Topic-Based Chinese Message Polarity Classification System at SIGHAN8-Task2

Chun Liao, Chong Feng, Sen Yang, Heyan Huang
School of Computer Science
and Technology, Beijing
Institute of Technology
{cliao, fengchong, syang, hhy63}@bit.edu.cn

Abstract

This paper describes the topic-based Chinese message polarity classification system submitted by LCYS_TEAM at SIGHAN8-Task2. The system mainly includes two parts: 1) a graph-based ranking model integrating local and global information is adopted to represent the classification ability of words towards different topics. In construction of graph model, a new weighting approach and a PMI-based random jumping probability selection method is proposed. 2) For sentimental features, word embedding is employed for acquiring expanded topical words and syntactic dependency is adopted for getting topic-related sentimental words. Experiment results demonstrate the effectiveness of our system.

1 Introduction

Sentiment analysis, which is to identify or determine the implied emotional orientation, attitude and opinion when people express something, is becoming more and more important for network monitoring with its application on microblog. In the traditional sentiment analysis， unsupervised methods were adopted in Ku(2005), Shen(2009), Vasileios(2000) and Turney(2002), and the limitation of such approaches based on semantic dictionary mainly is unable to solve the problem of Out-of-Vocabulary words. Supervised methods were employed with model of machine learning, such as Naive Bayes, Max Entropy, Support Vector Machine in Pang(2002), Dasgupta(2009), and Li(2011).

Hashtags, in the form of “＃topic＃”, are widely used as topics in Chinese microblogs. For the topic-related work, Wang(2011) and Jakob(2010) made research on hashtag-level sentiment classification in twitter. In the traditional sentiment analysis, the object people express sentiment on is not taken into consideration. And these methods are mostly topic-ignored and cannot perform the accurate sentiment analysis in many topic-related messages. We summarize such kind of difficult cases into two categories.

1) Microblogs with multiple candidate topics

For example, “#三星 galaxy s6##华为P8##mate8#”三星 galaxy s6真没什么亮点，华为P8就可以秒它了，更不用说mate8[拜拜]”. This sentence conveys negative sentiment towards topic of “三星 galaxy s6”, but positive sentiment towards topic of “华为P8” and “mate8”.

2) Microblogs with topic specific sentimental words

For example, “#股票#前天刚入手一支股票，一直在升，股价越来越高” and “#三星#三星手机电量明显不够用，耗能高”. The word “高” carrys positive sentiment orientation in the first sentence towards topic “股票” and negative sentiment orientation in the latter towards topic “三星”.

Considering the importance of topical information in microblogs, this paper studied topic-based Chinese message polarity classification. Given a message from Chinese Weibo Platform (Such as Sina, Tencent, NetEase etc.) and a topic, classify whether the message is of positive, negative, or neutral sentiment towards the given topic. For messages conveying both a positive and negative sentiment towards the topic, whichever is the stronger sentiment should be chosen.

Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing (SIGHAN-8), pages 158–163,
Beijing, China, July 30-31, 2015. ©2015 Association for Computational Linguistics and Asian Federation of Natural Language Processing

The rest of this paper is organized as follows. In Section 2, we briefly present the topic-based Chinese message polarity classification system from two aspects of graph-based ranking feature and topic-related sentimental feature. Evaluation results are presented in Section 3. Finally, the last section summarizes this paper and describes our future work.

2 System Architecture

In topic-based Chinese message polarity classification, our system is mainly composed by two parts: topic-related keyword feature selection and topic-related sentimental feature selection. In detail, topic-related keyword feature is acquired by a novel graph-based ranking algorithm of LT-IGT, and topic-related sentimental feature is obtained by topical words expansion based on word embedding and syntactic parsing according to the expanded topical words. The architecture of our system is illustrated in Figure 1.

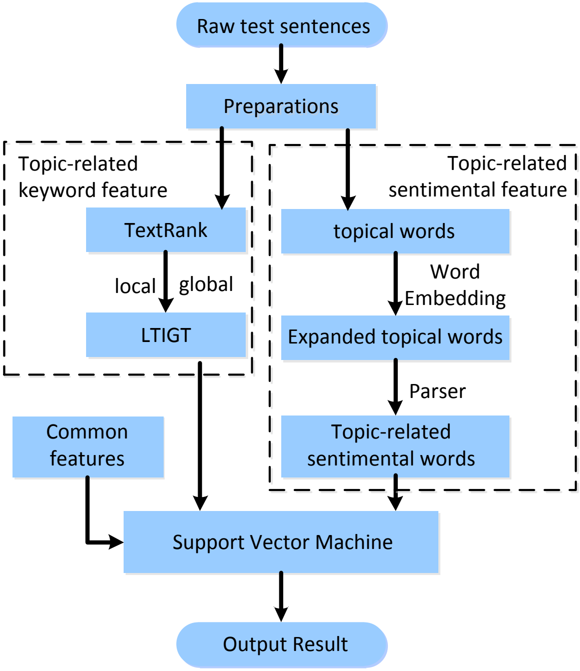

Figure 1. System architecture

2.1 Preparations

To evaluate the performance of method proposed in this paper for topic-based Chinese microblog polarity classification, we carry out experiments on dataset offered by SIGHAN8-Task 2 called Topic-Based Chinese Message Polarity Classification. This dataset is obtained from Chinese Weibo Platform, such as Sina, Tencent, NetEase etc. It contains 5*1000 manually annotated microblogs which cover 5 topics, such as "三星S6", "央行降息", etc. In experiments, we randomly select 800 microblogs of each topic for training and 200 for testing, and finally get training set of 4000 microblogs and testing set of 1000 microblogs to perform classification.

Considering the non-standard feature of microblog, the corpus is firstly normalized by following three rules.

Rule 1: Turn over the microblog with "//" to ensure the forwarding relationship and guarantee the latter sentence is analyzed based on the front sentence.

Rule 2: Delete structures like "@+username", "http://xxx" to reduce noises caused by username and website.

Rule 3: Replace the consecutive punctuations with the first one to normalize the structure of expression.

Through filtration by these rules, this paper conducts experiments on the preprocessed dataset and accesses them with traditional Precision(P), Recall(R) and F-measure(F) under Micro-average and Macro-average.

2.2 Selection of topic-related keyword feature

Inspired by TF-IDF(Salton et al., 1975,1983), words own higher local importance and lower global importance are more significant for classification. But for topic-based Chinese message polarity classification, it is obviously insufficient to extract keywords based on frequency information merely. For example, in the sentence of "GALAXY S6一改三星此前"万年大塑料"的形象，采用了前后玻璃面板和金属框组合的机身设计，为了支撑更纤薄的机身，不惜牺牲microSD卡槽和电池更换，即使如此，仍然无法与拥有完美外观的iphone媲美。", the conventional TF-IDF method tends to extract "三星、GALAXY S6、iphone、机身、卡槽、电池、外观" as keywords, but in this topic-based task, topic-related words such as "三星、GALAXY S6、卡槽、电池、外观" are expected to be selected as the keywords feature for the topic "三星". To better solve the problem of microblogs with multiple candidate topics introduced in section 1, this paper proposes a novel LT-IGT(illustrated in Figure 2) algorithm which integrates topic, position and co-occurrence information, its function is designed as follows.

$$LTIGT = LT \times IGT = TR_{lt}(v_i) \times \frac{1}{TR_{gt}(v_i)} \quad (1)$$

where $TR_{lt}(v_i)$ and $TR_{gt}(v_i)$ represent for ranking score of vertex v_i under local and global TexTRank.

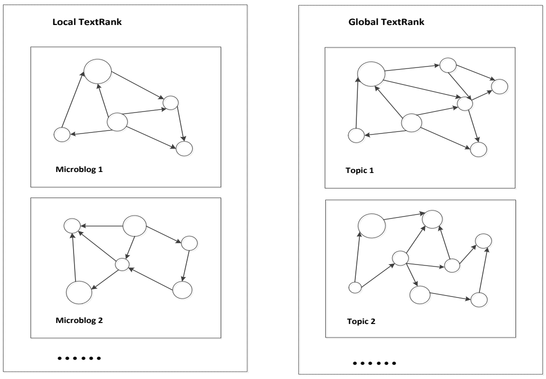

Figure 2. Graph Model of LT-IGT

The idea of TextRank(Mihalce,2004) derives from PageRank, which is achieved by dividing the text into several units to build graph model and exploiting voting mechanism for ranking. This method can model the relationship between the current word and contextual information, and the contextual related words can be recommended reciprocally. Considering the importance of a word is related to both itself and its relevant words, TextRank overcomes the independence of words in traditional "bag-of-words" model and characterizes the importance of a word more accurately.

- CST: A novel weighting method of graph-based ranking model

For each vertex in the graph, its importance ranking score benefits from adjacent nodes, and on the other hand, its own ranking score can also be transferred to the neighboring vertexes. According to the above assumptions, the indicator of vertex importance can be divided into following three parts: Coverage Importance, Semantic Similarity Importance and Topic-Related Importance. For two vertexes v_i and v_j, the influence of v_i to v_j can be transferred by the directed edge $e = < \mathrm{v_i}，\mathrm{v_j} >$. In this paper, we assign w_{ij} as the weight between v_i and v_j, α, β, γ as the proportions of these three indicators. Consequently, the weight value between two vertexes can be defined as follows:

$$w_{ij} = \alpha w_{cov}(v_i, v_j) + \beta w_{ss}(v_i, v_j) + \gamma w_{tr}(v_i, v_j) \qquad (2)$$

Where α+β+γ=1.

a) $w_{cov}(v_i, v_j)$ represents for coverage importance of v_i, it can be calculated by

$$w_{cov}(v_i, v_j) = \frac{1}{|\mathrm{Out}(v_i)|} \qquad (3)$$

Where $|\mathrm{Out}(v_i)|$ is the out-degree of vertex v_i. This formula expresses the coverage importance of v_i can be transmitted to its neighboring vertexes uniformly.

b) $w_{ss}(v_i, v_j)$ is regarded as semantic similarity importance from v_i to v_j. It can be expressed as

$$w_{ss}(v_i, v_j) = \frac{\mathrm{PMI}(v_i, v_j)}{\sum_{v_t \in \mathrm{Out}(v_i)} \mathrm{PMI}(v_i, v_t)} \qquad (4)$$

$$\mathrm{PMI}(v_i, v_j) = \log(\frac{p(v_i \,\&\, v_j)}{p(v_i)p(v_j)}) \qquad (5)$$

Where $\mathrm{PMI}(v_i, v_j)$ is the point mutual information between v_i and v_j. The larger the PMI value is, the higher the semantic similarity is. $p(v_i \,\&\, v_j)$ is the co-occurrence probability of v_i and v_j in sentences. $\mathrm{p}(v_i)$ and $\mathrm{p}(v_j)$ respectively represent for the independent occurrence probability of v_i and v_j. This function suggests that words with higher mutual information can substantially influence each other mutually.

c) $w_{tr}(v_i, v_j)$ shows the topic-related importance value of v_i. It can be computed by

$$w_{tr}(v_i, v_j) = \frac{\mathrm{P}(v_j)}{\sum_{v_t \in \mathrm{Out}(v_i)} P(v_t)} \qquad (6)$$

Where $\mathrm{P}(v_j)$ is the position importance score of v_j which can be designed with different strategies. Considering the importance of topical words in measuring position importance score, this paper assigns words occurring in topic or existing dependency with topical words a higher score than others. If we assign "vertex v occurring in topic or existing dependency with topical words" as X, the function is

$$\mathrm{P(v)} = \begin{cases} \lambda, & \mathrm{v} \in \mathrm{X} \\ 1, & others \end{cases} \qquad (7)$$

Where λ >1. We set λ =1.5 through investigation and evaluation in experiments.

- Selection of Random Jumping Probability

In the traditional graph model of TextRank, each vertex jumps to others randomly with an equal probability, which is shown in the function of $p_{rj}(w_i) = \frac{1}{|V|}$. But this method will bring about the problem of local optimization for its negligence of topical information. Considering the importance of topical words in charactering the main idea of an article, we assign topic-related words with a higher random jumping probability to get a larger score in ranking of graph model. Consequently, this paper adopts

PMI value between current word and topical word as the random jumping probability, and the function is as follows.

$$p_{rj}(w_i) = \frac{PMI(v_i,\text{topic})}{\sum_{j=1}^{|V|} PMI(v_j,topic)} \qquad (8)$$

where $PMI(v_i,\text{topic})$ denotes the point mutual information value between current word v_i and topical word topic, $|V|$ is the number of vertexes in graph model. Moreover, the calculation of co-occurrence probability for PMI is performed in unit of sentence in global TextRank, but in unit of window in local TextRank. The size of the window is assigned as 5 through experiments.

Consequently, in the construction of graph model $\mathrm{G} = (\mathrm{V}, \mathrm{E})$, vertexes, directions and weights of the links are three important points which should be considered. In this graph model, we denote the vertexes set as $\mathrm{V} = \{v_1, v_2, v_3 \ldots \ldots v_n\}$ which is combined of nouns and adjectives. Furthermore, the direction of a link between two vertexes is determined by a method of sliding window which adds links from the first word pointing to other words within the window. And the size of the sliding window is assigned as 10 through experiments. And the weight of a link is set by method of CST proposed in this paper. The basic formula of TextRank is performed for calculating the final ranking scores of each vertex. Finally, we can acquire two ranking scores for a vertex under global and local TextRank separately.

2.3 Selection of topic-related sentimental feature

In recent years, the method of word embedding based on neural network shows its outperformance in semantic expression and has attracted widespread attention paid to it(Tomas, 2013). The task of word embedding is to represent each word in corpus with a real vector, and establishing a mapping between discrete vocabulary and the feature vectors in real fields. Considering the semantic similarity between two words can be characterized by cosine value of the vectors, we propose a novel approach of topic-related sentimental word embedding which integrates syntax with semantics in this paper. This method expands topical words with word embedding first, and then performs parsing in center of these topical words to extract topical-related sentiment words based on the dependencies with them. Finally we cluster the topical-related sentiment words using K-means clustering algorithm and select the number of words belonging to a category in a microblog as the dimension values to finish the feature selection of this part.

- Expansion of Topical-words

For example, “三星S6的外观不错，但电池不行。”. Its dependency analysis result is illustrated in Figure 3 as follows.

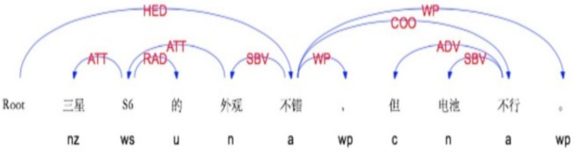

Figure 3. Example of dependency analysis result

As we can see in Figure 3, the sentimental words “不错”, “不行” do not exist dependencies with topical words “三星”, “S6”, but exist dependencies with words “外观”, “电池” of SBV(外观,不错), SBV(电池,不行). And these relationships also occupy a necessary place in topic-based sentiment analysis of Chinese mimicroblog. So we should obtain “外观”, “电池” as the expanded topical words from topical words “三星”, “S6”.

There are many approaches to expand the topical words such as PMI(Turney, 2003), and Synonyms-based method(Wang, 2009). For its better consideration of contextual information, we adopt word embedding to calculate the semantic similarity with topical words to expand the topical words. After getting word vectors, we calculate the cosine similarity between topical words and nouns under each topic, and select the highest N words as the expansion of topical words to fulfill the expansion of topical words.

- Extraction of topic-related sentimental words

As we all know, people usually express emotions towards a specific topic or object, and the emotional words often exist dependency relationship with topics or objects in syntactic analysis. Consequently, we mainly take following three dependency relations into consideration:

1) VOB

“VOB” represents for the relation between verbs and objects. Sentimental words are verbs and topical words are the objects of verbs. For example, “我喜欢三星。”. It exits “VOB” relation between “喜欢” and “三星”.

2) SBV

“VOB” represents for the relation between subjects and predicates. Sentimental words are predicates and topical words are the subjects of sentimental words. For example, “三星很漂亮。”. It exits “SBV” relation between “三星” and “漂亮”.

3) ATT

“ATT” represents for the relation of attributes. Sentimental words are attributes and topical words are the modified center of sentimental words. For example, “无与伦比的三星设计！”. It exits “ATT” relation between “无与伦比” and “三星”.

Therefore, we design an algorithm of topical-related sentimental words extraction towards dependency analysis result of microblogs. The process of this algorithm is described as below.

```
Algorithm1: Topical-related Sentimental Words Extraction
Input: Dependency analysis result(DP), Expanded Topical Words(ETW)
Output: Topical-related Sentimental Words (TSW)
for word in DP:
    if word in ETW and word.relate in ‘SBV’, ‘VOB’, ‘ATT’:
        TSW+= word.parent;
    if word.parent in ETW and word.relate in ‘SBV’, ‘VOB’, ‘ATT’:
        TSW+= word;
return TSW
```

3 Experiments

In SIGHAN8-Task2, we select emoticons, basic sentiment lexicon, dependency relation of “SBV”, “VOB”, “ATT” as common features(C), LT-IGT Ranking score as topic-related keyword feature(TK) and dependency parsing of topical words with word embedding for expansion as topic-related sentimental feature(TS).

Table 1 shows the evaluation results of our system with different groups of features.

By attempting different groups of feature for topic-related Chinese microblog sentiment classification, the performance of sentiment classification is notably improved after adding topic-related keyword feature(TK) and topic-related sentimental feature(TS). This is mainly because these two features explore both the syntactic and semantic information for classification compared with the other features. Consequently, this experiment not only demonstrates the effectiveness of LT-IGT algorithm, but also reveals the importance of topical word expansion to topic-related Chinese microblog sentiment classification.

Method	Precision	Recall	F-measure
C	0.6113	0.5572	0.5830
C+TK	0.6458	0.5982	0.6211
C+TK+TS	0.6863	0.6081	0.6448

Table 1: results of topic-based Chinese message polarity classification using SVM with different groups of features

4 Conclusion

In this paper we proposed a novel method for topic-based Chinese microblog sentiment classification, and put forward two novel feature generation approaches of LT-IGT and topic-related sentimental word embedding, with other kinds of features together, for addition to SVM classifier to perform the final polarity determination. The experimental results demonstrated the effectiveness of these two proposed features, which reminds us deep processing on syntax and semantics might be helpful for traditional regarded shallow works.

To further improve the performance of our system, we will try to extend our work in the following aspects: 1) Perform phrase structure analysis on microblog to excavate the relation between topical and sentimental words; 2) Investigate the impact on other classifiers other than SVM classifier.

Acknowledgments

We would like to thank SIGHAN8 for offering the dataset. We would also acknowledge the help of HIT-IR-Lab for proving the Chinese dependency parser(Che, 2010).

References

Lun-wei Ku and Tung-ho Wu and Li-ying Lee and Hsin-hsi Chen. 2005. Construction of an Evaluation Corpus for Opinion Extraction. In *Journal of NTCIR*, pages 513--520, Taipei, Taiwan.

Yang Shen, Shuchen Li, Ling Zheng, Xiaodong Ren and Xiaolong Cheng. 2009. Emotion mining research on microblog. In *Proceedings of Computer Science & Education (ICCSE)*, pages 477-480, LanZhou, China.

Vasileios Hatzivassiloglou and Janyce M. Wiebe. 2000. Effects of Adjective Orientation and Gradability on Sentence Subjectivity. In *Proceedings of the the 18th conference on Computational linguistics - Volume 1. Association for Computational Linguistics*, pages 299-305, New York.

Turney P D. 2002. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In *Journal of Proc An-*

nual Meeting of the Association for Computational Linguistics,pages 417--424.

Xiaolong Wang Y and Furu Wei Z. 2011. Topic Sentiment Analysis in Twitter: A Graph-based Hashtag Sentiment Classification Approach. In *International Conference on Information & Knowledge Management Proceedings*(2011).

Pang B, Lee L, Vaithyanathan S. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. Proceedings of Emnlp, pages: 79-86.

Dasgupta, S., & Ng, V. 2009. Mine the Easy, Classify the Hard: A Semi-Supervised Approach to Automatic Sentiment Classification. In *Meeting of the Association for Computational Linguistics*, pages 701-709.

Li F, Liu N, Jin H, et al. 2011. Incorporating Reviewer and Product Information for Review Rating Prediction. *In Proceedings of the twenty-second international joint conference on artificial intelligence*, pages 1820-1825.

Jakob N, Darmstadt T U, Gurevych I. 2010. Extracting opinion targets in a single and cross-domain setting with conditional random fields. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 427.

Salton G, Yu C T. 1975. On the construction of effective vocabularies for information retrieval. In *Proceedings of Acm Sigplan Notices*, pages 48-60.

Salton G, Fox E. 1983. Extended Boolean information retrieval. In *Journal of Communications of the Acm*, 26(11), pages:1022-1036.

Mihalcea R, Tarau P. 2004. TextRank: Bringing Order into Texts In *Proceedings of Unt Scholarly Works*.

Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Turney P D, Littman M L. 2003. Measuring Praise and Criticism: Inference of Semantic Orientation from Association. In *Journal of* Acm Transactions on Information Systems , 21(4), pages:315-346.

Wang S G, De-Yu L I, Wei Y J, et al. 2009. A Synonyms Based Word Sentiment Orientation Discriminating. In *Journal of Chinese Information Processing*, pages:68-74.

Wanxiang Che, Zhenghua Li, and Ting Liu. 2010. LTP: A Chinese Language Technology Platform. In Proceedings of the Coling 2010: Demonstration Volume, pages 13-16, Beijing, China.