A Cyclic Cascaded CRFs Model for Opinion Targets Identification Based on Rules and Statistics

Hengxun Li, Chun Liao, Guangjun Hu, Ning Wang

A Cyclic Cascaded CRFs Model for Opinion Targets Identification Based on Rules and Statistics

Hengxun Li[1], Chun Liao[2], Guangjun Hu[1], Ning Wang[1]

1 First Research Institute of the Ministry of Public Security of PRC
capital gymnasium south road NO. 1 ,Haidian District
Beijing 100048, China
2 Institute of information engineering, Chinese Academy of Sciences
minzhuang Road No. 89 ,Haidian District
Beijing 100091, China
DerekLee1985@126.com, liaochun@iie.ac.cn, cityof93@qq.com, wn_1209@163.com

Abstract. Opinion sentences on e-commerce platform, microblog and forum contain lots of emotional information. And opinion targets identification plays an import role in huge potential commercial value mining, especially in sales decision making and development trend forecasting. Traditional CRFs-based method has achieved a pretty good result to a certain extent. However, its discovery ability of out-of-vocabulary words and optimization of the mining model are both insufficient. We propose a novel cyclic cascaded CRFs model for opinion targets identification which incorporates rule-based and statistic-based methods. The approach acquires candidate opinion targets through part-of-speech, syntactic and semantic rules, and integrates them in a cyclic cascaded CRFs model for the accurate opinion targets identification. Experimental results on COAE2014 dataset show the outperformance of this method.

Keywords: opinion targets identification, cyclic cascaded CRFs model, rule-based, statistic-based

1 Introduction

With the development of the Internet, social platform has gradually integrated into people's lives, resulting in the increasing expansion of mass information. More and more opinion sentences on the Internet are generating. For the government, business or individual, the study of these opinion words is of great significance. Compared with regular grammar and news text, opinion sentences on social plat-form are more colloquial, interactive, and also contain a large number of advertisements and junk in-formation. These bring new challenge to opinion targets identification, and how to effectively extract the useful information has become more and more important.

Sentiment analysis, also called opinion mining, is to process, induce and infer the subjective texts[1]. Sentimental elements extraction is the basis of sentiment analysis. Sentimental elements extraction is to extract the opinion elements in the sentence, including opinion words (such as “好”), opinion targets (such as “三星手机”), opinion holder (such as “张三” in the sentence “张三认为……”). In this paper, we mainly study opinion targets identification.

Traditional CRFs-based method has achieved a pretty good result to a certain extent.

However, its discovery ability of out-of-vocabulary words and optimization of the mining model are both insufficient. We propose a novel cyclic cascaded CRFs model for opinion targets identification which incorporates rule-based and statistic-based methods. The approach acquires candidate opinion targets through part-of-speech, syntactic and semantic rules, and integrates them in a cyclic cascaded CRFs model for the accurate opinion targets identification.

Existing opinion targets identification methods cannot comprehensively discovery the out-of-vocabulary words, and do not optimize the mining model. To address these shortcomings, it is intuitive to consider the combination of rule-based and statistic-based methods, and at the same time take special features of opinion sentences on social platform into consideration. In this paper, we propose a novel cyclic cascaded CRFs model for opinion targets identification which incorporates rule-based and statistic-based methods. The approach acquires candidate opinion targets through part-of-speech, syntactic and semantic rules, and integrates them in a cyclic cascaded CRFs model for the accurate opinion targets identification. 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

The methods of opinion targets extraction are mainly divided into two categories: unsupervised and supervised methods. In the unsupervised methods, Hu and Liu[2] used association rules to excavate opinion targets and regarded the top-frequency words as opinion targets. Li and Zhou[3]extracted tuples like <emotional words, opinion targets> based on emotional and topic-related lexicons. Popescu and Nguyen[4] extracted properties of products with mutual information. Yao[5] used domain ontology to extract the topics and their attributes from a sentence, and summed up the subject-predicate structure rules based on syntactic analysis for opinion targets identification. Liu[6] used syntactic analysis to obtain the candidates , and then combined PMI with noun pruning algorithm to decide the final opinion targets. Besides, in the supervised methods, Zhuang[7] proposed a multi-knowledge-based approach which integrated WordNet, statistical analysis and movie knowledge. Jakob[8] modelled the task as a sequence labelling question and employed CRFs for opinion targets extraction. Wang[9] proposed a method of opinion targets identification based on CRFs, and selected morphology, dependency, relative position and semantics as features.

However, existing opinion targets extraction methods only took lexical-related features into account. Consequently, considering the specific features of Chinese microblog, we propose a new method for opinion targets extraction towards microblog using syntax and semantics in which we adopt a new approach of PDSP for domain lexicon construction and select groups of features for CRFs.

3 Candidate Opinion Targets Identification Based on Rules

The task of candidate opinion targets identification is automatically extracting the opinion targets using rule-based methods. Considering the importance of syntax and

semantics in opinion targets identification, we propose a method of candidate opinion targets identification which incorporates POS, dependency structure and semantic role. Opinion targets are usually nouns or noun phrases. Through statistics on corpus, we design six templates based on Part-of-Speech which are shown in Table 1 where n, adj, adv, aw,cmp and OTrepresents for noun,adjective, degree adverb, advocating word, comparative word and opinion target. Here we get adv and aw from Hownet , and acquire cmp from [10].

Template	Example	Template	Example
n+adv+adj	屏幕/OT 很好	adj+的+n	轻薄的机身/OT
n+adj	外观/OT 漂亮	n+cmp+n	iphone/OT 不如三星/OT
aw+n	认为蒙牛/OT	n+n	蒙牛牛奶/OT

Table 1. *Part-of-Speech sequence templates*

As we all know, when we express opinions towards a product, we need some opinion words which usually have strong semantic relation with the opinion targets. Therefore, we collect opinion words from Hownet and NTUSD[1]andperform HIT-LTP[2] for dependency parsingto discuss the relation "ATT" and "SBV" between opinion words and opinion targets, relation "COO" between already known opinion targets and unknown opinion targets.

As a necessary part of shallow semantic parsing, sematic role[11] occupies an important position in lexical and semantic analysis. People usually express opinions through opinion words in opinion sentences. And adjective andverb are two main forms of opinion words. Through investigation, we find when the opinion word are adjective, A0(agent) is opinion target. Furthermore, when the opinion word are verb, A1(patient) is opinion target.

4 CCCRFs: A Cyclic Cascaded CRFs Model for Opinion Targets Identification

In this paper, we propose a novel cyclic cascaded CRFs model for opinion targets identification which incorporates rule-based and statistic-based methods. The approach first acquires candidate opinion targets through method in 3. And then, this model adopts two-layer cascaded CRFs. In the first layer, we select opinion words and manual features for opinion words identification. And in the second layer, the outputs of the first layer are added as input and we select opinion words, candidate opinion targets and manual features for opinion targets identification. In each iteration, we choose sentences whose confidence value is larger than C as training data. And the remaining sentences are regarded as testing data.

4.1 Cascaded CRFs Model

CRFs(Conditional Random Fields, CRFs) is proposed by Lafferty[12] in 2001. Its chain structure is shown in Figure 1.

[1] http://www.datatang.com/data/11837
[2] http://www.ltp-cloud.com/

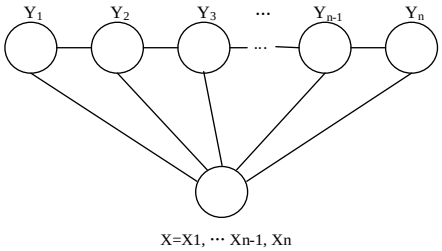

Fig.1. CRFs Model

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. CRFs are often used for sequence labelling tasks, such as part-of-speech tagging, named entity recognition and so on.

However, in the actual labelling process, we find it exists nesting phenomenon. For example, in Chinese named entity recognition, other named entities elements, such as names, places will be included in organization name. And this situation has led to incorrect identification. To solve this problem, you need to refine these tasks, step by step and gradually completed. In this paper, we adopt CCRFs(Cascaded Conditional Random Fields) as shown in Figure 2 to solve the above problems.

CCRFs reduced the coupling relationship between different layers of the model. Each layer of the model can be built independently, and each sub task can be done independently without interfering with each other. The complexity has linear relationship with the number of the model layers. Before the high-level model, some of the necessary pre-treatment can be carried out in the output of the underlying model and filter some errors. Consequently, CCRFs can avoid error propagation and diffusion and further improve the performance.

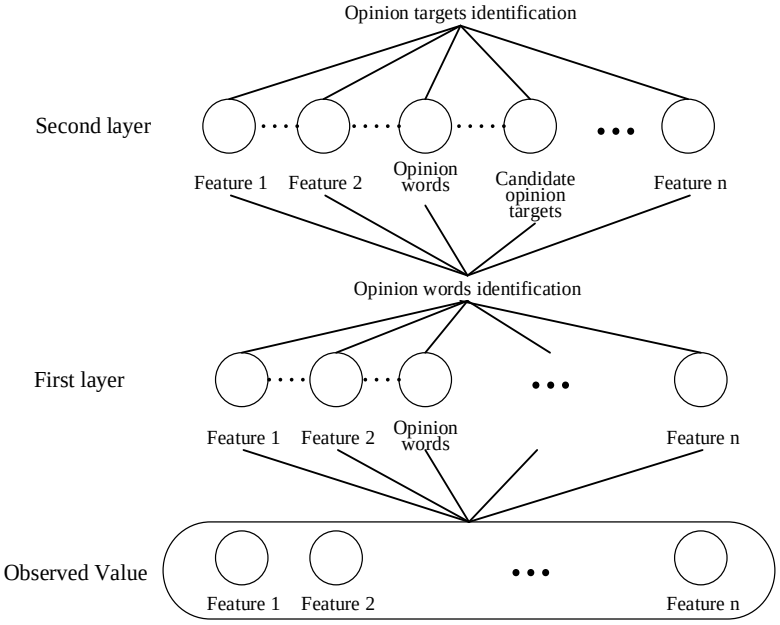

Fig.2. Cascaded CRFs Model

4.2 Cyclic Model

The selection of training data is always a focus of machine learning methods. In order to improve the recall rate of this method, we add cyclic method under CCRFs through screening each experimental results. Figure 3 is the flowchart of Cyclic Cascaded CRFs Model. If the confidence C is greater than threshold, we add them into training corpus circularly. If not, we treat them as testing data. Then, this model loops as this until the iteration number reaching a certain value N. In cyclic model, for every opinion targets identification results, its confidence degrees C, can be calculated as follows.

$$C = C_1 \times C_2 \quad (1)$$

C_1 is the confidence value of first layer in CCRF, and C_2 is the confidence value of second layer in CCRFs. These can be acquired from CRFs tool.

We selected sentences whose confidence value larger than M into training set, and the remainder continued as the testing data. Then we re-trained the model, and extract the results for N times iteration to get the final results.

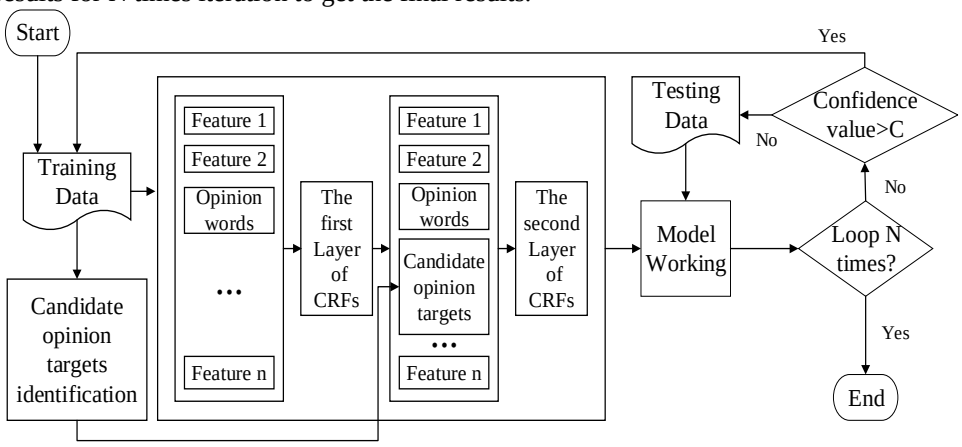

Fig.3. Cyclic Cascaded CRFs Model

4.3 Feature Selection

In feature selection, we refer to the features which are employed by Jakob[8] and Lu[13] 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.

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. Dependency parsing reflects the semantic dependency relations between core word and its subsidiaries words[14]. Consequently, 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. As we all know, since words which appear around emotional words are more likely to be opinion targets, we determine the

boolean value by judging whether the distance between current word and emotional word is less than 5. Considering there 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 CRFs.

5 Experiments and Analysis

In experiments, we firstly obtained candidate opinion targets through method in section 3, and then we employed CCCRFs with the candidate opinion targets and features in section 4 together to extract opinion targets.

5.1.Dataset

Through filtration, we finally obtain 5,000 normalized sentences with opinion orientation from COAE2014. And we perform segmentation, part-of-speech, syntactic parsing and semantic role labelling through LTP[15].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.

5.2. Parameter selection

In this section, we make comparing experiment on parameter selection of cyclic cascaded CRFs model. And the experiment results between loop number N and confidence value C in cyclic cascaded CRFs model are in Figure 4.

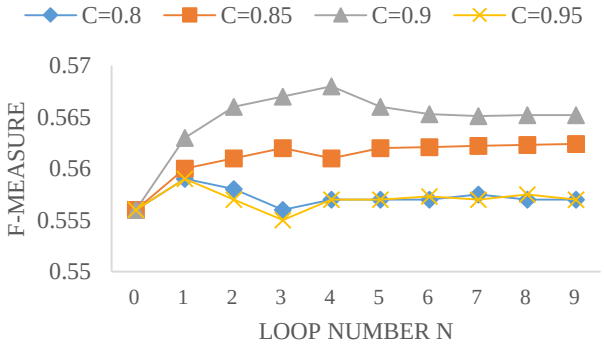

Fig.4.Comparing experiments between loop number and confidence value in cyclic cascaded CRFs model

From this figure we find that, it performs best when the confidence value C=0.9 and loop number N=4. We also find the curve is slowly tend to be stable under four confidence value C with the increase of loop number N. As the confidence value C increases, opinion sentences have been added to the training corpus completely and the training corpus is no longer increased, so the model tends to a stable state.

Consequently, this experiment not only demonstrated the effectiveness of cyclic cascaded CRFs model, but also revealed the importance of parameter selection to opinion targets identification.

5.3. Comparing results with different methods of opinion targets identification

In this section, we compare different methods of opinion targets identification, rule-based method, CRFs-based method, cascaded CRFs-based method and cyclic cascaded CRFs-based method. To verify the effectiveness of cyclic cascaded CRFs model for opinion targets identification, we conduct experiments under different methods of opinion targets identification and obtain the experimental results as shown in Table 2. And the rule, CRFs, CCRFs and CCCRFs respectively represent for rule-based method, CRFs-based method, cascaded CRFs-based method and cyclic cascaded CRFs-based method.

Method	Strict Evaluation			Lenient evaluation		
	Precision	Recall	F-measure	Precision	Recall	F-measure
Rule	0.5870	0.3206	0.4147	0.6025	0.3971	0.4787
CRFs	0.6780	0.4325	0.5281	0.7115	0.4600	0.5490
CCRFs	0.6985	0.4595	0.5403	0.7674	0.4935	0.5729
CCCRFs	0.7085	0.4752	0.5689	0.7803	0.5025	0.6113

Table 2.Results of opinion targets identification with different methods

It can be seen that the effect of opinion targets identification is highly improved after adopting cyclic cascaded CRFs model, which is mainly because this method not only uses candidate opinion targets identification in section 3to obtain candidate opinion targets, but also adopts machine learning method of CCCRFs to make up for the defect of rule-based method and so as to reach a higher precision, recall and F-measure. So this experiment strongly demonstrates the effectiveness and applicability of cyclic cascaded CRFs model.

6 Conclusions and Future Work

In this paper we propose a cyclic cascaded CRFs model for opinion targets identification which takes rules and statistics into consideration. We combine the candidate opinion targets extraction into a cyclic cascaded CRFs model to get the final opinion targets. The experimental results show that it performs better than other baseline approaches.

In the future work, we will take the following points into consideration:

– Considering the various expressions of Chinese microblog, we should excavate more rules and extract the kernel sentence for opinion targets extraction.

– In this paper, we perform opinion targets extraction on sentence level. We will investigate the effect of opinion targets extraction on corpus level.

References

1. Zhao Y,Qin B, Liu T. Sentiment Analysis[J]. Journal of Software, 2010, 21(8): 1834-1848.
2. Hu M, Liu B. Mining and summarizing customer reviews[C]// Tenth ACM

SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, Usa, August. 2004:168-177.
3. Li, Binyang, Zhou, Lanjun, Feng, Shi, et al. A unified graph model for sentence-based opinion retrieval[C]// Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2010:1367-1375.
4. Popescu A M, Etzioni O. Extracting Product Features and Opinions from Reviews[M]// Natural Language Processing and Text Mining. Springer London, 2007:9-28.
5. Yao T, Lou D. Research on Semantic Orientation Analysis for Topics in Chinese Sentences[J]. Journal of Chinese information processing, 2007, 21(5): 73-79.
6. Liu, H., Zhao, Y., Qin, B. & Liu, T.(2010).Comment Target Extraction and Sentiment Classification, Journal of Chinese information processing, 24(1): 84-88.
7. Zhuang, L., Jing, F., & Zhu, X. Y. (2006, November). Movie review mining and summarization. In Proceedings of the 15th ACM international conference on Information and knowledge management, pages:43-50.
8. Jakob, N., &Gurevych, I. (2010, October). Extracting opinion targets in a single- and cross-domain set-ting with conditional random fields. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages:1035-1045.
9. Wang Rongyang, JuJiupeng, Li Shoushan, Zhou Guodong.Feature Engineering for CRFs Based Opinion Target Extraction[J]. Journal of Chinese information processing, 2012, 26(2): 56-61.
10. Zhang, C., Feng, C., Liu, Q., Shi, C., Huang, H. & Zhou, H.(2013). Chinese Comparative Sentence Identification Based on Multi-feature Fusion,Journal of Chinese information processing, 27(6):110-116.
11. Hacioglu, K. (2004, August). Semantic role labelling using dependency trees. Computational Linguistics, pages: 1273.
12. John D. Lafferty, Andrew McCallum, Fernando C. N. Pereira. Conditional Random Fields: Probabilistic Models For Segmenting And Labelling Sequence Data[C]// 2001:282--289.
13. Lu, B. (2010, June). Identifying opinion holders and targets with dependency parser in Chinese news texts. In Proceedings of the NAACL HLT 2010 Student Research Workshop, pages: 46-51.
14. Li, X., & Roth, D. (2002, August). Learning question classifiers. In Proceedings of the 19th international conference on Computational linguistics, pages: 1-7.
15. Che, W., Li, Z., & Liu, T. (2010, August). Ltp: A chinese language technology platform. In Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations, pages: 13-16.