A Semi-supervised Method for Opinion Target Extraction

Tao Ge
MOE Key Laboratory of Computational Linguistics, School of Electronics Engineering and Computer Science, Peking University, Beijing, China
getao@pku.edu.cn

Wenjie Li
MOE Key Laboratory of Computational Linguistics, School of Electronics Engineering and Computer Science, Peking University, Beijing, China
lwj@pku.edu.cn

Zhifang Sui*
MOE Key Laboratory of Computational Linguistics, School of Electronics Engineering and Computer Science, Peking University, Beijing, China
szf@pku.edu.cn

ABSTRACT

This paper proposes a semi-supervised self-learning method, which is based on a Naive Bayes classifier exploiting context features and PMI scores, to extract opinion targets. The experimental results indicate our bootstrapping framework is effective for this task and outperforms the state-of-the-art models on *COAE2008 dataset2*, especially in precision.

Categories and Subject Descriptors

I.2.7 [**Computing Methodologies**]: Artificial Intelligence—*Natural Language Processing*

Keywords

Opinion target extraction, Self-learning, Bootstrapping

1. INTRODUCTION

Identifying opinion targets is an important subtask for sentiment analysis, attracting much attention in recent years. In general, approaches for this task can be divided into two categories: supervised and unsupervised methods. Supervised methods regard opinion target extraction as a sequence labeling task but they usually require a large amount of annotated data for training and thus are less applicable compared with unsupervised methods which are based on co-occurrence of opinion words. The state-of-the-art method is an unsupervised algorithm named *Double Propagation* [1] exploiting dependency information to extract the relational pairs of opinion words and opinion targets. However, this algorithm has some limitations: it largely depends on the parsing performance. In a large corpus, parsing is time-consuming and might introduce many parsing errors especially for on-line reviews including shorthand and typos. Hence, the performance of parsing would be undesirable and the parsing errors would be accumulated in bootstrapping phase, leading to a low precision.

For tackling the problems of *Double Propagation*, this paper exploits linguistic context features which are important in supervised methods but are usually ignored by unsupervised methods, and proposes a semi-supervised bootstrapping framework to extract opinion targets. In our method, a handful of manually annotated pairs of noun/noun phrase and adjective which co-occur in the corpus are used for training an initial binary Naive Bayes classifier which exploits the distributional context of each noun and adjective pair as features. In the bootstrapping phase, we use the prediction of the Naive Bayes classifier combined with PMI score to identify the most reliable predictions and add the instances to the corresponding class for enlarging the training set. By repeating this process, the classifier becomes stronger. According to experimental results, our method outperforms the methods based on either dependency parsing or PMI.

2. THE PROPOSED ALGORITHM

2.1 Bootstrapping Framework

As a prerequisite for extracting opinion targets, we first parse the review corpus and extract all noun/noun phrase and adjective pairs which co-occur in one sentence within a window size K. For each pair, we use all the occurrences of their context words and POS tags in the corpus as its global features. To construct a training set L, we randomly sampled a small set of pairs and manually annotated them. The pairs which are actually constituted by opinion words and opinion targets are labeled as positive instances and the others are labeled as negative. The rest of unlabeled pairs U are regarded as the candidates.

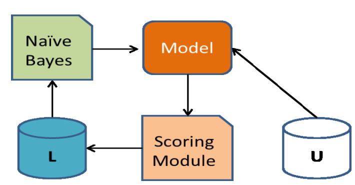

Figure 1: The framework of bootstrapping

The framework is shown in figure 1. In the scoring module, the prediction of Naive Bayes classifier and PMI score are combined to compute the confidence of each candidate pair

*Corresponding author

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2577337.

	Camera			Car			Laptop			Phone		
	P	R	F	P	R	F	P	R	F	P	R	F
Hu	0.63	0.65	0.64	0.62	0.58	0.60	0.51	0.67	0.58	0.69	0.6	0.64
DP	0.71	0.70	0.70	**0.72**	0.65	0.68	0.58	0.69	0.63	0.78	0.66	0.72
Zhang	0.71	0.78	0.74	0.69	**0.68**	0.68	0.57	**0.8**	0.67	0.8	0.71	0.75
Ours	**0.74**	**0.79**	**0.76**	**0.72**	**0.68**	**0.70**	**0.72**	0.74	**0.73**	**0.84**	**0.75**	**0.79**

Table 1: Experiments on COAE2008 dataset2

as shown in (1).

$$score(p) = \lambda \times PMI(p) + (1 - \lambda) \times P(positive|p) \qquad (1)$$

For the candidate p whose score is larger than a threshold t_1, we add it as a positive instance to the training set L. For the one whose score is less than a threshold t_2, it is added as a negative instance to L.

In the training phase, we use context of pairs as features. According to the analysis of the corpus, it is found that the context is usually similar when opinion words modify the opinion targets in reviews. Thus, linguistic context features are useful for capturing opinion relational pair of opinion target and opinion word. For a noun/noun phrase and adjective pair instance $<W_1,W_2>$, features shown in table 2 are used in the Naive Bayes classifier.

Position	Feature	Description
Before	B_Null	no word before W_1
	B_Word	K words before W_2
	B_POS	K word POS tags before W_1
Middle	M_Null	no word between W_1 and W_2
	M_Word	K word between W_1 and W_2
	M_POS	K word POS between W_1 and W_2
After	A_Null	no word after W_2
	A_Word	K word after W_2
	A_POS	K word POS tags after W_2

Table 2: Context features used by the Naive Bayes classifier

2.2 Candidate Estimation

In order to select the final opinion targets from the resulting pairs of noun/noun phrase and adjective in section 2.1, we use a graph-based algorithm to compute the confidence of each opinion target candidate. Like [2], we use PageRank to compute confidence of each node on a bipartite graph consisted of nouns/nouns phase nodes and adjective nodes. The weight on edges is estimated by (1). Specially, we set the confidence of nodes initially labeled as positive to 1.

2.3 Pattern-based Extraction

For improving recall, our work also uses patterns to extract more opinion target candidates as [2] did. PMI between the candidates and their corresponding class concept words is computed as confidence. A candidate with a higher PMI score is more likely to be an opinion target.

3. EXPERIMENTS

3.1 Experimental Setting

We test our approach on *COAE2008 dataset2* which contains 473 reviews of four different products: Camera, Car, Laptop and Phone. For each kind of reviews, we extract all pairs of adjacent noun/noun phrase and adjective and label 10% of them as input. For the parameters, we empirically set the window size $K = 4$, the interpolation $\lambda = 0.5$, positive threshold $t_1 = 0.6$ and negative threshold $t_2 = 0.4$.

3.2 Experimental Results

We compare our method to some typical methods. Hu [3] is the pioneering method using association rule mining to extract opinion targets. Zhang [2] is the method enhancing DP[1] by utilizing some additional linguistic patterns. The experimental results are reported in table 1.

According to table 1, the performance of our method is better than the previous work, demonstrating that our semi-supervised self-learning framework is effective for opinion target extraction. In table 1, Hu achieves a worse performance than other methods due to its limitation of association rules. In contrast, bootstrapping framework, which has an iterative learning strategy with a few seed opinion targets and opinion words, is appropriate for this extraction problem and leads to a better result. In terms of precision, our method achieves a significant improvement. The reason, we believe, is that the Bayesian classifier with PMI can filter the bad candidates in the bootstrapping steps. Also, the classifier using the distribution context information contributes to the selection of more reliable candidates and can alleviate the error propagation caused by bootstrapping.

4. CONCLUSION

This paper presents a novel and effective bootstrapping framework to extract opinion targets, which outperforms the state-of-the-art models on *COAE2008 dataset2*.

Ackownledgements

This paper is supported by National Key Basic Research Program of China 2014CB340504 and NSFC project 61375074.

5. REFERENCES

[1] Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. Opinion word expansion and target extraction through double propagation. *Computational linguistics*, 37(1):9–27, 2011.

[2] Lei Zhang, Bing Liu, Suk Hwan Lim, and Eamonn O'Brien-Strain. Extracting and ranking product features in opinion documents. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 1462–1470. Association for Computational Linguistics, 2010.

[3] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM, 2004.