignment direction in the alignment process. Then calculate the opinion association OA (Wt, Wo) between wt and Wo as follows:

$$OA(Wt,Wo)=(\alpha*P(Wt|Wo)+(1-\alpha)P(Wo|Wt))\text{^}-1 \qquad (5.6.2)$$

Where α is a constant, a harmonic factor used to combine these two alignment probabilities. In this paper, α= 0.5.

Estimating Candidate Confidence [1] With Graph Co Ranking. Confidence of candidates related to neighbour words. Confidence of a candidate (opinion target or opinion word) determined by its neighbours according to the opinion associations among them.

DOI 10.17148/IJARCCE.2017.61136

Candidate Confidence

Calculate the confidence of each candidate (Opinion target and Opinion word) through standard random walk with restart algorithm.

Confidence of each candidate is,

$$C_t^{k+1} = (1-\mu)* M_{to} *Co + \mu * I_t \qquad (5.6.3)$$

$$C_o^{k+1} = (1-\mu)* M_{to} *C_t + \mu * I_o \qquad (5.6.4)$$

Where C_t^{k+1} *and* C_o^{k+1} are the opinion target candidate and opinion word candidate confidence, in the (k+1) iteration. Co and Ct are the confidence of an opinion target candidate and opinion word candidate in the k iteration. Mt0 is opinion associations between opinion target candidate and opinion word candidate. mij ϵ Mto means the opinion association between the i-th opinion target candidate and the j-th opinion word candidate. where M_{to}*Co and M_{to} *C_t are the confidence of an opinion target or opinion word candidate is obtained through aggregating confidences of all neighbouring opinion word (opinion target) candidates together according to their opinion associations. The other ones are It and Io, which denote prior knowledge of opinion target candidate and opinion word candidate.

g. Methodology

The main framework of the method is extracting opinion targets/words as a co-ranking process. Assume that all nouns/noun phrases in sentences are opinion target candidates, and all adjectives/verbs are regarded as potential opinion words, which are widely adopted by previous methods. Each candidate will be assigned a confidence, and candidates with higher confidence than a threshold are extracted as the opinion targets or opinion words. If a word is likely to be an opinion word, the nouns/ noun phrases with which that word has a modified relation will have higher confidence as opinion targets.

The confidence of a candidate (opinion target or opinion word) is collectively determined by its neighbours according to the opinion associations among them of each candidate. Existing system on opinion mining have applied various methods for extracting opinion targets and opinion words. Extracting opinion targets and opinion words using word alignment model using partially supervised word alignment model. The proposed method contains three main modules. They are pre-processing, opinion target and word extraction and opinion word classification. The overall diagram of the proposed method is shown in Fig.1. In pre-processing the given comment is processed and eliminates stop word and stemming. Extracting opinion targets/words as a co-ranking process. Assume all nouns/noun phrases in sentences are opinion target candidates, and all adjectives/verbs are regarded as potential opinion words. And then the opinion word is classified as good or bad.

1. Pre-Processing:

This is the first step of the proposed method. Several pre-processing steps are applied on the given comment to optimize it for further experimentations. The proposed model for data pre-processing is shown in Fig. Tokenization process splits the text of a document into sequence of tokens. The splitting points are defined using all non-letter characters. This results in tokens consisting of one single word (unigrams). The movie review data set was pruned to ignore the too frequent and too infrequent words. Absolute pruning scheme was used for the task. Length based filtration scheme was applied for reducing the generated token set. The parameters used to filter out the tokens are the minimum length and maximum length. The parameters define the range for selecting the tokens. Stemming defines a technique that is used to find the root or stem of a word. The filtered token set undergoes stemming to reduce the length of words until a minimum length is reached. This resulted in reducing the different grammatical forms of a word to a single term.

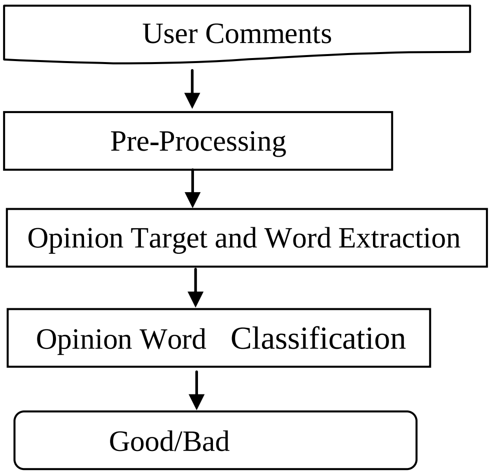
User Comments
Pre-Processing
Opinion Target and Word Extraction
Opinion Word Classification
Good/Bad

Figure (5.7.1) Pre-Processing

2. Opinion Targets and Opinion Words Extraction

The modified word alignment model assumes that all nouns/noun phrases in sentences are opinion target candidates, and all adjectives/verbs are regarded as potential opinion words. A noun/noun phrase can find its modifier through word alignment.

The proposed word alignment model applies a partially supervised modified word alignment model. It performs modified word alignment in a partially supervised framework. After that, obtain a large number of word pairs, each of which is composed of a noun/noun phrase and its modifier. And then calculate associations between opinion target candidates and opinion word candidates as the weights on the edges.

The modified word alignment model assumes that all nouns/noun phrases in sentences are opinion target candidates, and all adjectives/verbs are regarded as potential opinion words. A noun/noun phrase can find its modifier through word alignment.

The proposed word alignment model applies a partially supervised modified word alignment model. It performs modified word alignment in a partially supervised framework. After that, obtain a large number of word pairs, each of which is composed of a noun/noun phrase and its modifier. And then calculate associations between opinion target candidates and opinion word candidates as the weights on the edges.

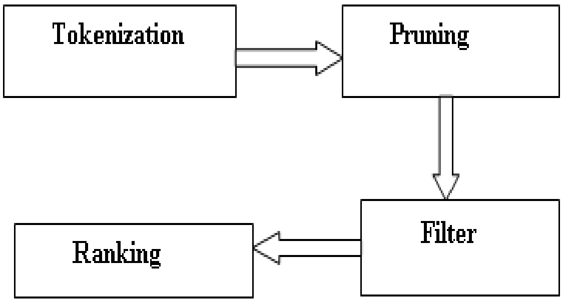

Figure (5.7.2) Process Flow Diagram of Pre-processing

3. Opinion Words Classification

After extraction opinion word and target the next step is to classify the opinion word.

In this process the opinion word is classified as good or bad. The knn classifier is used to classify the opinion word. Once it is classified as bad then the comment is removed. In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbours, with the object being assigned to the class most common among its k nearest neighbours (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbour.

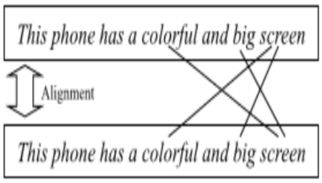
This phone has a colorful and big screen
Alignment
This phone has a colorful and big screen

Figure (5.7.3) Opinion Relations between Words and Targets using Modified Word Alignment Model

VI. RESULTS

Methods		Phone			TV			Camera	
	P	R	F	P	R	F	P	R	F
Word Alignment Model	0.84	0.75	0.79	0.73	0.82	0.77	0.72	0.84	0.78
Partially Supervised Word Alignment Model	0.86	0.75	0.80	0.78	0.83	0.80	0.75	0.85	0.81

TABLE. 1.RESULTS OF OPINION TARGET EXTRACTION ON DATA SETS

We select various types of datasets such as camera,Phone,TV and other reviews to test our method Word Alignment model. In our method we are currently limited to English language but we can also try other languages as the input to the method. In our method the reviews are first segmented into sentences according to punctuation. In the Table 1, “P” denotes Precision, “R” denotes Recall, and “F” denotes F-measure.

WAM uses an unsupervised Word Alignment model to mine the association between words. PSWAM uses an partially supervised Word Alignment model to mine the opinion relation between words. Next a graph based co-ranking algorithm is used to extract opinion target and opinion words.

VII. CONCLUSION

It proposes a method for co-extracting opinion targets and opinion words by using a word alignment model. The main contribution is focused on detecting opinion relations between opinion targets and opinion words. Compared to previous methods based on nearest neighbor rules and syntactic patterns, in using a word alignment model, the method captures opinion relations more precisely and therefore is more effective for opinion target and opinion word extraction. Next, to construct an Opinion Relation Graph to model all candidates and the detected opinion relations among them, along with a graph co-ranking algorithm to estimate the confidence of each candidate. The items with higher threshold are extracted out. The experimental results for the datasets with different sizes prove the effectiveness of the proposed method

REFERENCES

[1] F. Li, S. J. Pan, O. Jin, Q. Yang, and X. Zhu, “Cross-domain co extraction of sentiment and topic lexicons,” in Proc. 50th Annu. Meeting Assoc. Comput. Linguistics, Jeju, Korea, 2012, pp. 410–419.

[2] G. Qiu, L. Bing, J. Bu, and C. Chen, “Opinion word expansion and target extraction through double propagation,” Comput. Linguistics, vol. 37, no. 1, pp. 9–27, 2011.

[3] G. Qiu, B. Liu, J. Bu, and C. Che, “Expanding domain sentiment lexicon through double propagation,” in Proc. 21st Int. Jont Conf. Artif. Intell., Pasadena, CA, USA, 2009, pp. 1199–1204.

[4] J. M. Kleinberg, “Authoritative sources in a hyperlinked environment,” J. ACM, vol. 46, no. 5, pp. 604–632, Sep. 1999.

[5] K. Liu, L. Xu, and J. Zhao, “Opinion target extraction using word based translation model,” in Proc. Joint Conf. Empirical Methods Natural Lang. Process. Comput. Natural Lang. Learn., Jeju, Korea, Jul. 2012, pp. 1346–1356.

[6] K. Liu, H. L. Xu, Y. Liu, and J. Zhao, “Opinion target extraction using partially-supervised word alignment model,” in Proc. 23rd Int. Joint Conf. Artif. Intell., Beijing, China, 2013, pp. 2134–2140.

[7] L. Zhang, B. Liu, S. H. Lim, and E. O’Brien-Strain, “Extracting and ranking product features in opinion documents,” in Proc. 23th Int. Conf. Comput. Linguistics, Beijing, China, 2010, pp. 1462–1470.

[8] M. Hu and B. Liu, “Mining and summarizing customer reviews,” in Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, Seattle, WA, USA, 2004, pp.168–177.

[9] M. Hu and B. Liu, “Mining opinion features in customer reviews,” in Proc. 19th Nat. Conf. Artif. Intell., San Jose, CA, USA, 2004, pp. 755–760.

[10] N. Jakob and I. Gurevych, “Extracting opinion targets in a single- and cross-domain setting with conditional random fields,” in Proc. Conf. Empir. Meth. Nat. Lang.Process.2010, pp. 1035–1045.

[11] R. C. Moore, “A discriminative framework for bilingual word alignment,” in Proc. Conf. Human Lang. Technol. Empirical Methods Natural Lang. Process., Vancouver, BC, Canada, 2005, pp. 81–88.

[12] T. Ma and X. Wan, “Opinion target extraction in chinese news comments.” in Proc. 23th Int. Conf. Comput. Linguistics, Beijing, China, 2010, pp. 782–790.

[13] W. Jin and H. H. Huang, “A novel lexicalized HMM-based learning framework for web opinion mining,” in Proc. Int. Conf. Mach. Learn., Montreal, QC, Canada, 2009, pp. 465–472.

[14] X. Ding, B. Liu, and P. S. Yu, “A holistic lexicon-based approach to opinion mining,” in Proc. Conf. Web Search Web Data Mining, 2008, pp. 231–240.

[15] Y. Wu, Q. Zhang, X. Huang, and L. Wu, “Phrase dependency parsing for opinion mining,” in Proc. Conf. Empirical Methods Natural Lang. Process., Singapore, 2009, pp. 1533–1541.

DOI 10.17148/IJARCCE.2017.61136