Extracting the Comparative Relations for Mobile Reviews

Shuo He Fang Yuan Yu Wang
College of Mathematics & Computer Science, Hebei University,
Baoding, Hebei, China
heshuo_1226@163.com, yuanfang@hbu.cn, wy@hbu.cn

***Abstract*—This paper devotes the research on the comparative reviews in the field of mobiles. For the comparative reviews, the study builds a comparative feature lexicon and a superlative feature lexicon, and then classifies the words in the comparative feature lexicon according to their polarities. During the extracting, an identifying method based on the naming characteristics is proposed for the comparative entities. For the comparative results, the study constructs 6 extracting patterns to apply to the comparative reviews. Finally, a formula is given to calculate the comparative result. The experimental results demonstrate the feasibility of this method.**

Keywords-comparative reviews; comparative feature lexicon; superlative feature lexicon; naming characteristics; extracting patterns

I. INTRODUCTION

Comparative sentences refer to the sentences which contain comparative formats or comparative vocabularies. As a common way of expression, comparative sentences can help individuals enunciate their views conveniently. A number of them are contained in product reviews. Users prefer to compare similar products of different brands, and make their own reviews. The main task of mining the comparative reviews is to extract the comparative entities from sentences, judge the opinion polarities, and then feedback the mining results to producers and users.

This paper studies the comparative relations in the mobile reviews. On the one hand, two lexicons are built in order to judge the polarities of reviews and identify the comparative type respectively. On the other hand, this paper defines the target of our extracting. Then, it extracts the relations according to the naming characteristics and the pattern matching method. The experimental results show that this method can get a satisfactory effect.

II. RELATED WORKS

By now, some scholars have made basic studies related to comparative semantics. Linguists have discussed comparative structures from the perspective of syntactic and semantics. It sums up 28 kinds of syntactic structures which are used to express comparison in Chinese in [1]. Reference [2] discusses the characteristics and denotations of the Chinese comparative sentences, and uses SVM classifier to divide Chinese sentence into two categories, "comparative" and "non-comparative". It has achieved effective classification results. However, there are limited studies on comparative relation mining of Chinese reviews. Reference [3] proposes the LSR (Label Sequential Rules) algorithm for extracting the comparative relations. This algorithm generates sequential data and label sequential rules based on the proper nouns of the reviews, and then mines the comparative relations by the rules. It assumes that there is only one relation in each review. But usually, the user may compare several products in the same review. What's more, the expressions in Chinese emphasis on semantics, and there is no fixed structure for expressions. So the mining algorithm of [3] has certain limitations. Huang Gaohui [4] defines a series of extraction rules for each comparative word and sequence pattern. He combines the CRF (Conditional Random Fields) and rules-based method together to complete the extraction of the two comparative entities. Then he builds a lexicon about the features for a particular field in order to extract the feature words. Though it has a higher precision and recall, the definition of the extraction rules and the establishment of lexicon require lots of human intervention which is not conducive to the realization of automated mining.

In this study, we set 5-items as the mining targets. After building a comparative relation lexicon and a superlative relation lexicon, a mining method for the 5-items is proposed. Experiments prove the validity of this method.

III. MINING THE COMPARATIVE RELATION

For comparative reviews, this paper builds a comparative relation lexicon and a superlative relation lexicon. The former one summarizes the words with comparative relations in Chinese, and then classifies them according to their polarities. The latter one sums up the words ("最", "唯一" etc.) with superlative relations, which are used to judge the superlative relation reviews. During the mining, first identify comparative entities with the help of the mobile-named characteristics and similarity calculation. Second, extract the comparative features according to the POS (Part-Of-Speech) tagging. Then, we propose 6 opinion-matching patterns for comparative reviews. Finally, utilize these patterns to mine opinions, and then compute the result value. By now, the process is completed and feedback the 5-items to users and producers.

A. Definition of 5-items in Comparative Reviews

Xu Ming [5] divides comparative sentences into six components, which are subject, object, comparative point, comparative words, comparative attributes and comparative dispersion. Though his study is detailed on sentence division, it doesn't realize the comparative type would also influence the meaning of sentence. So, we add the type identification into exacting targets. According to this, we define the reviews into subject, object, comparative attributes, comparative result value and comparative type five ingredients.

- Subject *P1*: the entity that the user wants to compare.
- Object *P2*: the entity that is used for comparison.

978-1-4577-1415-3/12/$26.00 ©2012 IEEE

- Comparative attributes *A*: the feature attributes of entities.
- Comparative result value *R*: the attitude of the user, which indicates the comparative result. In order to express the result more intuitively, the study references the symbol expressions proposed in [6]. If *P1* is better than *P2,* then mark *R* as ">"; If *P1* is worse than *P2,* then mark *R* as "<"; If *P1* and *P2* are similar, then use the symbol "=". When *R* is labeled as "=", it means this is an equative review.
- Comparative type *T*: types of comparison. Different comparative types convey different meanings. Linguist Ma Jianzhong classifies Chinese comparative sentences into three types, which are "superlative", "equative" and "non-equal gradable" [7]. This paper holds that they are also appropriate for comparative reviews. Table Ⅰ gives examples for each type.

TABLE I. TYPES OF THE COMPARATIVE REVIEWS AND EXAMPLES

Comparative types	Examples
Superlative	步步高手机的外形比其他任一款手机都要漂亮。
Equative	诺基亚的性能跟三星手机不相上下。
Non-equal gradable	三星 i9000 的配置可以超过 ihpone4。

Define these five ingredients as 5-items, and regard them as the exacting targets. For example, the 5-items of "E71 手机的屏幕相对 N97 来说，好像更舒服一些" exacted are shown in Table Ⅱ.

TABLE II. EXAMPLES OF THE EXTRACTING RESULTS

5-items	P1	P2	A	R	T
Results	E71	N97	屏幕	>	non-equal gradable

B. *Mining the 5-items*

1) *Building of lexicons*

a) *Building of the comparative relation lexicon*

To identify the comparative reviews accurately and comprehensively, a comparative relation lexicon must be built before extracting. When building the lexicon in this paper, according to the method of [7] and based on the summary of comparative syntactic structures in [1], we finish the first stage of building by adding the comparative words in [2]. Moreover, we introduce the Tongyici Cilin [8] made by Harbin Institute of Technology to extend the lexicon. The words in lexicon are called comparative words.

From the aspect of semantics, comparative words should be divided into three types, which are positive, negative and neutral. In previous studies, it has not been discussed. Therefore, the study proposes a new point that divides the words in lexicon. For example, the words such as "强于", "高过" belong to positive polarity, which means the subject is better than object. So, they are put into the positive class. In addition, the words such as "不比", "逊色于", "没有那么" belong to negative polarity, which means the subject is worse than the object, so they are put into the negative class. What's more, there are a few neutral words, e.g., "相比", "相对" etc. In order not to affect the judgment of result value, put them into the positive class, too. The authors give the polarity label W_{cp} for each word and set W_{cp} a value. W_{cp} equals to +1 when the word is in positive class. W_{cp} equals to -1 when the word is in negative class.

b) *Building of the superlative relation lexicon*

In Chinese there are not any fixed formats of grammar and morphology to express the superlative relation. However, Chinese has a class of words which have exclusiveness. For example, the meanings of these words such as "最", "其他", "唯一" are equivalent to the superlative in English. So far, there is not any relevant work about the summary of this class of words. In view of this situation, a superlative relation lexicon is constructed in this paper by summarizing this class of Chinese words. The words in lexicon are called superlative words. The reviews that contain superlative words are superlative reviews. The object of superlative review is null.

2) *Exacting the subject and object*

Compared with common comparative sentences, the comparative reviews of mobiles have their own characteristics. There are only two aspects for subject and object, which are mobile brands or mobile models. This particularity provides the basis for mining of subject and object. Models are usually combined with English letters and Arabic numbers, which are easily to be recognized. Combinations which are in keeping with this feature can basically be the alternatives of comparative entities. Moreover, the brands can be distinguished by a mobile brand set.

The mobile brands or models which users comment on are subjects, which often appear on the same webpage with this review. Figure 1 is a part of the mobile comment page from www.IT168.com. On the upper left corner of this screenshot, we can see that the entity evaluated is Nokia N8.

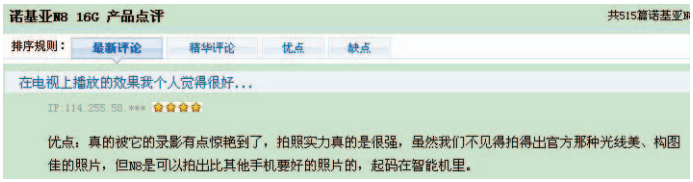

Figure 1. A page of mobile reviews from www.it168.com

It is necessary to indentify the entities exacted being the subject or the object. In this paper, we calculate the similarity between the entities exacted from reviews and the entity got from webpage. The entity with higher similarity is subject, and the lower one is object. Jaccard coefficient [9] shown in formula (1) is the calculation method.

$$Jaccard = \frac{a}{a + b_1 + b_2} \quad (1)$$

3) *Exacting the comparative attributes*

Feature attributes are often made up by nouns or noun phrases [10]. The study takes it as the basis for extracting the comparative attributes. Firstly, use ICTCLAS for POS tagging before extracting. Find clauses containing comparative words

according to comparative relation lexicon. Through the analyzing of expression habits in Chinese comparative sentences, it is easy to draw a conclusion that the attributes always appear in the vicinity of the comparative words. Therefore, we limit the range of extracting to the clauses containing comparative words and their neighbor clauses. Furthermore, there are some reviews such as "诺基亚的手机比联想的要好些", in which the user does not comment on any features. Therefore, set the attribute of 5-items to null.

4) Judging the comparative result value

During product reviews of Chinese, the emotional words that express opinions are always the nouns, adjectives, adverbs and verbs [10]. Reference [10] adopts dependent relationship to extract the opinion words. Distinguishing the dependent relationship needs third-party tools. However, the third-party tools may make mistakes because of the arbitrary expression in natural language. In order to overcome this weakness, the study adopts the pattern-matching method which has higher precision on extracting opinion words.

Pattern matching is often used to mine the opinion words. Currently, the existing patterns always follow the form as features + opinions. However, they require that the splice between features and opinions is compact and the interval is not too big. Consequently, these patterns are not suitable for comparative reviews because of their loose structures. Through analyzing lots of comparative reviews, the paper proposes six models shown in Table III. Moreover, based on the analysis, it can be concluded that opinion words only appear in the clauses containing comparative words or their neighbor clauses on the right. According to this notion, extract opinions from reviews by patterns. Then judge the polarities of opinions with the help of the emotional lexicon built before. If it is positive, set the polarity label W_{ep} as +1; if it is negative ,set W_{ep} as -1; else set W_{ep} as 0.

TABLE III. MATCHING PATTERNS AND EXAMPLES

Patterns	adv.+adj.	adv.+v	adv.+n.	adj.+m.	v.+m.	v.+n.
Examples	相当划算	更加 满意	更加 人性	实惠 一点	提升 不少	存在 差距

The negative prefixes that modify opinion words may affect the opinion polarity. During the process, the negative words are extracted together with opinion words. If there is a negative prefix, set the label W_n as -1; otherwise, set it as +1.

At present the methods of judging the opinion polarity only involve negative words and opinion words. But in the comparative reviews, the polarity of comparative words can also express the users' attitudes. Therefore, the judgmental method in comparative reviews is different from that in non-comparative ones. This paper gives a formula for computing the result value R shown in (2).

$$R=W_{cp}\times W_{ep}\times W_n \tag{2}$$

C. Describing the Mining Method for 5-items

According to the idea mentioned above, the specific mining method is described as follows.

- Step1: Search the whole review. If the review contains a superlative word, then set its type as superlative and turn to step 2. Conversely, it is equative or non-equal gradable and turn to step 3.
- Step2: Extract the entity in the review as the subject.
- Step3: Extract the entities in the reviews, and calculate the similarity between the entity and the product model from the web page. It is the subject if its similarity is higher, and lower one is the object.
- Step4: Find the clause that contains comparative words and extract nouns or noun phrases in it, which are the comparative attributes. If the clause doesn't contain, then extract them from its neighbor clauses.
- Step5: Find the clauses that contain comparative attributes and match them with the patterns to extract the emotional words and negatives. If the clause does not contain these patterns, then find them in its neighbor clauses on the right.
- Step6: Search the emotional lexicon and calculate the value of R. If the value not equals to 0 and the type is not superlative, then set it as the non-equal gradable review. It is equative if the value equals to 0.

IV. EXPERIMENTS

A. Experimental Data Sources

Extract some reviews about mobiles from four E-commerce websites (360buy, Zol, IT168 and Amazon) which contain lots of reviews. Then filter 300 reviews containing comparative relations artificially. In order to test the ability of the method in this paper for the type identifying, we constitute the experimental set with mixing the 3 types of reviews together. The set contains 35 superlative reviews, 49 equative reviews and 216 non-equal gradable reviews.

B. The Experimental Results and Analysis of the Polarity Classification of Lexicon

In the previous studies, the judgment of polarity is only according to the negative words and emotional words. We hold a view that the comparative words can also express the user's opinion. Therefore, the study takes the polarities of comparative words as a factor that influences the final opinions. An experiment is taken to check the validity and the rationality of this idea. During the experiment, we still use the patterns for extracting. Table IV shows the accuracy of the results.

TABLE IV. THE EFFECTS OF THE POLARITIES OF COMPARATIVE WORDS ON OPINIONS

	Positive opinions	Negative opinions
Considering the polarities	78.34%	73.58%
Without considering the polarities	66.75%	58.86%

For the judgment of the opinion polarity, the accuracy that added the polarity of comparative words is obviously improved. The experimental results prove our idea is available.

C. *Experimental Results and Analysis of Mining Algorithm for 5-items*

Table V shows the result of our exacting method based on the experimental set. Comparing to other components, the characteristics of subject and object are much more obvious, so they are easier to be recognized. The judgment of comparative type is accomplished with the help of emotional lexicon and superlative relation lexicon. Therefore, the experimental effects of these three components are much superior to that of the comparative attributes and result value.

TABLE V. THE EXPERIMENTAL RESULTS OF THE METHOD

5-items	Man-ual	Experi-mental	Correct	Prec	Rec	F-value
P1	300	291	285	97.84%	95.00%	96.40%
P2	312	295	273	92.54%	87.50%	89.95%
A	305	303	231	76.24%	75.73%	75.98%
R	308	319	244	76.49%	79.22%	77.83%
T	312	289	277	95.85%	88.78%	92.18%

Currently, there aren't many studies on comparative relation mining of Chinese reviews. Reference [3] proposes the LSR (Label Sequential Rules) algorithm. Table VI shows the results of the extraction among the same experimental set through the LSR.

TABLE VI. THE EXPERIMENTAL RESULTS OF LSR ALGORITHM

5-items	Man-ual	Experi-mental	Correct	Prec	Rec	F-value
P1	300	291	268	92.10%	89.33%	90.69%
P2	312	295	256	86.78%	82.05%	84.35%
A	305	286	205	71.68%	67.21%	69.37%
R	308	287	189	65.85%	61.36%	63.53%
T	312	283	259	91.52%	83.01%	87.06%

After analyzing the results of two experiments, it can be concluded that the precision, recall and F-value have been significantly improved compared with the LSR method based on the label sequence rules. The work of extracting these entities is not only to identify them correctly but also to distinguish the subject and the object. At this point, the LSR is easy to make mistakes. But the method proposed in this paper based on the judgment of the extraction results can reduce such errors. In addition, the [3] extracts the comparative properties and comparative results by the window mechanism. Due to the loose structures of reviews, the size of the window produces a great effect on precision and recall of the results. The new method overcomes this disadvantage. It extracts the comparative relations according to pattern matching and adopts the patterns to match the clause containing the comparative words and their neighbor clauses, which increases the precision and recall in a certain degree.

V. CONCLUSION

This paper proposes a method for extracting the comparative relations in mobile reviews. First of all, we build a comparative relation lexicon and a superlative type lexicon, and then divide the words in the comparative relation lexicon according to their polarities. After identifying the comparative entities based on the naming characteristics, six matching patterns are proposed, which are suitable for the opinion words in comparative reviews. Then, use these patterns to match the clause containing the comparative words and their neighbor clauses in order to extract opinion words. Finally, compute the result value with the formula we proposed.

Although the method in this paper has improved the precision and recall to a certain extent, but there are still some defects. We only consider the nouns when extracting the attributes. However, some verbs may also be the features. Our method is of no effect on extracting them. Reference [4] proposes to build a feature lexicon artificially. It can ensure the quality of extraction though needs lots of human interventions. In the further study, researchers could borrow ideas from it and add a few human interventions. Moreover, the research on naming rules is limited to the mobiles. How to extend this idea to other electronic products can be taken as a future project.

ACKNOWLEDGMENT

Thanks to the ICTCLAS made by Chinese Academy of Sciences and the Tongyici Cilin made by Harbin Institute of Technology, which are used during our study. Moreover, this research is supported by the Natural Science Foundation of China (No. 61170039).

REFERENCE

[1] J. Chen, X. B. Zhou. "The Selection and Arrangement of Grammatical Items concerning Comparative Sentences," Language Teaching and Linguistic Studies, No. 2, pp. 22-33, 2005(In Chinese).

[2] X. J. Huang, X. J. Wan, J.W. Yang, J.G. Xiao, "Learning to Identify Chinese Comparative Sentences," Journal of Chinese Information Processing, vol. 22, No. 5, pp. 30-38, September, 2008(In Chinese).

[3] N. Jindal, B. Liu. "Mining comparative sentences and relations," Proceedings of 21st National Conference on Artificial Intelligence. Boston, Massachusetts, USA, 2006, pp. 1331-1336.

[4] G. H. Huang，T. F. Yao，Q. S. Liu. "Mining Chinese comparative sentences and relations based on CRF algorithm," Application Research of Computers, vol. 27, No. 6, pp. 2061-2064, June, 2010(In Chinese).

[5] M. Xu. "Study on the resultant item of "Bizi" sentence and its related questions in modern Chinese," Anhui Normal University, May, 2003(In Chinese).

[6] K. Q. Xu, S. Liao, J. X. Li, Y. X. Song. "Mining comparative opinions from customer reviews for Competitive Intelligence," Decision Support Systems, vol. 50, No. 2011, pp. 743-754, 2011.

[7] R. Song, H. F. Lin, F. Y. Chang. "Chinese Comparative Sentences Identification and Comparative Relations Extraction," Journal of Chinese Information Processing, vol. 23, No. 2, pp. 102-107, March, 2009(In Chinese).

[8] W. Che, Z. Li, T. Liu. "LTP: A Chinese Language Technology Platform," Proceedings of the Coling 2010:Demonstrations. Beijing, China, August, 2010, pp.13-16.

[9] J. W. Han, M. Kamber, M. Fan, X. F. Meng. "Data Mining: Concepts and Techniques (the Second Edition)," Beijing: Mechanical Industry Publishing House, pp. 255-256, 2007.

[10] P. Li. "Research on Opinion Extraction and Classification Technologies for Product Review Mining," Chongqing University, Chongqing. 2009(In Chinese).