PolyU at MOAT in NTCIR-8

Wenping Zhang, Dehong Gao, Wenjie Li

Department of Computing, The Hong Kong Polytechnic University, Hong Kong

{cswzhang, csdgao, cswjli}@comp.polyu.edu.hk

ABSTRACT

In this paper, we briefly summarize our experience in participating in the Multilingual Opinion Analysis (MOAT) tasks in NTCIR-8 and present our preliminary experimental analysis of the effects of the opinion lexicons employed in Chinese opinion mining.

Keywords

Opinion mining, sentiment analysis, opinion lexicon, machine learning.

1. INTRODUCTION

With the growing availability and popularity of opinion-rich resource, such as movie and product review sites, online forum and personal blogs, new opportunities and challenges arise as people now can, and do, actively use information technologies to seek out and understand the opinions of others. As a result, opinion mining and sentiment analysis become a hot spot in Natural Language Processing (NLP). It aims to find out the opinion or attitude of a speaker or a writer to a specific topic. For instance, the opinion analysis of the product reviews can give the customers a guidance to select a good or suitable product. It can also collect important information for the producers or the retailers to improve their products and service. Owning to its practical use in the modern life, it has attracted more and more attentions in both research and industry communities.

In general, the basic task of opinion mining is to judge if the speaker or a writer expresses an opinion or not (i.e., Opinioned or Not Opinioned) in a sentence (or a document or a text snippet). If it does, the next task is sentiment analysis which determines the polarity of the opinion (i.e. Positive, Negative or Neutral). After that, it is the task to recognize the holder and the target of the opinion. Multilingual Opinion Analysis Task (MOAT) at NTCIR-8 includes 6 subtasks described as follows: (1) opinion judgment, (2) polarity judgment, (3) opinion holder identification, (4) opinion target identification, (5) scenario-based evaluation using opinion question, and (6) cross-lingual opinion analysis. Since this is our first time participating at MOAT, we only get involved in the first two subtasks and experiment on simplified Chinese and English.

2. Simplified Chinese Opinion Analysis Tasks

Both opinion judgment and polarity judgment can be cast as the binary or multi-class classification tasks. We compare two classification models, i.e., the Support Vector Machines (SVM) and Conditional Random Fields (CRF) models. The following three tools are utilized.

(1) Lib-SVM is a classification tool provided by Chih-Chung Chang and Chih-Jen Lin [1]. It is designed for easy use and is able to support multi-class classification.

(2) SVM-multiclass is another SVM-based classification tool provided by Thorsten Joachims [2]. With certain optimization algorithm, it performs much better in multi-class classification, as reported in [3].

During our work, we find that if a sentence is a strong opinioned sentence, it is more likely that the sentences surrounding it are opinioned sentences as well. If this contextual information could be taken into consideration, it may lead to the performance improvement. So the CRF model is also examined.

(3) CRF a probabilistic framework for labeling and segmenting sequential data, based on the conditional approach and heavily motivated by the principle of maximum entropy [4]. It is the model that can take the contiguous information of the features into consideration.

The features examined in our work include text *N*-gram features (more precisely, Uni-grams and Bi-grams of Chinese characters) and lexicon features. In has been concluded in previous work that the orientation and gradability of the opinion words, especially the adjective words, has an important effect on the attitude of the sentence [5, 6]. For example, the performance of lexicon-based methods is proved fairly good in [7, 8]. Four Chinese opinion lexicons are collected from HowNet. They are positive opinion words, negative opinion words, positive emotional words and negative emotional words. Table 1 summarizes the statistics of the words contained in the four lexicons. Single character words account for small percentage compared with the others while most words are 2-character words.

Table 1. Statistics of Four Lexicons

Numbers	Positive opinion words	Negative opinion words	Positive emotional words	Negative emotional words
1-character words	438	410	138	181
2-character words	1891	1323	518	667
N-character words (*N*>2)	1402	1384	181	407
Total	3731	3117	837	1255

These lexicons are used to define the features for machine learning, though in the different ways in the SVM model and the CRF model.

2.1 Comparison of Different Models

The evaluation is conducted on the MOAT-7 simplified Chinese data set, with 4/5 of all the documents in each given topic used for training classifiers, and the rest 1/5 for test.

Table 2 below presents the overall opinion and/or polarity classification accuracies on all the topics based on text Uni-gram features only. "Opinion" classifies sentences into opinioned (OP) ones and non-opinioned (NOP) ones and "Polarity" classifies the identified opinioned sentences into positive (POS), negative (NEG) and neutral (NEU) of polarity. "Opinion + Polarity" combines opinion and polarity classification into one task which attempts to differentiate the sentences among four classes, i.e., NOP, OP-POS, OP-NEG and NOP-NEU. The results clearly demonstrate the improvement of the integration strategy overall the sequential processing of opinion and polarity. It also shows that SVM-multiclass is more effective than LIB-SVM on our test data in feature selection and thus lead to better performance. Note that CRF does not work very well on "Opinion" and "Polarity", compared with SVM, but it significantly outperforms SVM on "Opinion + Polarity".

Table 2. Opinion/Polarity Classification based on Text Uni-gram Features

Overall accuracy on all topics	Opinion	Polarity	Opinion + Polarity
Lib-SVM	65.47%	53.72%	60.10%
SVM-multiclass	67.04%	56.38%	63.68%
CRF	66.59%	51.87%	67.50%

Table 3. Opinion/Polarity Classification based on Bi-gram plus Lexicon Features

Overall accuracy on all topics	Opinion only	Polarity only	Opinion + Polarity
Lib-SVM	67.26%	54.26%	60.54%
SVM-Multiclass	67.49%	53.19%	60.76%
CRF	54.30%	47.68%	52.69%

Table 3 then presents the classification accuracies based on Bi-gram features. This time, the lexicons have a role to play. For the SVM models, only the character Bi-gram that appears in a lexicon word is considered to be a feature and is endowed with a weight. If the Bi-grams are sub-matched with the lexicon words, they are endowed with a weight that is half of the fully matched ones which is measured by its occurrence in the sentence. These lexicon-based features are organized into four types of features, corresponding to the four different lexicons respectively. Meanwhile, the characters that are neither fully matched nor partially matched are still handled as the Uni-grams.

For the CRF model, due to the limitation of feature space, currently, we use the statistics of lexicon words, rather than the lexicon-matched Bi-grams, to represent the features. We choose the number of opinion words and their categories as the features. Since we have four different opinion lexicons, we have four lexicon features for each sentence. In order to make the features more operable, the feature weight is set to 0 if no lexicon word occurred in the sentence. It is set to 1 or 2 if the number of the lexicon words occurred in the sentence is in the range of (0, 5) or [5, ∞). The Uni-gram features are used in the same way as in the SVM models.

Therefore, the results shown in Table 3 are actually obtained based on the combination of Uni-grams, Bi-grams and lexicon words. Comparing Tables 2 and 3, the only improvement is observed with Lib-SVM and the improvement is not significant. This contradicts with our assumption and previous conclusions that lexicons play an important role in opinion mining in movie and product reviews. It seems that lexicon-based features do not work very well in news opinion mining. This might because that news is different from movie and product review.

The opinion expressions in movie and product reviews are rather explicit, such as "I like this movie" and "This camera is good." It is easily to determine the opinion and the polarity of these sentences through opinion words or their statistics. On the contrary, the expressions in news are inclined to objectivity. For example, in the following (E1) there are 7 opinion words: "全面 (comprehensive)", "公正 (fair-minded)", "持久 (lasting)", "和平 (peace)", "稳定 (stability)", "繁荣 (prosperity)", "发展 (development)". However, it is a NOP sentence actually.

(E1) 中东实现全面、公正、持久的和平将有利于本地区的稳定、繁荣与发展, 应当在联合国有关决议和"土地换和平"原则的基础上实现中东问题的公正、合理解决。(If the Middle East can achieve a comprehensive, fair-minded and lasting peace, it would be conducive to the stability, prosperity and development of this region. It should base on the United Nations' relevant decision and the principle of "Land for Peace" to solve the Middle East business justly and reasonably.)

On the other hand, two positive words (i.e., "同意 (agree)", "放宽 (relax)") and three negative words (i.e., "减缓 (mitigate)", "紧缩 (deflation)", "负面 (negative)") are included in (E2). If we count on opinion word statistics, it will be judged as a negative polarity sentence. However, it is an obvious positive polarity sentence as we all can see.

(E2) 国际货币基金组织12日表示,同意放宽对泰国贷款的一些条件,以减缓紧缩政策给泰国带来的负面效应。(The International Monetary Fund said on the 12th that it agreed to relax certain conditions of loans to Thailand to mitigate the negative effects the deflation policy that had brought to Thailand.)

We thus further conduct the following experiments and analysis, and hope to figure out whether the lexicon features are necessary in opinion mining and why they are not effective in our models.

2.2 Further Examination of Opinion Lexicon Features

First of all, we randomly sample some OP and NOP sentences from the given MOAT-7 data set and count the number of the opinion words contained in these sampled sentences, as shown in Table 4.

Table 4. Distributions of Opinion Words in OP and NOP Sentences

Size of Sampling	Number of opinion words	
	in OP sentences	in NOP sentences
500 sentences	1794	642
400 sentences	1540	574
300 sentences	1031	296
200 sentences	522	144

The value of t in t-test is 2.6 > 2.45 (p=0.05). So it can be concluded that the distribution of the opinion words in OP sentences is significantly different from the distribution of the opinion words in NOP sentences, and in this regard the lexicons should have a role to play in opinion mining. Seven follow-up experiments are carried out to further exam the effect of lexicons on opinion classification and polarity classification using Lib-SVM. The following table shows the results.

Table 5. Effect of Lexicons

Overall accuracy on all topics	Opinion	Polarity
Uni-gram	65.47%	53.72%
Uni-gram + Lexicon	65.49%	53.72%
Bi-gram	45.40%	31.91%
Bi-gram + Lexicon	67.26%	54.26%
N-gram(N > 3)	33.20%	21.68%
Lexicon	49.73%	43.24%
Lexicon word statistics	24.09%	53.52%

From the above table, we can see that the use of lexicons can improve the classification performance no matter we use Uni-gram or Bi-gram to represent the features. In the case of Uni-gram, the lexicons can help us to enhance the importance of the opinion words. However, since the 1-chracter opinion words only take a small part of the lexicons, the improvement is quite limited. Meanwhile, although the number of N-character (N>2) words is greater than that of 1-chracter words, the probability that they occur in a sentence is quite limited. Thus many sentences cannot be represented effectively. Of course, the result of it is not very good. So, only (Bi-grams + Lexicon) shows the improvement.

If we only extract the words that appear in the lexicons and use the statistics of them as features, feature space can be largely reduced and the importance of the opinion words can be emphasized. However, the thematic information is also important in the news's opinions. The loss of this information would be very disadvantageous to opinion and polarity classification. This has been illustrated in Table 5.

Sine the opinion lexicons we use are general purpose ones, it may bring noise in classification. Building domain-dependent opinion lexicons (or refining a domain-dependent opinion lexicon into a domain-dependent one) and lexicon domain adaptation are of our particular interest in our future work.

3. English Opinion Analysis Tasks

In addition to the Chinese opinion analysis tasks, we also participate in the English opinion analysis tasks. The approach is built on the SVM model (with Lib-SVM) and the features concerned include English word Uni-grams and the lexicon-based features derived from SentiWordNet [9] and Wilson lexicon [10].

The word Uni-grams can be weighted using either TF*IDF or binary (1 for present, 0 for absent), while the weight of lexicon-based features are binary.

The MOAT-7 English data set covers varieties topics, such as "Yasukuni Shrine", "regenerative medicine" and "uranium bullets" etc. In order to avoid the domain divergence problem and meanwhile to retain sufficient training samples, we consider merging the topics into five domains, including biography, politics, war, medicine and technique. Then a SVM classifier is trained on each domain. Alternatively, of course we can take the data in all domains to train a single overall classifier.

Lib-SVM provides four basic kernel functions: RBF, Linear, Polynomial, and Sigmoid. In general the RBF kernel is a reasonable first choice. However the RBF kernel function maps data to a higher dimensional space. It makes the process of training and test extremely time-consuming. Taking the training data, feature weighting and learning kernel into consideration, we design three learning plans, as shown in Table 6 in the next page.

Table 6. Learning Plans

Plan	Feature	Training data set	Kernel function
Plan-I	Lexicon + Uni-gram (TF-IDF)	Each domain	RBF
~~Plan-I'~~	~~Lexicon + Uni-gram (TF-IDF)~~	~~All domain~~	~~RBF~~
Plan-II	Lexicon + Uni-gram (binary)	All domains	RBF
Plan-III	Lexicon + Uni-gram (binary)	All domains	Linear

Actually, we have four plans originally. The purposes of these plans are to compare the effects of training data scope, feature design and kernel selection. However, Plan-I' is finally excluded because it requires a large amount of time to train the classifiers. Tables 7 and 8 present the performance evaluations, where L and S denote Lenient and Strict respectively, and P/R/F are calculated according to the formulas provided by MOAT-8.

Table 7. Results of Opinion Classification

Plan	L/S	Opinion		
		P	R	F
Plan-I	L	27.83%	40.69%	33.05%
Plan-II	L	57.82%	27.78%	37.53%
Plan-III	L	20.13%	35.41%	25.67%
Plan-I	S	11.32%	58.54%	18.97%
Plan-II	S	30.58%	43.53%	35.92%
Plan-III	S	10.89%	50.54%	17.92%

Table 8. Results of Polarity Classification

Plan	L/S	Polarity		
		P	R	F
Plan-I	L	23.26	13.79	17.32
Plan-II	L	48.43	12.06	19.31
Plan-III	L	20.01	30.51	24.17
Plan-I	S	08.14	17.07	11.02
Plan-II	S	17.19	12.71	14.67
Plan-III	S	10.12	48.12	16.72

4. Conclusion

In this paper we briefly summarize our experience in participating in the Multilingual Opinion Analysis (MOAT) tasks in NTCIR-8 and present our preliminary experimental analysis of the effects of the opinion lexicons employed in Chinese opinion mining and the training strategies for English opinion mining. We feel that the domain or topic dependent lexicon refinement and domain or topic dependent training sample selection are worth further investigation in our future studies.

5. Acknowledgements

The work presented in this paper is supported by the Hong Kong RGC grant (Project No. PolyU5230/08E) and the Hong Kong Polytechnic University internal grant (Account No. G-YH53).

6. References

[1] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[2] http://svmlight.joachims.org/svm_multiclass.html

[3] Crammer K. and Singer Y. 2001. On the Algorithmic Implementation of Multi-class SVMs, *Journal of Machine. Learning Research*, 2, 265-292.

[4] Clifford P. 1990. Markov Random Fields in Statistics. In Geoffrey Grimmett and Dominic Welsh, editors, *Disorder in Physical Systems: A Volume in Honour of John M. Hammersley*, 19-32. Oxford University Press.

[5] Hatzivassiloglou V. and Wiebe J. 2000. Effects of Adjective Orientation and Gradability on Sentence Subjectivity, in *Proceedings of the International Conference on Computational Linguistics*, 299-305.

[6] Hu M. and Liu B. 2004. Mining and Summarizing Customer Reviews. In *Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 168-177.

[7] Andreevskaia A. and Bergler S. 2006. Mining WordNet for Fuzzy Sentiment: Sentiment Tag Extraction from WordNet Glosses. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, 209-216.

[8] Kim S. and Hovy E. 2004. Determining the Sentiment of Opinions. In *Proceedings of the 20th international conference on Computational Linguistics*, 1367-1373.

[9] http://sentiwordnet.isti.cnr.it/

[10] http://www.cs.pitt.edu/mpqa/opinionfinderrelease