Mining Biomedical Entity from Literature Based on CRF

Lejun Gong[1*] ,Ronggen Yang[2] ,Jiacheng Feng[1] and Geng Yang[1]

[1]Jiangsu High Technology Research Key Lab for Wireless Sensor Networks

College of Computer Science & Technology, Nanjing University of Posts and Telecommunications, Nanjing, 210003,China

[2]Faculty of Information Technology, Jinling Institute of Technology, Nanjing 211169, China

*Corresponding author: glj98226@163.com

Keywords: biomedical entity recognition, CRF, Feature selection, text mining

Abstract. With the rapid expansion of biomedical literatures, it provides an opportunity for mining biomedical knowledge from the huge amount of biomedical text. Entity recognition is a challenging task of biomedical text mining. In this work, we described a method to identify biomedical entity based on Conditional Random Fields(CRF). In the test dataset, the performance of the submitted method obtained the relatively satisfied performance . At the same time, we also develop a system with identified six class entities using different color representation Taken together, our method is promising for developing the technology of biomedical entity recognition.

1. Introduction

With the rapid expansion of biomedical literatures, for example, over 25 million published papers collected on Pubmed on Septemper, 2015, the huge amount of electronic biomedical text offers a fine opportunity for biomedical text mining, while Entity Recognition(ER) is a fundamental task of biomedical text mining, referring to the mentions of biomedical entitie. The ER task remains challenging due to the irregularities and ambiguities in biomedical entities nomenclature, such as synonyms, the nomenclature of short or long words, domain-specific jargon. Thus, great efforts dedicated to the task are necessary in that it is a fundamental technique in biomedical knowledge extracting from literature. At the same time, the performance of its recognition has direct effects on further biomedical knowledge discovery or biomedical application database.

The named entity recognition is concerned by the several international competitions on ER shared tasks at the JNLPBA [1] ,BioNLP[2], BioCreative[3]. Some entity recognition tools usually focus on genes and proteins for example, ABGene [4] is a tool for tagging gene and protein named entities trained on Medline abstracts AIMed at a randomly selected set of full text in the biomedical domain. ABNER[5] and BNNER[6]. The bio-entity recognition attracts the attention of most research groups. Our research group also focus on the entity recognition field. The bio-entity recognition is similar to a sequence segmentation task, while the Conditional Random Fields (CRF) is a family of conditionally trained undirected graphical models for sequence labeling. We applied the CRF model to identify the bio-entities due to the characteristic based selected features based on Java program in Linux OS. In this paper, we also developed a system to identify biomedical entity for biomedical entities researchers with identified six class entities using different color representation .

2. Methods and Materials

We use the Conditional Random Field-based (CRF) model to identify biomedical entity with the extracted features. The core of approach is conditional random fields and the feature sets. The following would describe the two details.

2.1Conditional random fields

Entity recognition is treated as a sequence segmentation task. each word makes up a unit of a sequence to be labeled corresponding category. Conditional Random Fields is a linear chain involving with a conditionally trained finite-state machine based on undirected statistical graphical models. Thus the model is used to act as sequence analysis.

Let Y denote the label sequence and X denote the corresponding observation sequence. The line chain conditional field define the conditional probability of a state sequence given an observation sequence to be :

$$P(Y \mid X) = \frac{1}{Z_0} \exp(\sum_{i=1}^{n} \sum_{j=1}^{m} \lambda_j f_j(Y_{i-1}, Y_i, X, i)) \quad (1)$$

Where $f_j(Y_{i-1}, Y_i, X, i)$ is one of m functions that describes a feature, and λ_j is a learned weight for feature function. The aim of training process is for the weights that maximize the log likelihood of instances in the training data. Z_0 is a normalization factor of all sequences defined as the followings:

$$Z_0 = \sum_Y \exp(\sum_{i=1}^{m} \sum_j \lambda_j f_j(Y_{i-1}, Y_i, X, i)) \quad (2)$$

The complete detail about CRF model is present in the [7].In the work, we used a CRF toolkit, called CRF++ [8] which be applied to biomedical entity recognition task.

2.2 Feature sets

(1) Part-Of-Speech(POS) features: Every biomedical entity is composed by word sequences, and every word is assigned to the POS. Each biomedical entity is corresponding with the word-POS pairs. The POS features could be important to identify biomedical entities.

(2) Surface clue features: Words themselves' clues are helpful to find biomedical entity names, for example, we use regular expressions to extract surface clue features including: uppercase, lowercase, digitalization, underline, hyphen, the combination both digital and character, alles beginhoofdletter.

(3) Boundary detection: Biomedical entity include both word and multi-words. We used the IOB tagging to detect both phrases and entities' boundary. When the word is the outset of phrase, it would be tagged as B-NP. When the word is among the phrase, it would be tagged as I-NP. When the word is not part of the phrase, it would be tagged as O. Similarly, the entity would be tagged as B-entity_ category, I-entity_category, O. For example, when the word is part of protein, it would be tagged as B-protein, I-protein, O.

3. Results and Discussions

To evaluate our approach, we used three popular measurement to measure the performance: precision,recall,and F-measure using AIMed[9] corpus to measure the performance of protein recognition by 10-cross-validation as shown in Table 1.

Table 1 Identified performance aiming at the AIMed corpus using 10-cross-validation(%)

Project.	All	POS	Num	Lett_Num	Unde	Hyphen	Init	Allcap	Lowe
precision	82.00	83.61	87.60	82.42	87.19	87.11	83.01	84.86	84.83
recall	77.33	63.26	47.32	64.63	45.95	46.10	59.33	50.30	64.24
F-measure	79.50	71.93	61.27	72.28	60.01	60.10	69.15	62.99	73.06

All: the combination of all features; POS: part-of-speech; Num: number; Lett_Num: letter and number;Unde:underline; Hyphen:hyphen; Init: initial letter, Allcap:all capital,Lowe:lowercase letter

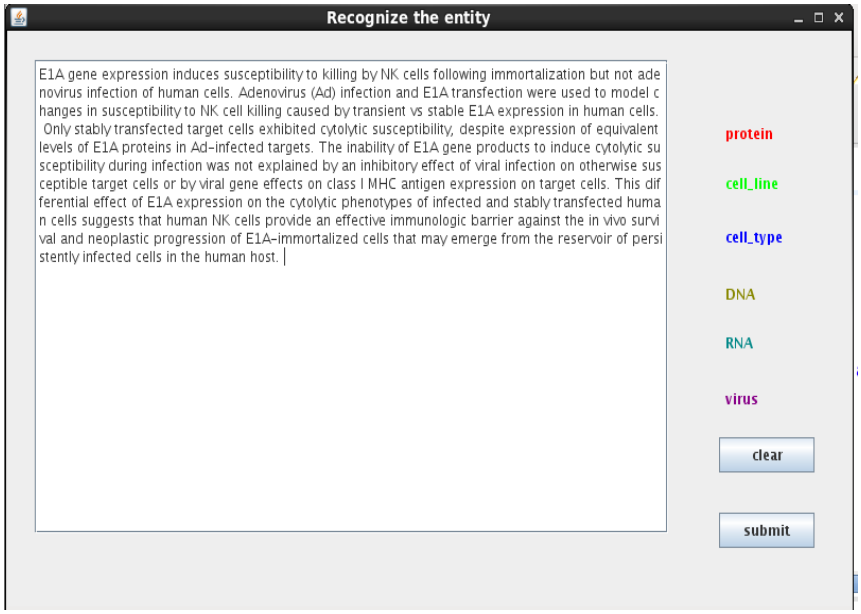

Fig.1 Interface for entity recognition

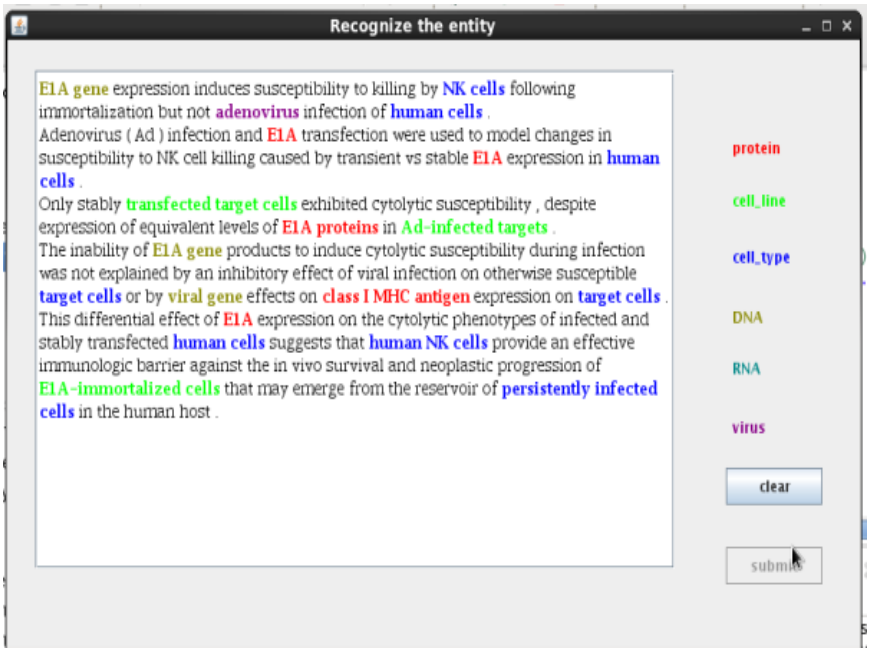

Fig. 2 Identified six biomedical entities

The precision obtain over 82 percent in all features, and the highest performance is in the number feature. The highest performance of recall is 77.33 percent in the combination of all features. The highest F-measure is 79.50 percent in the combination of all features. Thus, combined features is more helpful to identify biomedical entities. To validate our above approach, we also develop a system as shown in Fig. 1 and Fig.2 which could identify six kind of biomedical entities.

In Fig. 2, the system identified six class biomedical entities using different color. Every color represents a kind of biomedical entity. Red indicates protein. Green is cell_line. Blue means cell_type. Brown denotes DNA. Purple is virus, and indigo signifies RNA.

4. Conclusions

In this work, we described an approach of biomedical entity based on CRF. Aiming at the test data, our approach obtained the performance with 82 percent of precision, 77.33 percent of recall, and 79.50 F-measure. Moreover, our developed system could identify six class biomedical entities

using different color representation. Usually some existed entity recognition tools could extract five class entity including: protein cell_line, cell_type, DNA, RNA. Our system could also identify virus entity. Thus our identified approach is promising for developing the technology of biomedical entity recognition.

Acknowledgments

This research is supported by the National Natural Science Foundation of China(Grant No. 61272084, 61300240,61572263, 61502251 and 61502243), Natural Science Foundation of the Jiangsu Province (Grant No. BK20130417, BK20140875), Jiangsu province postdoctoral Science Foundation(Grant No. 1501072B), and Nanjing University of Posts and Telecommunications' Science Foundation (Grant No. NY214068 and NY213088).

References

[1]J.D. Kim, T. Otna, Y. Tsuruoka Y, et al.Introduction to the bio-entity recognition task at JNLPBA,Proceeding JNLPBA '04 Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications.Pages 70-75.

[2]J.D. Kim, T. Otna , S. Pyysalo, et al. Overview of BioNLP'09 shared task on event extraction. Proceeding BioNLP '09 Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task Pages 1-9.

[3]L. Hirschman L, A. Yeh, C. Blaschke, A. Valencia . Overview of BioCreAtIvE: critical assessment of information extraction for biology.BMC Bioinformatics. 2005;6 Suppl 1:S1.

[4]L. Tanabe L, W.J. Wilbur.Generation of a large gene/protein lexicon by morphological pattern analysis. J Bioinform Comput Biol. 2004 Jan;1(4):611-26.

[5]B. Settles. ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. Bioinformatics. 2005 Jul 15;21(14):3191-2. Epub 2005 Apr 28

[6]R. Leaman ,G. Gonzalez. BANNER: an executable survey of advances in biomedical named entity recognition.Pac Symp Biocomput. 2008:652-63.

[7]J.Lafferty, A. McCallum, F. Pereira.Conditional random fields: Probabilistic models for segmenting and labeling sequence data.In Proc. of ICML, pp.282-289, 2001,282-289.

[6]F. Sha, F. Pereira. Shallow parsing with conditional random fields, In Proc. of HLT/NAACL 2003,1-8

[7]R. Bunescu , R. Ge R, R.J. Kate, E.M. Marcotte, R.J.Mooney, A.K. Ramani , Y.W. Wong .Comparative experiments on learning information extractors for proteins and their interactions. Artif Intell Med. 2005 Feb;33(2):139-55.