Fine-grained Product Feature Extraction in Chinese Reviews

Hanqian WU
School of Computer science and technology
Southeast University
Nanjing, China
E-mail: hanqian@seu.edu.cn

Tao LIU
School of Computer science and technology
Southeast University
Nanjing, China
E-mail: liutao_seu@163.com

Jue XIE
School of Information Technology
Southeast University-Monash University Joint Graduate School
Suzhou, China
E-mail: jue.xie@monash.edu

***Abstract*—Fine-grained product feature extraction is the most important task in opinion mining. To realize the fine-grained product feature extraction in Chinese reviews, three main tasks have been solved in this paper. Firstly, we propose a dependency parsing based method to directly extract the explicit feature-opinion pairs. Then, by analyzing the characteristics of two synonyms features and the relations with opinion words, we calculate the similarities to cluster features. Finally, we propose a novel implicit feature extraction method by combining review context information and two kind opinions to extract implicit features. Experiments show that the dependency parsing based method can get high precision, by considering verbs as product feature can improve the recall obviously. Besides, several proven pruning strategies can improve the accuracy. The comparison demonstrates that our implicit feature extraction method outperforms existing method, and feature clustering before implicit feature mining can get better results.**

Keywords-dependency parsing; explicit feature; feature clustering; implicit feature

I. INTRODUCTION

With the development of Internet, online shopping has become a common way of consumption. The number of online reviews on the electronic business platform is increasing with the huge amount of trading orders. These most intuitive user data are of great potential values. How to automatically extract the useful product features from the massive user comments has become a popular research field, a lot of research efforts have been down on opinion mining.

Previous researches mainly focus on document or sentence level opinion mining, but it is necessary to get the fine-grained product features with the associated opinions. The main task of feature-based opinion mining is feature extraction, both explicit feature and implicit feature. In a given review, feature that appears in the comments is explicit feature. For example, the feature "价格"(price) is an explicit feature in sentence "电脑的价格便宜"(The price of the computer is cheap). In contrast, a feature that does not appear in review sentence but can be implied by some indicator words is called implicit feature. For example, in sentence"很便宜，值得推荐" (Very cheap, worth recommending), the indicator word"便宜" (cheap) implies the implicit feature "价格"(price). Besides, a feature can be expressed in different words in reviews. For example, "外观", "外表"and "样子"mean the same feature appearance. Therefore, it is an important task to cluster the synonymous features.

In fine-grained feature based opinion mining, existing feature extraction approaches mainly use language rules, sequence models or topic models. Each approach has advantages, but for the purpose of directly extracts product features with its associated opinion, we construct several language rules based on dependency parsing to extract explicit features and related opinions. Because the lack of standard lexicon in Chinese, in order to perform feature clustering we calculate the similarities between two features in literal content and the modified relationship among feature and opinions. In implicit feature extraction, based on the result of clustering, we propose a method combine the context information and two kind opinions to extract the implicit features.

II. RELATED WORK

In explicit feature mining, Hu and Liu [1] firstly use the association rule method to extract frequent words as product features. This method is simple and useful, but only extracts frequency nouns and noun phrase as features is incomplete. Infrequent words and some verbs may also be product features. Besides, this method ignores implicit feature. Popescu and Etzioni [2] improve the method in [1] by calculate the Point-wise Mutual Information (PMI) score between the phrase and class discriminators to judge whether the phrase is a product feature or not. However, this method computes the PMI by searching on the Internet is time-consuming. By using the modifying relationship of opinion words and features, Zhuang, Jing and Zhu [3] first employ the dependency relation to extract feature-opinion pairs from movie reviews. Qiu, Liu, Bu and Chen [4] have proposed a double propagation method, which extracts certain syntactic

978-1-5386-3886-6/17 $31.00 © 2017 IEEE
DOI 10.1109/CIIS.2017.53

relations between opinion words and features by using dependency grammar. But it only extracts noun features and for large corpora, this method will introduce a lot of noise data, for small corpora, it will miss certain features. Besides, sequence models like Hidden Markov Models (HMM) and Conditional Random Fields (CRF), topic models like Latent Dirichlet Allocation (LDA) are used to extract product features. However, when corpus is small, these statistical methods may not be reliable. Topical model has difficulty in finding fine-grained features, and needs extra operations to get associated opinions. In this paper, we build several association rules using the dependency parsing tools in Chinese to conduct a more complete explicit feature extraction.

In feature clustering, Huang, Liu and Peng [5] have proposed a feature clustering methods in English domain, they calculate the feature similarities based on WordNet. But in Chinese, there has no authoritative dictionary. In Chinese, Yang, Ma and Lin [6] compute the similarity between different feature expressions in feature-opinion bipartite graph, and then adopt a Bayesian classifier to get better classification result. Xi [7] introduces some must-links and cannot-links, and has proposed a constrained hierarchical clustering algorithm to accomplish feature clustering. Our feature clustering method combines several strategies in their methods to realize clustering.

In implicit feature mining, Su et al. [8] propose a novel mutual reinforcement approach that exploits the hidden sentiment association between product feature category and opinion word group to extract implicit features. Hai, Chang and Kim [9] propose a novel two-phase co-occurrence association rule mining approach to identifying implicit features. They first extract a set of rules from explicit feature and opinion matrix, and then cluster the explicit features to get more robust rules. Our implicit feature mining method also clusters explicit features. Liu, Lv and Wang [10] propose a method using opinions and candidate features to calculate important score to identify the implicit features, the opinion words are divided into two categories, vague opinion words and clear opinion words. Zhang and Zhu [11] not only concentrate on the associations between feature words and opinion words, but also utilize the associations between feature words and the rest of the notional words in the clause. Our method is similar to this method, but we cluster the explicit feature and we use some ideas in [10]. Schouten and Frasincar [12] propose a novel implicit feature mining method by directly constructing the implicit feature and the notional words co-occurrence matrix. This method will greatly reduce the size of co-occurrence matrix and better performance in implicit feature mining. But this method is domain dependent and need manually annotated of implicit features.

In this paper, we propose a dependency parsing method to extract explicit feature, and then we calculate the similarities of features to conduct feature clustering, finally we use the review content information and consider the opinion types to realize implicit feature mining.

III. METHOD

A. Explicit Feature Extraction Based on Dependency Parsing

According to the analysis of dependency parsing for review sentence, we can straight extract the explicit feature and the associated opinion. In this paper, we construct a set of rules to extract explicit feature-opinion pairs in Table I.

TABLE I. EXPLICIT FEATURE EXTRACTION RULES

	Rules	Feature	Opinion
1	$n \xleftarrow{SBV} (d \xleftarrow{ADV}) a$	n	a
2	$n \xleftarrow{ATT} n \xleftarrow{SBV} (d \xleftarrow{ADV}) a$	n+n	a
3	$v \xleftarrow{ATT} n \xleftarrow{SBV} (d \xleftarrow{ADV}) a$	v+n	a
4	$a \xleftarrow{ATT} n$	n	a
5	$v \xrightarrow{CMP} a$	v	a

In Table I, we not only extract nouns and noun phrase as features but also extract verb and verb+noun phrase as fine-grained features. The dependency relation of ADV may exist in the review sentence or may not exist.

In order to improve the precision of explicit feature extraction, we adopt several pruning strategies. Firstly, we remove the stop words in review sentence. Besides, we use the common sense in Chinese, single Chinese character cannot describe a product feature. Lastly, we remove some common verbs such as "没有"(without), "应该"(should). For the purpose of removing the noise data, we simply apply a threshold to remove low frequency noise data. Experiment show that these strategies are effective.

Finally, we get the explicit feature opinion co-occurrence matrix C.

B. Feature Clustering

Due to the limitation of Chinese thesaurus, we not use it like other scholars. The feature clustering method we adopted is based on the similarities computing among features.

1) Sharing words similarity

Feature expressions sharing some words are likely to belong to the same cluster. We introduce a similarity calculated method based on sharing words.

$$Sim_{sharing}(f_i, f_j) = \frac{Num_{sharing}}{Num_i + Num_j - 2Num_{sharing}} \quad (1)$$

Where Num_i and Num_j represent the number of words feature f_i and f_j contained, and $Num_{sharing}$ represents the number of words they both contained.

2) Associated opinions similarity

Features that belong to a same cluster have a high concordance with their associated opinions. Based on the explicit feature opinion co-occurrence matrix *C* that we get

from explicit feature extraction, we calculate the similarity by Cosine distance of two features.

$$Sim_{opinion}(f_i, f_j) = C\cos ine(v_i, v_j) = \frac{v_i \cdot v_j}{\| v_i \| \cdot \| v_j \|} \quad (2)$$

Where v_i and v_j represent the opinion vector of feature f_i and f_j in the matrix C.

Considering these two factors, the similarity of features is calculated in (3).

$$Sim(f_i, f_j) = \alpha Sim_{sharing}(f_i, f_j) + \beta Sim_{opinion}(f_i, f_j) \quad (3)$$

$\alpha + \beta = 1$. After repeatedly experiment, the values of them are 0.6 and 0.4.

3) A clustering constraint rule

To our prior knowledge and observation, features appeared in the same product review cannot belong to the same cluster. We introduce this constraint rule to improve feature clustering. If two features appear in a sentence, the similarity of them is zero.

4) Clustering step

To cluster the high similarities features into groups, we construct a weighted undirected graph $G = (V, E)$. V is the product feature set. Each feature represents a graph node. E represents the weight edges between two features, calculated by (3).

Algorithm 1:feature clustering

Input:product feature $Set = \{f_1, f_2, ...f_n\}$ *,feature opinion co-occurrence matrix C,end threshold* e

Output: $\mathrm{Re}sultSet = \{Set_1, Set_2, ...\}$

Step:

1. Calculate the feature similarity matrix M among features for Set.

2. Select the MAX $Sim(f_i, f_j)$ *in M, if MAX >=* e *,go Step 3, else go Step 4.*

3. Cluster f_i *and* f_j *:merge the sets contain* f_i *or* f_j *in* $\mathrm{Re}sultSet$ *,if the set is absent, then created. Select the one with the high frequency as the center of the cluster. Merge the associated opinion in C. Delete the feature with low frequency, update C. return step 1*

4. Put the remaining features into $\mathrm{Re}sultSet$ *, feature itself is a cluster.*

5. Return $\mathrm{Re}sultSet$ *and exit*

C. Implicit Feature Extraction

In order to extract product features more completely, it is necessary to consider the implicit features. People like to describe the familiar product features with feature indicators. Previous researches mainly use opinion words as feature indicators, but it failed to conclude the vague opinions in implicit sentence. For example, in implicit sentence “这电脑看着不错” (The computer looks good), the implicit feature is “外观”(appearance) , but we cannot conclude it by the vague opinion word “不错”(good), the verb “看”(look) can help conclude the implicit. In this paper, we not only consider the two kinds of opinions but also use review context information to extract implicit feature.

Our method includes follow steps:

Step 1: Construct candidate feature context information matrix

For every review sentence, we extract the context information for every explicit feature and merge the information according to feature cluster. Then we construct a candidate feature context information matrix S, row is the center of a feature cluster, columns are the context information. Fig.1 shows a small example of matrix S.

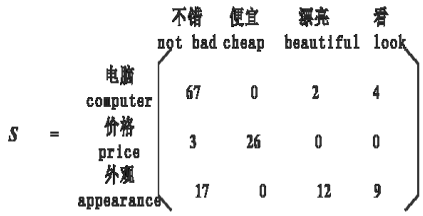

Figure 1. Candidate feature context information matrix.

Step 2: several statuses of implicit features

Implicit features in review sentence have different status, previous work almost ignores these differences and the implicit feature extract is incomplete. The common status of implicit is as follows:

A review sentence has only one implicit feature. Besides, it is common to see an implicit feature appears with some explicit feature. And there may appear multiple implicit features in one sentence. For example, “这个电脑买了好几台了，便宜好用” (this computer has bought several times, cheap and easy to use), the implicit feature are “价格”(price) and “使用”(use).

In this paper, we try to consider these statuses and extract the implicit feature as many as possible.

Step 3: implicit feature extraction method

During the implicit feature extraction process, we ignore the explicit features and associated opinions. We find the implicit sentences by opinion words.

- If a sentence only contains vague opinions, no other verbs or nouns, we use the product itself as implicit.
- If the sentence only contains clear opinions, no other verbs or nouns, we use the clustered feature opinion co-occurrence matrix to extract the implicit for every clear opinion. In which, we select the one with the max $con(f_i)$ in the matrix as implicit feature. This

strategy can extract multiple implicit features and also reduce the extract time.

$$con(f_i) = n_{fo} / n_{f_i} \quad (4)$$

Where n_{fo} is the weight of candidate feature f_i and clear opinion in the clustered feature opinion co-occurrence matrix. And n_{f_i} represents the total number of opinions co-occurrence with candidate feature f_i.

- Otherwise, we use the implicit feature extraction calculate method in [11]. Given a review sentence R contains implicit indicators, we extract the context information, denoted as set $W = \{w_1, w_2, ..., w_v\}$. v is the number of the word in R. For each candidate feature f_i in S, we calculate the occurrence probability between candidate feature and W. We choose the one with the max $T(f_i)$ value as the implicit feature.

$$T(f_i) = \sum_{j=1}^{v} P(f_i \mid w_j) / v \quad (5)$$

$$P(f_i \mid w_j) = \frac{n_c}{n_{w_j}} \quad (6)$$

n_{w_j} represents the times word w_j appears in whole reviews. n_c represents the times word f_i and w_j appear together in a sentence.

IV. EXPERIMENTS

This section evaluates the effectiveness of our fine-grained product feature extraction method.

A. Data Sets

In Chinese review opinion mining, there is no standard open experiment data. We capture the data from a famous commercial web (www.jd.com) in China. After data cleaning, we select 1500 ThinkPad computer reviews and 1500 Huawei mobile phone reviews as experiment data. We ask three students to label the data in fine-grained.

B. Evaluation Metrics

We use Precision, Recall and F-measure as evaluation metrics for both explicit feature and implicit feature mining.

We do not evaluate the feature clustering task because we do not set the cluster number when feature clustering, one cluster only contains several features, but we can see that feature clustering can improve the implicit feature extraction.

C. Experimental Results and Analyses

Firstly, we carry on the explicit feature extraction based on dependency parsing. In this paper, we compare our method with different pruning strategies. The average results of the two kind data are presented in Table II. Test1 only filters the stop words, Test2 continue to filter the single word features, Test3 adds verb extraction task. Finally, Test4 sets a threshold to filter noise features. The threshold is 2.

From Table II, we can see the proposed explicit method is effective, especially add verbs as product features can improve the Recall obviously from 0.63 to 0.77.

For implicit feature mining, we compare our method with Liu, Lv and Wang [10] and Zhang and Zhu [11] in Table III. Our method learns from [10] and [11], experiment results show the use of context information in Test5 is better than paper [10], which only use opinion words to identify implicit feature. Besides, we consider the two kinds of opinions and try to extract multiple implicit features in one sentence help improve the Recall compared with paper [11]. After cluster the features, Test6 performs better than no cluster Test5, the main reason dues to the aggregation of co-occurrence weight between features and indicator words enhance the confidence to conclude correct implicit features.

TABLE II. EXPLICIT FEATURE EXTRACTION RESULT

	Precision	Recall	F-measure
Test1	0.74	0.64	0.69
Test2	0.76	0.63	0.69
Test3	0.72	0.77	0.74
Test4	0.86	0.69	0.77

TABLE III. IMPLICIT FEATURE EXTRACTION RESULT

	Precision	Recall	F-measure
paper[10]	0.69	0.62	0.65
paper [11]	0.76	0.72	0.74
Test5 (no cluster)	0.78	0.77	0.77
Test6 (cluster)	0.81	0.76	0.78

V. CONCLUSIONS

In review opinion mining, fine-grained product feature extraction is the most significant task. This paper dedicates to accomplish the three main task of feature extraction in Chinese reviews. We construct a set of rules based on dependency parsing to extract explicit features. Several pruning strategies have been proven effectively. Besides, we cluster the features by calculate the similarities between features. Finally, we combine context information and two kinds of opinion words to extract implicit features.

However, our explicit feature extraction method depends on dependency parsing, it will be affected by the precision of dependency parsing tool. The implicit feature extraction

needs to establish two kind opinion dictionaries. Besides, we use the opinion word to judge whether a sentence includes implicit feature or not, but some sentence have no opinion words can also imply some implicit features. In the future, we will care more about how to extract implicit features in no opinion word sentences.

ACKNOWLEDGMENT

This work is supported by National High-tech R&D Program of China (863 Program) under Grant No.2015AA015904.

REFERENCES

[1] Minqing Hu and Bing Liu, "Mining opinion features in customer reviews," in *Proceedings of the National Conference on Artificial Intelligence*, pp. 755–760. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2004.

[2] Popescu, Ana-Maria and Oren, Etzioni. 2005, "Extracting product features and opinions from reviews," In Proceedings of EMNLP, 2005.

[3] Zhuang L, Jing F, Zhu X Y, "Movie review mining and summarization," ACM, Conference on Information and Knowledge Management. DBLP, 2006, pp. 43-50.

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

[5] Huang S, Liu X, Peng X, et al, "Fine-grained Product Features Extraction and Categorization in Reviews Opinion Mining," International Conference on Data Mining Workshops. IEEE, 2012,pp.680-686.

[6] Yuan Yang, Yunlong Ma, Hongfei Lin, "Clustering Product Features in Opinion Mining," Journal of Chinese Information Processing, vol.26(3),pp.104-109, May 2012.

[7] Yahui Xi, "Recognizing the Feature Synonyms in Product Review," Journal of Chinese Information Processing, vol.30(4), pp.150-158, July 2016.

[8] Su Q, Xu X, Guo H, et al, "Hidden Sentiment Association in Chinese Web Opinion Mining," International Conference on World Wide Web, WWW 2008, Beijing, China, April. DBLP, 2008,pp.959-968.

[9] Hai Z, Chang K, Kim J, "Implicit Feature Identification via Co-occurrence Association Rule Mining," International Conference on Computational Linguistics and Intelligent Text Processing. Springer-Verlag, 2011,pp.393-404.

[10] Liu, Lizhen, Zhixin Lv, and Hanshi Wang, "Extract Product Features in Chinese Web for Opinion Mining," Journal of Software (1796217X), vol. 8(3), pp. 627-632, March 2013.

[11] Zhang, Y., and Zhu, W, "Extracting implicit features in online customer reviews for opinion mining," International Conference on World Wide Web Companion , vol.30, pp.103-104, May 2013.

[12] Schouten K, Frasincar F, "Implicit Feature Extraction for Sentiment Analysis in Consumer Reviews," Natural Language Processing and Information Systems.2014.pp.228-231.