

A Novel Approach of Extracting Opinion Targets and Opinion Words from Online Review

Mingshuang Tang^{1, a}, Hongrong Cheng^{1, *}, Yanwei Guo¹, Jiaqing Luo¹ and Shijie Zhou¹

¹ School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China.

*hrcheng@uestc.edu.cn, ^amstang13@163.com

Keywords: opinion mining, sentiment analysis, opinion target, opinion word.

Abstract. In opinion mining, the precision of extracting opinion words and opinion targets from online review dataset can severely influence the precision of opinion mining systems. In this paper, we propose a graph-based filtering approach to extract effective opinion targets and opinion words from the candidate words, which can be mined by word-based translation model in a monolingual scenario. Experimental results on four real Chinese review datasets show that our approach is promising performance.

Introduction

Extracting opinion targets and corresponding opinion words is the kernel tasks in opinion mining, attracting much attention in recent years. In general, opinions and sentiments in textual reviews can be expressed at document, sentence, even aspect (or word) levels. In aspect-level sentiment analysis, the term indicating opinion, sentiment or attitude of customers is called opinion word, and the evaluated entity in a review is called opinion target [7]. In Table 1, the words marked in green color are opinion targets which correspond to the opinion words in red color respectively.

Table 1 Two Chinese hotel review samples to show the opinion relationship

The *environment* of the hotel is *so so*, and the *price* is *expensive*, but the *transportation* is *convenient*.

The *price* of this hotel is *expensive*, but the *service* is *nice*.

In this paper, firstly we apply a word-based translation model (WTM) [1] into monolingual scenario to mine the candidate opinion words and opinion targets and then evaluate the opinion associations between them. Then we propose an original algorithm to identify the term indicator of candidate opinion targets and candidate opinion words respectively. Thirdly, we calculate the energy of each candidate by using a graph-based algorithm. After ranking the candidates by the energy, the candidates with higher energy score are extracted as opinion words/targets. Our contributions are as follows: we propose a novel approach to calculate the term indicator of opinion targets and opinion words respectively, which combine the term distribution and term frequency. In contrast, the existing researches, to our knowledge, just take into account the opinion relations [2, 3] or the term distribution only [4].

Preparation

Generally, adjectives/verbs are extracted as candidate opinion words, and nouns/noun phrases are extracted as candidate opinion targets [3, 5, 6]. Liu [2, 3] formulate extracting candidate opinion words and opinion targets as a monolingual word alignment task.

After gathering all word pairs from the review sentences, the alignment probabilities between noun/noun phrase w_t and its modifiers w_o is $p(w_o|w_t) = \frac{\text{count}(w_o, w_t)}{\text{count}(w_t)}$. Similarly, the alignment probabilities $p(w_t|w_o) = \frac{\text{count}(w_o, w_t)}{\text{count}(w_o)}$. Thus, the opinion association can be calculated

$$\text{Association}(w_o, w_t) = \left(\frac{u}{p(w_o|w_t)} + \frac{(1-u)}{p(w_t|w_o)} \right)^{-1}. \quad (1)$$

Where u is the harmonic factor to combine these two alignment probabilities, which we set 0.5.

Methodology

Calculate Opinion Target's Indicator. If a candidate target t_i has higher frequency and homogeneous distribution in the corpus, it will have higher probability to be an opinion target, we apply Information Entropy to formulize this theory. Given the review corpus with n documents $D = \{d_1, d_2, \dots, d_n\}$, every document d_j regarded as independent category, if he candidate target t_i in the document d_j frequent and distribute more homogeneous, the Entropy score $IE(t_i)$ will be higher.

$$IE(t_i) = - \sum_{j=1}^n p(d_j, t_i) \log p(d_j, t_i) \quad (2).$$

Where $p(d_j, t_i) = \frac{tf_{ij}}{\sum_{j=1}^n tf_{ij} + \varepsilon}$ is the probability that t_i in d_j . We set the constant factor $\varepsilon = 0.0001$, which make sure $IE(t_i) \neq 0$. Since $IE(t_i)$ score tends to mine the high-frequency candidates, to address this problem, we argue that opinion targets are usually domain-specific, thus m domain-independent corpus are applied to further improve the accuracy of opinion targets.

$$Ds(t_i)_j = \begin{cases} \frac{\log(1 + df_{in})}{\log(1 + df_{out,j})} & \text{if } df_{out,j} \neq 0 \\ \alpha \times \log(1 + df_{in}) & \text{otherwise} \end{cases} \quad (3).$$

Thus the target's important indicator $I(t_i)$ can be denoted as

$$I(t_i) = IE(t_i) \times \frac{1}{m} \sum_{j=1}^m Ds(t_i)_j \quad (4).$$

df_{in} is the document frequency in review corpus of t_i , $df_{out,j}$ is the document frequency in the j th domain-independent corpus of t_i , the harmonic parameter $\alpha > 1$, we set $\alpha = 2$. In order to guarantee the fairness of document frequency, the scale of m domain-independent corpus must be equal to our review corpus.

Calculate Opinion Word's Indicator. Different from opinion target, most of the opinion words are domain-independent, like "good", "love", "hate", etc. Meanwhile, there are a few domain-dependent opinion words. So we use a method which combines the term frequency and distribution to evaluate the important indicator of opinion words.

$$I(o_i) = \log(1 + df_i) \times D_i \quad (5).$$

Where df_i is the document frequency in the review corpus of the i th candidate opinion word in the candidates list. The candidates' distribution denotes as $D_i = \frac{\bar{tf}_i}{\sqrt{\frac{\sum_{j=1}^n (tf_{ij} - \bar{tf}_i)^2}{(n-1)}}}$, $\bar{tf}_i = \frac{1}{n} \times \sum_{j=1}^n tf_{ij}$ means

the average frequency across all documents for o_i .

Estimate The Energy Of Candidate. To model the mined informations between target candidates and opinion word candidates, we employ Random Walking with Restart algorithm in a weighted bipartite graph like Liu's method [3].

$$\begin{aligned} E(t) &= \lambda \times R \times E(o) + (1-\lambda) \times I_t \\ E(o) &= \lambda \times R^T \times E(t) + (1-\lambda) \times I_o \end{aligned} \quad (6)$$

Where $E(t)$ and $E(o)$ represent the energy of opinion targets and opinion words respectively, where R is the association matrix and each item in R is computed using Eq. (1). I_t is the target candidates' Indicator vector in which each item is computed by Eq. (4), I_o means the opinion word

candidates' Indicator vector and each item in \mathbf{I} is calculated using Eq.(6). $\lambda \in [0, 1]$ is the proportion of candidate indicator in the candidate energy. The candidates with higher energy scores than the threshold are confirmed as the opinion targets / words.

Experiments

Data Preparation and Evaluation Metrics. We crawled four different domain Chinese reviews from four web sites to evaluate our method, all documents from our review corpus as well as the domain-independent corpus were tokenized and parsed using a Chinese Segmentation System NLPIR [8]. And we randomly selected 3000 sentences from each review corpus to generate testing dataset. Then opinion targets/words expressed in every review sentence are manually annotated as the gold standard for evaluations. We apply precision (P), recall (R) and F-measure (F) as evaluation metrics in this paper.

Experiment of Extraction Opinion Targets. As space is limited, we just compare our results with some state-of-the-art methods to prove the effectiveness of filtering opinion targets. *Liu* [2, 3] is the pioneering method used unsupervised WTM to mining the associations between words and use tf-idf calculate importance of words. *EDR*, *IDR* and *IEDR* are three methods described in [4]. The experimental results are reported in table 2. From the results, we can see that for all datasets, *Ours* outperforms others. Though *Liu* took opinion association into account as well as the importance of words, they straightforwardly calculated the importance of candidate opinion targets by tf-idf and didn't compute the candidate opinion words' importance. We can also observe that *IEDR* performance improvement over *IDR* and *EDR*. This indicates that the opinion targets are not only closely related to the given corpus, but also have domain independence. In *Liu*'s method, calculate the tf-idf score need a large-scale corpus, which is time consuming. In *EDR* and *IEDR* the size and topic of the domain-independent corpus are significantly affect extraction performance. However, in our method, the domain independent corpus' size are same as the given review corpus, and we collect 5 corpus which include multiple topics, so we avoid discussing the size and the topic of domain independent corpus.

We show the Precision, Recall and F-measure of *Ours* on four corpus when different λ in Figure 1. From the shape of the curves, we can see that the best performance is obtained when λ is around 0.4. It turned out that candidates' indicator and candidates' opinion association are both important for candidates' energy.

From the results, we can see that for all the four domains, our method outperforms other methods. This indicates that our method combined opinion association and the indicator of words is effective for extraction words.

Table 2 Experimental results of opinion target extracting on four datasets

	Restaurant			Hotel			Book			Camera		
	P	R	F	P	R	F	P	R	F	P	R	F
<i>Liu</i>	0.60	0.85	0.71	0.67	0.78	0.72	0.60	0.85	0.71	0.54	0.77	0.63
<i>IDR</i>	0.51	0.77	0.61	0.61	0.76	0.78	0.57	0.78	0.66	0.62	0.57	0.59
<i>EDR</i>	0.53	0.63	0.58	0.65	0.80	0.71	0.59	0.77	0.67	0.63	0.61	0.63
<i>IEDR</i>	0.55	0.81	0.66	0.67	0.84	0.75	0.60	0.82	0.69	0.67	0.60	0.64
<i>Ours</i>	0.63	0.85	0.72	0.70	0.87	0.78	0.64	0.91	0.76	0.60	0.76	0.67

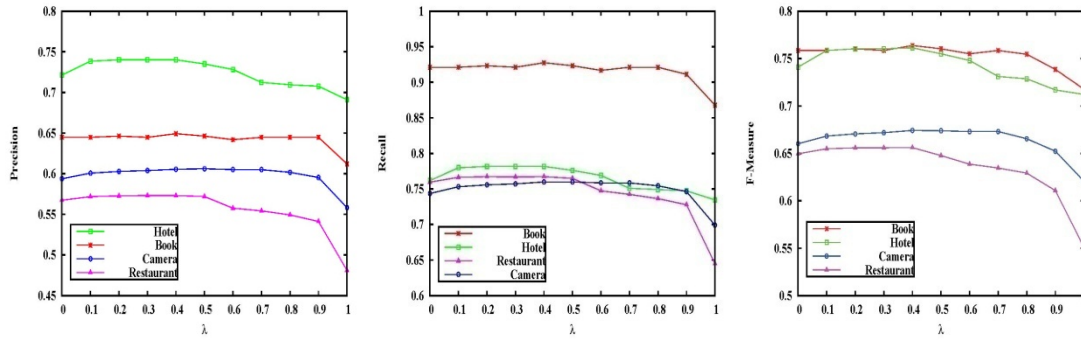


Fig. 1 The Experimental Results Influenced By λ

Conclusion

This paper presents a novel and effective work to extract opinion targets and opinion words, which outperforms the pioneering models on four Chinese reviews.

Acknowledgements

This research is supported by National Nature Foundation under Grant 61133016, 61300094 and the Fundamental Research Funds for the Central Universities under Grant ZYGX2013J083.

References

- [1] Peter E. Brown, Stephen A. Della Pietra and etc. The Mathematics of Statistical Machine Translation: Parameter Estimation. In Proceedings of ACL, 1993.
- [2] K. Liu, L. Xu, etc. Opinion Target Extraction Using Partially-Supervised Word Alignment Model. In Proceedings of IJCAI, China, 2013.
- [3] K. Liu, L. Xu, and J. Zhao. Opinion target extraction using word-based translation model. In Proceedings of EMNLP, Korea, 2012.
- [4] Z. Hai, K. Chang, and etc. Identifying Features in Opinion Mining via Intrinsic and Extrinsic Domain Relevance. IEEE TKDE, 2014.
- [5] K. Liu. Extracting Opinion Targets and Opinion Words from Online Reviews with Graph Co-ranking. In Proceedings of ACL, USA, 2014.
- [6] M. Hu, B. Liu. Mining and summarizing customer reviews. In Proceedings of KDD, USA, 2004.
- [7] L. Fang; M. Huang and X. Zhu. Exploring Weakly Supervised Latent Sentiment Explanations for Aspect-level Review Analysis. In Proceedings of CIKM, 2013.
- [8] <http://ictclas.nlp.ir.org/>.