

Biological Text Classification: BioCreAtIvE II Challenge sub-task 1

Man LAN¹

lanman@comp.nus.edu.sg

Chew Lim TAN¹

tancl@comp.nus.edu.sg

Jian SU²

sujian@i2r.a-star.edu.sg

¹ School of Computing, National University of Singapore, 3 Science Drive 2, Singapore 117543

² Institute for Infocomm Research, 21 Hen Mui Keng, Terrace, Singapore 119613

Abstract

The BioCreAtIvE II PPI IAS is a biomedical text classification task which concerns whether a given abstract contains protein interaction information. In order to improve the performance of text classification, we examined ways to represent text from the term type and term weighting aspects. In addition, we also combined different classifiers by simple majority voting technique.

Keywords: biological text classification, text representation, name entity, term weighting method

1 Introduction

For general text classification task, researchers usually adopted the vector space model to represent the text. Thus, there are two issues of text representation, i.e. (1) what should a term be and (2) how to weight a term. In this challenge, we investigated different text representations for biological text categorization from the above two aspects. That is, we use a protein name entity-based representation and a new effective term weighting method based on our two previous studies in [1] and [2]. So far no such work has been done on biomedical text categorization from the two representation aspects. Moreover, we also explored several machine learning algorithms to construct the text classifier. The experiments showed the term weighting method slightly improved the performance while term type based on name entity failed. We state that to significantly improve the performance of text categorization, more techniques in NLP for text representations need to be developed.

2 Methodology and Results

2.1 Text Preprocessing

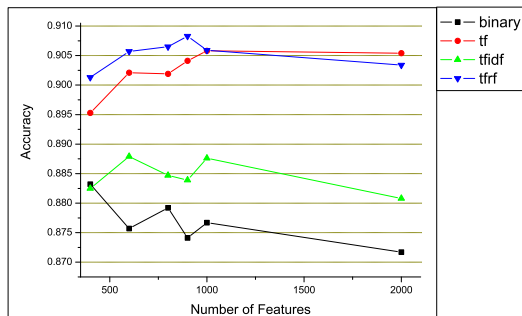
The BioCreAtIvE II PPI IAS training corpus consists of 3536 positive documents and 1959 negative documents on which this constructed system has based on. Before we constructed the classifier system, we need to preprocess these documents. This text preprocessing step is of great important for the performance of this system.

The Porter's stemming was performed to reduce words to their base forms. Stop words (513 stop words), punctuation and numbers were removed. The threshold of the minimal term length is 3 (since many biomedical keywords contain 3 letters, such as acronym). Null vectors (i.e. vectors with all attributes valued 0) were removed from the data set. The resulting vocabulary has 24648 words (terms or features). By using the χ^2 statistics ranking metric for feature selection, the top p features per category were selected from the training sets. In our experiments, we set $p = \{200, 300, 400, 450, 500, 1000\}$ respectively.

2.2 Preliminary Results on the Training Corpus

2.2.1 Performance of Different Term Weighting Methods for Text Categorization

Based on our previous work [2], we choose four methods, i.e. *binary*, *tf*, *tf.idf* and *tf.rf*. Figure 1 and Table 1 show the results of different term weighting methods using SVM based on 2-folder cross validation with respect to different evaluation measures.



Scheme	Micro-P	Micro-R	Micro-F1
<i>binary</i>	92.55 ± 0.94	91.99 ± 4.05	92.22 ± 1.77
<i>tf</i>	92.34 ± 1.23	94.26 ± 3.26	93.17 ± 1.43
<i>tf.idf</i>	92.19 ± 1.01	94.48 ± 3.69	93.28 ± 1.64
<i>tf.rf</i>	92.23 ± 1.24	95.11 ± 2.79	93.63 ± 1.23

Table 1: Best results of four term weighting schemes.

Figure 1: Results of four term weighting methods.

It is clear to find that *tf.rf* and *tf* consistently performed better than *tf.idf* and *binary*. *tf* has shown good performance even though sometimes it had a bit lower accuracy than *tf.rf*. On the other hand, the widely-used *tf.idf* method only performs better than *binary* method. These findings are consistent with our previous work [2]. Meanwhile, the best performance has been obtained (using *tf.rf* and *tf*) with between 800 and 1000 features. Therefore, we chose *tf.rf* and 900 features used for indexing in the following test experiment.

2.2.2 Performance of Name Entity-based Representation for Text Categorization

Based on the consideration that protein name entity-based representation may capture more information than bag-of-words approach, we conducted experiment using alternative term type on the BioCreAtIvE II corpus.

The noticing phenomena of these name entities are sparse and skewed distribution. First, most of the name entities are in the positive category (76.7%) and only 23.3% are in the negative category. This is reasonable since the positive documents are relevant to protein protein interaction articles and thus they must contain more protein name entities than those in the negative category. Second, most of the name entities occur only once or few times in the corpus. For example, 25740 name entities (83.7%) occur only once in the corpus. 2529 named entities (8.2%) occur more than three times and only 380 named entities (1.2%) occur more than ten times in the whole corpus. This sparse distribution problem make the indexing of documents difficult since many documents will be represented as null vectors when the number of name entities used for indexing is quite small. Based on this consideration, we also combined name entity-based representation with bag-of-words approach based on different term weighting methods. Table 2 shows the results of these combined different representations, where NE denotes name entity and BOW means bag-of-words approach. Based on the results from Table 1 and Table 2, we can find that name entity-based representation was the most disappointing. It only achieved 78.56% F_1 score. When combined with bag-of-words approach based on different term weighting methods, the name entity-based representation has not increased the performance of text categorization.

Scheme	Micro-P	Micro-R	Micro-F1	Classifier	Accuracy
NE	68.03 \pm 0.81	92.98 \pm 2.76	78.56 \pm 1.28	LibSVM	0.9083 \pm 0.0011
NE+BOW(<i>binary</i>)	91.51 \pm 0.94	92.98 \pm 4.05	92.20 \pm 1.77	kNN	0.7821 \pm 0.0013
NE+BOW(<i>tf</i>)	91.90 \pm 1.19	94.74 \pm 2.76	93.27 \pm 1.35	AdaBoost	0.8667 \pm 0.0094
NE+BOW(<i>tf.rf</i>)	91.97 \pm 1.19	95.16 \pm 2.76	93.52 \pm 1.35	Voted Perception	0.8917 \pm 0.0080
				Majority Voting	0.9099 \pm 0.0023

Table 2: Results of different combined represents on the BioCreAtIvE II corpus.

Table 3: The results of classifier committee.

2.3 Different Classifiers

Generally, SVM has been confirmed to perform best among many promising machine learning algorithms. In addition, since different high-quality classifiers make at least partially uncorrelated errors, and when combined with a simple majority voting, they are expected to lead to higher performance. We also tried majority voting technique in this work. Table 3 lists the performance (accuracy) using different algorithms based on 900 features and *tf.rf* scheme based on two-folder cross validation.

2.4 System Configuration

According to the above experimental results, we conducted test experiments based on the following three sets of system configuration:

Table 4: The system configuration for test corpus.

Run	#_features	weighting method	Classifier(s)	Accuracy
1	900	<i>tf.rf</i>	LibSVM	0.9083 \pm 0.0011
2	900	<i>tf.rf</i>	Classifier Committee	0.9099 \pm 0.0035
3	800	<i>tf.rf</i>	Classifier Committee	0.9060 \pm 0.0011

3 Concluding Remarks

Name entity-based representation has no improvement over bag-of-words approach. This supports the general conclusion that significant advances must be made before NLP techniques can be used to improve text classification. On the other hand, our proposed *tf.rf* method shows classification power in biomedical text classification as well as in previous newswire classification. We should point out that the observations above are made based on the controlled experiments and the accuracy of extracted name entities also has an effect on the result. We believe that to significantly improve the performance of text categorization, more techniques in NLP for text representations need to be developed.

References

- [1] Guodong Zhou, Jie Zhang, Jian Su, Dan Shen and Chewlim Tan. Recognizing names in biomedical texts: a machine learning approach. In *Bioinformatics* 20(7):1178–1190, 2004. Oxford University Press, UK.
- [2] Man Lan, ChewLim Tan and HweeBoon Low. Proposing a New Term Weighting Scheme for Text Categorization. 2006. In *the Proceedings of AAAI2006*, page 763-768.