Proceedings of the 2013 International Conference on Machine Learning and Cybernetics, Tianjin, 14-17 July, 2013

COMPARATIVE OPINION SENTENCES IDENTIFICATION AND ELEMENTS EXTRACTION

CHENGXIANG LIU, RUIFENG XU*, JIE LIU, PENG QU, HE WANG, CHENGTIAN ZOU

Key Laboratory of Network Oriented Intelligent Computation, Shenzhen Graduate School,
Harbin Institute of Technology, China
E-MAIL: xuruifeng@hitsz.edu.cn

Abstract:

This paper presents a subsystem for analyzing the comparative opinion sentences. This subsystem consists of three modules. Firstly, rule-based and CSR-based method is developed to identify comparative sentences, respectively. Secondly, in opinion element extraction module, CRFs model is applied to label elements. Finally, in the polarity determination module, some heuristic rules are complied to classify different types of comparative sentences. The evaluations on COAE2012 dataset show that the proposed analysis subsystem achieves a good result.

Keywords: Comparative sentences; Opinion analysis; Comparative relations extraction

1. INTRODUCTION

With the popularity of Web 2.0, a large amount of user comments and reviews have been published on various public communication platforms, such as paste bars, forums, blogs and micro-blogging. The analysis of such kind of subjective information is important to public opinion analysis, hot topic tracking, and user experience improvement. Accordingly, opinion analysis has attracted more attentions.

Comparative opinion structure is commonly used in user's comments. It usually contains two or more opinion targets in one sentence. Since these structures are obviously different from the regular opinion expression, the identification and analysis of comparative sentence is one of the key issues in opinion analysis.

Jindal et al. [1] studied the identification of comparative sentences in English by employing the pattern discovery and supervised machine learning method. Based on this, they investigated how to extract comparative elements from a comparative sentence [2]. Huang Xiaojiang et al. [3] treated comparative sentences identification as a binary classification. They applied machine learning method to comparative sentence identification while some linguistic features and statistical features were evaluated. Song Rui et al. [4] constructed a Chinese comparative pattern database. Furthermore, a CRFs model is trained with selected features for comparative relations extraction.

In this paper, we present a subsystem for identifying and analyzing the comparative opinion sentences. Firstly, the rule-/pattern based method and machine learning based method are investigated to recognize the comparative sentences, respectively. Secondly, CRFs model is employed to extract comparative elements from the identified comparative sentences. Finally, based on some rules, the comparative sentences are classified into 3 comparative relation types for polarity determination. The evaluations on comparative opinion analysis subset in Chinese opinion analysis evaluation (COAE 2012) show that the proposed subsystem achieves encouraging performance.

The rest of this paper is organized as follows. Section 2 presents the development of comparative sentences identification and opinion elements extraction subsystem. Evaluation and discussion are given in Section 3. Section 4 concludes.

2. DEVELOPMENT OF COMPARATIVE OPINION SENTENCE ANALYSIS SUBSYTEM

The comparative opinion sentence analysis subsystem consists of two major modules: comparative sentences identification and opinion element extraction.

2.1. Comparative sentences identification

Comparative sentences identification may be treated as a binary-classification problem in this study. The rule-based and class sequential rules-based (CSR-based) classification methods are investigated, respectively. Rule-based identification method exploits comparative words, comparative content words, the relative degree adverbs and syntactic patters to identify comparative sentences. Once the identification rule is matched to the input sentence, this sentence is considered comparative. The CSR-based

978-1-4799-0260-6/13/$31.00 ©2013 IEEE

identification method combines class sequential rules mining and machine learning techniques. Firstly, the comparative words are used as a clue to transform training corpus into sequential data sets, and then multi-minimum support strategy is employed for CSR mining. The result sequence set is used as baseline feature set for training classifiers to identify comparative sentences.

2.1.1 Rule-based comparative sentence identification

There are many differences between comparative and non-comparative sentences in terms of vocabulary, syntax mode and some certain collocations. Based on the observation and analysis of comparative sentences, some rules are manually compiled. These rules are divided into the following categories:

(1)Identification rules based on comparative words

In comparative sentences, some certain words can be treated as a discriminative feature before comparison standard, which called the comparative word. Such as:

世纪星，侧面不如前面大气好看，似尼桑风度。

(*The side face of Shijixing is not as nice as its front face, just like Nissan*)

Rule1='不如|相比|较之|.....'

(2)Identification rules based on comparative content words

Some comparative sentences use the content words' semantic meaning rather than function words or patterns. Such as:

超值版很值，除内饰不一样其它和新中华一样。

(*The special edition is very valuable, it`s all the same to new CMC except interior*)

Rule2='一样|差不多|无法匹敌|'.....'

(3)Identification rules based on relative degree adverbs

A majority of comparative sentences have a comparative attribute or result word (adjective) for comparison. In English, we change these words' form (comparative level, highest level) for adjectives and adverbs' comparison. There is no words morphological change in Chinese, but there are some modifications adverbs. For example, the relative degree adverbs in the "最 (most)" category (最、极、顶), equivalent to the highest level in English, and "更 (more)" category (更／更加、越发／愈发), equivalent to the comparative level in English . Such as:

奥迪是世界上质量最次的车.

(*Audi is the worst car of the world in quality*)</DOC1035>

Rule3='最棒|最差|最轻|.....'

(4)Identification rules based on specific syntactic pattern

In many cases, some specific syntactic patterns are adopted in comparative sentences. Thus, pattern matching is investigated to identify comparative sentences. Such as: "比……好|强|专业|便宜", "和|跟|与……比较", "没有……好|舒服|耐用" . Such as:

可那个发动机的噪音跟咱小三比那真是差啊！

(*But the engine noise is worse than Xiaosan*)

Rule4='比\s.* (好|强|贵|...)'

For each set of above identification rules, the corresponding regular expressions are complied. These rules are applied to match the sentences in test corpus.

2.1.2 CSR-based comparative sentence identification

Although many comparative sentences have obvious word or structure features, the sentence which contains a comparative word may not be always a comparative sentence. For example, "在 A、B、C 中，B 的性能最佳 (*Among A, B and C, B got the best performance*)"is a comparative sentence, while "A 这款跑车的最佳性能体现在速度 170-200km/h (*The best performance of racing car A can be reached at a speed between 170km/h and 200km/h*)" is not. Besides, many Chinese words can have different part of speech and different meanings. It makes different results in comparative sentence identification. Our preliminary studies show that the method based on the rules always achieves high recall, but low precision. Therefore, the CSR based comparative sentence identification is investigated.

The CSR-based method is the combination of CSR and machine learning methods. Firstly, we summarized comparative words which show discriminative capacities manually, and then a sequential dataset is constructed with the use of comparative words. Then we apply CSR to mine appropriate rules through following multi-minimum supports and one fixed confidence strategy. The final CSR dataset served as the baseline feature set for classifier training.

(1) Comparative words list

A list of comparative words is built manually. The list is composed by both comparative words and comparative result words, such as "比", "对比起来", "相比", "无与伦比", "旗鼓相当", "媲美", "领先", "佼佼者" (*than, compared to, perfect, the same as*)

(2) Sequential dataset construction

The sentence dataset is firstly transferred into sequential dataset since the direct use of original sentences leads to data sparse. In this study, clause rather than sentence is selected as the observation boundary for analysis while the [-5, +5] words of the comparative word is regarded as the observation context. Meanwhile, considering that part of speech is useful for the identification, so the comparative word and its POS are combined as one element, while only part of speech is

reserved for the remaining words.

(3) Multi-minimum supports CSR mining

For a given input sequence, the task of pattern mining is to find out all sequential patterns that satisfy user-specified minimum support. A sequential pattern is a sub-sequence whose frequency exceeds the minimum support threshold. For each class sequential rule (CSR), the left side is a sequence pattern, while the right side is a class label. CSR mining aims to find out sequential pattern highly correlated with the category.

Sequence S is an ordered set of items, denoted by $S=<a_1, a_2 \ldots a_r>$, where a_i is an item set, also known as an element of S. Given sequence $S_1=<a_1, a_2 \ldots a_r>$, $S_2=<b_1, b_2 \ldots b_r>$, if there are integers $1 \le j_1 \le j_2 \le \ldots \le j_{r-1} \le jr$, which makes $a_1 \subseteq b_{j1}$, $a_2 \subseteq b_{j2} \ldots a_r \subseteq b_{jr}$, S_2 contains S_1.

The sample dataset $D\{(S_1,C_1), (S_2,C_2), (S_m,C_n)\}$, where S_i is the sequence, C_i is the class label. A sequence rule can be formalized as $X->C$, while X is a sequence mode and C is a class label. For one sample (S_i,Y_i) in the sample data set D, if the sequence S_i contains sequence X, the sample (S_i,Y_i) is regarded covering class sequence rule $X->C$. Furthermore, for a sample (S_i,Y_i) in D, if the sequence S_i contains a sequence of X and $Y_i=C$, the sample (S_i,Y_i) is regard satisfying class sequence $X->C$. The support of a class sequence rule is defined as the percentage of the satisfied samples by this rule within the whole data set D. The confidence of a class sequence rule is defined as the percentage of the sample data covered and satisfied by this rule within the whole data set D.

CSR is a supervised mining approach. Here, we use the Prefix-Span deformation algorithm to generate the initial class sequence rule set. The details of this algorithm are given in [5].

Since there are various frequencies of elements in the sequence, using one single global minimum support threshold for sequence-rule mining is not appropriate. For example, to discover the sequence of low-frequency words, we had to set the support threshold very low, but this will also brings a large number of noise class sequence rules generated by the high-frequency words which affects to the classifier. Thus, we adopt the multi-minimum support strategy here. For each rule, the minimum support is calculated as follows:

$$MinSup(r) = max(\alpha * min(f_i), s) \quad (1)$$

Where f_i is the dataset frequency of the *i-th* element from rule r, α is a factor in (0,1), s is a minimum threshold used to filter out low $\alpha * min(f_i)$.

For the initial class sequence rule set mined out by CSR, a multi-pass minimum supports and confidence filtering are applied to generate the final rule set.

(4) Comparative sentence classifier

Regarding the satisfaction status of each rule as features, the satisfaction status of each clause forms a classification example. If a clause satisfied a rule, the attribute value for this rule is set to 1, otherwise set to 0. It is observed that the comparative sentences normally have more objects and attributes in the sentence. It motivates the use of information related to object and attribute as the appended features. Here, a simple ontology containing various product names and attributes is constructed. A string matching is used to determine whether a product name or an attribute appeared in the sentence or not. Two new features, i.e. number of matched products and number of matched attributes, are added into the rule-based feature set.

During the estimation stage, a variety of classifiers are tested, including Decision Tree, SVM and Naive Bayes. In which, Naïve Bayes achieve the best performance.

2.2. Element extraction and polarity determination

2.2.1 Element extraction

The purpose of this module is to extract attributes and product names in comparison from comparative sentences, and then determine the polarity of sentence. For example,

诺基亚 N8 的屏幕不如 iPhone 的好

(*The screen of Nokia N8 is not as good as that of iPhone*)

The names of two comparative products are "诺基亚 N8 (Nokia N8)" and "iPhone", respectively. The target attribute is "屏幕 (screen)" .The sentence want to express the opinion that "iPhone 的屏幕 (*the screen of iPhone*)" is better than "诺基亚 N8 的屏幕 (*the screen of Nokia N8*)".

There are some complex cases. For example, comparative elements do not appear in some cases.

音质不如 Sony

(*The tone quality is not good as that of Sony*)

In this sentence the first product name is hidden. So we use "NULL" to substitute it.

Our subsystem adopts a CRFs (Conditional Random Fields) based classifier to identify the product name and target attribute. A complete comparative sentence contains four elements, namely, comparative subject product name, comparative subject product attribute, comparative object products name and comparative object product attribute. These four parts are usually constructed by the phrase such as"森海塞尔的 MX360 (*MX360 from Sennheiser*)", "低音的音质 (*the tone quality of bass*)", "屏幕做工 (*screen quality*)"and so on. If these phrases are directly used to mark in CRFs, the inconsistency between training set and test set will make this task complex and difficult. Considering that in most cases, the ending word is the head-word of Chinese

phrase, we simplified the CRFs based classification in which only the ending word in every element phrase were classified. For example, In this phrase"森海塞尔 MX360", only "MX360" is marked.

There were four kinds of features used in comparative elements extraction section: word, POS, location, and domain knowledge features. The word features including words themselves and their neighboring words. The POS features include the POS of current words and their neighboring word. The location feature describes the distance between the current word and the comparative word. The domain knowledge features determine whether current word is in domain knowledge database. Table 1 gives the list of these four kinds of feature.

TABLE 1. FEATURE LIST FOR ELEMENT EXTRACTION

Feature category	Description
word feature	Word itself the front of the word the following word is comparative word or not is comparative result word or not
POS feature	The current word's POS the front word's POS the following word's POS
Location feature	Before comparative word within 5 words After comparative word within 5 words around comparative word within 2 words
Domain knowledge feature	is product attribute word or not is product name word or not

After comparative elements classification is completed, the four types of elements are obtained. Consider that the experimental data is derived from online product reviews in which the words used in on-line reviews are more flexible, our system used a method based on combination to form the requested instances. *SetAI* is the subject product name set. *SetAA* is the subject product attribute set. *SetBI* is the object product name set. *SetBA* is the object attribute set. If there is nothing in SetAI, we put "NULL" in it. *SetAA* and *SetBA* consist attributes set (*SetA*). We choose an element from the three sets (SetAI, SetBI, SetA), respectively, to construct an instance. The system generates all possible constructions in the final set of instances. For example,

同价位，做工，性能，操控感都比佳能好！
(*The workmanship, performance, and control are better than those of Canon at the same price*)

SetAI = {NULL}，*SetAA* = {做工 (*workmanship*)，性能 (*performance*)}，*SetBI* = {佳能 (*Canon*)}，*SetBA* = {NULL}，

The result: *SetAI*={NULL}，*SetBI*={佳能 (*Canon*)}，*SetA*={做工 (*workmanship*)，性能 (*performance*)}

2.2.2 Polarity Determination

Based on the extracted elements, the system classifies the comparative sentences to three classes, namely equal comparative, most comparative and different comparative. Polarities of equal comparative sentences and most comparative sentences are easy to identify, while polarities for different comparative need further analysis due to the importance of the relative positions from subject and object to comparative feature words. For complex sentences, many product attributes may be compared. To determine the appropriate attributes for the comparison, the attribute appears in the same clause within the comparative feature words (or comparative mode) is prioritized, followed by the attributes appears before sentence, and the property after the clause is considered at last.

1. Polarity determination for equal comparative

If a clause satisfies the following patterns and comparison elements appear in the first and the second * position, this clause is identified as equal comparative. It`s easy to determine the polarity, which could be marked as 0 for all such instances. The example compare patterns are:

(1) *[^不没无](像|似|和|如|和|于|跟|同|于|较).*[^不没无](相同|一样|一致|相仿|相差无几|同样|).*

(2) *[^不没无](像|似|和|如|和|于|跟|同|于|较).*[^不没无](有|存在).*(差异|出入|区别|差别).*

2. Polarity determination for most comparative

If a clause satisfies the following patterns, the clause will be recognized as most comparative. According to whether there is negative word before and the polarity of word follows, the polarity can be set to 1 or -1. Such as:

[^不](作为|是|属于|成为|位居|算).(第一|老大|王者|之王|顶尖|一流).*

3. Polarity determination for different comparative

(1) Different comparative formed by single word, Ex. 比不上，不及，不如，超，超过，超越，反衬，高过，弱于，胜过，完胜，逊色于，逊于，亚于，优于

There is a general pattern of "A + Compare_Word + B" for different comparative sentences formed by single word, and A, B maybe the subject (or object) or its attributes. Set the polarity value -1 to words like "弱于", "不及", "比不上" (*worse than*), and set 1 to "超过", "高于", "好于"(*better than*). The polarity value of A is always the same to the one of these comparative words, and B is always on the opposite.

(2) Different comparative formed by "比+a", while "a" refer to a result adjective. e.g. "比…霸气", "比…可靠", "比…差", "比…难看", "比…逊色" (*than*).

General pattern "A 比 B a" may be easily found. We also divide the words of comparative result into two groups:

positive and negative, responding to 1 and -1. Polarities of A is always the same to the one of these comparative result words and B is always on the opposite. For example, in "iPhone 的屏幕比 Nokia 先进", A is "iPhone 的屏幕", B is "Nokia", we get:

iPhone 屏幕 1

Nokia 屏幕 -1

Note that even when pattern and adjective are the same, the polarity may change with different compare attributes. For example, " A 比 B 性能高 (*A has higher performance than B*)", " A 比 B 油耗高 (*A consumes more fuel than B*)", apparently the polarities of A(or B) are completely different, we summarized words like "油耗 (*fuel consumption*)", "磨损 (*abrasion*)", " 耗电 (*power consumption*)", " 散热 (*radiating*)", "噪音 (*noise*)" which make the polarity reverse. When matching these attributes in the set, the polarity will be reversed.

3. EXPERIMENTAL RESULTS

3.1. Comparative Opinion Sentence Identification

The experiment data comes from Chinese Opinion Analysis Evaluation (COAE) 2012. The dataset is comments and reviews in digital and car area. The distribution of the training sample is shown in table 2:

TABLE 2. SAMPLE DISTRIBUTION OF COMPARATIVE AND NON-COMPARATIVE SENTENCES

Comparative	Non-comparative
200	1000

Firstly, Rule-based method is evaluated. The performances achieved by adopting different types of rules are given in Table 3, respectively. It is shown that Rule set 4 achieves much better performance. It means that the identification rules based on syntactic pattern play a very important role in comparative sentence identification.

TABLE 3. RESULT OF COMPARATIVE SENTENCE IDENTIFICATION BASED ON EACH CLASS OF RULES(RULE-BASED)

Digital	**Rule1**	0.2880	0.6901	0.4064
	Rule2	0.0883	0.4354	0.1469
	Rule3	0.0474	0.5918	0.0878
	Rule4	0.7446	0.8010	0.7718
Car	**Rule1**	0.3327	0.8644	0.4805
	Rule2	0.1109	0.7727	0.1940
	Rule3	0.0538	0.4583	0.0963
	Rule4	0.6231	0.7860	0.6951

Secondly, we evaluate the CSR-based identification algorithm. Furthermore, we evaluate the influence of the two new features, namely number of products and number of attributes. The achieved performances are listed in Table 4. It is shown that CSR-based algorithm achieved good performance. Meanwhile, the two new features are shown effective to comparative sentence identification, especially for comments of digital products. Especially, the number of products is more useful. This is because the omission of comparative attributes is more common than that of comparative objects

TABLE 4. RESULT OF COMPARATIVE SENTENCE IDENTIFICATION BASED ON CSR

	Feature set	Precision	Recall	F1-score
Digital	**Rule-based feature set**	0.8272	0.7676	0.7963
	Add number of products	0.8428	0.7987	0.8202
	Add number of attributes	0.8579	0.7709	0.8121
	Both added	0.8707	0. 7938	0.8305
Car	**Rule-based feature set**	0.7962	0.7455	0.7700
	Add number of products	0.7868	0.7586	0.7724
	Add number of attributes	0.7842	0.7471	0.7652
	Both added	0.7896	0.7651	0.7771

Table 5 gives the overall performance of Rule-based method and CSR-based method.

TABLE 5. RESULT OF COMPARATIVE SENTENCE IDENTIFICATION: RULE-BASED AND CSR-BASED

		Recall	Precision	F-measure
Digital	**Rule-based**	0.8814	0.7676	0.8206
	CSR-based	0.7938	0. 8707	0. 8305
		Recall	**Precision**	**F-measure**
Car	**Rule-based**	0.8140	0.7481	0.7797
	CSR-based	0.7651	0. 7896	0.7771

It is observed that both the rule-based method and CSR-based method achieved relatively good performance.

It`s difficult to rule-based method for increasing recall without bringing in noise. The low recall of CSR-based method is partially caused by the following three reasons: (1) The list of comparative feature words is incomplete; (2) Semantic comparative sentences are too complex and ambiguous, such as "H180 在天上，索尼的在地上 (*H180 is much better than Sony*)", "A 和 B 有一拼 (*A is not worse than B*)" and "没什么两样 (*nothing different*)". (3) Wrong spelling and non-standardized punctuations: such as: punctuation "。" is typed to be ".", causing wrong clause segmentation. And "不比" is wrongly spelled as "不逼", , and so on.

3.2. Element extraction and polarity determination

The achieved performance on element extraction is given in Table 6. In this table, S1 classifies all words while S2 only classifies the headwords. It is shown that the

headword classification strategy is proved effective in the increase of precision. The system achieved a relatively good F1. Furthermore, the result in the car area is better than that of digital area. This may be caused by the fact that the vocabularies about the car provide more domain knowledge.

TABLE 6. PERFORMANCE CONTRAST IN DIGITAL CORPUS

		Precision	Recall	F1
Digital	S1	0.4094	0.0813	0.1357
	S2	0.5302	0.0885	0.1517
Car	S1	0.3517	0.1895	0.2463
	S2	0.5743	0.1918	0.2875

The performance of the polarity determination is shown in table 7. Experimental data show that our method achieved an acceptable performance.

TABLE 7. PERFORMANCE OF OPINION ANALYSIS

	Precision	Recall	F1
Digital	0.4612	0.0770	0.1319
car	0.4222	0.1410	0.2114

5 CONCLUSIONS

This paper presents an analysis subsystem for comparative sentences. Firstly, rule-based and CSR-based method is developed to identify comparative sentences, respectively. Rule-based method achieves better recall and CSR-based method achieves better precision. In opinion element extraction module, CRFs model is used to label elements. Finally, in the polarity determination module, some heuristic rules are complied to classify different types of comparative sentences. A relatively high precision is achieved, while the recall is low due to limited word list. The future work including expanding the list of words and rules, and improving the feature set of CRFs.

Acknowledgements

This research is supported by MOE Specialized Research Fund for the Doctoral Program of Higher Education20122302120070, Shenzhen Foundational Research Funding JCYJ20120613152557576, and Shenzhen International Cooperation Research Funding GJHZ20120613110641217.

References

[1] Jindal. N, Bing. Liu, "Identifying comparative sentences in text documents", Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York: ACM Press, pp. 244-251, August 2006.

[2] Jindal. N, Bing. Liu, "Mining comparative Sentences and Relations", Proceedings of the 21st National Conference on Artificial Intelligence, Boston, pp. 1331-1336, July 2006.

[3] Xiaojiang Huang, Xiaojun Wan, Jianwu Yang, Jianguo Xiao, "Learning to Identify Chinese Comparative Sentences", Journal of Chinese Information Processing, Vol. 22, No. 5, pp. 30-37, 2008.

[4] Rui Song, Hongfei Lin, Fuyang Chang, "Chinese Comparative Sentences Identification and Comparative Relations Extraction", Journal of Chinese Information Processing, Vol. 23, No. 2, pp. 102-122, February 2009.

[5] Bing. Liu, Web Data Mining: Exploring hyperlinks, Contents, and Usage Data, Springer, 2006/2007.

[6] J. Pei, J. Han, B. Mortazavi-Asl, J. Wang, H. Pinto, Q. Chen, U. Dayal, M.C. Hsu, "Mining sequential patterns by pattern-growth: the Prefix-Span approach", Transactions on Knowledge and Data Engineering, Vol. 16, No. 11, pp. 1424-1440, November 2004.