Chinese Microblogs Sentiment Classification using Maximum Entropy

Dashu Ye, Peijie Huang, Kaiduo Hong, Zhuoying Tang,
Weijian Xie, Guilong Zhou

College of Mathematics and Informatics South China Agricultural University
Guangzhou 510642, Guangdong, China *pjhuang@scau.edu.cn

Introduction

Polarity classification of Chinese microblogs is still an open problem today, since the difficulties like the out-of-vocabulary Internet words and emoticons. In our system, Maximum Entropy (MaxEnt) is employed, which is a discriminative model that directly models the class posteriors, allowing them to incorporate a rich set of features. Moreover, oversampling approach is used to handling the unbalance problem. Evaluation results demonstrate the utility of our system, showing an accuracy of 66.4% for restricted resource and 66.6% for unrestricted resource.

The Proposed System

Figure 1 shows the flowchart of our CMSC system.

The system is can be separated mainly into four parts: sentence container, language model, emoticon corpus and sentiment corpus.

1. A given sentence is required to put into the sentence container, and only the Chinese characters and some specific notation remain in the sentence after this phase.

2. Extracting structured feature using the feature functions.

$$f_1(x,y)=\begin{cases}1 & CPE(x)>CNE(x),\\ -1 & CPE(x)<CNE(x),\\ 0 & otherwise\end{cases}$$

where$CPE(x)$ calculate the number of positive emoticons, while $CNE(x)$calculate the number of negative emoticons.

$$f_2(x,y)=WS(x)$$

where WS (x) return a vector of words derived from the word segmentation of given sentence x.

$$f_3(x,y)=\begin{cases}1 & CPW(x)>CNW(x),\\ -1 & CPW(x)<CNW(x),\\ 0 & otherwise\end{cases}$$

where CPW(x) return the appearances of positive word of a given sentence and CNW(x) counts the negative one.

3. In this training phase, Maximum Entropy (MaxEnt) regarded as discriminative model yields a satisfactory performance.

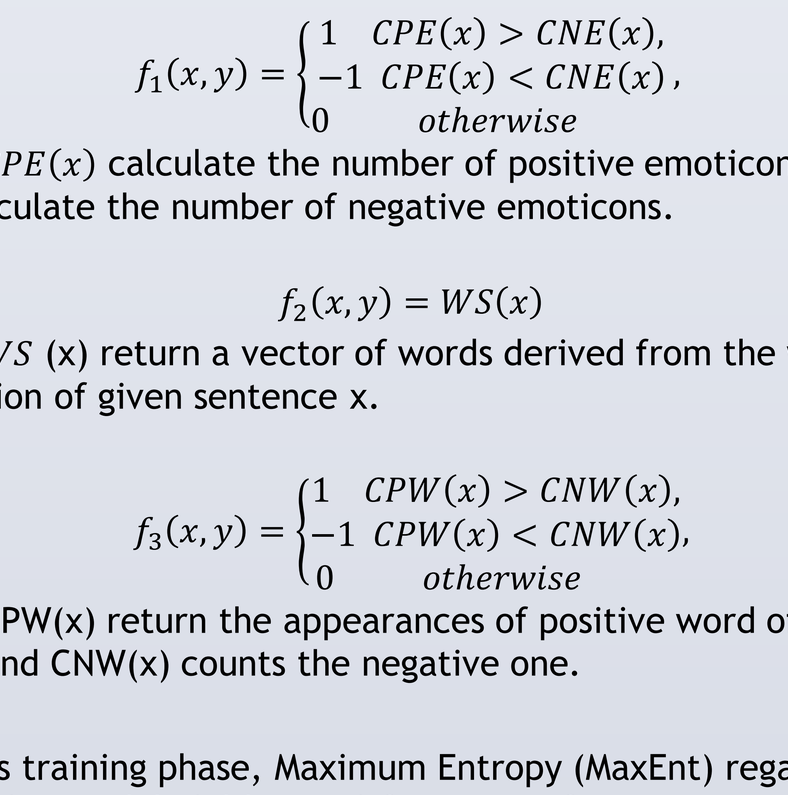

Figure 1. flowchart of the CMSC system

Result and Conclusion

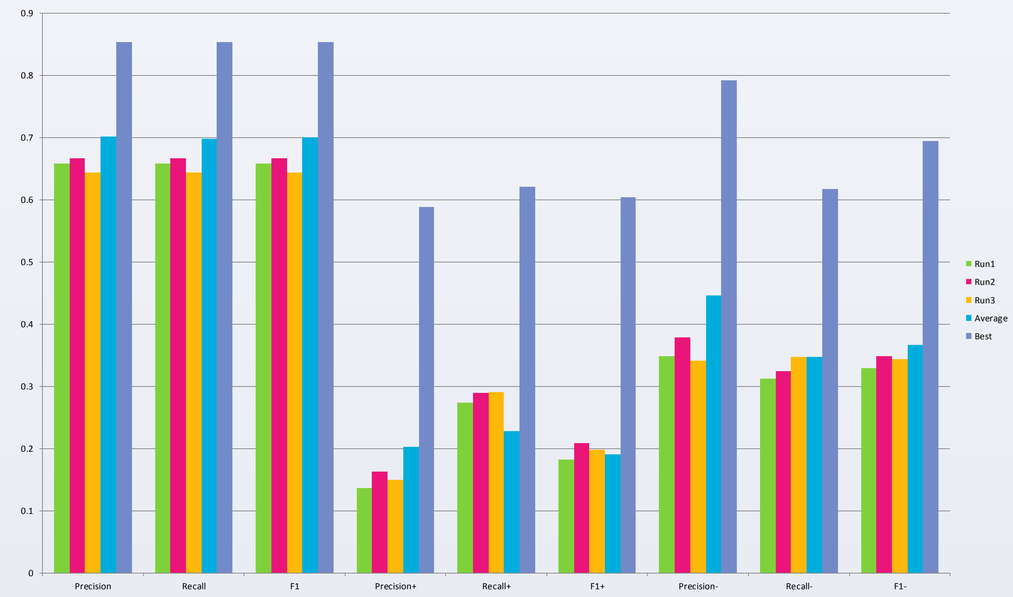

Figure 2. Evaluation score of unrestricted resources

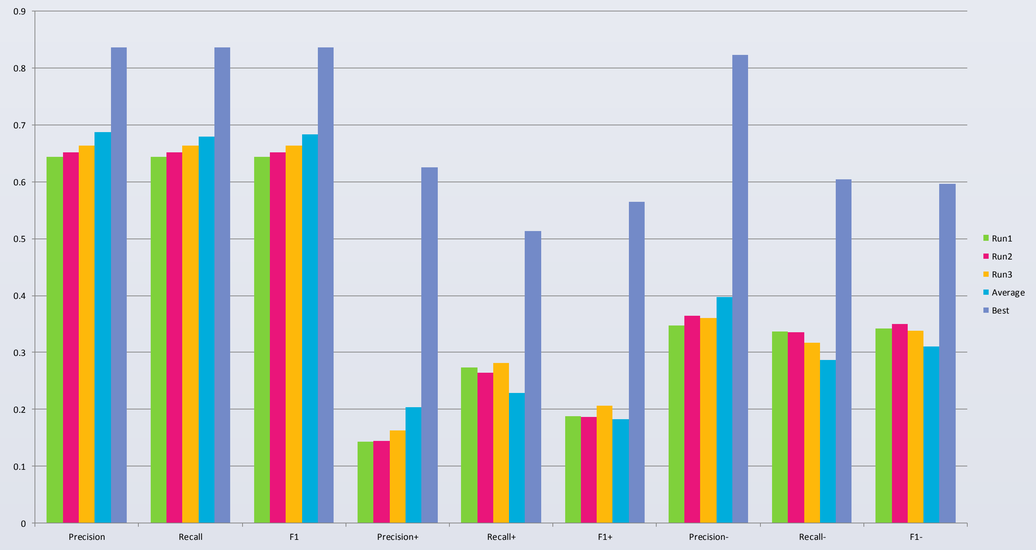

Figure 3. Evaluation score of restricted resources

The “Best” indicates the highest score of each metric achieved in the task. “Run” is the evaluation score of our system. And the “Average” represents the average score of all participants. As we can see from Figure 2 and Figure 3, we achieve a result close to the average level, even better than average one at some point.

Future Work

There are many possible and promising enhancements in the coming future. More appropriate features can be added to the system for a better modeling. Besides, existing sentiment corpuses and lexicons are filled with "book words" (literary, abstract and technical terms), while microblogs are usually in much less formal forms, with a significant amount of using of colloquial phrases, network language and even emoticons and pictures. Long distance relation and adverting detection are also a challenging research topic.

Selected References

- Berger A, Della Pietra S.D, Pietra V.D. 1996. A Maximum Entropy Approach to Natural Language Processing. Computational Linguistics, Vol. 22, No.1, pp. 5-9.
- Chawla, N.V., Japkowicz, N., and Kołcz, A., editors 2004. SIGKDD Special Issue on Learning from Imbalanced Datasets.
- Jiang L., Yu M., Zhou M., et al. 2011. Target dependent Twitter Sentiment Classification. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011), pp. 151-160.