Advances in Computer, Signals and Systems (2016) 1: 23-27
Clausius Scientific Press, Canada

One subjective feature extraction method of sentiment analysis based on dependency grammar

Xinkai Yang

College of Information, Mechanical and Electronical Engineering, Shanghai Normal University, Shanghai, China, 200234
E-mail: xkyang@shnu.edu.cn

Keywords: sentiment analysis; feature extraction; dependency grammar

Abstract: In this paper, we propose one subjective sentimental feature extraction algorithm based on dependency grammar. Several dependency features and relations are selected for construction of dependency graph. The effectiveness of this approach is verified by well designed experiments.

1. Introduction

Sentiment analysis has become a very active research area in recent years. It also has many slightly different names such as opinion mining, subjectivity analysis, affect analysis, emotion analysis, etc. The purpose is to analyze people's opinions, attitudes and emotions towards products or services. Sentiment analysis applications cover many domain, such as consumer products or services, healthcare, financial services, social events and even politics.

Sentiment feature extraction from original text resources is an important step and essential process in sentiment analysis [1]. In the context of sentiment analysis, some specific features can facilitate the extraction process. For example, there always has a target when one wants to express some opinion. The target is often the topic to be extracted from a sentence. Thus, it is important to recognize each opinion expression and its target from a sentence. There are several approaches to extract features from sentences: extraction based on frequent nouns and noun phrases, extraction by exploiting opinion and target relations, extraction using supervised learning, and extraction using topic modeling.

In this paper we develop one subjective feature extraction method of sentiment analysis based on dependency grammar, which can be used to improve the performance of sentiment analysis algorithm. The rest of this paper is organized as follows. In Section II, some related research works on feature extraction are discussed. In section III we propose a general framework. Some metrics defined in Section IV. We discuss our evaluation process in Section V and conclude this paper in section VI.

2. Related Works

There is lots of sentiment information to be extracted from original text sources. Many research works have been launched on this research field and can be summarized as the follow.

2.1 Extraction and discrimination of sentiment words

Extraction and discrimination of sentiment words is an integrative process which is mainly divided into corpus-based approach and dictionary-based approach. Andreevskaia et al. realize fuzzy sentiment recognition method by identifying the polarity of sentiment words in WordNet [2]. Wiebe

et al. develop one clustering algorithm based on similarity distribution to obtain the sentiment words [3]. But these two methods confine the sentiment words within adjectives ones, ignoring the words with other part of speech. The dictionary-based approach mainly uses the lexical and semantic relationship between words in the dictionary to identify the polarity of sentiment words. The dictionary here generally refers to WordNet or HowNet, etc. In [4] some sentiment words are manually collected as word seeds and these seeds are expanded to obtain a large number of sentiment words. This method is easy to introduce noises because it depends on the seeds selection and some sentiment words have polysemy in some cases. In order to avoid the usage of ambiguous words, the work in [5] uses the annotation information in the dictionary to judge the polarity of sentiment words. Kamps et al. in [6] identify sentiment words by calculating the correlation value between adjectives and the seed representative of positive words and negative words. The advantage of dictionary-based method is that it can get a considerable scale of sentiment words. But sentiment dictionary often has many ambiguous sentiment words due to polysemy effect.

2.2 Extraction of subjective expression

A subjective expression refers to a word or phrase that represents the subjectivity of a sentiment text unit. Wiebe and Wilson construct an expression library by mining a large number of subjective expressions. Then the objective classification and polarity identification are launched based on the expression library. Specifically, n-gram words/phrases ($1 <= n <= 4$) in the corpus are extracted as the candidates of subjective expressions. The probability of each candidate subjective expression is calculated by comparison with the standard subjective expressions in training set. Finally, these subjective expressions are obtained through the analysis of these probability values. J. Wilson and T. Wiebe introduce the concept of "subjective expression density" to coordinate the subjective expression selection. They also use syntactic analysis to mine the syntactic subjective expression. C. Whitelaw and N. Garg use a variety of features and machine learning methods to identify the sentient level from a large number of subjective expressions.

2.3 Extraction of sentiment target

The full definition of sentiment target may be complex and may cover a lot of sentences or paragraph. But on its narrow sense it usually refers to the topic discussed in a section of reviews about which an opinion has been expressed, such as an event in the news comment. Many existing research mainly focus on the extraction of sentiment target from product reviews. They are mostly confined to the category of noun or noun phrase. The rule based method is popular to extract sentiment target. Yi extract the real sentiment target from the candidate evaluation objects by using 3 rules of restricted grade components. Some other methods try to find out the sentiment target based on syntactic analysis and association rule mining. But the rule based method has poor scalability, heavy workload and high cost.

3. Subjective Feature Extraction Method based on Dependency Grammar

Obviously, the identification of subjective sentiment words in the text is one of the most important problems to be solved in sentiment analysis. In order to avoid noise, some researches only extract the adjectives in the sentence to act as the subjective words. However, verbs often express the author's opinion in many cases. For example, "I like this book." The subjective word in this sentence is "like" which is a verb. The only extraction of adjectives as subjective words may lead to loss or misjudgment on sentence sentiment tendency.

Because the emotional words often have dependency relationship with related topics or objects in syntactic analysis, here we propose a model based on dependency grammar rules. The following three dependency relations are taken into consideration: (1) VOB: "VOB" represents for the relationship between verbs and objects. Sentimental words are verbs and topical words are the objects of verbs. (2) SBV: "SBV" represents for the relation between subjects and predicates. Sentimental words are predicates and topical words are the subjects of sentimental words. (3) ATT: "ATT" represents for the relation of attributes. Sentimental words are attributes and topical words are the modified center of sentimental words.

Several sentiment features are selected in the sentiment words extraction rules. Then the grammar dependency tree is constructed based on these features. Each node represents a word, and the tree is composed of a number of binary relations between different words. Each relationship has a word that is used as a parent node and another one as a child node. Each word has one, and only one parent node, and one word can have more than one child. Features for construction of extraction rules can be classified into eight types in table 1.

Table 1. Features for construction of extraction rules

Feature name	Description
EM	Set as the number of (positive–negative) emoticons
NDA	If word in negative verbs, set as 1;or set as 0
BSL	(positive–negative) of basic sentiment lexicon
VOB	(positive–negative) of sentimental words having “VOB” relation with topical words
SBV	(positive–negative) of sentimental words having “SBV” relation with topical words
ATT	(positive–negative) of sentimental words having “ATT” relation with topical words

Subjective words extraction rules based on dependency grammar:

1) When the dependency grammar relation in the sentence is VOB, to extract the adjective or the verb as the subjective word; 2) When the dependency grammar relation in the sentence is ATT, only the adjective is taken as the subjective word; 3) When the dependency grammar relation in the sentence is SBV, only the verb is taken as the subjective word.

4. Evaluation Indexes and Algorithm

Firstly we describe the general metric of precision and recall - two evaluation indexes of text classification system. Precision is the ratio of the number of correct classified text to all. The precision of the formula is expressed as follows:

Precision = TP/TP+FP, where TP denotes the number of true positives and FP the number of false positives;

Recall = TP TP+FN, where FN denotes the number of false negatives.

Then we adapt two popular statistical measures as complementary ones. These measures were applied to determine the information value of sentimental items and thus allow us to discriminate them between positive or negative.

Therefore, we propose one subjective sentimental feature extraction algorithm based on dependency grammar analysis (EDG) in figure 1, in which the process of this algorithm is described in details.

ALGORITHM: Topic-related sentimental features extraction based on dependency grammar
Input: Dependency analysis result (DP), Expanded Topical Words (ETW)
Output: topic-related features (TRF)

for word in DP do
if word in ETW and word.relate in 'SBV', 'VOB', 'ATT' then
TRF += word.parent
if word.parent in ETW and word.relate in 'SBV', 'VOB', 'ATT' then
TRF += word
return TRF.

Figure. 1 Topic-related sentimental features extraction

5. Experimental Result Analysis

Experimental results of IG, MI and EDG algorithms in KNN classifier are listed in table 2. Text dataset is divided into six categories, including economy, education, politics, and sports, military. The training set has 872 samples and the test set has 430 samples.

We preprocess all texts through word segmentation system processing. Then we extract the text features in accordance with the number of the text. The number of extraction features are 1000, 500, 300, 100 and 50. The purpose of this experiment is to compare the accuracy of different feature extraction methods in the KNN classifier with the same amount of training set.

Table 2. Results of the Evaluation

Feature number		**1000**	**500**	**300**	**100**	**50**
Feature Extraction Method	IG	79.1	84.2	87.8	88.5	86.3
	MI	72.2	65.8	59.5	45.6	44.7
	EDG	80.1	85.2	87.0	83.7	85.3

Table 2 lists the experimental results of IG, MI and EDG feature extraction algorithms in KNN under different feature vector space. With the decrease of the number of features, the accuracy of different feature extraction methods increases firstly and then decrease. The accuracy of Mutual information (MI) feature extraction method decreases quickly with the reduction of the number of features. The classification performance of MI is the worst. The dimension of feature space is 1000 when the accuracy of IG feature extraction method reaches the largest. And the performance of EDG feature extraction method is best when the dimension of feature space is 300. From this experiment, it is not difficult to find that classification accuracy decrease when the feature vector space is too large or too small.

6. Conclusions

In this paper, we propose one subjective sentimental feature extraction algorithm based on dependency grammar to extract sentimental words from text resources. Three dependency relations are selected for construction of dependency graph. The proposed procedure allows the fast extraction of sentimental words properly adapted to different contents. Our results suggest that the proposed approach might be a useful facility for future research.

We employed a large labeled data set to test the effectiveness of this approach and two statistical measures to calculate the sentiment score. The results on the test data confirmed that the accuracy of EDG increases when compared with other benchmarks.

The experimental results remind us that deep processing on syntax and semantics might be helpful for future sentiment analysis works. Such as investigation on phrase structure analysis and more general models will be investigated and evaluated.

Acknowledgements

This work was financially supported by the Shanghai Natural Science Foundation (0666666), Innovation Program of Shanghai Municipal Education Commission (060000) and Shanghai Leading Academic Discipline Project of Shanghai Municipal Education Commission (0555555).

References

[1] Bing Liu, Xiaoli Li, Wee Sun Lee, Philip S. Yu. Text classification by labeling words[C]. Proceedings of the 19th national conference on Artifical Intelligence, AAI'04, 2004, 55-65.

[2] A. Andreevskaia and S. Bergler. Mining WordNet for a fuzzy sentiment: Sentiment tag extraction from WordNet glosses[C]. Proceedings of the European Chapter of the Association for Computational Linguistics (EACL), 2006.

[3] E. Riloff, S. Patwardhan, and J.Wiebe. Feature subsumption for opinion analysis[C]. Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2006. 274-286.

[4] Kim SM, Hovy E. Identifying and analyzing judgment opinions[C]. Proceedings of the Joint Human Language Technology/North American Chapter of the ACL Conf. (HLT-NAACL). Morristown: ACL, 2006, 200-207.

[5] Esuli A, Sebastiani F. Determing term subjectivity and term orientation for opinion mining[C]. Proceedings of the European Chapter lf the Association for Computational Linguistics(EACL). Morristown: ACL, 2006, 193-200.

[6] Kamps J, Marx M, Mokken RJ. Using WordNet to measure semantic orientation of adjectives[C]. Proceedings of the LREC. 2004, 1115-1118.