A Combined Approach for the Extraction of the Multi-word and Nested Biomedical Entity

Lejun Gong
School of Computer Science & Technology
School of Software,Nanjing University of Posts and Telecommunications, Nanjing, People's republic of china

Jiacheng Feng
School of Computer Science & Technology,
School of Software, Nanjing University of Posts and Telecommunications, Nanjing, People's republic of china

Ronggen Yang
Faculty of Information Technology
Jinling Institute of Technology
Nanjing 211169 People's republic of china

Geng Yang
School of Computer Science & Technology,
School of Software, Nanjing University of Posts and Telecommunications, Nanjing, People's republic of china

***Abstract*—Name entity recognition is the fundamental task in text mining area. This work focuses on the problems of multi-word and nested entity names. A combined approach is proposed for identifying multi-word and nested bio-entity names, which achieve an F-measure of 80.8% in extracting the total of bio-entity names and an F-measure of 82.2% aiming at nested entities. Experimental results show the combined approach is promising for developing text mining technology.**

Keywords—bioinformatics; text mining; name entity recognition; VSSWA;

I. INTRODUCTION

With the increasing expansion of biomedical literature, the demand for efficiently extracting biomedical information from the large volume of literature resources offers an excellent opportunity for biomedical text mining. Biomedical named entity recognition (bio-NER) is a prerequisite for automatic extraction of knowledge from literature [1]. In the biomedical domain, name entities refer to protein, disease, gene, virus, etc. Due to the irregularities and ambiguities in biomedical entities, named entity recognition might be considered as a solved problem in some domains, but it still poses a significant challenge in the area of biomedicine.

Biomedical entities are often long and include common words, multi-word names, nested names, spelling variation. Further, the use of capitalization, parenthesis, hyphen and abbreviation does not follow a well-defined convention. These factors make bio-NER in the biomedical domain difficult. In JNLPBA2004[2], the best system achieved an F-measure of 72.6%[3]; and in BioCreative 2004, the best system obtained an F-measure of 83.2% using relax matching and this score reduced to 74.3% using exact matching.

There are three main approaches to bio-NER, namely dictionary based, rule based and machine learning based. Dictionary-based approach matches dictionary entries against text, which need to build a large dictionary of collections of name. Tsuruoka and Tsujii [4] tagged proteins with a combination of dictionary and Naïve Bayes Classifier, achieving an F-measure of 66.6%. Cohen [5] achieved an F-measure of 75.6% in gene and protein through building the dictionaries from online genomics resources.

Rule-based systems use rules that describe common naming structures for certain term classes, based on morphological, orthographic and syntactic characteristics. Tanabe and Wilbur [6] developed a system called AbGene which achieved a precision of 85.7 % at a recall of 66.7 % for recognizing gene and protein using rule-based approach based on lexical-statistical characteristics. Chang et al. [7] created the GAPSCORE system which assigns a numerical score to each word within a sentence by examining the appearance, morphology and context of the word and then applying a classifier trained on these features with sloppy matches(F-measure 77 %) and exact matches(F-measure 54%).

Machine-learning approach depends heavily on the quality and quantity of the training set and the selection of feature set. It is a time-consuming and costly work to build a large and qualified training set. In JNLPBA2004, Settles [8] achieved an F-measure of 69.8 percent using conditional Random Fields (CRFs) with only several kinds of features and no external resource. Zhou et al. [9] trained a Hidden Markov Model (HMM) on a set of features based on word formation morphology, POS, semantic triggers and intra-document name aliases with an overall precision of 66.5 % at a recall of 66.6 % on the GENIA corpus.

978-1-4799-8058-1/15/$31.00

Most similar works don't focus on the problem of nested entity names. Our work aims to exploit the solutions to the problem of multi-words and nested entity names. This paper presents a combined approach for identifying multi-word and nested bio-entity by pattern matching and Variant Size Sliding Window Algorithm (VSSWA). Experimental results show our approach is promising for developing the text mining technology in biomedical domain.

The remaining part of this paper is organized as follows: Section 2 describes our methods. Section 3 presents the experiment results using GENIA corpus. Section 4 offers some concluding remarks.

II. METHODS

Biomedical literature is generally raw text which is digitally represented as a sequence of characters. Such plain text representation is usually processed to explicitly add structure in a machine-readable form. For identifying multi-word and nested bio-entity, we develop a combined approach which mainly includes three processing steps: Part Of Speech tagging (POS tagging), noun phrases recognition, multi-word and nested bio-entity recognition. An example is used to illustrate the steps in Figure 1.

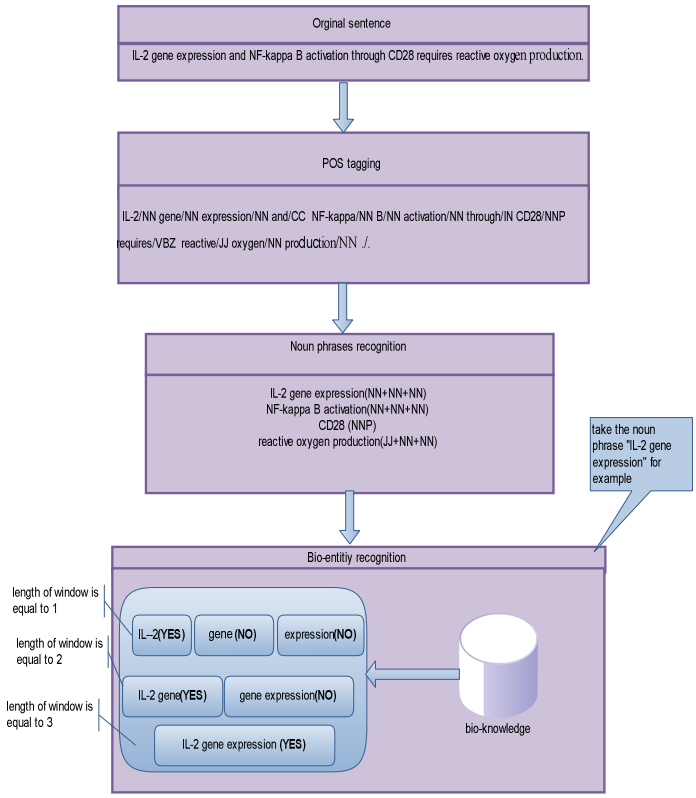

Figure 1. An example of how a sentence is analyzed step by step to extract entities

In Figure 1, an original sentence 'IL-2 gene expression and NF-kappa B activation through CD28 requires reactive oxygen production' is used as an example for clearly interpreting our approach. The sentence is first tagged by the tagger. The results of the POS tagger are then fed to finite state machine (FSM) for identifying noun phrases. The VSSWA is used to process these noun phrases for identifying multi-word and nested bio-entities. For example, noun phrase 'IL-2 gene expression' can be identified three bio-entities by the VSSWA including: 'IL-2', 'IL-2 gene' and 'IL-2 gene expression'. More details will be discussed in the following sections.

A. *POS tagging*

Experimental corpus is a kind of unstructured raw text such as abstracts. These abstracts are first broken into constituent sentences. Each sentence is then tokenized, which means it is broken into its constituent tokens. The tokens are then processed by POS tagging. POS tagger assigns a part of speech tag to each taken. The POS tags are the word classes based on its context and the definition belonging to a certain grammar. For instance, in Penn Treebank, POS tag 'VB' means verb, present tense or past tense. The POS tagger in our work, Stanford POS Tagger [10] with the Maximum Entropy is used to map the words to the POS tags from biomedical texts. The tagger achieves superior performance principally by enriching the information sources used for tagging with an accuracy of 96.86% on the Penn Treebank [11].

B. *Noun phrases recognition*

The results of the POS tagger are analyzed by the pattern matching based on certain grammar for recognizing noun phrases. We consider noun phrase is composed of several components including: adjective, noun, quantifier, participle and verbal noun (IL-2 gene expression, reactive oxygen production). The above several combinations of POS are up 96 to percent in the GENIA corpus. Therefore, noun phrase patterns are the set for candidate bio-entities by a Finite State Machine (FSM) defined as in [12] in accordance with the combination of the POS tags of noun phrases.

C. *Multi-word and nested bio-entity recognition*

Noun phrases extracted are considered as candidate maximum boundary bio-entities contained multi-word or nested entities. An algorithm called the VSSWA is developed to extract these bio-entities. The VSSWA algorithm is described as follows:

```
1. input: Noun phrases (max-boundary bio-entity)
2. initialize:hashset h=dictionary,    arraylist entity
3. for each noun phrase{
4.       string [] str_Noun=noun.split(" ")
5.       for (int win=1; win<=len (noun phrase); win++){
6.            for (int indices=0;indices<len-win;indices++){
7.                 compare the element of h with str_Noun[indices to win+indices]
8.                 if the sequence  of the window exists in the dictionary
9.                      entity.add(sequence)
10.           }
11.      }
12.}
13. output: entity
```

Figure 2 Algorithm VSSWA

Noun phrases act as the input of the VSSWA. Aiming at each noun phrase, algorithm finds the potential entities according to variant size windows with ascending order. The potential entities are filtered noise data by bio-knowledge dictionary, which obtains real bio-entities as the output results. The flow chart of the VSSWA is shown in Figure 3.

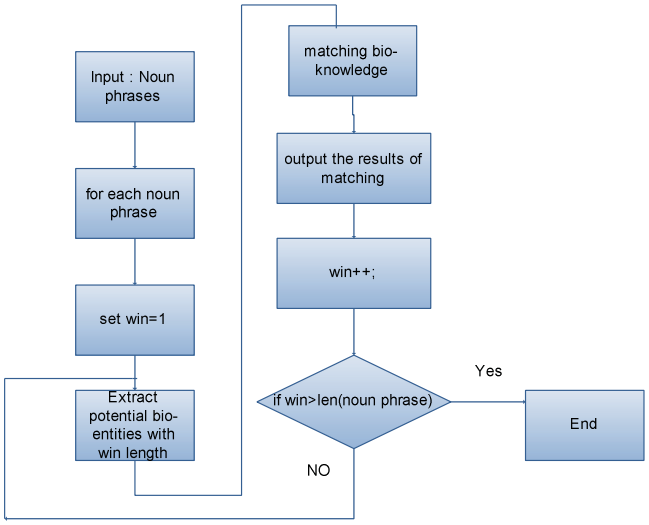

Figure 3. Flow chart of the VSSWA

III. EXPERIMENTS AND RESULTS

A. Datasets

We conducted experiments using GENIA corpus version 3.02. It is a time-consuming and costly work to build GENIA corpus which is the largest corpus of its type currently available, comprising 2000 abstracts with 18,545 sentences. This corpus contains 36 class entities annotated by manual analysis including: DNA, RNA, protein, cell-type, and cell-line, etc. Two experiments, extracted the total entities and nested entities, are conducted for the task. In this corpus, we have extracted 34 287 annotated bio-entity names which include 14 823 nested names. It is used to test the performance of our combined approach as gold standard dataset. Our test dataset is the same 2000 original abstracts download from MEDLINE as the GENIA corpus. Compared to obtain 18045 bio-entity names only using FSM approach, we obtained 23 273 bio-entity names with 10 305 nested names by our combined approach.

B. Evaluation

The F-measure is used to measure the performance for evaluating our proposed combined approach, which is the weighted harmonic mean of precision and recall. Precision is the number of correctly extracted entities divided by the total number of entities extracted, while recall is the number of correctly extracted entities divided by the gold standard data.

The general expression for measuring the f-measure is as follows:

$$F\text{-}measure=\frac{(1+\beta^2)(precision\times recall)}{(\beta^2\times precision+recall)} \quad (1)$$

Here the value of β is taken as 1.

We also conducted the experiment using FSM (containing bio-knowledge to filter noise data like the VSSWA) to extract bio-entity names for test dataset for contrasting the performance of our above combined approach. Experimental results are shown in Table 1.

Table1. Performance of different approaches

NO.	Project	F-measure
1	FSM (total entity names)	68.9%
2	FSM+VSSWA(total entity names)	80.8%
3	FSM+VSSWA(nested entity names)	82.2%

Due to the use of bio-knowledge, it also ensures the correctness of entities extracted in the above approaches. The approach of FSM can achieve 68.9% F-measure, while the approach of FSM and VSSWA can achieve 80.8% F-measure in extracting total entity names. In addition, the combination of FSM and VSSWA can achieve 82.2% F-measure in nested entity names extracted.

C. Discussion

The combination of FSM and VSSM can effectively extract multi-word bio-entity names including nested entities such as '<Protein><DNA>kappa 3</DNA>binding factor</Protein>'. Our approach recognizes not only 'kappa 3' DNA but also 'kappa 3 binding factor' Protein. Experimental results achieve 80.8% F-measure in extracting the total bio-entity names and 82.2% F-measure aiming at nested entities in

2000 abstracts of the GENIA corpus 3.02 versions. Compared to the similar works, Settle achieved 69.8% F-measure on GENIA corpus using CRF. Saha[13] obtained an F-measure of 58.63 using maximum entropy. Moreover, the problem of nested bio-entity isn't solved by them. In addition, the best system only achieved an F-measure of 72.6% in JNLPBA2004 which extracts 2000 abstracts as training set and 404 abstracts as test set from GENIA corpus version 3.02 from identifying five class entities. The performance of our combined approach is higher than the best system in JNLPBA2004. These evidences show our proposed approach is promising for developing text mining technology. Further, comparing extracted data with gold standard, there are bio-entities of spelling variation, for example, the variations of 'EGR-1' including 'EGR 1' and 'Egr-1', 'malignant tissues' and 'malignant tissue'. Our proposed approach can't tackle the problem of this kind of spelling variation, which also reduces the performance of our approach. Consequently, future work will improve our approach and solve the problem of spelling variation.

IV. CONCLUSIONS

This paper presents a combined approach for identifying multi-word and nested bio-entity. The combined approach uses pattern matching to find maximal boundary of bio-entity, applying the VSSWA to recognize multi-word and nested entity names. Experimental results achieve 80.8% F-measure in extracting the total bio-entity names and 82.2% F-measure aiming at nested entities in 2000 abstracts of the GENIA corpus 3.02 versions. The performance is higher than the best system in JNLPBA2004. The evidences of experimental results show our combined approach is promising for developing text mining technology.

ACKNOWLEDGMENT

This work is supported by the Natural Science Foundation of the Jiangsu Provice (Project No. BK20130417), and Scientific Research Foundation for the introduction of talent of Nanjing University of Posts and Telecommunications (Project No. NY213088) and NUPTSF (Project No. NY214068).

REFERENCES

[1] B. de Bruijin, J. Martin," Getting to the (c)ore of knowledge: mining biomedical literature," Int J Med Inform. 2002 Dec 4;67(1-3):7-18.

[2] J.D. Kim, T. Ohta, Y. Tateisi, et al. "Introduction to the bio-entity recognition task at JNLPBA," In: Proceedings of the International Workshop on Natural Language Processing in Biomedicine and its Applications(JNLPBA-04): 70-5, 2004.

[3] J.D. Kim, T. Ohta, Y. Tateisi, J. Tsujii "GENIA corpus--semantically annotated corpus for bio-textmining," Bioinformatics. 2003;19 Suppl 1:i180-2.

[4] Y. Tsuruoka, J. Tsujii , "Boosting precision and recall of dictionary-based protein name recognition," In: Proceedings of the ACL-03 Workshop on Natural Language Processing in Biomedicine. 2003.41-8.

[5] A.M. Cohen, " Unsupervised gene/protein entity normalization using automatically extracted dictionaries. In Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics," Proceedings of the BioLINK2005 Workshop. 2005.14-24.

[6] L. Tanabe, W.J. Wilbur, " Tagging gene and protein names in biomedical text," Bioinformatics. 2002 Aug;18(8):1124-32.

[7] J.T. Chang, H.Schütze, R.B. Altman , "GAPSCORE: finding gene and protein names one word at a time," Bioinformatics. 2004 Jan 22;20(2):216-25.

[8] B. Settles, " Biomedical named entity recognition using conditional random fields and rich features sets," In: Proc. JNLPBA-2004. 104-107.

[9] G.D. Zhou, J. Su , "Exploring deep knowledge resources in biomedical name recognition," In: Proc. JNLPBA-2004,96-99.

[10] K. Toutanova, C.D. Manning," Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger," In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000): 63-70.

[11] M. Marcus, B. Santorini, M.A. Marcinkiewicz, " Building a Large Annotated Corpus of English: The Penn Treebank," Computational Linguistics, 1994,19(2): 313-330.

[12] L.J. Gong, Y.Y. Yan, X. Sun. Target Identification and Target-centered Network Construction from Biomedical Literature. Journal of Software, Vol 8, No 2, 316-319, 2013

[13] S. K. Saha, S. Sarkar, "Mitra P. Feature selection techniques for maximum entropy based biomedical named entity recognition," J Biomed Inform. 2009 Oct;42(5):905-11. Epub 2009 Jan 23.