ISBN 978-981-11-3671-9
Proceedings of 2017 the 7th International Workshop on Computer Science and Engineering (WCSE 2017)
Beijing, 25-27 June, 2017, pp. 410-416
doi: 10.18178/wcse.2017.06.070

A Mixed Approach of Expanding Sentiment Lexicon Based on word2vec and Phrase Dependency Tree

Junxia Wang [1+], Pu Zhang [1] and Yinghao Wang [1]

[1] Department of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing, China

Abstract. As one of the basic and crucial tasks in sentiment analysis, sentiment lexicon generation has great significance. Due to diversities of sentiment words, automatic expansion of sentiment lexicon has become a challenging research topic in recent years. This paper proposes a new approach to expand sentiment lexicon. At first, the approach selects some sentiment seed words, then uses word2vec to train word embeddings and find out the words which have potential similarities with seed words to generate a candidate sentiment lexicon, and then finds out the words which have conjunctive relations with seed words based on the phrase dependency tree to generate another candidate sentiment lexicon; Finally, we take the words appearing in the two candidate sentiment lexicons as final expanded sentiment lexicon. Experimental results demonstrate the effectiveness of our approach and the advantages against state of-the-art baselines.

Keywords: sentiment analysis, sentiment lexicon, word2vec, phrase dependency tree

1. Introduction

With the rapid development of e-commerce, more and more people tend to shop online. It's essential to view the products' reviews before consumers deciding whether to buy a product. By analysing the sentiment polarities (e.g., positive and negative) of the product reviews, we not only can provide more comprehensive reference information for consumers' purchase decisions, but also can provide decision support for businesses to understand customer needs and improve products' qualities [1]. The integrity and accuracy of sentiment lexicon can greatly affect the result of the sentiment analysis. At present, some mature methods such as expanding sentiment lexicon based on WordNet has some errors. For example, for the expansion of positive seed words (such as beautiful), which will eventually get the negative sentiment words (such as bad). Issues, such as poor portability, a great deal of manual labeling effort or the classification of sentiment words is wrong, can significantly influence the performance of sentiment lexicon and restrict the development of sentiment analysis. Aiming at the above problems, this paper proposes a mixed approach of expanding sentiment lexicon based on word2vec and phrase dependency tree. It mainly uses word2vec tool to train word embeddings and uses phrase dependency tree to analyze the relationship between words in order to obtain words' similarities, then according to the similarities to automatically expand sentiment lexicon.

2. Expanding Sentiment Lexicon Based on Word2vec and Phrase Dependency Tree

The general idea of the approach is as follows: First, select a certain number of positive and negative sentiment seed words, and use word2vec to train word embeddings, then calculate similarities between the words in the embeddings and the seed words, the word whose similarity is greater than a certain threshold will be added in the candidate sentiment lexicon. Second, find out the sentiment words that have conjunctive relations with the seed words through the analysis of phrase dependency tree, then add them to another

[+] Corresponding author. Tel.: +8618716319589.
E-mail address: 1263038889@qq.com.

candidate sentiment lexicon. Finally, take the sentiment words both appearing in the two candidate sentiment lexicons as the final expanded sentiment lexicon. The workflow of our approach is shown in Fig. 1:

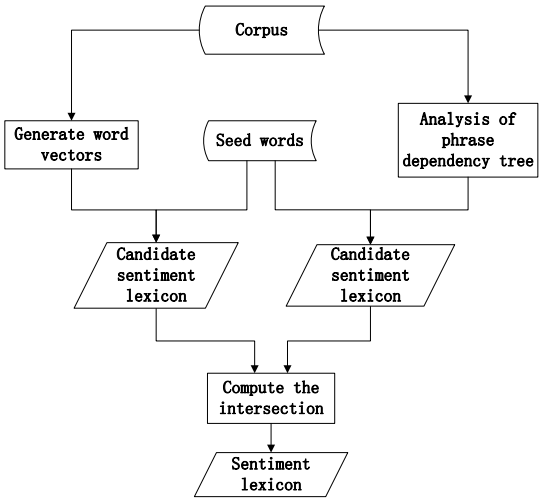

Fig. 1: The workflow of the approach

2.1. Seed words selection

In the expansion of sentiment lexicon, the seed words usually have clear emotional tendencies, and the seed words selection directly determines whether or not the result of sentiment lexicon is good. We select some strongly subjective words in SentiWordNet [2] as seed words. The judgment of "strongly subjective" is based on SentiWordNet, it scores the sentiment words in the various categories of SentiWordNet respectively, the scores are arranged from high to low. We select common sentiment words those have high scores as seed words. If the same word appears in various categories, we retain the word in the category with highest score. We put the positive sentiment words into the set of positive seed words (we name the set of positive seed words as "Pos" set), we put the negative sentiment words into the set of negative seed words (we name the set of negative seed words as "Neg" set). For example, "good" and "happiness" are put into "Pos" set, "bad" and "sad" are put into "Neg" set. Some selected seed words are shown in Table I:

Table 1: Sets of some seed words

Pos	**Neg**
good	bad
happiness	sad
better	poor
nice	worst
…	…

2.2. Expanding candidate sentiment lexicon based on word2vec

Word2vec is an efficient tool to characterize words as real numbers. It can be used to simplify the processing of text into vectors computing on K-dimensional vector space. The similarities in vector space can be used to represent the similarities in text semantics. This expression is superior to one-hot representation (It constructs a thesaurus, the vector dimension is equal to the vocabulary size and the word is expressed as the corresponding dimension of 1), K-dimensional vector not only contains the similarities between words, but also avoids the dimension disaster [3]. Therefore, we use word2vec to expand candidate sentiment lexicon.

2.2.1 The introduction of word2vec

Word2vec uses distributed representation [4] to represent word vectors. The basic idea is using deep learning to map each word to multi-dimensional real vector, and then through the distance between words (such as cosine similarity) to determine the semantic similarities between them. The word2vec tool includes two training models, continuous bag-of-words (CBOW) model and skip-gram model. CBOW model predicts the current word based on the context; skip-gram model predicts the context based on the current word.

Since the skip-gram model is extremely efficient for training a large number of unstructured texts [5], we select it as training model. Use the generated word vectors, we can calculate the similarities between the words specified by the user. The skip-gram model is shown in Fig. 2:

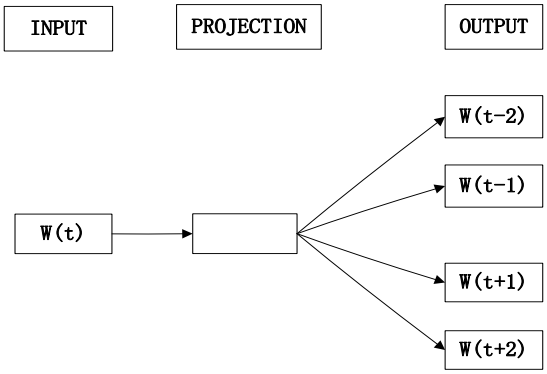

Fig. 2: The skip-gram model architecture in word2vec

Assuming the existence of the phrase sequence w_1, $w_2 \ldots w_t$, the purpose of skip-gram model is to maximize the objective function F, the function F can refer to (1):

$$F = \frac{1}{T}\sum_{t=1}^{T}\sum_{-d\le j\le d, j\neq 0} \log P(w_{t+j} \mid w_t) \tag{1}$$

where d is a constant, the greater the d, the more accurate the result, however, the corresponding training time also increases. Since negative sampling can speed up training and represent words more accurately [6], we use it to train the skip-gram model.

2.2.2 Expanding candidate sentiment lexicon

In this paper, we first use word2vec tool to train word embeddings to get the word vector of each word, and then expand the positive and negative seed words respectively. Then we add the words whose similarities with the seed words are larger than threshold into the candidate sentiment lexicon, and then we get the positive and negative candidate sentiment lexicon respectively. Finally for the candidate words appearing in both lexicons, we retain them in the lexicon which has higher similarity.

2.3. Expanding candidate sentiment lexicon based on phrase dependency tree

The phrase dependency tree can effectively get dependencies and connection relations between words. In general, the adversative conjunctions in the text will change the polarity of sentiment word before and after the conjunctions, and the coordinating conjunctions will not change the polarity of sentiment word. Therefore, we first transform sentence into phrase dependency tree, and then use the change of polarity before and after the conjunction to carry on the expansion of sentiment lexicon.

2.3.1 The construction of phrase dependency tree

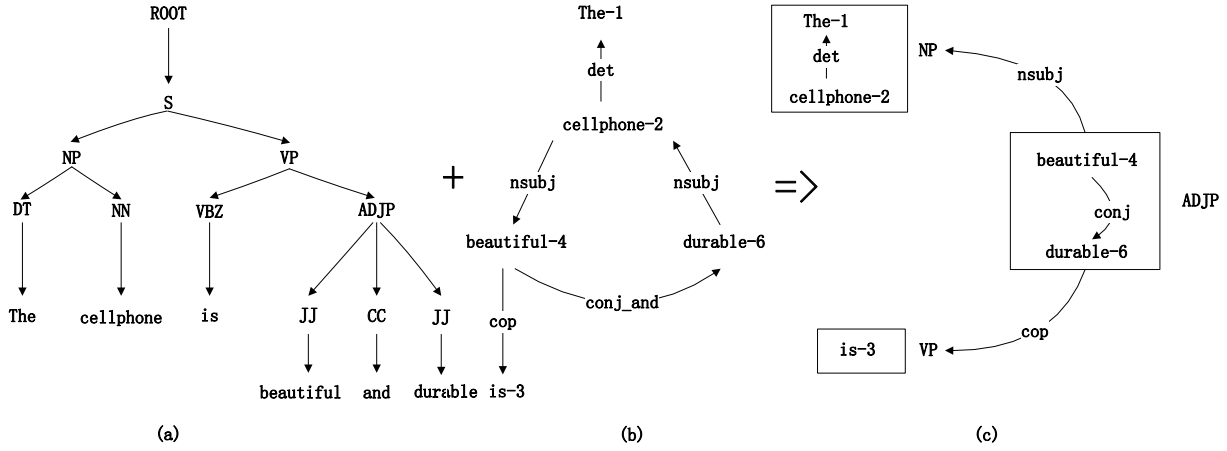

Fig. 3: The construction process of phrase dependency tree

We first use the Stanford Parser to parse the sentence and get the phrase structure tree (as shown in Fig. 3(a)), universal dependencies (as shown in Fig. 3(b)), and then transform them into a phrase dependency tree (as shown in Fig. 3(c)) [7]. Each chunk node in phrase dependency tree is used to preserve the words that have dependency relations. Fig. 3 describes the construction process of phrase dependency tree for sentence "The cellphone is beautiful and durable", each word in the phrase structure tree will be added into the phrase dependency tree in turn.

In the Fig. 3, "The" and "cellphone" belong to a same chunk node, "beautiful" and "durable" belong to another same chunk node, "is" alone is in another separate chunk node, so the entire phrase dependency tree is divided into three chunk nodes.

2.3.2 Expanding candidate sentiment lexicon

Using the phrase dependency tree to carry out the expansion of candidate sentiment lexicon, we need seed words to start the propagation learning. The selection of seed words as described in section 2.1. In general, two words that have coordinating relationship have the same polarity, and two words that have adversative relationship have the opposite polarities. Therefore, we divide the conjunctive relations into two types: the coordinating relationship and the adversative relationship. The coordinating relationship is used to extract sentiment words that have the same polarity with seed words, and the adversative relationship is used to extract sentiment words that have the opposite polarity with seed words. The conjunctions with coordinating relationship are defined in this paper as follows: and, neither ... nor, either ... as, as well as, not only ... but also ..., the conjunctions with adversative relationship are defined as follows: but, yet, however, still, while, on the contrary. The product reviews with coordinating relationship such as Example 1) and Example 2), the product reviews with adversative relationship such as Example 3) and Example 4).

Example 1) "The cellphone is beautiful and durable."
Example 2) "The phone is neither cheap nor beautiful."
Example 3) "The phone's performance is good, but the price is too expensive."
Example 4) "The introduction of this phone is very good, yet in fact it is disappointing."

When the sentiment word in the text and the seed word are in a chunk node, and they have coordinating relationship, we add the sentiment word into candidate sentiment lexicon which has the same polarity with the seed word. If the sentiment lexicon has the word, the amount of the word will increase by one. When the sentiment word in the text and the seed word are in a chunk node, and they have adversative relationship, we add the sentiment word into candidate sentiment lexicon which has opposite polarity with the seed word. If the sentiment lexicon has the word, the amount of the word will increase by one. Like this, we use conjunctive relations to expand the sentiment words for each seed word until no candidate sentiment word can be expanded. For example, "beautiful" and "durable" in Example 1) are in a chunk node and they have a coordinating relationship, so their polarities are the same; "good" and "expensive" in Example 3) are in a chunk node and they have a adversative relationship, so their polarities are opposite.

For the candidate words that appear in positive and negative sentiment lexicons, we retain them in the lexicon which has higher amount.

2.4. Mixed expansion of sentiment lexicon

For the expanded candidate sentiment lexicon based on word2vec and the expanded candidate sentiment lexicon based on phrase dependency tree, they both have some noise. So we take the sentiment words appearing in both sentiment lexicons as final sentiment lexicon. On the basis of the above discussions, the algorithm for the expansion of sentiment lexicon is illustrated as follows.

Algorithm 1: The expansion of sentiment lexicon

Input: Review corpus (RC); Positive seed words (SD_{pos}), negative seed words (SD_{neg}); Parameter wv_{pos} and wv_{neg} are used as the threshold of expanded positive and negative sentiment lexicon based on word2vec respectively; Positive candidate sentiment lexicon (SL'_{pos}) and negative candidate sentiment lexicon (SL'_{neg})

Output: Positive sentiment lexicon (SL_{pos}) and negative sentiment lexicon (SL_{neg})

1) Initialize the set SL'_{pos}, SL'_{neg}, SL_{pos} and SL_{neg} as empty set, $SL'_{pos} = \varnothing$, $SL'_{neg} = \varnothing$, $SL_{pos} = \varnothing$, $SL_{neg} = \varnothing$
2) Perform part of speech tagging on the corpus RC

3) For each $SD_i \in SD_{pos}$:
   - Calculate the similarity s_i of SD_i and the word SW_i in RC, if $s_i > wv_{pos}$, then $SL_{pos} = SL_{pos} \bigcup SW_i$
   - Analyze corpus RC using phrase dependency tree, if SD_i and the word SW_j have coordinating relationship, then $SL'_{pos} = SL'_{pos} \bigcup SW_j$; if SD_i and the word SW_j have adversative relationship, then $SL'_{neg} = SL'_{neg} \bigcup SW_j$
4) For each $SD_j \in SD_{neg}$:
   - Calculate the similarity s_j of SD_j and the word SW_i in RC, if $s_j > wv_{neg}$, then $SL_{neg} = SL_{neg} \bigcup SW_i$
   - Analyze corpus RC using phrase dependency tree, if SD_j and the word SW_j have coordinating relationship, then $SL'_{neg} = SL'_{neg} \bigcup SW_j$; if SD_i and the word SW_j have adversative relationship, then $SL'_{pos} = SL'_{pos} \bigcup SW_j$
5) End for, $SL_{pos} = SL_{pos} \bigcap SL'_{pos}$, $SL_{neg} = SL_{neg} \bigcap SL'_{neg}$
6) Return SL_{pos} and SL_{neg}.

3. Experiments

3.1. Corpus description and evaluation metrics

In order to evaluate the performance of our proposed approach, we selected 194185 cell-phone reviews [8] from Stanford University on the Amazon site as experimental corpus. In this paper, we evaluate the accuracy of the approach based on human annotations [9].

3.2. Baseline methods

For convenience, we name our proposed approach as WP. In order to evaluate the performance of the approach, we compare the proposed approach with several baseline methods as follows:

1) A method of expanding sentiment lexicon based on WordNet (WN): This method uses WordNet to extract the synonyms of sentiment seed words directly and take them as final sentiment lexicon.

2) A method of expanding sentiment lexicon based on word2vec (W2V): This method uses word2vec to train word embeddings, then calculate the semantic similarities between the words in the embeddings and the sentiment seed words, and the words whose similarities are greater than a certain threshold as final sentiment lexicon.

3) A method of expanding sentiment lexicon based on phrase dependency tree (PDT): This method uses the phrase dependency tree described in section 2.3 to expand sentiment lexicon.

3.3. Experimental results and discussion

In this paper, we respectively select 20, 30, 50 seed words for experimentation. The experimental results are shown in Table 2, Table 3, and Table 4 respectively.

Table 2: The experimental results of 20 seed words

Methods	***Pos （%）***	***Neg （%）***	***Accuracy （%）***
WN	81.6	60.0	70.8
W2V	74.0	76.6	75.3
PDT	79.0	62.8	70.9
WP	***80.0***	***72.0***	***76.0***

Table 3: The experimental results of 30 seed words

Methods	***Pos （%）***	***Neg （%）***	***Accuracy （%）***
WN	80.0	52.0	66.0
W2V	66.0	79.6	72.8
PDT	80.0	62.0	71.0
WP	***80.0***	***73.5***	***76.8***

Table 4: The experimental results of 50 seed words

Methods	*Pos （%）*	*Neg （%）*	*Accuracy （%）*
WN	62.0	40.0	51.0
W2V	70.0	74.0	72.0
PDT	82.0	66.0	74.0
WP	***79.5***	***75.0***	***77.3***

From the Table 2, 3, and 4, we can see that the overall performance of W2V is high in baseline methods, especially W2V achieves high accuracy in the expansion of the negative sentiment lexicon, but the results are not very good in the expansion of the positive sentiment lexicon. The reason might be that some of the words those have potential relationships with positive seed words are not sentiment words, which introduced some noise. The results of PDT are just opposite, it has high performance in the positive sentiment lexicon, but low performance in the negative sentiment lexicon. We can also see WN has the worst performance, mainly because WordNet is a manual construction of sentiment lexicon, the number of sentiment words is limited and it can't expand for the sentiment words that are not in WordNet. Therefore, the WN has some limitations on the expansion of sentiment words.

In this paper, we combine the advantages of W2V and PDT, and calculate the intersection of the two methods to filter out noise introduced by the respective method. The above comparisons of methods also demonstrate the view. For the expanded sentiment lexicon, the accuracy of WP is superior to other baseline methods. With the number of seed words increasing, the accuracy of our proposed approach is also slightly improved from 76.0% to 76.8%, and then improved from 76.8% to 77.3%. The reason why the accuracy increased slowly is that as the number of seed words increased, the expanded words increased, which introduced more noise, so the accuracy increases slightly.

Fig. 4 shows the results with varying the number of seed words. We can see that the accuracy curve of WP lies well above the other curves with the variations of the number of seed words, which indicates that our proposed approach works well and robust. With the number of seed words increases, we also observe that both the accuracy of WP and PDT have increased, and the accuracy of W2V and WN have declined, especially WN declines greatly.

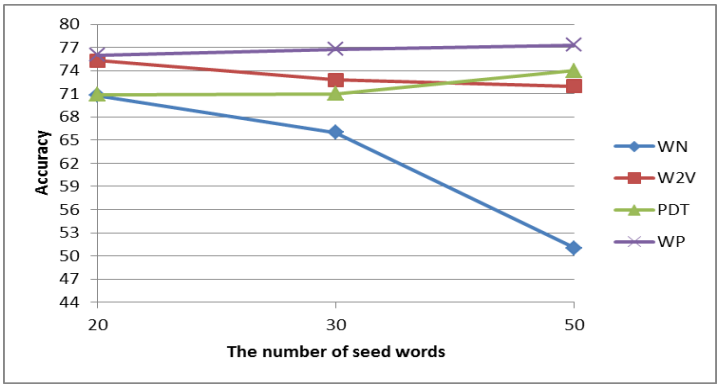

Fig. 4: The results with varying the number of seed words

4. Conclusion

In this paper, we propose a novel approach to expand sentiment lexicon. The approach combines word2vec and phrase dependency tree to analyze the relationship between words, and get expanded sentiment lexicon, it solves the problems of poor portability and a great deal of manual labeling effort. The expanded sentiment lexicon can be used in many aspects such as judging the polarity of text, sentences and words. In the future, we will try to combine with other methods to distinguish expanded words' polarities, such as label propagation algorithm, and further improve the accuracy of sentiment lexicon.

5. Acknowledgements

This work is supported by Chongqing research and innovation project of graduate students (No.CYS15171), the Basic and Frontier Research Program of Chongqing (No. cstc2015jcyjA40025, No.

cstc2015jcyjA40036), the Scientific and Technological Research Program of Chongqing Municipal Education Commission (No. KJ1600440), the Doctoral Program Foundation of CQUPT (No.A2016-02).

6. References

[1] J. Z. Du, J. Xu, Y. Liu. Research on Construction of Feature-Sentiment Ontology and Sentiment Analysis[J]. New Technology of Library and Information Service, 2014, 5: 74-82.

[2] S. Baccianella, A. Esuli, F. Sebastiani. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining[C]//LREC. 2010, 10: 2200-2204.

[3] Y.Yang, L. F. Liu, H. F. Lin, et al. New methods for extracting emotional words based on distributed representations of words[J]. Journal of Shandong University (Natural Science) , 2014, 49(11): 51-58.

[4] A. Paccanaro, G. E. Hinton. Learning distributed representations of concepts using linear relational embedding[J]. IEEE Transactions on Knowledge and Data Engineering, 2001, 13(2): 232-244.

[5] T. Mikolov, K. Chen, G. Corrado, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013..

[6] T. Mikolov, I. Sutskever, K. Chen, et al. Distributed representations of words and phrases and their compositionality[C]//Advances in neural information processing systems. 2013: 3111-3119.

[7] Y. Wu, Q. Zhang, X. Huang, et al. Phrase dependency parsing for opinion mining[C]//Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3. Association for Computational Linguistics, 2009: 1533-1541.

[8] J. McAuley, C. Targett, Q. Shi, et al. Image-based recommendations on styles and substitutes[C]//Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2015: 43-52.

[9] A. Neviarouskaya, H. Prendinger, M. Ishizuka. SentiFul: A lexicon for sentiment analysis[J]. IEEE Transactions on Affective Computing, 2011, 2(1): 22-36.