Extracting Product Features and Sentiments from Chinese Customer Reviews

Shu Zhang, Wenjie Jia, Yingju Xia, Yao Meng, Hao Yu
Fujitsu Research and Development Center
Dong Si Huan Zhong Rd, Chaoyang District, Beijing and 0086, China
E-mail: {zhangshu, wj_jia, jyxia, mengyao, yu}@cn.fujitsu.com

Abstract

With the growing interest in opinion mining from web data, more works are focused on mining in English and Chinese reviews. Probing into the problem of product opinion mining, this paper describes the details of our language resources, and imports them into the task of extracting product feature and sentiment task. Different from the traditional unsupervised methods, a supervised method is utilized to identify product features, combining the domain knowledge and lexical information. Nearest vicinity match and syntactic tree based methods are proposed to identify the opinions regarding the product features. Multi-level analysis module is proposed to determine the sentiment orientation of the opinions. With the experiments on the electronic reviews of COAE 2008, the validities of the product features identified by CRFs and the two opinion words identified methods are testified and compared. The results show the resource is well utilized in this task and our proposed method is valid.

1. Introduction

In recent years, there is a growing interest in opinion mining from web data. With more and more people express their opinions on almost anything in Internet, forums, blogs and community websites, the web contains a wealth of opinions. For example, the rapidly increased product reviews left by the customers. They contain lots of useful information both for the potential buyers and the merchants. It is hard and time-consuming for people to read hundreds of reviews on a product, so how to extricate people from this work, how to analyze and extract valuable information automatically, which have been receiving a lot of attention from researchers. The task is not only technically challenging, but also very useful in practice.

According to the processing grain, opinion mining could be divided into three levels: document level, sentence level and feature level. For the opinion mining on document and sentence level, the task is to classify either positively or negatively in a review. However, the sentiment orientation of a review is not sufficient for many applications. Opinion mining begins to focus on the finer-grained features level mining. The task is to find not only the sentiment orientation but also the commented features. This information could be used to deeply analyze prevalent attitudes or generate various types of opinion summaries.

This paper focuses on the feature level product reviews mining. Given a review, the task is to extract product feature associated with its sentiment orientation. The task is typically divided into three main subtasks: (i) identifying product features, (ii) identifying opinions regarding the product features, and (iii) determining the sentiment orientation of the opinions. This paper focuses on Chinese customer reviews, introduces the resource utilized in the task, and probe into performance of supervised method on solving this problem.

The remainder of the paper is organized as follows: section 2 describes the related work. Section 3 gives the task description and resources description. Section 4 presents our methods and techniques in three subtasks, and how to utilize the lexicon information. Section 5 gives the experiments and results. Finally section 6 summarizes this paper.

2. Related Work

A series of topic symposiums and evaluation sessions on opinion mining have appeared in TREC and NTCIR.

The techniques on identifying product features are primarily based on unsupervised mining. The most representative research is that of (Hu and Liu, 2004). They adopt association rule mining for extracting nouns as frequent features. Compactness pruning and redundancy pruning are used to filter the incorrect features. Popescu and Etzioni (2005) utilize relation-specific extraction patterns with web PMI assessor to assess feature candidates. For the identification of opinion words, the method of the nearest vicinity match is simple and effective. Further, Researchers (Su et al., 2008; Du and Tan, 2009) focus on the association between the features and opinion words by mutual reinforcement approach. Ding and Liu (2007) combines the linguistic rules and opinion aggregation function to determine the semantic orientation.

Different from the previous work, we adopt supervised method to identify product features. In this stage, we utilize some language resource, such as domain feature lexicon, opinion lexicon, factor words lexicon. The importing of domain knowledge is aimed to improve the quality of opinion analysis. With the manually tagged training corpus, we transfer the task of identifying product features into information extraction task using CRFs. In the stage of identifying opinions regarding the product features, we compare the performances of the nearest vicinity match based and syntactic tree based methods. And we also import multi-level analysis module to determine the sentiment orientation of the opinions.

3. Resources Description

3.1 Task Description

Given a sentence S, extract each tuple (f, o) in S, where f is the opinionated product feature in sentence S, o is the sentiment orientation of the features f. Here, feature set is

an open set, sentiment orientation labels set is ({positive, negative}).

For example:

Sentence S: 相机的像素不错，照片很清晰，可惜内存不够大。(The camera has the good pixel, photo is very clear, unfortunately, the memory is not big enough.)

The tuples (像素(pixel), positive), (照片(photo), positive), (内存(memory), negative) would be assigned as the right result.

The features extracted are only evaluated ones. The features set include the following types: properties, parts, features of product parts, related concepts, parts and properties of related concepts. Table 1 shows the examples of such features in Camera domain.

Product feature	Feature sample
Properties	Size
Product Parts	Screen
Features of Product Parts	Screen Resolution
Related Concepts	Photo
Properties of Related Concepts	Photos' Definition

Table 1: Examples of Product Features

3.2 Resource Description

The knowledge resource is useful for improving the performance of the opinion mining. We adopt three lexicons in this task: opinion words lexicon, factor words lexicon and domain feature lexicon. The details of the resources will be introduced in the following.

3.2.1. Opinion Words Lexicon

General opinion words lexicon, which contains about 5,000 comment and sentiment words. They are oriented to all the fields. Each opinion words have a sentiment orientation and extent with it.

For example, <安全(safety) +2>. Here, +, - and 0 are used to represent positive, negative and neutral orientation.

The extent is ranged from 1 to 3 with the increasing sentiment degree.

3.2.2. Factor Words Lexicon

Factor words lexicon, which contains about 300 extent words. These words could strengthen, weaken the surrounding opinion words' extent or even transit its sentiment orientation.

For example, <不(not) -1>. If it is appeared nearly to the opinion word "good" in a sentence, then it transits the sentiment orientation from positive to negative.

3.2.3. Domain Feature Lexicon

Domain feature lexicon, which contains not only the features in the given domain but also the common opinion words associated with the features.

For example, <[CCD 尺寸(CCD size)] 大(big)/1 小(small)/-1 …>. This information is valuable, since general opinion words may not reflect the real orientation for some domain features. For example, the word "small" associated with the feature "CCD size" has a negative orientation. However when it is associated with the feature "camera size", it has a positive orientation.

For Chinese opinion mining, the language resources given above are valuable, which have been constructed in recent years.

4. Our Methods

We adopt the traditional way to divide the task into three subtasks, firstly identify the product features, then identify opinion words associated with the opinionated features, and determine the sentiment orientation of the opinions at last.

Knowledge resources are imported in all three stages. Three lexicons are all adopted in the stage of identifying product features. Opinion words lexicon is adopted in the stage of identifying opinions regarding the product features. Opinion words lexicon and factor words lexicon are adopted in the stage of determining the sentiment orientation of the opinions.

4.1 Identifying Product Features

The product features are mostly noun or noun phrases, so we regard this subtask as an entity recognizing process, and hope to transfer the effective NER techniques to solve this problem. We adopt the Conditional Random Fields module (Lafferty et al, 2001) to implement this subtask, which has been proved well performance in information extraction field.

CRFs modules has the advantages of relaxing strong independence assumptions made in HMM (Rabiner, 1989), and avoiding the label bias problem existed in MEMM (McCallum et al, 2000).

In the CRFs modules, we import word, POS and semantic information as tokens. The semantic information not only includes its character as a product feature, but also contains the character about opinion expression. The opinion information is a good indicator, because people like to express their opinions around the product features. All the semantic information is captured dependent on the above language resources. In this stage, we not only tag the product features but also tag the opinion words as attachment.

Another reason for us to adopt the supervised method to implement this subtask is that the unsupervised frequency-based methods are dependent on the statistic of the corpus, when given a single sentence, they couldn't execute effectively.

4.2 Identifying Opinions Regarding the Product Features

In this subtask, nearest vicinity match based and syntactic tree based methods are proposed to confirm the associated opinion word.

As observed, the opinion words mostly appear around the features in the review sentences. They are highly dependent on each other. So we hypothesize that opinion words appear around product features. If an opinion word

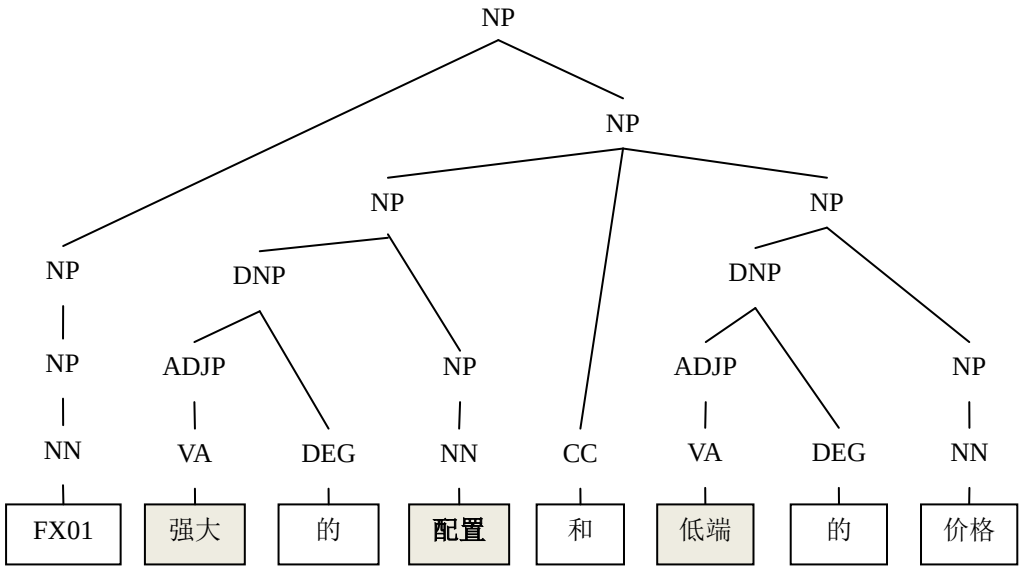

Figure 1: Example of Parsing Tree

co-occurs with a product feature within a given distance in a sentence, this opinion word is regarded to be associated with this product feature. Otherwise they are considered to be unrelated.

Nearest vicinity match based method has two steps to identify the opinion words. First, it takes the product feature as the center to find opinion word tagged by CRFs in the given distance. If there is no opinion word tagged by CRFs, then it secondly looks at the opinion lexicon for the further search. If there is also no opinion word found, the product feature is considered to have no sentimental meaning, which will be deleted.

Dependent on the plane distance to capture the opinion words is not sufficient. So we adopt syntactic tree based method to capture the relation. Here, we compute the distance of two items based on the syntactic parsing tree, and measure it by the shortest path.

Figure 1 shows the example of a review. It could not be determined by nearest vicinity match based method which opinion word is associated with the feature "配置". As in the sentence, word "配置" has the same plane distance with both the opinion word "强大" and opinion word"低端". Nearest vicinity match based method is dependent on the distance of the text string to judge the relative extent of two terms. It has no consideration for grammar information of the sentence. In fact, the grammar and syntactic structure contains more associated information between the terms.

So we utilize the distance of the two terms in the parsing tree to measure their relation. The distance of two leaf nodes are calculated with the shortest path of the two nodes in parsing tree. The distance of word "配置" with opinion word "强大" is 7, and that with opinion word"低端" is 9. The opinion word "强大" is more associated with feature "配置".

In the process, we calculate the nodes' shortest path on parsing tree between each opinion words and the feature. Then, we choose the opinion word with the shortest length as the associated opinion words.

4.3 Determining the Sentiment Orientation of the Opinions

The orientation of the product features are judged from multi-levels: sentence level, context level and opinion word level.

Sentence level judgment considers whether a sentence express any opinion information. The features in non sentiment sentence should not be extracted. Context level judgment considers the sentiment transition by emotional adverbs or phrases, such as word "不(no)". Opinion word level judgment considers the opinion word associated with product feature.

Since sentiment orientation is only positive or negative in this task, the results are combined of three level judgments with product. It is defined in the following way.

$$O(f_i)=IsO(S)*IsN(a_i)*\ Sign(o_i) \qquad (1)$$

Here, IsO(S) is a two value function, which judges the orientation of the feature by sentence level. If the sentence is judged to have no sentiment, then its value is 0, else is 1. At present, we only consider assumption sentence. We think this type of sentence doesn't express any opinion information, so the features in it should not be tagged.

$IsN(a_i)$ is also a two value function, which considers the sentiment transition by emotional adverbs or phrases. It is on the context level to consider the orientation. If there is such a word around the feature in a region, the value of the function is -1, else is 1.

$Sign(o_i)$ is directly the orientation of associated opinion word for the feature. Its value is defined as -1 for negative, 0 for neutral and 1for positive orientation.

Both the information about a_i and o_i are determined by looking at the opinion words lexicon and factor words lexicon.

5. Experiments

5.1 Corpus and Metric

The data used in the experiment is provided by the COAE (The first Chinese Opinion Analysis and Evaluation), which was held in 2008, aims to enable researchers to participate in large-scale experiments and evaluations, make each researcher's result comparable and promote the related technique in Chinese opinion analysis.

Here, we use the data of electronic domains, which includes Camera, Phone and Notebook domain. It has about 1,500 sentences. All sentences have been annotated by human beforehand.

The precision, recall and F-measure will be presented for the task. Both the strict match and lenient match are adopted for evaluating the correct units. Strict match means the results submitted by systems are exactly same with the human labels. Lenient match means if the result is only part of the standard, it is also thought to be the right one. The correct one is not only giving the right opinionated features but also giving its sentiment orientation correctly.

5.2 Results and Analysis

Table 2 and Table 3 present the whole performance of the task in the following. Table 2 shows the strict metric-based result and Table 3 shows the lenient metric-based result.

The difference of the two RunID is that they adopt different methods in stage of identifying opinions regarding the product features. RunID Near adopts nearest vicinity match based method to confirm the associated opinion word. RunID Tree adopts syntactic tree based method.

RunID	Strict		
	Precision	Recall	F-measure
Near	0.4041	0.3693	0.3859
Tree	0.4226	0.3489	0.3822

Table 2: Results by Strict metric

RunID	Lenient		
	Precision	Recall	F-measure
Near	0.5257	0.4805	0.5020
Tree	0.5484	0.4527	0.4960

Table 3: Results by Lenient metric

There are 13 participants in this task in COAE 2008, our results of nearest vicinity match based method gets rank 1 in strict metric and rank 2 in lenient metric, which proves that the proposed methods in this task is feasible and valid. And it also proves the nearest vicinity could perform well in finding the associated opinion words for the product feature.

Compared the nearest vicinity match based and syntactic tree based methods for identifying opinions regarding the product features, we find that they have the similar performance. Syntactic tree based method is better in precision with the loss of recall. Though importing more syntactic information, syntactic tree based method has no obvious improvement. These maybe have the relation with the low performance of the syntactic parsing technique, or the simple distance measurement chosen in calculating the relative associated extent.

6. Conclusion

In this paper, we probe into the problem of product opinion mining. We describe the details of our language resources, and import them into the task of extracting product feature and sentiment task. With the experiments on the electronic reviews, the validities of the product features identified by CRFs and the two opinion words identified methods are testified and compared. The results show that our method is valid.

7. References

Ding, X. W., Liu, B. (2007). The Utility of Linguistic Rules in Opinion Mining. In *Proceedings of Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 811--812.

Du, W.F., Tan, S.B. (2009). An Iterative Reinforcement Approach for Fine-Grained Opinion Mining. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 486--492.

Hu, M.Q., Liu, B. (2004). Mining and Summarizing Customer Reviews. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, pp. 168--177.

Hu, M.Q., Liu, B. (2004). Mining Opinion Features in Customer Reviews. In *Proceedings of the American Association for Artificial Intelligence*, pp. 775--760.

Popescu, A., Etzioni, O. (2005). Extracting Product Features and Opinions from Reviews. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 339--346.

Su, Q., Xu, X.Y. and et al. (2008). Hidden Sentiment Association in Chinese Web Opinion Mining. In *Proceedings of the Seventeenth International Conference on World Wide Web*, pp. 959--968.

Lafferty, J., McCallum, A., Pereira, F. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of International Conference on Machine Learning*, pp. 282--289.

Rabiner, L. R. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. In *Proceedings of* IEEE, 77 (2), pp. 257--286.

McCallum, A., Freitag, D., Pereira, F. (2000). Maximum Entropy Markov Models for Information Extraction and Segmentation. In *Proceedings on International Conference of Machine Learning*, pp. 591--598.