A Word Vector and Matrix Factorization Based Method for Opinion Lexicon Extraction

Zheng Lin
Institute of Information Engineering, CAS
No.91(A), Minzhuang Road
Beijing, China
linzheng@iie.ac.cn

Weiping Wang
Institute of Information Engineering, CAS
No.91(A), Minzhuang Road
Beijing, China
wangweiping@iie.ac.cn

Xiaolong Jin
Institute of Computing Technology, CAS
No.6, Kexueyuan South Road
Beijing, China
jinxiaolong@ict.ac.cn

Jiguang Liang
Institute of Information Engineering, CAS
No.91(A), Minzhuang Road
Beijing, China
liangjiguang@iie.ac.cn

Dan Meng
Institute of Information Engineering, CAS
No.91(A), Minzhuang Road
Beijing, China
mengdan@iie.ac.cn

ABSTRACT

Automatic opinion lexicon extraction has attracted lots of attention and many methods have thus been proposed. However, most existing methods depend on dictionaries (e.g., WordNet), which confines their applicability. For instance, the dictionary based methods are unable to find domain dependent opinion words, because the entries in a dictionary are usually domain-independent. There also exist corpus-based methods that directly extract opinion lexicons from reviews. However, they heavily rely on sentiment seed words that have limited sentiment information and the context information has not been fully considered. To overcome these problems, this paper presents a word vector and matrix factorization based method for automatically extracting opinion lexicons from reviews of different domains and further identifying the sentiment polarities of the words. Experiments on real datasets demonstrate that the proposed method is effective and performs better than the state-of-the-art methods.

Categories and Subject Descriptors

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

Keywords

Opinion Word; Matrix Factorization; Word Vector

1. INTRODUCTION

Opinion lexicon is a crucial resource for sentiment analysis. Although there are several opinion lexicons publicly available, it is hard to maintain a universal opinion lexicon to cover all domains, as sentiment polarities of words may vary significantly from domain to domain. For example, the opinion word, *unpredictable*, is likely to be positive in a movie review but negative in a car review. Therefore, it is attractive to automatically identify the sentiment polarities of opinion words for different domains.

Many existing studies on opinion lexicon extraction heavily rely on broad-coverage dictionaries (e.g., WordNet). However, dictionary based methods fail to deal with the domain dependency problem, because the entries in a dictionary are often domain-independent. Recently, corpus-based graph models for automatic opinion lexicon extraction have emerged and prevailed, where the polarities of opinion words are inferred by the sentiment labels of seed words. However, these methods are very sensitive to seed words and improper seed words may lead to poor performance [5]. Yu et al. [5] proposed a method that utilizes the sentiment labels of documents instead of seed words. However, this method ignores the semantic association between words in the documents.

In [2], Liang et al. developed a model, CONR, that takes both the sentiment labels of documents and the semantic relationships between words into accoun. Inspired by CONR, we develop a Word Vector and Matrix Factorization (WVMF) based method that improves CONR from two aspects: First, CONR captures the semantic relationship between opinion words through pointwise mutual information, which suffers from the sparsity problem. To overcome this problem, WVMF employs pre-trained word vectors for similarity measurement; Second, WVMF adds more features including inverse document frequency to calculate the sentiment contribution of a given word.

2. THE PROPOSED METHOD

Let $D = \{d_1, d_2, ..., d_m\}$ denote a set of m documents, and $L = \{l_i\}_{i=1}^{m}$ denote the corresponding sentiment labels, where $l_i = +1$ if the corresponding document d_i is positive; Otherwise, $l_i = -1$. Let $W = \{w_1, w_2, ...w_n\}$ denote the vocabulary. We can then define an $m \times n$ matrix R to indicate the relationships between documents and words: $r_{ij} = 1$, if $w_j \in d_i$; Otherwise, $r_{ij} = 0$. C is defined as an $m \times n$ sentiment contribution matrix, where c_{ij} denotes the sentiment contribution of w_j to d_i. We define S as a $n \times n$ influence matrix, where s_{ij} characterizes the semantic similarity between w_i and w_j.

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2742713.

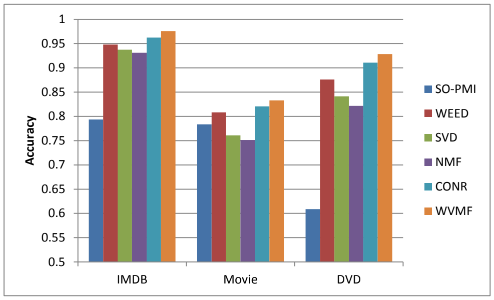

Figure 1: The accuracy of different methods in extracting opinion words.

Since it is noticed that words with high frequencies are more important and words only occur in positive or negative documents are more informative, we can define c_{ij} as

$$c_{ij} = TF^i(w_j) \cdot IDF(w_j) \cdot \left(\frac{F^{(pos)}(w_j)}{F^{(neg)}(w_j)} \right)^{l_i}, \quad (1)$$

where $TF^i(w_j)$ is the term frequency of w_j in d_i; $IDF(w_j)$ is the inverse document frequency of w_j; $F^{(pos/neg)}(w_j)$ is the frequency of w_j occurring in the positive/negative corpus.

In this paper, the similarity between two words is measured by the cosine distance with the word vectors that are trained on Google News and publicly available [3].

Therefore, the matrix factorization based method that combines both the document-word relationship and the word-word relationship can be formulated as

$$\begin{aligned}\min_{U,V} \mathcal{J}(C,U,V) = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n} r_{ij}(c_{ij} - U_i^T V_j)^2 + \\ \frac{\alpha}{2}\sum_{j=1}^{n} \|V_j - \sum_{k\in\mathcal{K}(j)} s_{jk}V_k\|_F^2 + \frac{\beta}{2}\|U\|_F^2 + \frac{\gamma}{2}\|V\|_F^2,\end{aligned} \quad (2)$$

where U is a $k \times m$ latent document feature matrix, V is a $k \times n$ latent word feature matrix, $k < min(m,n)$ and $\alpha, \beta, \gamma > 0$. $\mathcal{K}(i)$ denotes the neighbors of w_i. Here, we consider two words with high similarity as neighbors.

Finally, the sentiment polarity of w_j can thus be derived as follows:

$$\omega_j = \frac{1}{|\mathcal{D}^{(+)}|}\sum_{i\in\mathcal{D}^{(+)}} c_{ij} - \frac{1}{|\mathcal{D}^{(-)}|}\sum_{i\in\mathcal{D}^{(-)}} c_{ij} \quad (3)$$

where $\mathcal{D}^{(+)}$ and $\mathcal{D}^{(-)}$ represent the positive and negative documents in the corpus, respectively; w_j is considered as a positive word, if $\omega_j > 0$, and a negative one, if $\omega_j < 0$.

3. EXPERIMENTAL VALIDATION

We carried out experiments on three publicly available datasets from different domains, namely, IMDB[1], Movie reviews[2], and DVD reviews[3] The opinion words for testing were obtained from MPQA[4]. We adopted five representative methods, SO-PMI, WEED, SVD, NMF and CONR, as the baselines. SO-PMI [4] is a typical seed word based method and thus we randomly selected 20% seed words for it from MPQA; WEED [5] is an optimization based method; SVD, NMF [1] and CONR [2] are all matrix factorization based.

Figure 1 presents the accuracy of all methods in extracting opinion words from the datasets of different domains. It can be seen that WVMF outperforms all baseline methods. Table 1 presents the accuracy of all methods in identifying the sentiment polarities of the top k opinion words. It is observed that the matrix factorization based methods including WEED, SVD and NMF consistently outperform the seed word based method SO-PMI. Particularly, the proposed WVMF method exhibits consistent better performance than the best state-of-the-arts method CONR.

[1]http://ai.stanfor.edu/amaas/data/sentiment/
[2]http://www.cs.cornell.edu/people/pabo/movie-review-data/
[3]http://www.datatang.com/data/44115/
[4]http://mpqa.cs.pitt.edu/

Table 1: The accuracy of different methods in identifying sentiment polarities of opinion words.

Datasets	Methods	Top10	Top20	Top50	Top100	Top200
IMDB	SO-PMI	0.5121	0.5533	0.5083	0.5187	0.5267
	WEED	0.8944	0.8613	0.8507	0.8288	0.7768
	SVD	0.8848	0.8361	0.8147	0.7904	0.7342
	NMF	0.8919	0.8704	0.8140	0.7950	0.7433
	CONR	0.9383	0.9171	0.8782	0.8466	0.7930
	WVMF	**0.9633**	**0.9300**	**0.9017**	**0.8825**	**0.8450**
Moive	SO-PMI	0.5121	0.5537	0.5289	0.5187	0.4879
	WEED	0.7448	0.6951	0.7083	0.6687	0.6475
	SVD	0.6341	0.6511	0.6085	0.5937	0.6375
	NMF	0.6814	0.5609	0.5833	0.5812	0.6113
	CONR	0.8333	0.7804	0.7625	0.7353	0.6694
	WVMF	**0.8733**	**0.8650**	**0.8433**	**0.8100**	**0.7666**
DVD	SO-PMI	0.5625	0.5238	0.4901	0.4455	0.4404
	WEED	0.8064	0.7380	0.7745	0.7326	0.6487
	SVD	0.8489	0.7727	0.7843	0.7178	0.6959
	NMF	0.8333	0.7857	0.7884	0.7128	0.6717
	CONR	0.9085	0.8809	0.7841	0.7623	0.7233
	WVMF	**0.9254**	**0.8961**	**0.8430**	**0.8038**	**0.7625**

4. CONCLUSION

This paper has presented a word vector and matrix factorization based method for opinion lexicon extraction. Experiments on real datasets have demonstrated that the proposed method performs better than the state-of-the-art methods.

5. ACKNOWLEDGEMENT

This work was funded by the 973 Program of China (2012CB31-6303 and 2014CB340401), National High-Tech Research and Development Program of China (2013AA013204), National HeGaoJi Key Project (2013ZX01039-002-001-001) and Strategic Priority Research Program of the Chinese Academy of Sciences (XDA060-30200).

6. REFERENCES

[1] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[2] J. Liang, X. Zhou, Y. Hu, L. Guo, and S. Bai. Conr: A novel method for sentiment word identification. In *Proceedings of the 23rd CIKM*, pages 1943–1946. ACM, 2014.

[3] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013.

[4] P. D. Turney and M. L. Littman. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346, 2003.

[5] H. Yu, Z.-H. Deng, and S. Li. Identifying sentiment words using an optimization-based model without seed words. In *ACL (2)*, pages 855–859, 2013.