# A Term Investigation and Majority Voting for Protein Interaction Article Sub-task 1 (IAS)\*

 $Man LAN^1$ 

Chew Lim  $TAN^1$ 

Jian  $SU^2$ 

lanman@comp.nus.edu.sg

tancl@comp.nus.edu.sg

sujian@i2r.a-star.edu.sg

- School of Computing, National University of Singapore, 3 Science Drive 2, Singapore 117543
- <sup>2</sup> Institute for Infocomm Research, 21 Hen Mui Keng Terrace, Singapore 119613

#### Abstract

The BioCreAtIvE II PPI Interaction Article Sub-task 1 (IAS) is a biological 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 majority voting technique.

**Keywords:** biological text classification, text representation, named entity, term weighting

#### 1 Introduction

For general text classification task, vector space model is usually adopted to represent the text. Thus, there are two issues of text representation involved, i.e. (1) what should a term be and (2) how to weight a term. In this work, we investigated different text representations for biological text classification from the above two aspects. That is, we adopted a protein name-based representation and a new effective term weighting method based on our two previous studies in [1] and [2]. Based on our knowledge, so far no such work has been done on biological text classification from the two representation aspects. Moreover, we also explored several machine learning algorithms to build the classifier.

# 2 Methodology and Results

## 2.1 Text Preprocessing

The BioCreAtIvE II PPI IAS training corpus consists of 3536 positive and 1959 negative documents on which this constructed system is based on. The Porter's stemming was performed to reduce words to their base forms. Stop words (513 stop words), punctuations and numbers were removed. The threshold of the minimal term length is 3 (many biological keywords contain 3 letters, such as acronym). The resulting vocabulary has 24648 words (terms or features). By using the  $\chi^2$  statistics ranking metric for feature selection, the top  $p = \{200, 300, 400, 450, 500, 1000, 1500\}$  features per positive and negative category were selected from the training set.

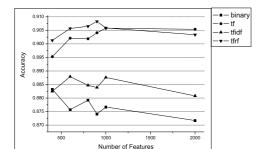
## 2.2 Preliminary Results on the Training Corpus

#### 2.2.1 Performance of Different Term Weighting Methods for Text Classification

Different features have different importance in a text and thus an important indicator represents how much the feature contributes to the semantics of document. Our proposed weighting method, i.e. tf.rf [2], is based on the idea that the more concentrated a high frequency term is in the positive category than in the negative category, the more contributions it makes in selecting the positive documents from among the negative documents. We name it rf (relevance frequency) because only the frequency of relevant documents (i.e. those which contain this term) are considered and it is calculated as the ratio of relevant documents in the positive and negative category (in stead of using whole document distribution in the corpus). It has shown consistently better performance than other traditional methods based on crossmethod, cross-classifier and cross-corpus validation. In this work, we chose four methods, i.e. binary,

<sup>\*</sup>The work is partially supported by a Specific Targeted Research Project(STREP) of the European Union's 6th Framework Programme within IST call 4, Bootstrapping Of Ontologies and Terminologies STrategic Project(BOOTStrep).

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 performed better than binary method. These findings are consistent with those in our previous work [2]. Meanwhile, since the best performance has been achieved using tf.rf with 900 features (bag-of-words), we chose these settings in the following test experiment.



| Scheme | Micro-P          | Micro-R          | $\text{Micro-}F_1$ |
|--------|------------------|------------------|--------------------|
| binary | $92.55 \pm 0.94$ | $91.99 \pm 4.05$ | $92.22 \pm 1.77$   |
| tf     | $92.34 \pm 1.23$ | $94.26 \pm 3.26$ | $93.17 \pm 1.43$   |
| tf.idf | $92.19 \pm 1.01$ | $94.48 \pm 3.69$ | $93.28 \pm 1.64$   |
| tf.rf  | $92.23 \pm 1.24$ | $95.11 \pm 2.79$ | $93.63 \pm 1.23$   |

Table 1: Best results of four term weighting schemes.

Figure 1: Results of four term weighting methods.

#### 2.2.2 Performance of Named Entity-based Representation for Text Classification

The protein entity names in biology domain are more complex than those in other domains like newswire due to the long descriptive naming convention, non-standardized naming convention and ambiguous abbreviation, etc. Based on the consideration that protein named entity may capture more information left out of the bag-of-words approach, we conducted experiment using this representation on this corpus. We adopted an existing named entity recognition system named PowerBioNE [1], where various evidential features are integrated through a Mixture Markov Model(HMM)-based named entity recognizer.

The noticing phenomena of these extracted named entities are sparse and skewed distribution. First, most named entities are in the positive documents (76.7%) and only few (23.3%) are in the negative documents. This is reasonable since the positive documents are relevant to protein interaction articles and thus they must contain more protein names than negative documents. Second, most of the named entities occur only once or few times in the corpus. For example, 25740 named entities (83.7%) occur only once, 2529 entities (8.2%) occur more than three times and only 380 entities (1.2%) occur more than ten times in the whole corpus. This sparse distribution problem make the document indexing difficult since many documents will be represented as null vectors when the number of named entities used for indexing is quite small. Therefore, we also combined named entity-based representation with the bag-of-words approach. Table 2 shows the results of these combined different representations, where NE denotes

| Scheme         | Micro-P          | Micro-R          | Micro-F1         |
|----------------|------------------|------------------|------------------|
| NE(tf)         | $68.03 \pm 0.81$ | $92.98 \pm 2.76$ | $78.56 \pm 1.28$ |
| NE+BOW(binary) | $91.51 \pm 0.94$ | $92.98 \pm 4.05$ | $92.20 \pm 1.77$ |
| NE+BOW(tf)     | $91.90 \pm 1.19$ | $94.74 \pm 2.76$ | $93.27 \pm 1.35$ |
| NE+BOW(tf.rf)  | $91.97 \pm 1.19$ | $95.16 \pm 2.76$ | $93.52 \pm 1.35$ |

| Classifier       | Accuracy            |
|------------------|---------------------|
| ${ m LibSVM}$    | $0.9083 \pm 0.0011$ |
| kNN              | $0.7821 \pm 0.0013$ |
| AdaBoost         | $0.8667 \pm 0.0094$ |
| Voted Perception | $0.8917 \pm 0.0080$ |
| Majority Voting  | $0.9099 \pm 0.0023$ |

Table 2: Results of different combined represents on the BioCre-AtIvE II corpus.

Table 3: The results of classifier committee.

named entity and BOW means bag-of-words approach. Based on the results from Table 1 and Table 2,

we can find that named entity-based representation was the most disappointing. It only achieved 78.56%  $F_1$  score. When combined with the bag-of-words approach based on different term weighting methods, the named entity-based representation has not increased the performance of text classification.

#### 2.2.3 Performance of Different Classifiers

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

#### 2.3 Results on the Test Data and Error Analysis

According to the above experimental results and system settings on the training data, Table 4 lists the three sets of system configuration and the corresponding different evaluation scores on the 750 test documents (350 positive and 350 negative documents), where AUC means the area under the ROC curve.

| Run | #_features | weighting | Classifier(s)        | Accuracy | $F_1$  | AUC    |
|-----|------------|-----------|----------------------|----------|--------|--------|
| 1   | 900(BOW)   | tf.rf     | LibSVM               | 0.7467   | 0.7775 | 0.8141 |
| 2   | 900(BOW)   | tf.rf     | Classifier Committee | 0.7453   | 0.7761 | 0.8105 |
| 3   | 800(BOW)   | tf.rf     | Classifier Committee | 0.7373   | 0.7685 | 0.8019 |

Table 4: The system configuration and results on the test data.

To further evaluate our system and explore possible improvement, we have implemented an error analysis. The average error rate of our system is 0.2569. First, the reason for protein name-based representation failing to improve the performance may be caused by the precision of named entities extracted. Although PowerBioNE achieved 77.8%  $F_1$  score on the "protein" class of GENIA V3.0 which is higher than other systems [1], the accuracy of extracted named entities is still not high. The system performance can be improved further by using more annotated corpora and incorporating more effective features based on the domain knowledge. On the other hand, the accuracy performance on the training corpus using tf.rf is above 90% while it only achieved 74% on the test data. The possible reason may be caused by the different category distribution in the training and test data, i.e. the ratio of positive and negative documents in the training corpus is almost 2:1 while it is 1:1 in the test corpus. The term weighting is calculated based on the distribution of the training corpus and used for the test corpus. This may cause the significantly different performance in the training and test corpora.

## 3 Concluding Remarks

Our proposed tf.rf method showed classification power in biological text classification while named entity-based representation has not yet succeeded in improving text classification performance over the bag-of-words approach. We should point out that the observations above are made based on the controlled experiments and the accuracy of extracted named entities also has an effect on the result. We believe more advanced NLP techniques and advanced ways of incorporating NLP output could further improve the performance of text classification, for example, high performance coreference resolution to normalize the protein names through different variations, nominal or pronominal expressions could generate more occurrences of the same protein names to facilitate the further text classification.

## References

- [1] GuoDong Zhou and Jian Su. Exploring deep knowledge resources in biomedical name recognization. *Proceedings of JNLPBA shared task*, 99–102, 2004.
- [2] Man Lan, Chew Lim Tan and HweeBoon Low. Proposing a New Term Weighting Scheme for Text Categorization. *Proceedings of the*  $21^{st}$  *AAAI*, 763–768, 2006.