NECLC at Multilingual Opinion Analysis Task in NTCIR8

Bo Chen
NEC Laboratories China
Beijing 100084, P.R. China
chen_bo@nec.cn

Saike He
Beijing University of Posts and Telecommunications
Beijing 100876, P.R. China
hsk000@gmail.com

ABSTRACT
In this paper, we briefly describe our machine-learning based method used in the NITCIR8 MOAT task, particularly, the opinioned sentence judgment subtask on both English side and Chinese side. We view this subtask as a binary classification problem and build a supervised-learning based framework. To extract meaningful sentiment features, we propose several n-gram patterns to assemble basic words and part-of-speech tags. Meanwhile, our basic classifiers are trained merely on the previous NTICR annotated corpus, in which samples are inadequate and unbalanced. Thus, we adopt a few self-learning strategies to utilize the NTCIR8 testing corpus to adjust our basic classifiers. Using the same learning framework in both language sides, we get similar performances.

Categories and Subject Descriptors
H.3.1 [**Information Storage and Retrieval**]: linguistic processing

General Terms
Experimentation

Keywords
Opinion Analysis, Sentiment Classification, Maximum Entropy

1. INTRODUCTION
Sentiment Analysis has been a research focus in recent years. Since 2006, the Blog track has been introduced into the NIST's TREC series. The most important aspect of this track is to retrieval opinioned text about some specified topics among millions of blog posts [3], most of which were written in English. Meanwhile, a new task, called MOAT (Multilingual Opinion Analysis Task) has been introduced into the NTCIR workshop. Different from the Blog track, the MOAT task focused on opinionated text analysis across languages, including English, Chinese and Japanese.[6, 7, 8]

In NTCIR8 MOAT, there are five subtasks: opinionated/relevance sentence judgment, polarity judgment, and opinion holder/target identification [8]. We only participate in the opinion sentence judgment subtask, in both English side and Simplified Chinese side.

In the remainder of this paper, we will firstly introduce our methods in section 2, and display our evaluation results in section 3. Finally, some discussions are made in section 4.

2. SYSTEM DESIGN
In participating in the NTICR8 MOAT opinion sentence judgment subtask, we focus on developing a common sentiment classification framework that can be effectively used in our real applications. We view this task as a binary classification problem, and adopt the Maximum Entropy (ME) model as the basic supervised classifier. The brief flow chart is illustrated in figure 1.

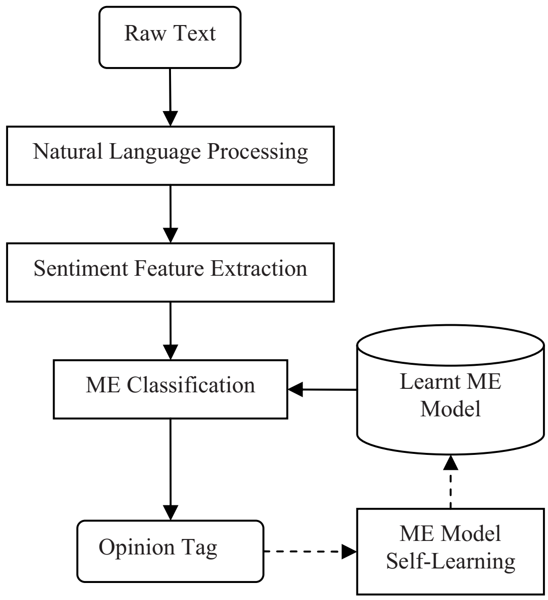

Figure 1. Opinion Judgment System Framework

2.1 Supervised Classification Model
Maximum Entropy model, as a classical supervised classification model used in traditional text categorization [2] and natural language processing [5], has been introduced into the sentiment classification area recently [4][1]. Maximum Entropy model is good at integrating kinds of heterogeneous information. Thus, we adopt it as our main learning model.

In Maximum Entropy model, a feature is typically Boolean, e.g.:

$$f(y,x)=\begin{cases}1 & \begin{array}{l} y: subjective, \\ x: word='like', pos='verb' \end{array} \\ 0 & \quad\quad else\end{cases}$$

In our system, we not only use such Boolean features, but also use real value features with positive weights such as normalized word frequency.

2.2 Sentiment Features

Sentiment classification differs from the traditional text classification in that, mostly, topics are explicitly presented by key words, while sentiments are implicitly expressed by latent linguistic features.

To extract meaningful sentiment features, we firstly apply a few basic natural language processing (NLP) procedures. We utilize the NLP toolkit provided by Stanford NLP Group[1]. For English side, we use the toolkit to perform Part-of-Speech (POS) tagging, while for Chinese side, we use it to perform Chinese word segmentation and POS tagging.

Similar to the work reported in [1], we design a set of n-gram pattern to extract features, including basic unigram tokens, bigram token pairs within fixed context window. Here, token refers either words or POS tags. That is, such n-gram patterns are applied on either word sequence, or POS tag sequence respectively. Further more, we introduce a new kind of cross-sequence bi-gram patterns, which are applied cross the word sequence and its corresponding word sequence. Figure 2. is an illustration of simple bi-gram patterns based sentiment feature extraction.

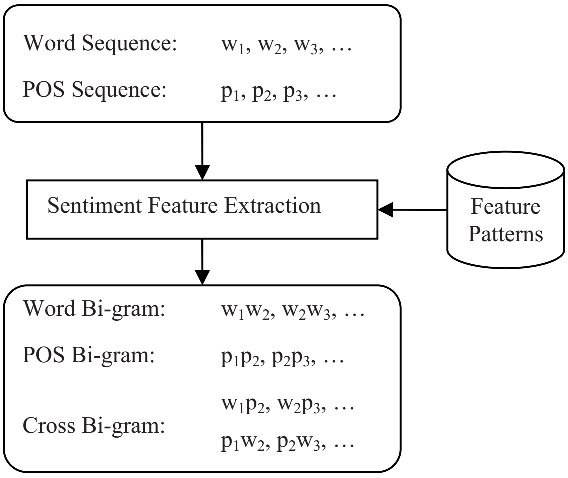

Figure 2. Simple bigram patterns based feature extraction

2.3 Self-Learning Strategies

In training the Maximum Entropy model, we use the previous NTCIR MOAT (i.e. NTCIR6 and NTCIR7) annotated corpus to train our basic classifier. Theses samples are far from adequate. On the other hand, the proportion of subjective sentences in these corpuses are low, this makes the trained classifier bias to the objective sentences. Thus, we adopt two self-learning strategies to adjust the trained classifier on the NTCIR8 test corpus. The basic idea is to use the basic classifier to judge each testing sample, and add these most confidential ones into the training set to retrain the ME model.

- The first strategy takes one-step iteration, and merely learns the subjective samples with high classification probabilities to make the training sets less unbalanced.
- The second strategy takes unlimited iterations, and simultaneously learns both subjective and objective samples. To smooth the unbalance, we take stricter confidence threshold on objective samples than on subjective ones. After each iteration, the ME model are retrained, and the new model are used to judge the test corpus again. The self-learning procedure terminates when few new training samples are added in one round.

[1] http://nlp.stanford.edu/software/index.shtml

3. EVALUATION RESULTS

We participated in the opinioned sentence judgment task in both English and Chinese sides. In these two sides, we used almost the same learning framework, and submitted 3 run results for each language side respectively. The official evaluation results are displayed in following tables.

Table 1. English opinioned evaluation results

RunID	P	R	F
bs0	25.85	51.44	34.41
bs1	21.79	69.8	33.21
bsf	26.5	52	35.11

Table 2. Simplified Chinese opinioned evaluation results

RunID	P	R	F
bs0	28.53	66.74	39.97
bs1	24.39	85.37	37.94
bsf	28.66	68.91	40.48

These runids correspond to 3 kinds of learnt models: 1) a basic model trained directly by using previous NTCIR corpus, marked as bs0; 2) an adjusted model learnt by using the one-step iteration self-learning strategy, marked as bs1; 3) a second adjusted model learnt by using the unlimited iteration self-learning strategy, marked as bsf.

Form the results, we can find that the absolute performances on both language sides are similar, which means that our methods are language independent. However, compared with other participants' performances, our relative ranks on the two sides are opposite. As for this writing, the other participants' reports are not available yet, thus we can not make any comparative analysis. What we can conclude is that our machine-learning based methods perform consistently cross languages.

Despite, some clues can be derived. 1) First, the one-step self-learning strategy tends to significantly improve the recall metrics but sacrifices the precision and overall F-measure. 2) Second, the unlimited iteration self-learning strategy can slightly and robustly improve all performances. 3) Self-learning can lead to a better classification model so long as appropriate strategies are adopted.

4. DISCUSSIONS

The corpus used in NTCIR MOAT tasks are mostly official news press, which are very ambiguous to be judged even by humans.

This can be seen from the low inter-annotator agreement Kappa values reported both in NTCIR6 [6] and NTCIR7 [7]. Therefore, it makes us have to give up deeply analyzing the samples and designing elaborate solutions. Still, to make the system perform well on new test corpus, we adopt two simple and intuitive self-learning strategies. Evaluation results show that the appropriate strategies do lead to overall improvement. The difference in performance between English and Chinese may caused by the languages themselves, e.g., preprocessing.

5. REFERENCES

[1] B. Chen, H. He, J. Guo. “Constructing maximum entropy language models for movie review subjectivity analysis”, Journal of Computer Science and Technology (JCST), 23(2), 2008, 231-239.

[2] Nigam, K., Lafferty, J. and McCallum, A. 1999. Using maximum entropy for text classification. In Proceedings of the International Joint Conference on Artificial Intelligence IJCAI-99 Workshop on Machine Learning for Information Filtering. 61-67.

[3] Ounis, I., Rijke, M,, Macdonald, C., Mishne, G. and Soboroff I. 2006. Overview of the TREC-2006 Blog Track. In Proceedings of the 15th Text REtrieval Conference. 17-31.

[4] Pang, B., Lee, L. and Vaithyanathan, S. 2002. Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of Conference on Empirical Methods in Natural Language Processing. 79-86.

[5] Ratnaparkhi, A. 1996. A Maximum Entropy Model for Part-Of-Speech Tagging. In Proceedings of Conference on Empirical Methods in Natural Language Processing. 133-142.

[6] Y. Seki, D. K. Evans, L. W. Ku, H. H. Chen, N. Kando, and C. Y. Lin. Overview of Opinion Analysis Pilot Task at NTCIR-6. In Proc. of the Sixth NTCIR Workshop Meeting on Evaluation of Information Access Technologies, pages 265_278, NII, Japan, May 2007.

[7] Y. Seki, D. K. Evans, L. W. Ku, L. Sun, H. H. Chen, and N. Kando. Overview of Multilingual Opinion Analysis Task at NTCIR-7. In Proc. of the Seventh NTCIR Workshop Meeting on Evaluation of Information Access Technologies, pages 185_203, NII, Japan, December 2008.

[8] Y. Seki, L. W. Ku, L. Sun, H. H. Chen, and N. Kando. Overview of Multilingual Opinion Analysis Task at NTCIR-8. In Proc. of the Seventh NTCIR Workshop Meeting on Evaluation of Information Access Technologies, 2010.