Opinion Targets Identification Based on Kernel Sentences Extraction and Candidates Selection

Hengxun Li[1(✉)], Chun Liao[2], Ning Wang[1], and Guangjun Hu[1]

[1] First Research Institute of the Ministry of Public Security of PRC,
capital gymnasium south road NO. 1, Haidian District, Beijing 100048, China
DerekLee1985@126.com, wn_1209@163.com, cityof93@qq.com

[2] Institute of Information Engineering, Chinese Academy of Sciences,
minzhuang Road No. 89, Haidian District, Beijing 100091, China
liaochun@iie.ac.cn

Abstract. With the developing of the Internet, communication becomes more and more frequent, and the traditional opinion mining technology has been unable to meet the people's needs, especially in the field of opinion targets identification. Therefore, how to do appropriate pre-processing and post-processing with opinion sentences to improve the quality of opinion sentence identification has become a hot issue in recent years. Researches on kernel information filtering and candidates screening of traditional opinion targets identification methods are insufficient. In this paper, we propose a novel opinion targets identification method which integrates kernel sentences extraction with candidates selection based on rules analysis and SVM screening. Experimental results on COAE2014 dataset show that this approach notably outperforms other baselines of opinion targets identification.

Keywords: Opinion targets identification · Kernel sentence extraction · Candidates selection · SVM

1 Introduction

With the widespread popularity of the Internet, Internet already becomes the main way for people to gain and share information. As an emerging platform for interaction and communication, microblog has become part of people's life gradually. According to the Thirty-seventh Statistical Report on The Development of China Internet Network published by the CNNIC, until December 2015, the total number of Internet users in China is Six hundred and eighty-eight million [1]. More and more people begin to pay attention to microblog, people share their moods and opinions and discuss the popular topics. Microblog has huge data and is time-limited. It can be dug out a lot of meaningful information. Therefore, it has attracted a large number of scholars to develop related research, one of the hottest direction of the research is the sentiment analysis about the microblog.

Sentiment analysis is also known as opinion mining. It refers to carry on the subjective analysis, the induction and the sentiment polarity judgment [2]. According to

Z. Shi et al. (Eds.): IIP 2016, IFIP AICT 486, pp. 152–159, 2016.
DOI: 10.1007/978-3-319-48390-0_16

the progressive level of the task of the sentiment analysis, the task of the sentiment analysis can be divided into three categories: the extraction of sentiment information, the classification of the sentiment information, and the retrieval and induction of the sentiment information. The extraction of sentiment information is the basic task of the sentiment analysis.

It means to excavate the structured information from unstructured text sentiment, including the opinion targets, opinion words, opinion tendency and opinion holders and so on. As the basic task of the sentiment analysis, it not only can serve the upper level of the sentiment analysis, such as the classification of the sentiment information, but also can be directly applied to the electronic commerce, information security and other fields. For example, in the statistics of commodity assessments, we can make other consumers understand the advantages and disadvantages of the goods clearly in all directions. It can also help to improve the marketing strategy and the performance of the goods.

Opinion targets is also known as sentiment targets or view targets. It mains the subject of discussion in a text. For example, “对三星手机彻底没好感了”, the opinion target is “三星手机”. Extracting the correct opinion targets means that we can make more accurate analysis and inference to a certain object, which also means great commercial and social value.

In this paper, we propose a novel opinion targets identification method which integrates kernel sentences extraction with candidates selection based on rules analysis and SVM screening. We first extract the kernel sentences of the oral opinion sentences, then we adopt the CRF-based method to perform opinion targets identification to get candidate opinion targets. Finally, we screen all the candidate opinion targets based on SVM classifier and acquire the final opinion targets identification results. In experiments on the COAE 2014 dataset we find that our method can substantially extract opinion targets more effectively under different evaluation metrics.

2 Related Work

Minqing and Bing [3] thought that the opinion targets was a noun or noun phrase. Gaining the candidate opinion targets by digging out the noun or frequent item sets of the noun phrase. Zhuang et al. [4] proposed a multi-knowledge-based approach which integrated WordNet, statistical analysis and movie knowledge. At the same time, it is considered that the nearest adjective to the opinion targets is the opinion words. Hongyu et al. [5] extracted the opinion targets by syntactic analysis, PMI and feature pruning. Li et al. [6] extracted tuples like <emotional words, opinion targets> based on emotional and topic-related lexicons. Fangtao et al. [7] and Tengfei and Xiaojun [8] transformed the opinion targets into the sequence labelling, and used conditional random fields model to extract the opinion targets. Xu et al. [9] used the shallow parsing information and the heuristic location information and other features in the training of conditional random fields model, so that the extraction effect of opinion targets has been improved. Jakob and Gurevych [10] modelled the task as a sequence labelling question and employed CRF for opinion targets extraction. Wang et al. [11] used the new feature of the semantic role labelling in the training of conditional random fields model, there are four features used to training the conditional random fields model: morphology, dependence, relative position and semantic. Song and Shi [12] gained the seed set by sample survey, then expanded the seed set by semi-supervised learning to extract more accurate opinion tar-gets. Xu et al. [13] used the syntactic analysis and random walk model to extract opinion targets.

It is not difficult to find whatever which way we use, statistics corpus will help a lot. Consequently, considering the specific features of Chinese microblog, we propose a new method for opinion tar-gets extraction towards microblog based on kernel sentence extraction and candidates selection.

3 Kernel Sentence Extraction

The key idea of kernel sentence extraction in this paper is mainly to delete redundancy, retain and evaluate the main components of the oral sentence. This paper aims to improve the accuracy of opinion targets identification by using the kernel sentence extraction. The principle of extracting the kernel sentences is to standardize the opinion sentences, and try not to lose the ingredients related to the original opinion sentences. Through statistics and observation of a large number of data, we sum up 10 kinds of rules, as shown in Table 1.

We perform kernel sentence extraction based on the rules in Table 1 and obtain a standardize corpus for opinion sentence identification.

4 Candidate Opinion Targets Identification and Selection

After kernel sentence extraction in Sect. 3, we perform candidate opinion targets identification using CRF model. CRFs (Conditional Random Fields, CRFs) is proposed by Lafferty et al. [14] in 2001. Its model structure is shown in Fig. 1. Given a set of input random observed variables, this conditional probability distribution model can generate another set of implicit output random variables by training the model.

In CRF-based method, the features we employed as input are of great importance. In this section, we refer to the features which are employed by Jakob and Gurevych [10] in English and meanwhile put forward some new features based on the specific grammar of Chinese. Generally, we think opinion targets extraction is primarily related with four kinds of features which are named as lexical features, dependency features, relative position features and semantic features. First, as words with the same Part-of-Speech usually appear around the opinion targets, we select the current word itself and the POS of current word as lexical features. Second, we select whether the dependency between current word and core word exists, the dependency type, parent word and the POS of parent word as the dependency features. Finally, considering here is a strong relationship between the sematic roles and POS of emotional words, we select the sematic role name of current word and POS of emotional word in this sentence for CRF.

Table 1. Rules of kernel sentences extraction

Rules	Examples	Explanation
Delete the English and interrogative sentences;	I am a researcher ……?	This paper only focuses on Chinese opinion targets identification
Turn over the sentences with "//"	电池也很耐用//三星手机不错	This measure is to ensure the forwarding relationship
Delete text sequence which contains link address in it	http: …… 网址: …… 地址: ……	Remove the link address content and the expression of the address, and the extraction result of the opinion targets is not affected
Delete text sequence which contains microblog symbols in it	@……. @…….转发微博 @…….回复 #…….#	"@" + user name, #……# indicates the topic of microblog. The deletion of them will not affect opinion targets identification
Replace the consecutive punctuations with the first one;	。。。 !!!! ??	This measure is to standardize the expression
Delete supplementary text sequence	【…】[…] (…) (…) 文章来源	Words in brackets are generally supplementary explanation of the main text, and the article source indicates the path, both will not influence opinion targets identification
Delete words of hypothetical tendencies in a sentence	如果…… 希望…… 假如…… ……	These words would cause noise for opinion targets identification
Remove sentences which are for introduction of the following passage	……的优点 ……的不足 ……的优势 ……	
Delete the degree-words in front of the sentence	尤其是…… 特别是…… 还…… ……	
Delete sentences which do not contain emotional words	三星Galaxy S4是三星电子在2013 年推出的一款手机,搭载的是Exynos 5410 双四核处理器。	We mainly make research on opinion sentences

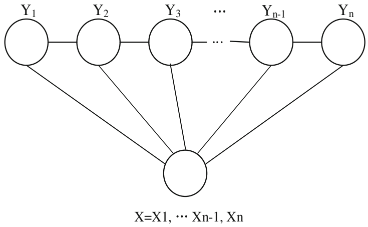

Fig. 1. CRFs model

Table 2. Feature description of candidates selection based on SVM

Feature name	Feature value
Semantic role	A0(agent): 1
	A1(patient): −1
	Others: 0
Minimum distance	−1, 1, 0, 1, 2, 3…
Word frequency	0, 1, 2, 3…

We perform candidate opinion targets selection after candidate opinion targets identification. Candidate opinion targets selection is to judge whether a candidate opinion target is the opinion target of this microblog. This paper considers that the process of screening candidate opinion targets is equivalent to a question of binary-classification, and that is to judge whether the candidate opinion target is the opinion target of this microblog. So this paper adopts SVM classifier for candidate opinion targets selection, and the feature combination is shown in Table 2.

We sum up three kinds of features for candidate opinion targets selection: semantic role, minimum distance and word frequency. For semantic role feature, we label the agent and patent of semantic role labelling result for candidate opinion targets selection. For example, a microblog of "相机很漂亮", through the semantic role analysis, "相机" is the agent, "漂亮" is patient. Thus, the agent or patient may be the opinion targets. In the experiments, we use the Language technology platform (LTP) [15] of Harbin Institute of Technology for semantic role labelling. For the minimum distance feature, we select the number of words that are nearest to the opinion targets as minimum distance value. In this paper, it is considered that each opinion word has an emotional word to modify. For the word frequency feature, we select the occurrence frequency as the feature value. In a number of microblogs, if the occurrence frequency of a noun or noun phrase is very high, then the noun or noun phrase is the main description of the text, that is to say, it is likely to be the object target.

Finally, we construct the training model based on the methods in Table 2 and finally complete the opinion targets identification.

5 Experiments and Analysis

In the experiment, we firstly obtained kernel sentences using methods in Sect. 3, and then perform CRF-based method for candidate opinion targets identification. Finally, we adopt SVM classifier for candidate opinion targets selection to complete the opinion targets identification. Its structure model is illustrated in Fig. 2.

For the experiment data, we adopt 7000 sentences which are provided by COAE2014, these sentences are acquired from microblog, forum and other social network platform. Considering the short, interactive, non-standard features of these sentences, we use rules in Sect. 3 to acquire kernel sentences. Through filtration by these rules, we finally obtain 4,500 normalized sentences with opinion orientation. In this paper, we conduct experiments on such a dataset and assess it with traditional Precision, Recall and F-measure under strict and lenient evaluations which respectively represents the extraction result is exactly the same or overlapped with the labelled one.

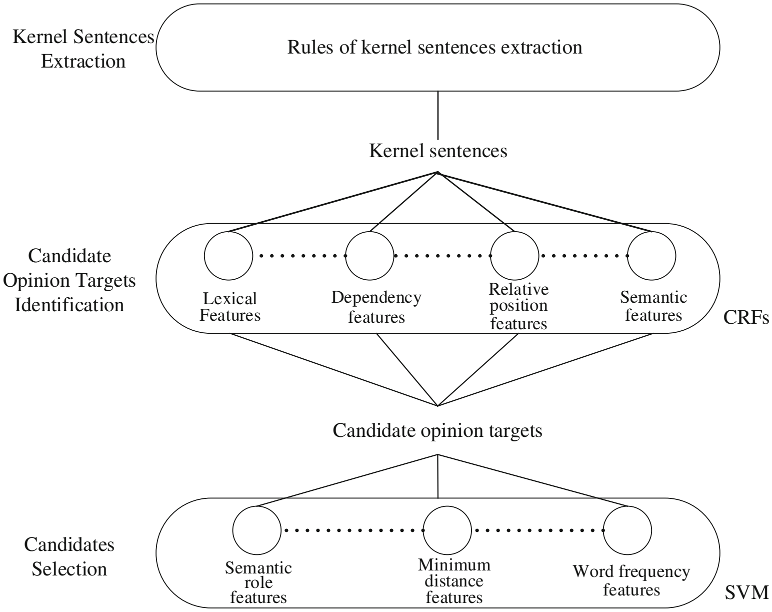
Kernel Sentences Extraction
Rules of kernel sentences extraction
Kernel sentences
Candidate Opinion Targets Identification
Lexical Features
Dependency features
Relative position features
Semantic features
CRFs
Candidate opinion targets
Candidates Selection
Semantic role features
Minimum distance features
Word frequency features
SVM

Fig. 2. Opinion targets identification based on kernel sentences extraction and candidates selection

As the quality of kernel sentences extraction and candidates selection greatly influences the result of opinion targets extraction, we make a comparison of different approaches of opinion targets identification in this section. We conduct experiments of opinion targets identification with methods of CRF-based method (CRF), performing kernel sentence before the CRF-based method (KSE + CRF), performing kernel sentence before the CRF-based method and conducting candidate opinion targets selection

Table 3. Comparing results of opinion targets identification using different methods

Method	Strict evaluation			Lenient evaluation		
	Precision	Recall	F-measure	Precision	Recall	F-measure
CRF	0.6864	0.4513	0.5446	0.7398	0.4712	0.5757
KSE + CRF	0.6925	0.4598	0.5527	0.7536	0.4873	0.5919
KSE +CRF + CS	**0.7035**	**0.4668**	**0.5612**	**0.7732**	**0.4963**	**0.6046**

after the CRF-based method (KSE + CRF + CS). The comparing results of these four approaches are represented in Table 3.

It can be seen that the effect of opinion targets extraction is highly improved after adding kernel sentences extraction and candidate opinion targets selection, which is probably because they can standardize the corpus and reduce the noise of opinion targets identification. This method not only uses kernel sentences extraction method in Sect. 3 to standardize the corpus, but also adopts machine learning method of SVM to screen the candidate opinion targets and so as to reach a higher precision, recall and F-measure. So this experiment strongly demonstrates the effectiveness and applicability of combination of kernel sentences extraction and candidates selection method.

6 Conclusions and Future Work

In this paper we propose a novel opinion targets identification method which takes kernel sentences extraction and candidates selection into consideration. We extract the kernel sentences of the oral opinion sentences through rules, and screen all the candidate opinion targets based on SVM classifier after CRF-based methods, finally acquire the opinion targets identification results. The experimental results show that it performs better than other baseline approaches.

In the future work, we will excavate more kernel sentence extraction rules and features for opinion targets identification.

References

1. CNNIC. Thirty-seventh statistical report on the development of China internet network [EB/OL]. http://cnnic.cn/gywm/xwzx/rdxw/2015/201601/W020160122639198410766.pdf
2. Zhao, Y., Qin, B., Liu, T.: Sentiment analysis. J. Softw. **21**(8), 1834–1848 (2010)
3. Minqing, H., Bing, L.: Mining opinion features in customer reviews. In: Proceedings of American Association for Artificial Intelligence, pp. 755–760. AAAI Press (2004)
4. Zhuang, L., Jing, F., Zhu, X.Y.: Movie review mining and summarization. In: Proceedings of the 15th ACM International Conference on Information and Knowledge Management, pp. 43–50 (2006)
5. Hongyu, L., Yanyan, Z., Bing, Q., Ting, L.: Comment target extraction and sentiment classification. J. Chin. Inf. Process. **24**(1), 84–88 (2010)

6. Li, B., Zhou, L., Feng, S., Wong, K.F.: A unified graph model for sentence-based opinion retrieval. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 1367–1375 (2010)
7. Fangtao, L., Chao, H., Minlie, H., et al.: Structure-aware review mining and summarization. In: Proceedings of the 23rd International Conference on Computational Linguistics, pp. 653–661 (2010)
8. Tengfei, M., Xiaojun, W.: Opinion target extraction in Chinese news comments. In: Proceedings of the 23rd International Conference on Computational Linguistics, pp. 23–27 (2010)
9. Xu, B., Zhao, T.J., Wang, S.Y., et al.: Extraction of opinion targets based on shallow parsing features. Zidonghua Xuebao/acta Automatica Sinica **37**(10), 1241–1247 (2011)
10. Jakob, N., Gurevych, I.: 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, pp. 1035–1045 (2010)
11. Wang, R., Jiupeng, J.U., Shoushan, L.I., et al.: Feature engineering for CRFs based opinion target extraction. J. Chin. Inf. Process. **26**(2), 56–61 (2012)
12. Song, H., Shi, N.S.: Comment object extraction based on pattern matching and semi-supervised learning. Comput. Eng. **39**(10), 221–226 (2013)
13. Xu, L., Liu, K., Lai, S., et al.: Mining opinion words and opinion targets in a two-stage framework. In: Meeting of the Association for Computational Linguistics, pp. 1764–1773 (2013)
14. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional random fields: probabilistic models for segmenting and labelling sequence data, pp. 282–289 (2001)
15. Che, W., Li, Z., Liu, T.: LTP: a Chinese language technology platform. In: Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations, pp. 13–16 (2010)