A Classification-based Approach for Implicit Feature Identification

Lingwei Zeng and Fang Li

Department of Computer Science and Engineering,
Shanghai Jiao Tong University, China
{chasuner,fli}@sjtu.edu.cn

Abstract. The explosive development of e-commerce has attracted a lot of people to shop online, and most of the online merchants allow users to post their reviews about various products or services. Now there are massive amounts of reviews on the web, these reviews are valuable resources for both potential customers and e-commerce companies. In recent years, sentiment analysis and opinion mining has grown to be one of the most active research areas. Automatically analyzing and processing these online reviews could be extremely helpful for a lot of web applications. To extract the details of each review, finer-grained opinion mining has received more and more attention. Most of the existing researches on feature-level opinion mining are dedicated to extract explicitly appeared features and opinion words on review sentences. However, among the numerous kinds of reviews on the web, there are a significant number of reviews that contain only opinion words and these opinion words implies product features. The identification of such implicit features is still one of the most difficult problems in opinion mining, and it's a even harder task on Chinese reviews due to complexity of Chinese language. In this paper, we focus on the task of implicit feature identification, which aims at identifying the implicit features without explicitly appearing in customer reviews. We propose a novel classification-based approach to deal with the problem of implicit feature identification. Firstly, By exploiting the word segmentation, part-of-speech(POS) tagging and dependency parsing, we propose a rule based method to extract the explicit feature-opinion pairs from customer reviews. Secondly, we cluster the feature-opinion pairs for each opinion word, and then construct the training document for each clustered feature-opinion pair from customer reviews. Finally, we formulate the identification of implicit feature into a classification-based feature selection problem, so we can identify the implicit feature by exploiting the text classification approach. Empirical evaluation demonstrates that our approach outperforms existing state-of-the-art methods significantly.

Keywords: Opinion Mining, Implicit Feature, Feature Extraction

1 Introduction

Recently, with the emerging and tremendous growth and popularity of Web2.0, more and more people tend to express their opinions about various kinds of things on the Internet. There are a large amount of user-generated contents in online forums, review websites and shopping websites, such as Yelp and Amazon. Such contents usually contains people's opinions and are extremely valuable resources for a lot of applications. Nowadays, an increasing number of people are buying products on the web. Product reviews on the web can provide a lot of helpful information for these potential customers. At the same time, the product manufacturers can also obtain useful feedbacks about products from users, so they can effectively improve their products according to these feedbacks. However, as the number of reviews that a product receives grows rapidly, sometimes the amount of comments of a product may exceed one thousand or even more, which makes it very hard for a potential customer to read all these reviews to obtain useful information conveyed by other users who have experienced it first-hand. In order to automatically process and analyze reviews

on the web, a lot of research efforts[4, 9, 5] have been done on opinion mining and sentiment analysis from the magnitude of reviews.

Previous researches on opinion mining usually deal with the task of mining a large scale document collection of customer reviews as a classification of either positive or negative sentiment. Document-level sentiment analysis[12] mainly classifies the whole review's emotional orientation, which determines whether the review expresses an overall positive or negative opinion about the product. Instead of identifying the whole review's sentiment orientation, sentence-level sentiment analysis determines whether each sentence expressed a positive, negative, or neutral opinion, which is closely related to subjectivity classification[19, 21, 20] that distinguish subjective sentences with opinions from objective sentences that only express factual information. However, both the document-level and sentence-level analysis can not discover what exactly people liked and did not like, thus simply judging the sentiment orientation of a review unit fails to detect many significant details, which is far from sufficient for many applications. In order to extract specific features and opinions from reviews, many researchers began to study the problem of finer-grained opinion mining, which is known as feature-level opinion mining[13, 4, 9, 5].

Example 1. 很漂亮，功能很多值得购买，价格有点贵，很喜欢白色，送货比较快！ (Very beautiful, there are plenty of functions that worth to buy, the price is a little expensive, really like the white color, the delivery is also very fast!)

When we read a customer review, we mostly concern the opinion word and its corresponding aspect or feature. In product review mining, feature is usually the component or attribute of the product. Example 1 is a digital camera review about *Nikon D90*. Explicit features such as "功能" (function), "价格" (price) and "送货" (delivery) can be extracted from the above comment. Except for explicit feature, there is another significant kind of feature that doesn't directly appear in the review sentences but is implied or can be deduced from the opinion word, which is known as implicit feature. In the above example reviews, the opinion word "漂亮" (beautiful) has implied that the feature which the user talked about is the camera's "外观" (exterior) although this feature doesn't explicitly appear in the sentence.

In feature-specific opinion mining, most of the existing researches[1, 24, 7, 23] mainly focused on the problem of extracting product features and opinions that explicitly appeared in review sentences. However, according to the observation, in our crawled Chinese reviews of five kinds of digit cameras, we statistically discover that at least 28 percent of the sentences are implicit sentences that imply implicit features, which is a considerable proportion. Although important, researches on the identification of implicit features are relatively few.

In this paper, we mainly focus on the identification of implicit features. We propose a novel classification-based approach to deal with the problem of implicit feature identification. Our approach is consisted of three main steps. The main purpose of the first step is to extract explicit feature-opinion pairs from customer reviews, we propose a rule based method to achieve this goal. In the second step, considering the situation that some different feature words are referred to the same feature, we cluster these feature-opinion pairs for each opinion word. Then we construct the training document for each clustered feature-opinion pair by collecting sentences labeled by the feature-opinion pair from customer reviews. In the third step, as the feature-opinion pair can be regarded as the sentence's topic or category, thus we formulate the identification of implicit feature into a text classification problem.

Our approach is very different from existing research works. The approach that former researches used is based on association rule mining. The core idea of this method is to use mined association rules to identify the implicit feature by finding the mapping of a specific feature for the opinion word. Although the association rule based approach is very useful and effective to identify the implicit feature for some kinds of opinion words that have relatively certain collocated features, for example, the opinion word "便宜" (cheap) is always used to describe the product's feature "价格" (price). But it fails to deal with many complex situations, for example, the opinion word "好"(good) are often used to describe a lot of features, such as "信号" (signal), "屏幕" (screen) and "摄像头" (camera), by using the mined rules it can only map the opinion word "好" (good) to a specific feature, which is not correct for many other different situations. By considering both the associated relation and the context of the opinion word, our classification-based approach is able to identify different implicit feature for the opinion word with different using situations. What's more, the rule-based approach usually need to set threshold parameters to prune rules, while our approach dose not have such trivial settings.

The remaining parts of this paper are organized as follows. In section 2, we introduce some related works. We introduce the details of our proposed classification-based approach in section 3. In section 4, the experimental results are evaluated and discussed. We present our conclusions and future work in section 5.

2 Related Work

Opinion mining has been extensively studied by many researchers in recent years. Most of these researches have focused on two main research directions: one is sentiment classification and the other direction is feature-based information extraction. Research efforts[10, 17, 11, 14] on sentiment classification deal with the task of classifying each customer review as positive, negative and neutral. While feature-based opinion mining[4, 6, 5, 8] focused on the task of extracting opinions consisting of information about features. In contrast to sentiment classification, opinion extraction aims at producing richer information and requires an in-depth analysis of reviews. The most representative researches in feature-level opinion mining are Hu and Liu's works[4, 5, 9]. The conception of implicit feature was first mentioned in their papers based on the analysis of English reviews. In contrast to explicit feature that directly appears in review sentences, implicit feature is the feature that does not occur in the comment, but can be deduced from opinion words and contexts based on the understanding of human language.

In [15], they attempted to infer the implicit features by using Point-wise Mutual Information(PMI) based semantic association analysis. They predefined a domain-specific feature set as candidate implicit features, and then take the mutual information approach to map a opinion indicator to a certain feature of the feature set. In [16], a clustering method was proposed to map implicit opinion words to their corresponding explicit features. They clustered product features and opinion words simultaneously and iteratively by fusing both their content information and association relation, and then construct the sentiment association set between the groups of features and opinion words by identifying their strongest n sentiment links.

In [3], a co-occurrence association rule mining (coAR) approach was proposed to identify implicit features. They firstly mined a set of association rules of the form

[opinion-word, explicit-feature] from review sentences based on the co-occurrence of the opinion-word and explicit-feature. And then they cluster the explicit features to generate more robust rules. When given a new opinion word with no explicit feature, they searched a matched list of rules, among which the rule with the highest frequency weight is fired to map the opinion word to its identified implicit feature. In [18], they proposed a hybrid association rule mining approach for the task of implicit feature identification. Their approach used several complementary algorithms to mine as many association rules as possible. They firstly extract candidate feature indicators and then compute the co-occurrence degree between the candidate feature indicators and the feature words. Each indicator and the corresponding feature word constitute a rule(feature indicator $\rightarrow$ feature word). They used such rules to identify implicit features.

3 Classification-based Approach

In this section, we first illustrate the problem of implicit feature identification and present some definitions we have used in this paper. The framework of our proposed classification-based approach has also been presented. Then we explain the main steps of our approach in detail.

3.1 Problem Statement

In this paper, we focus on the feature-level opinion mining of product reviews. On the online shopping websites, such as Amazon or Taobao Marketplace, each product can receive a large number of customer reviews that have been posted by people who have bought this product. The set of products can be represented as $P = \{P_1, P_2, P_3, ..., P_n\}$. For each product P_i, there is a set of customer reviews $R_i = \{r_1, r_2, r_3, ..., r_m\}$. The customer reviews can be regarded as text documents, although some of them may be very short and consisted of just a few sentences, but there are also many long reviews that can be as long as articles. We represent each review r_j as a sequence of sentences $r_j = \{s_1, s_2, s_3, ..., s_l\}$. Each sentence s_k may be consisted of several clauses $s_k = \{c_1, c_2, c_3, ..., c_h\}$.

Definition 1 ***implicit feature****:*

A product feature f is defined as the whole product, service or the attribute and component of the product. If a feature f appears in review sentences, then it is defined as *explicit feature*. If f does not appear in review sentences, but it is implied, which means that people who read the review can understand what feature has been talked about, then this feature f is regarded as *implicit feature*.

Definition 2 ***implicit sentence****:*

Implicit sentence is a sentence in a review that contains at least one implicit feature. *Explicit sentence* is defined similarly, a sentence that contains at least one explicit feature is called *explicit sentence*. It should be noted that a implicit sentence can also be a explicit sentence.

Definition 3 ***feature-opinion pair****:*

A feature-opinion pair is consisted of a feature and an opinion word, and the opinion word is used to modify the feature. If opinion word and its modified feature co-occur in a sentence, then such feature-opinion pair is defined as the sentence's *explicit feature-opinion pair*. The feature-opinion pair is denoted as $< feature, opinion >$.

3.2 Overview of the Approach

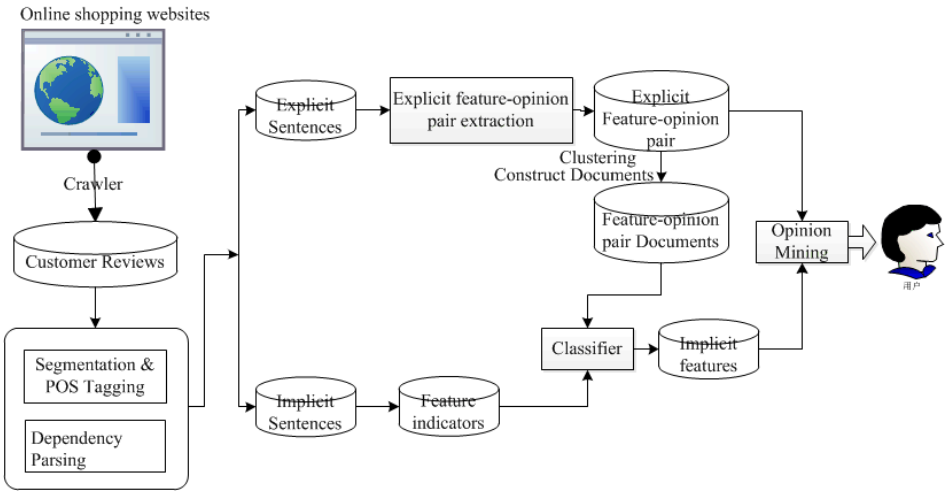

Fig. 1. Framework of Our Approach

In this subsection, we present an overview of our approach, including the flowchart of the approach and the introduction of the main steps, and then explain each step in details. As it has been shown in Fig. 1, it takes the corpus of customer reviews that have been crawled from the online shopping websties as input, and generates the opinion mining summary consisting of both explicit feature-opinion pairs and implicit feature-opinion pairs as the output.

Our approach is consisted of three main steps, including the explicit feature-opinion pair extraction, feature-opinion pair training document construction and implicit feature identification. The detail of each step is described in the following subsections.

3.3 Explicit feature-opinion pair Extraction

There are many existing research works on feature-level opinion mining that are dedicated to extract explicit feature-opinion pairs from customer reviews. Some supervised learning models, such as HMM(Hidden Markov Model) model and CRF(Conditional Random Fields) model, and topic models, such as the MaxEnt-LDA (a Maximum Entropy and LDA combination)[23] hybrid model, are widely used in this task. In this paper, we propose a rule based method to extract feature-opinion pairs from review sentences.

Rule Based Method. This method exploits Chinese dependency grammar to extract feature-opinion pairs. Firstly, we use Chinese dependency grammar to set several rules. Then we make use of these rules to extract candidate feature-opinion pairs. In order to improve the precision of feature-opinion pair extraction, we construct the candidate feature word set $\mathcal{CF}$ and the candidate opinion word set $\mathcal{CO}$ for each product.

After analyzing and studying the dependency parsing results of review sentences, we find that most of the discussed features are in subject-predicate(SBV) structure or DE (“的”) structure. Therefore, we mainly exploit the two kind of dependency

relation as rules to extract feature-opinion pairs. According to our observation, when the feature word appears before the opinion word, the feature usually satisfy the SBV relation with the opinion word, and when the feature appears after the opinion word, there usually exist a DE structure between the feature and the opinion word. Based on the above observations, we define three different rules to tackle different types of sentence structures to extract the explicit feature-opinion pairs. A summarized representation of these rules is presented in the following paragraphs.

Rule-1: In a dependency relation SBV, the dependency structure is denoted as $sbv(w_1, w_2)$, which means that word w_2 depends on word w_1 through SBV, if word w_2 belongs to the opinion word set $\mathcal{CO}$ and word w_1 belongs to the feature word set $\mathcal{CF}$, then $< w_1, w_2 >$ can be extracted as feature-opinion pair.

Rule-2: In a dependency relation SBV, the dependency structure is denoted as $sbv(w_1, w_2)$, which means that word w_2 depends on word w_1 through SBV, if word w_1 belongs to the feature word set $\mathcal{CF}$ and word w_2 doesn't belong to the opinion word set $\mathcal{CO}$, and after word w_2 exist a word w_3 that belongs to the opinion word set $\mathcal{CO}$, then $< w_1, w_3 >$ can be extracted as feature-opinion pair.

Rule-3: In a dependency relation DE, if exist a word w_1 belongs to the opinion word set $\mathcal{CO}$ before the word "的" and a word w_2 belongs to the feature word set $\mathcal{CF}$ after the word "的", then $< w_2, w_1 >$ can be extracted as feature-opinion pair.

The process of the rule based method is described in algorithm 1.

Algorithm 1 Rule based algorithm

Input: Review sentences in the corpus.
Output: A set of feature-opinion pair.

```
for each sentence s_k in corpus do
    for each rule r_i in the rule set R do
        if match the rule r_i then
            extract the feature-opinion pair fo =< f, o >, put fo into the sentence s_k's feature-opinion pair set FO_k
        end if
    end for
end for
return the set of feature-opinion pair.
```

3.4 Feature-opinion pair Training Document Construction

If we regard each explicit sentence as a training text, then the topic or category of this sentence can be labeled as the sentence's feature-opinion pair. For example, for the explicit sentence "送货很快，早上下单下午就送到了！" (Delivering is very fast, order in the morning and have received in the afternoon!), the feature-opinion pair <"送货", "快"> (<delivering, fast>) can be viewed as the sentence's labeled topic. If a sentence s_k contains more than one feature-opinion pair (FO_k denotes the sentence s_k's feature-opinion pair set), then the sentence can be classified into each feature-opinion pair topic of FO_k.

Feature-opinion pair Clustering. For each opinion word, there are usually more one feature-opinion pair that contains the opinion word. For a feature-opinion pair $< f, o >$, it means that the opinion word o is used to describe the feature word f. In review sentences, a opinion word generally can be used to describe several different features. For example, the opinion word "好" (good) is often used to describe a lot

of product features, such as "手机"(mobile phone), "屏幕" (screen), or "质量" (quality). In product reviews, many different feature words or phrases may be used to express the same feature. For example, features "音质" (vocality quality), "音乐" (music) and "音效" (sound effect) are all related to the same product feature "声音" (vocality). So for each opinion word o, we cluster the feature-opinion pairs $\mathcal{FO}(o) = \{< f_1, o >, < f_2, o >, ..., < f_n, o >\}$ that contains the opinion word based on the conceptual and semantical relation of these features $\mathcal{F}(o) = \{f_1, f_2, ..., f_n\}$. Our clustering method is based on [22], we mainly exploit the sharing words and the lexical similarity to cluster features. The size of the feature set $\mathcal{F}(o)$ is relatively small compared with the whole set of features $\mathcal{F}$, so it is much easier and more effective to the clustering of feature-opinion pairs.

After getting the set of clustered feature-opinion pair, we construct the training document for each clustered feature-opinion pair. For each clustered feature-opinion pair, we collect the sentences that contain the feature-opinion pair into a document, which is labeled by the clustered feature-opinion pair. The document constructing process is presented in algorithm 2.

Algorithm 2 Document construction

Input: The set of feature-opinion pair of explicit sentences.
Output: A set of feature-opinion pair document.

for each opinion word o_i in the set of opinion word O **do**
  get the $\mathcal{FO}(o) = \{< f_1, o >, < f_2, o >, ..., < f_n, o >\}$
  cluster the feature-opinion pairs
  get the clustered feature-opinion pairs $\mathcal{FO}_c(o_i) = \{< f_{c1}, o_i >, < f_{c2}, o_i >, ..., < f_{cm}, o_i >\}$
end for
for each explicit sentence s_k in corpus **do**
  for each feature-opinion pair $< f_i, o_i >$ in FO_k **do**
    put the sentence into the document $d(f_{ci}o_i)$ of the clustered feature-opinion pair $f_{ci}o_i =< f_{ci}, o_i >$ that includes $< f_i, o_i >$
  end for
end for
return the set of clustered feature-opinion pair document $\mathcal{D}$.

3.5 Implicit Feature Identification

By constructing the feature-opinion pair training set, we formulate the problem of identifying implicit features into a text classification problem. Thus many existing text classification approaches can be used to solve this problem. In this step, we mainly deal with the implicit sentences. For each implicit sentence $\mathcal{I}s_k$, the set of opinion word can be denoted as $\mathcal{I}o_k = \{o_1, o_2, ..., o_n\}$. The task of implicit feature identification is to find the implicit feature f_i for each opinion word o_i in $\mathcal{I}o$. The set of clustered feature-opinion pair that contains opinion word o_i can be denoted as $\mathcal{FO}_c(o_i) = \{< f_{c1}, o_i >, < f_{c2}, o_i >, ..., < f_{cm}, o_i >\}$. The key issue of this problem is to find the feature f_{ci} that the opinion word o_i in implicit sentence $\mathcal{I}s_k$ has modified from the feature set $\mathcal{F}_c(o_i) = \{f_{c1}, f_{c2}, ..., f_{cm}\}$. As the feature-opinion pair can be regarded as the sentence's topic or category, thus the problem of finding the implicit feature fc_i for opinion word o_i in implicit sentence $\mathcal{I}s_k$ has been transformed into a text classification problem.

In order to classify the implicit sentence with a opinion word o_i into the most probable feature-opinion pair $< f_i, o_i >$ topic, we design a topic-feature-centroid

classifier based on [2]. We modify the centroid construction and classification process to accommodate the situation in this problem.

Topic-feature-centroid construction: Different from the centroid-based approaches in [2] that uses all the words in the corpse to form the lexicon set, we use only a small set of feature-related discriminative words in the training set to construct the lexicon set. For instance, considering the collected training set of feature-opinion pair <"送货", "快"> (<delivery, fast>), there are over two hundred different words in this topic, while only a very small number of words contribute to the characteristic space of this topic, such as word "速度" (speed), "下单" (order) and so on. Many other words, such as "很" (very), "是" (is), "有" (have) and so on, even some of them with a very high frequency, hardly have any discrimination for this topic. Moreover, such irrelevant words could bring on a lot of noise in the representation of the topic. Therefore, in the construction of the lexicon set, we only consider nouns, adjectives and verbs in the training set, and we also construct a filter word set to remove the stop words and irrelevant words. The constructed lexicon set is denoted as $\mathcal{L} = \{wf_1, wf_2, ..., wf_L\}$, so the centroid for category $< f_j, o_j >$ can be represented by a word vector $Centroid_j = \{wf_{1j}, wf_{2j}, ..., wf_{Lj}\}$, where $wf_{kj}(1 \leq k \leq L)$ represents the weight for word wf_k.

In our topic-feature-centroid classifier, we derive a different formulation for the calculation of the weight for word wf_k. The weight for word wf_k of topic $< f_j, o_j >$ is calculated as following:

$$wf_{kj} = f_{w_k} \times \log(\frac{|C|}{|CF_{w_k}|}) \tag{1}$$

where f_{w_k} is the word w_k's frequency in the training document of feature-opinion pair topic $< f_j, o_j >$, $|c|$ is the total number of feature-opinion pair topics for the given opinion word o_j , $|CF_{w_k}|$ is the number of feature-opinion pair topics that contains the word w_k. When a word w_k occurs in every feature-opinion topic, the value of wf_{kj} is 0 because $\log(\frac{|C|}{|CF_{w_k}|})$ becomes 0, which means that word w_k has no discrimination for the topic. Thus our weight calculation method can produce more discriminative features for the feature-opinion topic.

Classification: After the centroid vector of each category is obtained, the implicit sentence is classified by using a denormalized cosine measure:

$$C' = arg\max_j(\overrightarrow{s_i} \bullet \overrightarrow{Centroid_j}) \tag{2}$$

where $\overrightarrow{s_i}$ is the word vector representation for the implicit sentence s_i, since the sentence is usually very short, so we only concern the word's appearance or not. We use the binary representation to denote the word's weight in $\overrightarrow{s_i}$. By using this denormalized cosine measure, it preserves the discriminative capability of feature-opinion pair topic's centroid vector.

The process of implicit feature identification is described in algorithm 3.

4 Experiments

In this section, we conducted several experiments and evaluate the performance of our approach. Firstly, we describe the data sets used in our experiments. Then we give the definition of several performance metrics. Lastly, we describe experiment

Algorithm 3 Implicit feature identification

Input: The set of implicit sentences.
Output: A set of implicit feature-opinion pair.
1: **for** each implicit sentence $\mathcal{I}s_k$ **do**
2: **for** each opinion word o_i in $\mathcal{I}o$ **do**
3: apply the topic-feature-centroid classifier for o_i
4: get the implicit feature fc_i
5: **end for**
6: **end for**
7: return the set of implicit feature-opinion pair.

results and corresponding analysis. Both the results of explicit feature-opinion pair extraction and implicit feature identification have been evaluated.

4.1 Date Sets and Evaluation Measures

Since there is no standard data set for our experiment, so we crawled the experiment data from the popular Chinese shopping website, Amazon.cn[1], the regional website of Amazon.com in China. Customer reviews are collected from two different domains: cell phone and digital camera. There are totally 4083 reviews and 12760 sentences in our data set. Both the explicit feature-opinion pair and implicit feature-opinion pair of each sentence are manually annotated by two research students in our lab. And to be fair, those sentences that are annotated inconsistently have been removed and the rest has been confirmed by the author. The details of the data sets are given in Table 1.

Table 1. Experiment Data

Data Sets	Reviews	Sentences	Explicit features	Implicit features
Cell Phone	2694	8305	4233	1449
Digital Camera	1389	4455	1817	798
Total	4083	12760	6050	2247

We use the traditional precision (P), and recall (R) and F-measure (F) to evaluate our experiment results of both explicit feature-opinion pair extraction and implicit feature identification. The F-measure is defined as follows:

$$F = \frac{2 * precision * recall}{precision + recall} \tag{3}$$

4.2 Evaluation of explicit feature-opinion pair Extraction

A important step of our approach is to extract the explicit feature-opinion pairs from explicit sentences. The construction of the training document is based on the result of the extracted explicit feature-opinion pairs. In this paper, we use LTP[2] to accomplish the Chinese word segmentation, part-of-speech(POS) tagging and dependency parsing.

Table 2 shows the result of explicit feature-opinion pair extraction by using our rule based method. As we can see from the table, our rule based method achieves

[1] `http://www.amazon.cn`
[2] `http://ir.hit.edu.cn/ltp/`

a comparatively satisfactory result in the extraction of feature-opinion pairs. In the construction of the candidate feature word set $\mathcal{CF}$ and the candidate opinion word set $\mathcal{CO}$ for each product, not only considering nouns as candidate features and adjectives as candidate opinion words, we also add some verbs into the feature set and opinion word set. For example, the verb "送货" (deliver) is frequently used as feature and the verb "喜欢" (like) is frequently used as opinion word in customer reviews. And we also construct a filter word list to remove many product-irrelevant nouns and adjectives from the candidate set, such as "朋友" (friends), "伤心" (sad) and so on.

Table 2. Result of explicit feature-opinion pair extraction

Data Sets	Precision	Recall	F-measure
Cell Phone	80.21%	79.99%	80.10%
Digital Camera	81.95%	83.43%	82.68%

4.3 Evaluation of implicit feature Identification

In the end, we give the final experimental results via using our proposed classification-based approach. Our classification-based approach is compare with the rule based approach coAR[3]. We implement the approach coAR proposed in [3]. The best results for each approach is listed in Table 3.

Table 3. Result of implicit feature identification

Data Sets	Our Approach			CoAR		
	Precision	Recall	F-measure	Precision	Recall	F-measure
Cell Phone	82.07%	68.48%	74.66%	67.88%	52.93%	59.48%
Digital Camera	85.59%	72.93%	78.76%	79.94%	66.91%	72.85%

CoAR mined the association rules of explicit feature-opinion pair based on the co-occurrence of opinion word and feature word, and then used the clustered rules to identify the implicit feature for a given opinion word.

As we can observe from Table 3, our approach outperforms coAR on all the evaluation metrics for corpora in both cell phone data set and digital camera data set. Co-AR used the mined association rules from the review corpus to identify the implicit feature for the given opinion word by rule matching, which means that their approach always map the opinion word to the same feature word without considering the context of the implicit sentence. While our classification-based approach not only exploit the association relations by extracting explicit feature-opinion pairs, but also take into account the context of the implicit sentence by using the category-feature-centroid classifier to map the opinion word in a specific implicit sentence to the most probable feature word. Thus our approach's precision is higher than coAR. In addition, our approach's recall is also higher than the coAR approach. This is because that the rule based coAR approach only adopted the association rules whose weight is greater than the threshold as robust rules. The higher threshold can weed out the lower-frequency association rules and promote the precision, but it would reduce the recall. No matter what threshold is selected, it can not capture

a significant number of uncommon association rules. This shortcoming of the rule based approach determines that the recall of coAR is limited to a certain extent.

5 Conclusion and Feature Work

In this paper, we propose a novel classification-based approach to deal with the problem of implicit feature identification. By constructing the document for the clustered feature-opinion pair, we can obtain the training document that has been labeled by the specific clustered feature-opinion pair. Then we formulate the problem of implicit feature identification into a text classification problem. We propose a rule based method to extract explicit feature-opinion pairs from customer reviews. In the phase of implicit feature identification, we design a topic-feature-centroid classifier to perform the classification task. It should be pointed out that other feature-opinion pair extraction methods and text classification methods can also be used in our approach. Compared with the rule based approach coAR, our approach overcomes the shortcomings and limitations of the rule based approach and achieves a much better performance on all the measure metrics. However, some undesirable errors still exist in the result of implicit feature identification. Some are caused by the incorrect classification, some are caused by the wrong identification of implicit feature indicators. In our future work, we will explore the performance of approach by using several other text classification approaches in the implicit feature identification step.

References

1. Sasha Blair-Goldensohn, Kerry Hannan, Ryan McDonald, Tyler Neylon, George A Reis, and Jeff Reynar. Building a sentiment summarizer for local service reviews. In *WWW Workshop on NLP in the Information Explosion Era*, 2008.
2. Hu Guan, Jingyu Zhou, and Minyi Guo. A class-feature-centroid classifier for text categorization. In *Proceedings of the 18th international conference on World wide web*, pages 201–210. ACM, 2009.
3. Zhen Hai, Kuiyu Chang, and Jung-jae Kim. Implicit feature identification via co-occurrence association rule mining. In *Computational Linguistics and Intelligent Text Processing*, pages 393–404. Springer, 2011.
4. Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM, 2004.
5. Minqing Hu and Bing Liu. Mining opinion features in customer reviews. In *Proceedings of the National Conference on Artificial Intelligence*, pages 755–760. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2004.
6. Minqing Hu and Bing Liu. Opinion feature extraction using class sequential rules. In *AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs, Palo Alto, USA*, 2006.
7. Niklas Jakob and Iryna Gurevych. Extracting opinion targets in a single-and cross-domain setting with conditional random fields. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1035–1045. Association for Computational Linguistics, 2010.
8. Nozomi Kobayashi, Kentaro Inui, and Yuji Matsumoto. Extracting aspect-evaluation and aspect-of relations in opinion mining. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1065–1074, 2007.
9. Bing Liu, Minqing Hu, and Junsheng Cheng. Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web*, pages 342–351. ACM, 2005.
10. Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 115–124. Association for Computational Linguistics, 2005.

11. Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.
12. Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics, 2002.
13. Ana-Maria Popescu and Orena Etzioni. Extracting product features and opinions from reviews. In *Natural language processing and text mining*, pages 9–28. Springer, 2007.
14. Ellen Riloff and Janyce Wiebe. Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 105–112. Association for Computational Linguistics, 2003.
15. Qi Su, Kun Xiang, Houfeng Wang, Bin Sun, and Shiwen Yu. Using pointwise mutual information to identify implicit features in customer reviews. In *Computer Processing of Oriental Languages. Beyond the Orient: The Research Challenges Ahead*, pages 22–30. Springer, 2006.
16. Qi Su, Xinying Xu, Honglei Guo, Zhili Guo, Xian Wu, Xiaoxun Zhang, Bin Swen, and Zhong Su. Hidden sentiment association in chinese web opinion mining. In *Proceedings of the 17th international conference on World Wide Web*, pages 959–968. ACM, 2008.
17. Peter D Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 417–424. Association for Computational Linguistics, 2002.
18. Wei Wang, Hua Xu, and Wei Wan. Implicit feature identification via hybrid association rule mining. *Expert Systems with Applications*, 2012.
19. Janyce M Wiebe, Rebecca F Bruce, and Thomas P O'Hara. Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 246–253. Association for Computational Linguistics, 1999.
20. Theresa Wilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. Opinionfinder: A system for subjectivity analysis. In *Proceedings of HLT/EMNLP on Interactive Demonstrations*, pages 34–35. Association for Computational Linguistics, 2005.
21. Theresa Wilson, Janyce Wiebe, and Rebecca Hwa. Just how mad are you? finding strong and weak opinion clauses. In *Proceedings of the National Conference on Artificial Intelligence*, pages 761–769. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2004.
22. Zhongwu Zhai, Bing Liu, Hua Xu, and Peifa Jia. Clustering product features for opinion mining. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 347–354. ACM, 2011.
23. Wayne Xin Zhao, Jing Jiang, Hongfei Yan, and Xiaoming Li. Jointly modeling aspects and opinions with a maxent-lda hybrid. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 56–65. Association for Computational Linguistics, 2010.
24. Li Zhuang, Feng Jing, and Xiao-Yan Zhu. Movie review mining and summarization. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 43–50. ACM, 2006.