A Morpheme-based Method to Chinese Sentence-Level Sentiment Classification

Xin Wang, Yanqing Zhao, Guohong Fu
School of Computer Science and Technology, Heilongjiang University
Harbin 150080, China
E-Mail: wangxincs@hotmail.com, yanqing_zhao@live.cn, ghfu@hlju.edu.cn

Abstract

Sentiment classification is a fundamental task in opinion mining. However, most existing systems require a sentiment lexicon to guide sentiment classification, which inevitably suffer from the problem of unknown words. In this paper, we present a morpheme-based fine-to-coarse strategy for Chinese sentence-level sentiment classification. To approach this, we first employ morphological productivity to extract sentiment morphemes from a sentiment dictionary and to calculate their polarity intensity at the same time. Then, we apply the acquired morpheme-level sentiment information to predict the semantic orientation of sentiment words and phrases within an opinionated sentence. Finally, we determine sentence-level semantic orientation by combining all the sentiment phrases and their relevant polarity scores. The experimental results on NTCIR-6 OAPT data set show our system can achieve state-of-the-art performance.

Keywords

Opinion mining, Sentiment classification, Sentiment morphemes, Dynamic polarity

1. Introduction

With the explosive growth of user-generated content on the web, opinion mining is getting considerable attentions from natural language processing community. As an active topic in opinion mining, sentiment classification aims at classifying an opinionated document or sentence as expressing a positive or negative opinion (Liu 2010), and plays a critical role in many opinion mining applications such as opinion summarization and opinion retrieval (Liu 2010; Pang and Lee 2008).

Although recent years have seen a great progress in sentiment classification, it is still challenging to develop a practical sentiment classifier for open applications. Lexical semantic orientation has proved to be a dominating indicator for sentiment classification (Turney 2002). However, it is not easy to determine the semantic orientation of a word in context. On the one hand, most existing systems require a sentiment lexicon to guide sentiment classification. But real text may contain sentiment words that are out of the sentiment lexicon. Therefore, an important issue to be concerned in sentiment classification is how to identify the semantic orientation of unknown sentiment words in real reviews. On the other hand, opinionated documents are usually expressed in a subtle manner. The semantic orientation of a subjective expression is often context and/or domain-dependent (Pang and Lee 2008). This makes it hard to explore informative cues for sentiment classification. In some cases, it needs to consider the synthetic effects of all sentiment units (e.g. sentiment words or phrases) within an opinionated sentence to determine its final semantic orientation. Therefore, both dynamic polarity identification and polarity aggregation are important factors that affect sentiment classification performance.

In this paper, we present a sentiment morpheme-based approach to Chinese sentiment classification at sentence level. We prefer morphemes to words as the basic tokens for sentiment classification because morphemes are much less numerous than words (Yuen et al. 2004; Fu et al. 2008). The relatively small number of morphemes makes it possible to construct a large-coverage morpheme lexicon for sentiment classification. Moreover, words are made of morphemes. In general, the lexical meaning of a word is closely related to its component morphemes. So we can derive the semantic orientation of an unknown sentiment word from its component morphemes (Yuen et al. 2004; Ku et al. 2009).

The rest of the paper is organized as follows: Section 2 provides a brief review of the related works on sentiment classification. Following an introduction to Chinese sentiment words and sentiment morphemes in Section 3, Section 4 details the sentiment morpheme-based approach to sentiment classification. Section 5 reports our experimental results on NTCIR-6 OAPT data (Seki et al. 2007). Finally, in Section 6 we conclude our work and discuss some possible directions for future research.

2. Related work

Over the past years, a variety of techniques have been developed for sentiment classification at word and phrase level. Turney (2002) presents a PMI-IR algorithm for detecting phrase-level semantic orientation. Following Turney's study, Yuen et al. (2004) try to infer sentiment orientation of Chinese words from their association with some strongly-polarized morphemes. More recently, Ku et al. (2009) consider eight morphological types that constitute Chinese opinion words, and further combine them in a machine learning based classifier. The study we present in this paper is an attempt to exploit a morpheme-based framework for Chinese sentiment classification. It is similar to the previous studies discussed above in that the focus is on the use of morphological cues to infer semantic orientations of words. It differs from (Yuen et al. 2004) and (Ku et al. 2009) in the way sentiment morphemes and their polar intensity are acquired.

With regard to sentence-level sentiment classification, several approaches are discussed (Yu and Hatzivassiloglou 2003; Kim and Hovy 2004; Wilson et al. 2009). Yu and Hatzivassiloglou (2003) exploit the main perspective expressed in opinionated sentences to recognize their semantic orientations. Kim and Hovy (2004) present a method that combines sentiment words to perform sentence-level sentiment classification. More recently, Wilson et al. (2009) propose a framework for sentiment analysis that distinguishes prior polarity and contextual polarity. To compute the polar intensity of an opinionated sentence, in the present study we propose a fine-to-coarse strategy, which can consider multiple granularity-level sentiments, from sentiment morphemes, sentiment words to sentiment phrases in a morpheme-based framework, and can thus handle both unknown lexical sentiments and contextual sentiments for sentence-level sentiment classification.

3. Sentiment words and morphemes in Chinese

Chinese sentiment words can be categorized into static sentiment words (SSWs) and dynamic sentiment words (DSWs). As shown in Table 1, the semantic orientation of a SSW remains unchanged. So their prior polarity in a sentiment dictionary can be identified as their real semantic orientation in opinionated text.

<table>
<tr><td colspan="2">Types</td><td>Definitions</td><td>Examples</td></tr>
<tr><td rowspan="3">Static sentiment words</td><td>positive</td><td>Sentiment words that express positive sentiments.</td><td>美丽 ‘beautyful’</td></tr>
<tr><td>negative</td><td>Sentiment words that express negative sentiments.</td><td>卑劣 ‘beggary’</td></tr>
<tr><td>neutral</td><td>Sentiment words that express neutral sentiments.</td><td>策划 ‘design’</td></tr>
<tr><td colspan="2">Dynamic sentiment words</td><td>Sentiment words whose polarity depends on their contexts.</td><td>高 ‘high’, 大 ‘big’</td></tr>
</table>

Table 1: Types of sentiment words in Chinese text

Types	Examples	Words composed by sentiment morphemes
Positive morphemes	美 ‘beauty’	精美 ‘exquisite’, 优美 ‘graceful’
	爱 ‘love’	喜爱 ‘like’, 爱慕 ‘adoration’
Negative morphemes	污 ‘dirty’	污染 ‘pollution’, 贪污 ‘corruption’
	败 ‘fail’	腐败 ‘corruption’, 败坏 ‘undermine’

Table 2: Types of sentiment words in Chinese text

However, a precompiled sentiment dictionary cannot cover all sentiment words in open opinionated documents. This raises an issue of predicting the semantic orientation of unknown sentiment words. To address this problem, we introduce sentiment morphemes. As illustrated in Table 2, we consider two types of sentiment morphemes, namely positive morphemes and negative morphemes, to infer semantic orientations of unknown sentiment words.

As for DSWs, we cannot determine their semantic orientations simply by consulting a sentiment lexicon because they may have different semantic orientations in different contexts. In the present study, we attempt to disambiguate the polarity of DSWs at phrase-level.

4. Approach

4.1 Sentiment morpheme extraction

Sentiment morphemes prove to be helpful in inferring the semantic orientations of unknown sentiment words (Yuen et al. 2004; Ku et al. 2009). In fact, Chinese sentiment words usually contain a key morpheme that determines their emotional tendency. Take the following two sentiment words for example, 败坏 'undermine' and 腐败 'corruption'. They share a same negative sentiment morpheme 败 'fail', and thus have the same negative orientation. However, there is not a dictionary of sentiment morphemes available for sentiment analysis. To avoid this, we employ morphological productivity proposed by Fu et al. (2008) to automatically extract sentiment morphemes from a sentiment lexicon.

For each word in the sentiment lexicon, we first segment it into morphemes, and further compute their morphological productivity in forming sentiment words. It should be noted that with a view to the coverage of the extracted morpheme lexicon and the reliability of the morphological productivity, in the present study we only consider mono-character morphemes.

The morphological productivity of a positive sentiment morpheme *m*, denoted by $MP_{PSM}(m)$, can be formulated as

$$MP_{PSM}(m) = \frac{Count(m, w_{positive})}{Count(m, w_{postive \cup negative})} \tag{1}$$

Where, $Count(m, w_{positive})$ denotes the total number of positive words that contain the morpheme *m*, and $Count(m, w_{positive \cup negative})$ denotes the total number of words in the sentiment lexicon.

Similarly, the morphological productivity of a negative sentiment morpheme *m*, denoted by $MP_{NSM}(m)$, can be formulated as

$$MP_{NSM}(m) = \frac{Count(m, w_{negative})}{Count(m, w_{postive \cup negative})} \tag{2}$$

Where, $Count(m, w_{negative})$ denotes the total number of negative words that contain the morpheme *m*.

A larger productivity value implies higher likelihood of an eligible sentiment morpheme. In fact, MP_{PSM} and MP_{NSM} also indicate the polar intensity of sentiment morphemes.

Therefore, they can be used to estimate the polar intensity of sentiment words. If a sentiment word contains both positive morphemes and negative morphemes, or several sentiment morphemes with a same orientation, it will have the same semantic orientation and polar intensity as the morpheme that has the highest MP_{PSM} or MP_{NSM} value. If a sentiment word does not contain any sentiment morphemes, then its polar intensity is zero.

4.2 Morpheme segmentation

To achieve the necessary morpheme information for inferring the semantic orientations of unknown sentiment words, a morpheme segmentation module is therefore needed to decompose a word into a sequence of morphemes. In view of system efficiency, in the present study we employ the forward maximum matching (FMM) word segmentation technique to perform this task. While FMM has the advantage of simplicity and efficiency in word decomposition, it does not work well for some words such as 非常规 *fei1-chang2-gui1* 'unconventionality' that involve multiple possible way of segmentation. To resolve such ambiguity, we use a set of rules to correct error morpheme segmentation yielded by FMM. The rules for correction have the following form: *Error FMM-segmentation* → *correct segmentation* (Wang et al. 2000).

4.3 Polarity determination for sentiment phrases

To detect sentiment phrases within sentences, we consider two-word phrases with four types of structures as shown in Table 3. Besides adjectives, verbs and idioms, negation is another important indicator of sentiment orientation. So in addition to the phrases concerned in (Turney 2002), we also take into account phrases whose initial word is a negation. Moreover, we reduce some function words like 的 'of' from the opinionated sentences under discussion before phrase-level polarity identification.

Structures	Definitions	Examples
Adjective ∪ #	Phrases containing an adjective.	成功率/高/ 'high success rate'
Verb ∪ #	Phrases containing a verb.	详细/讨论/ 'discuss in detail'
Idiom ∪ #	Phrases containing an idiom.	企图/掩人耳目/'attempt to deceive the public'
Negation and #	Phrases beginning with a negation.	不/好/ 'not good'

Table 3: Structures of sentiment phrases

Thus, we can identify the sentiment orientation of the extracted sentiment phrases and estimate their polar intensity scores using a set of rules defined in Table 4.

The initial word	The final word	Polarity score of a sentiment phrase
$Score(w_1) > 0$	$Score\ (w_2) > 0$	$Score(w_1) \times Score(w_2)$
$Score(w_1) < 0$	$Score\ (w_2) < 0$	$-1 \times Score(w_1) \times Score(w_2)$
$Score(w_1) = 0$	$Score\ (w_2) \neq 0$	$Score(w_2) \times \|Score(w_2)\|$
$Score(w_1) \neq 0$	$Score\ (w_2) = 0$	$Score(w_1) \times \|Score(w_1)\|$
$Score(w_1) = 0$	$Score\ (w_2) = 0$	0
$Score(w_1) > 0$	$Score\ (w_2) < 0$	$Score(w_1) \times \|Score(w_2)\|$, if $\|Score(w_1)\| > \|Score(w_2)\|$
		$Score(w_1) \times Score(w_2)$, if $\|Score(w_1)\| < \|Score(w_2)\|$
$Score(w_1) < 0$	$Score\ (w_2) > 0$	$Score(w_1) \times \|Score(w_2)\|$, if $\|Score(w_1)\| > \|Score(w_2)\|$
		$Score(w_1) \times Score(w_2)$, if $\|Score(w_1)\| < \|Score(w_2)\|$
w_1 is a DSW.	$Score\ (w_2) \neq 0$	$Score(w_1) \times Score(w_2)$
$Score(w_1) \neq 0$	w_2 is a DSW.	$Score(w_1) \times Score(w_2)$
w_1 is negation.	$Score\ (w_2) \neq 0$	$-1 \times Score(w_2)$

Table 4: Computing polarity scores for sentiment phrases

4.4 Polarity determination for dynamic sentiment words

As we have mentioned above, the semantic orientation of some sentiment words in real reviews may be different from their prior polarity defined in lexica.

(a) 英国/ns 国防部/n 一/m 份/q 内部/f 报告书/n 在/p 四/m 年/q 前/f 发出/v 警告/v 说/v ，/w 接触/v 到/v 贫铀弹/n 的/u 人/n 患上/v 癌症/n 的/u 风险/n 较/d 高/a 。/

An internal report issued by British Ministry of Defense four years ago warned that People exposed to depleted uranium bombs would be at a higher risk of cancer.

(b) 该所/r 科学家/n 认为/v ，/w 应用/v 这/r 一/m 技术/n 把/p 患者/n 的/u 体细胞/n 与/c 人类/n 胚胎/n 干/v 细胞/n 融合/v 在/p 一起/s 的/u 成功率/n 较/d 高/a 。/w

Scientists from the institute believe that the success rate would be higher if applying this technology to mix the patient's somatic cells and human embryonic stem cells.

Figure 1: Examples of sentences with the dynamic sentiment word 高 ‘high’

Figure 1 illustrates two different semantic orientations of the DSW 高 ‘high’ in two different opinionated sentences. In sentence (a), the DSW 高 ‘high’ expresses a negative orientation when collocating with the phrase 患上/v 癌症/n 的/u 风险/n ‘risk of cancer’. But in sentence (b), it will express a positive opinion when modifying the noun 成功率 ‘success rate’. Therefore, correct disambiguation of semantic orientation for DSWs is of great value to sentiment classification.

Rather than directly determining the polarity of DSWs, here we resolve the DSW problem by identifying the semantic orientation of phrases that involve DSWs. To achieve this, we use the rules in Table 4. Table 5 illustrates the process of calculating the polar intensity of two sentiment phrases that contain the DSW 高 'high'.

<table>
<tr><th colspan="4">Phrase</th><th rowspan="2">Score(phrase)</th></tr>
<tr><th colspan="2">First word</th><th colspan="2">Second word</th></tr>
<tr><td rowspan="4">风险</td><td>PSW(风)=0.636</td><td rowspan="4">高</td><td rowspan="2">PSW(高)=0.843)</td><td rowspan="4">Score(风险)
=-1*NSW('险')*PSW('高')
=-0.684 * 0.843
=-0.5766</td></tr>
<tr><td>PSW(险)=0.342</td></tr>
<tr><td>NSW(风)=0.374</td><td rowspan="2">NSW(高)=0.165</td></tr>
<tr><td>NSW(险)=0.684</td></tr>
<tr><td rowspan="6">成功率</td><td>PSW(成)=0.630</td><td rowspan="6">高</td><td rowspan="3">PSW(高)=0.843</td><td rowspan="6">Score(成功率)
=PSW('功')*PSW('高')
=0.771 * 0.843
=0.6500</td></tr>
<tr><td>PSW(功)=0.771</td></tr>
<tr><td>PSW(率)=0.630</td></tr>
<tr><td>NSW(成)=0.389</td><td rowspan="3">NSW(高)=0.165</td></tr>
<tr><td>NSW(功)=0.257</td></tr>
<tr><td>NSW(率)=0.407</td></tr>
</table>

Table 5: An example for polarity identification of the phrases 风险高 'high risk' and 成功率高 'high success rate' with the dynamic sentiment word 高 'high'

4.5 Sentence-level polarity identification

Given an opinionated sentence *S*, we can employ the above approach to detect sentiment phrases within *S*, and at the same time calculate their polar intensity. The average polar intensity of these phrases (denoted by *AveragePI*) can be further used to determine the semantic orientation of the opinionated sentence *S* (denoted by *O(S)*) under the rule defined in equation (3).

$$O(S) = \begin{cases} Positive & AveragePI \geq \lambda_P \\ Neutral & \lambda_N < AveragePI < \lambda_P \\ Negative & AveragePI \leq \lambda_N \end{cases} \quad (3)$$

Where, λ_P and λ_N denote two experientially determined thresholds.

Besides phrase-level polar cues, some syntactic features are also taken into consideration during sentence-level polarity identification. If an opinionated sentence contains an adversative conjunction like 但是 'but', we can only consider the part of the sentence that follows the adversative conjunction in calculating its polar intensity, because the author generally wants to emphasize this part.

Based on the above principle, we implement a system for Chinese sentence-level sentiment classification. Figure 2 presents the pseudo code for the classification algorithm.

Algorithm 1. Sentence-level sentiment classification.

Input: An opinionated sentence S

Output: The semantic orientation of the sentence $O(S)$

Preprocessing: word segmentation and part-of-speech tagging using the system in (Fu et al. 2008)

Detecting sentiment morphemes within S and calculating their polar intensity MP_{NSW} or MP_{PSW}.

$Score(S) = 0$

Extracting a set of sentiment phrases $\{P_i|1 \le i \le n\}$ from S

for each sentiment phrase P_i ($1 \le i \le n$) in S

  for each word w_j ($j \in [1,2]$) in P_i

    if $MP_{PSM}(m_k) \ge MP_{NSM}(m_k)$

      $Score(w_j) = \max\{MP_{PSM}(m_k), MP_{NSM}(m_k)\}$

    else

      $Score(w_j) = -1 \times \max\{ MP_{PSM}(m_k), MP_{NSM}(m_k)\}$

  endfor

  $Score(P_j) = Score(w_1) \times Score(w_2)$

  $Score(S) = Score(S) + Score(P_j)$

end for

Determining $O(S)$ with equation (3).

Figure 2: The algorithm for sentence-level sentiment classification

5. Experiments

To assess the effectiveness of our approach, we conducted experiments on the NTCIR-6 OAPT test set (Seki et al. 2007). This section reports the results of these experiments.

5.1 Experimental setup

The NTCIR-6 OAPT test set for Chinese consists of 700 opinionated documents and 9247 sentences, in which 62% are opinionated sentences under the lenient standard (Seki et al. 2007). For comparison, the performance is reported in terms of the same metrics as used in NTCIR-6 OAPT. They are F-score (F), recall (R), precision (P) under the LWK evaluation with the lenient standard. To conform to NTCIR-6 OAPT evaluation, the sentiment density-based naive Bayesian classifier in (Wang and Fu 2010) is also embedded in the present system to perform opinionated sentence detection.

The source sentiment lexicon is from (Xu et al. 2007). In our experiments we exclude some derived words like 不美丽 'not beautiful' from it before morpheme extraction, and finally obtain a total of 1382 positive morphemes and 2026 negative morphemes.

5.2 Experimental results

To test our system, we conducted three experiments. Firstly, to examine the effects of different sentiment granularities on sentiment classification, we consider three sentiment granularities, namely morpheme-level, word-level and phrase-level sentiments, respectively for sentiment-level sentiment classification. Secondly, to find out which type of basic tokens is more suitable for Chinese sentiment classification, we compare the performance achieved by using words and morphemes as the basic tokens, respectively. Finally, we make a comparison between our system and the best performing three systems at NTCIR-6 OATP. The results are summarized in Tables 6, 7 and 8, respectively. It should be noted that here we set λ_P = 0.1 and λ_N = -0.1.

Granularity	**Morphemes as basic tokens**			**Words as basic tokens**		
	P	**R**	**F**	**P**	**R**	**F**
Morpheme	0.357	0.442	0.395	-	-	-
Word	0.363	0.448	0.400	0.340	0.419	0.376
Phrase	0.364	0.450	0.402	0.347	0.427	0.383

Table 6: Results for sentence-level sentiment classification with different sentiment granularities and basic tokens

From these results, we can draw some conclusions: (1) By comparing Table 6 with Table 7, we can observe that the system with morphemes as the basic tokens consistently outperforms the system based on words, illustrating the benefits of using sentiment morphemes in dealing with unknown sentiment words. (2) The results in Table 6 and Table 7 demonstrate an improvement of performance after using phrase-level granularity information. The reason may be that under the fine-to-coarse framework, the use of sentiment phrases can handle both internal and external contextual cues, and thus result in performance improvement.

Systems	**P**	**R**	**F**
The CUHK system for NTCIR-6	0.522	0.331	0.405
The NTU system for NTCIR-6	0.335	0.448	0.383
The UMCP system for NTCIR-6	0.292	0.441	0.351
Our system	0.364	0.450	0.402

Table 7: Comparison of our system with the best three systems for sentiment classification at NTCIR-6 OAPT

(3) As can be seen from Table 8, our system performs better than the top three systems at NTCIR-6 OAPT with regard to recall. As for F-score, our system is slightly worse than the best system but better than the other two systems, showing in a sense the feasibility of

the proposed approach.

6. Conclusions

In this paper, we have presented a morpheme-based approach to Chinese sentiment classification at sentence level. Compared to previous works, our approach offers a fine-to-coarse framework for sentiment classification in a way that more features, including word-formation patterns, contextual features and even syntactic features, can be statistically explored to predict sentence-level sentiment orientations. The preliminary experiments on the NTCIR-6 OAPT test set for Chinese shows that our approach can achieve state-of-the-art performance.

The results of the present study suggest two possibilities for future research. First, we have shown the benefit of using morphemes as the basic tokens for sentiment classification, further research is still needed to exploit more deep morphological cues for unknown sentiment word prediction. Second, the present system relies on a set of rules, and is not applicable for open domains. Thus, more extensive research would be necessary to develop a robust sentiment classifier for real applications.

Acknowledgements

This study was supported by National Natural Science Foundation of China under Grant No.60973081 and Harbin Innovative Foundation for Returnees under Grant No.2009RFLXG007, respectively.

7. References

Fu, G., C. Kit and J. J. Webster, 2008, Chinese word segmentation as morpheme-based lexical chunking, Information Sciences, vol.178, no.9, pp.2282-2296.

Kim, S.-M. and E. Hovy, 2004, Determining the sentiment of opinions, Proceedings of COLING ’04.

Ku, L.-W., T.-H. Huang and H.-H. Chen, 2009, Using morphological and syntactic structures for Chinese opinion, Proceedings of EMNLP ’09, pp.1260-1269.

Liu, B., 2010, Sentiment analysis and subjectivity, Handbook of Natural Language Processing, (Second Edition, Editors: N. Indurkhya and F. J. Damerau).

Pang, B. and L. Lee, 2008, Opinion mining and sentiment analysis, Foundations and Trends in Information Retrieval, vol.2, nos.1-2, pp.1-135.

Seki, Y., D. K. Evans, L.-W. Ku, L. Sun, H.-H. Chen and N. Kando, 2007, Overview of opinion analysis pilot task at NTCIR-6, Proceedings of NTCIR-6 Workshop Meeting, pp.265-278.

Turney, P. D., 2002, Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews, Proceedings of ACL’02, pp. 417-424.

Wang, X. and G. Fu, 2010, Chinese subjectivity detection using a sentiment density-based naive Bayesian classifier, Proceedings of ICMLC'10, pp.3299-3304.

Wang, X., G. Fu, D. S.Yeung, J.N.K.Liu and R. Luk, 2000. Models and algorithms of Chinese word segmentation. Proceedings of ICAI'00, pp.1279-1284.

Wilson, T., J. Wiebe and P. Hoffmann, 2009, Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis, Computational Linguistics, vol.35, no.3, pp.99-433.

Xu, R., K.-F. Wong and Y. Xia, 2007, Opinmine: Opinion mining system by CUHK for NTCIR-06 Pilot Task, Proceedings of NTCIR-6 Workshop Meeting, pp.350-357.

Yu, H. and V. Hatzivassiloglou, 2003, Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences, Proceedings of EMNLP '03, pp.129-136.

Yuen, R.W.M., T.Y.W. Chan, T.B.Y. Lai, O.Y. Kwong and B.K.Y. T'sou, 2004, Morpheme-based derivation of bipolar semantic orientation of Chinese words, Proceedings of COLING '04, pp.1008-1014.