Journal of Biomedical Informatics 45 (2012) 156–164

Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Boosting performance of gene mention tagging system by hybrid methods

Lishuang Li [a,*], Wenting Fan [a], Degen Huang [a], Yanzhong Dang [b], Jing Sun [a]

[a] School of Computer Science and Technology, Dalian University of Technology, 116023 Dalian, China
[b] School of Management Science and Engineering, Dalian University of Technology, 116023 Dalian, China

ARTICLE INFO

Article history:
Received 28 December 2010
Accepted 20 October 2011
Available online 28 October 2011

Keywords:
Hybrid methods
Gene mention tagging
Named entity recognition
Bioinformatics
Biomedical literature

ABSTRACT

NER (Named Entity Recognition) in biomedical literature is presently one of the internationally concerned NLP (Natural Language Processing) research questions. In order to get higher performance, a hybrid experimental framework is presented for the gene mention tagging task. Six classifiers are firstly constructed by four toolkits (CRF++, YamCha, Maximum Entropy (ME) and MALLET) with different training methods and features sets, and then combined with three different hybrid methods respectively: simple set operation method, voting method and two layer stacking method. Experiments carried out on the corpus of BioCreative II GM task show that the three hybrid methods get the *F*-measure of 87.40%, 87.31% and 87.70% separately without any post-processing, which are all higher than those of any single ones. Our best hybrid method (two layer stacking method) achieves an *F*-measure of 88.42% after post-processing, which outperforms most of the state-of-the-art systems. We also discuss the influence on the performance of the ensemble system by the number, performance and divergence of single classifiers in each hybrid method, and give the corresponding analysis why our hybrid models can improve the performance.

1. Introduction

The target of biomedical named entity recognition (Bio-NER) is to automatically analyze lots of biomedical texts and it is a preliminary step for other steps such as protein–protein relation extraction and gene normalizing in biomedical text mining. Over the past years NER has made some progress in the biomedical field. Many algorithms have been proposed for NER task, however, with the flourishing development of biomedicine, new NEs (Named Entities) are emerging one after another. Irregular naming as well as new uses of old words has made Bio-NER a hard task, to some degree, influencing the development of research in biomedical domain. Therefore Bio-NER remains a challenging task and there is still a large gap between the best Bio-NER systems and the best algorithms in newswire domain.

It is more difficult for biomedical NER in the following facts [1]:

(1) New named entities continue to be created.
(2) The same word or phrase can refer to different entities depending upon their contexts. Conversely, many entities have various spelling forms.
(3) Some modifiers are often used before basic named entities, which highlight the difficulties for identifying the boundaries of named entities.
(4) Named entities may be cascaded.
(5) Abbreviations are frequently used in biomedical domain.

To tackle these problems, it is necessary to explore effective methods and rich features. A great number of research methods for Bio-NER have been presented and these methods can be mainly classified into three categories which are dictionary-based methods, rule-based methods and statistical machine learning methods. The dictionary-based methods are mainly used to recognize the terms in text through exact or partial matching [2–4], but because of the irregularities and ambiguities in bio-entities nomenclature, this traditional dictionary-based method cannot work effectively. In rule-based methods [5], rules are generated manually or heuristically. Although they can get higher precision than dictionary-based approaches, the recall is lower and what is more, they are difficult to recognize complex named entities and are not portable to other domains.

Compared with other methods, machine learning methods are more robust and there is an advantage that they can identify potential biomedical entities which are not previously included in standard dictionaries. Leveraging lots of training data, which is luckily available for the specific task of protein/gene named entity recognition, the statistical machine learning methods have become a better choice in bio-NER domain. So far there have been many attempts to develop machine learning techniques to identify biomedical entities. These techniques include Support Vector Machine (SVM) [6], Hidden Markov Model (HMM) [7], Maximum Entropy (ME) [8] and Conditional Random Fields (CRF) [9–11], etc.

* Corresponding author. Fax: +86 041184708140.
E-mail addresses: lilishuang314@163.com (L. Li), fanwent333@163.com (W. Fan), huangdg@dlut.edu.cn (D. Huang), yzhdang@dlut.edu.cn (Y. Dang), katrinasunjing@126.com (J. Sun).

1532-0464/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2011.10.004

A lot of recent research on machine learning-based Bio-NER focused on incorporating effective features for different classifiers including local features and external resources features into the powerful machine learning frameworks [8,12]. However, the performance of Bio-NER systems is still not as good as that of common NER systems. To tackle these problems, currently biological named entity recognition systems not only use single technology, but also combine a variety of treatment methods, that is, hybrid approaches. The approaches which combine the results of single models have been presented and become a trend in machine learning-based Bio-NER. The ensemble methods can overcome the possible local weakness of single classifiers and produce more robust performance.

In previous work, we have combined the results of multiple biomedical entity recognition systems that used rich and diverse feature representations based on union and intersection operations, whose results are better than those of single classifiers [13]. In this paper, we analyze several best performing systems in BioCreative II, and to achieve a better performance, a total of six divergent models are implemented and combined with union and intersection operations, voting and stacking in our hybrid methods. Experiments show that our best performing combination model can achieve an *F*-score of 88.42% on the test corpus of BioCreAtIvE II, which is fairly good performance.

The remaining part of this paper is organized as follows: Section 2 introduces current Bio-NER systems which used hybrid methods. Our gene mention tagging methods with hybrid models are described in Section 3. Section 4 shows the experiments and results. Section 5 gives comparison and analysis. The discussion and error analysis are presented in Section 6. Finally, conclusions are given in Section 7.

2. Related work

The second BioCreative challenge (BioCreative II) [14] is a recent competition for biological literature mining systems. It took place in 2006 and followed by a workshop in April 2007. There were three tasks in the challenge, namely, gene mention tagging (GM), gene normalization (GN) and protein–protein interaction (PPI) tasks. BioCreative II GM task [15] was built on the similar task from BioCreative I [16], in which participants were given a labeled training corpus to develop their systems and an unlabeled testing corpus to apply their systems for evaluation. The training corpus of BioCreative II GM task contains 15,000 sentences, including the training and testing corpora from the previous task, and the testing corpus consists of an additional 5000 sentences which were held 'in reserve' from the previous task. In BioCreative II, each participant was allowed to submit up to three runs, and each run was evaluated based on performance measures precision, recall and *F*-score:

$$P = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \quad R = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \quad F\text{-score} = \frac{2*P*R}{P+R},$$

where *P* is precision, *R* is recall, TP is true positives, FP is false positives and FN is false negatives.

For BioCreative II GM task, we pay attention to the following four systems:

One of them is Kuo et al.'s system [17], which is the best performing system based on CRF (ranked 2nd) in BioCreative II at the time of 2006. The contributions of their system include the application of a rich feature set and the combination of bidirectional parsing models based on likelihood scores and dictionary -filtering. Their combination algorithm runs as follows:

(1) Parse the input sentences in both directions to obtain the top ten solutions for each direction with their output scores.
(2) Compute the intersection of bidirectional parsing models and return the solution in the intersection that minimizes the sum of its output scores. If the intersection is empty, return the top solution of the backward model and the top solution of the forward model.
(3) From the other unselected solutions, return the labeled terms appearing in a dictionary with its length greater than three.

They applied MALLET[1] to build their CRF models and the top ten solutions were obtained by MALLET's *n*-best option. Besides, the second step of their algorithm was derived from the optimal model integration [18], and the dictionary used for dictionary-filtering consisted of the aliases and approved gene symbols obtained from HUGO [19].

Another is Huang et al.'s system [20], in which SVM was combined with CRF model and achieved one of the best F-scores (ranked 3rd) in BioCreative II 2006. Huang et al. considered the GM task as a classification problem and applied SVM to solve it. To further improve the performance, they tried to construct divergent but high performance models and combine them into an ensemble. Firstly, two backward parsing SVM models were trained by YamCha[2] with different multi-class extension methods, i.e., one vs. all and one vs. one. And then, a backward parsing CRF model was trained by MALLET with the same features employed in SVM models, to increase the divergence of the ensemble. Finally, the outputs of the three models were combined with simple set operations, union and intersection. Their experiments proved that integrating divergent but high performance models can improve the performance.

The third is Hsu et al.'s system [21] which is better than both of the presented systems above. In their system, a high dimensional feature set that includes most of information was designed and Hsu et al. also trained bi-directional CRF models, one applied forward parsing and the other backward, then integrated them based on the output results and dictionary filtering. Hsu et al. found that due to different feature settings, CRF is asymmetric and the feature setting will not only produce different results but also give backward parsing models slight but constant advantage over forward parsing models. To fully explore the potential of integrating bidirectional parsing models, they applied many bi-directional paring models and integrated them based on the output scores.

The last is Li et al.'s system [13], which combined six divergent models (FMALLET, BMALLET, CRF++BIO, CRF++BIOEW, SVM one vs. one and SVM one vs. all) constructed by different machine learning algorithms and dissimilar feature sets and the result was improved greatly. They also applied some post-processing on the results at last.

The most prominent feature of the above four systems is the use of backward parsing. In their experiments, backward parsing models constantly outperformed forward parsing models. However, Chang et al. [22] found that backward parsing was not always superior to forward parsing. They trained two groups of bidirectional parsing models and found that MALLET tagger performed better applying backward parsing than forward parsing, but on the contrary, CRF++[3] tagger performed worse applying backward parsing. They supposed that the benefit of applying bidirectional parsing was the creation of a wider variety of complementary models.

Another prominent feature of their systems is the combination of divergent but high performance models. Kuo et al. [17] succeeded in improving the performance of bi-directional parsing MALLET models by combining them based on likelihood scores

[1] MALLET, http://mallet.cs.umass.edu/.
[2] YamCha, http://chasen.org/taku/software/yamcha/.
[3] CRF++, http://crfpp.sourceforge.net/.