A Hybrid model for Named Entity Recognition in Biomedical Text

Dr. D. Ramesh, Dr. Suresh Kumar Sanampudi

Abstract— The phenomenal growth of medical information raises the need for mining the knowledge from biomedical text. Majority of biomedical information exists in unstructured text form. Named Entity Recognition (NER) from biomedical text is one of the basic tasks whose purpose is to recognize the name of some specified type of biological entities such as proteins, DNAs, RNAs, cells etc. Several attempts were made by the researchers to develop techniques for identification of named entities from biomedical literature. These methods are broadly classified into three categories such as, heuristic rule-based, dictionary-based, and statistical machine learning-based. But each approach has a drawback of either performance or complexity. This paper introduces a framework for Named Entity Recognition system in biomedical text that had shown improved performance and complexity. The foremost task of the framework is to identify and classify biomedical entities into one of these five classes, namely DNA, RNA, protein, cell-line and cell-type.

Index Terms— Bio medical text, Named Entity Recognition, Entity classification, DNA, RNA, Protien, Cell-line, Cell-type.

1 INTRODUCTION

With the outburst of data in the biomedical domain, there is a robust need for automated biomedical information extraction techniques. Identification and classification of named entity (NE) namely proteins, DNAs, RNAs, cells etc. has become one of the most essential tasks in the biomedical knowledge discovery. Several algorithms were developed to achieve the task of named entity recognition from biomedical text documents. NER for biological content still lies as a challenge due to its large gap of performance when compared with named entity recognition of other domains. Different methods were developed to identify NE from biomedical text. These methods are broadly classified into three categories such as rule based, dictionary based and statistical machine learning based. On the other hand, the state-of-the-art techniques for biomedical NER do not accomplish satisfactory results.

In this paper, a framework has been developed to explore various features that helped to identify and classify the NER from biomedical text. For example consider the following sentence

(1) (1) My medical documents contains the monocytes, piRNA, Elastin, Cadherin, Viral genome etc**.**

The framework developed aims to identify and classify monocytes as CELL-TYPE, piRNA and Viral genome as RNA, Elastin and Cadherin as PROTEIN from sentence 1. The experiments shown that our system achieves encouraging performances.

Variety of techniques were developed for biomedical named entities recognition from unstructured text. All these techniques broadly fall into one of these three categories namely Rule-based, dictionary-based, and machine learning approaches.

In rule-based approaches were built by defining specific rules by noticing the common features of biomedical entities in a biomedical text[1]. A huge amount of text collection has to be processed to construct the accurate rules. One of the example of rule based approach is PROPER (Protein Proper-noun phrase Extracting Rules)[2] which is used for protein names extraction from biomedical documents. The rules were framed based on orthographic features, such as presence of upper cases or special characters etc. several other models of rule based approaches were constructed to build this system[3][4].Rule based systems were found promising for small data sets. As the dataset size increases proportionately the number of rules increased. As the rules are mostly developed manually it found to be costly for adapting these system to new entity classes. Furthermore, these systems are incompetent to recognize new named entities that have not been grasped earlier. Since new entity names are often coined in the biomedical domain, this is a significant drawback.

In dictionary-based approach, formerly prepared lexicon list is compared though a given text to recover chunks containing the biological word [5]. Given a collection of strings (dictionary) this approach involves to compute word similari-

- *Dr. D Ramesh, Assoc Prof. Dept of C.S.E, J.N.T.U H College of Engineering Jagtial, Karimnagar, Telangana, INDIA E-mail: dantamr@yahoo.com*
- *Dr. S. Suresh Kumar, Asst Prof. Dept of I.T, J.N.T.U H College of Engineering Jagtial, Karimnagar, Telangana, INDIA E-mail: sureshsanampudi@jntuh.ac.in.*

ty among the word in the input text with that of dictionary content[6]. Certain post-processing is required in this approach to enrich the original dictionary with more concise name variations.

Machine learning based methods have a benefit that they do not require manual framing of rules and can also recognize new named entities not included in standard dictionaries [7]. Extracting of features plays a significant role in this approach. A diverse set of features such as orthogonal features, word class features, POS features, morphological features, key word features, frequency features etc., were used to improve the recognition efficiency of biomedical named entities. This can be achieved with methods like Hidden Markov Model [8][9], the Support Vector Machine (SVM)[10], the Maximum Entropy Markov Model[11], and Conditional Random Fields (CRF)[12][13].

The remaining part of this paper is organized as follows: Section 2 introduces a framework that can identify and classify NE. Experiments and results are described in section 3. Conclusion and future directions are drawn at last.

2 ARCHITECTURE

This paper proposes a method that provided a combination of dictionary based and machine learning approached to extract the features that consist of biomedical domain knowledge. This framework consist of two sub tasks namely named entity identification and named entity classification.

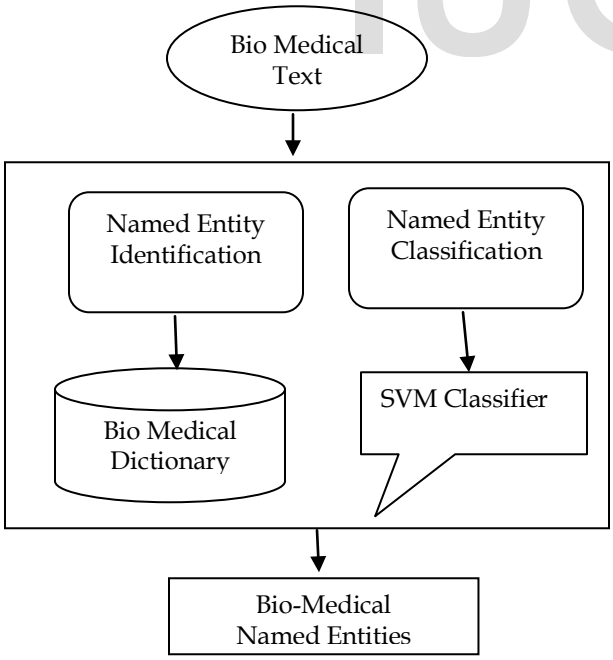

Figure 1: Architecture for Named Entity Recognition in Bio-Medical Text

In the named entity identification of architecture a paragraph of text containing biomedical information is given as input. Pre-processing steps are performed on the input in order to generate tokens. For this token generation process rules were defined to generate tokens. The generated tokens may or may not contains the named entities. The next step after generating the tokens is Named Entity token separation from all the available tokens. For this a specially designed dictionary which contains all the biomedical named entities are used. Each and every token in the tokens list is compared with the dictionary. If the token is available in the dictionary then that token is considered as a named entity and it separated from the other tokens. But this will not work all times because the named entity may be one word or more than one word. Sometimes instead of using original names the alias names will be used. The system should be in a position to identify the tokens even in this special situations. To achieve this feature sets were developed which will facilitate the identification of Named Entities. The procedure is repeated to identify all the named entities present in the given text.

For example consider the following text

(2) I and my sister Lully have same amount of Cadherin, Ependymin, Integrin and Serum Albumin.

Named entity Identification phase receives text as input and identify "Cadherin", "Ependymin", "Integrin" and "Serum Albumin" as named entities.

The next phase after the identification of the named entities is Named Entity Classification. In this phase the identified named entities were subjected for the classification by the analyzer. Support vector machine classifier is used to classify the named entities into its respective types [10]. Support vector is found to be an efficient method among the other classifiers than can classify the biomedical named entities. In this phase the output obtained from named entity identification is received as input and classify Cadherin, Ependymin, Integrin as "PROTEIN" and Serum Albumin as "RNA".

3 RESULTS

Experiments were performed on the data set from GENIA corpus which is a collection of Medline abstracts. Precision, recall and F measure were calculated for the results obtained through the proposed architecture. Comparative study of these results for the experiments conducted on same data set with respect to different methods are shown in table [8],[9],[10],[11].

Table 1. Performance Comparision

Methods	Precision	Recall	F Score
Proposed System	80.76%	84.24%	80%
Hidden Markov Model	69.41%	62.98%	66.04%
Conditional Random Fields	72%	69.9%	70%
Maximum Entropy Model	76.3%	77.6%	76.9%

The proposed method found to outperform when compared with other classification methods. This advantage is due to the mixture of advantages of both the dictionary based and machine learning based approaches. The proposed architecture found to obtain encouraging performance.

4 Conclusion

In this paper, we propose a hybrid method using support vector machines and dictionary/rule-based methods. Dictionary is only used in the Named entities identification stage as a preprocessing activity. Rules were developed to recover alias names as named entities were not identified by dictionary based method. In the next stage support vector classification is used to classify the type of named entity namely "protein", "DNA" etc. Experiments were conducted on the GENIA corpus consisting of MELINE abstracts. Precision, recall and F-score are calculated to evaluate the performance of the proposed system. When the results are compared with other methods it is observed that the results of the proposed methods were encouraging.

References

[1] L.J. Gong, and X. Sun, "ATRMiner: A system for Automatic Biomedical Named Entities Recognition," ICNC 2010, pp.3842-3845.

[2] Fukuda K, Tsunoda T, Tamura A, Takagi T. (1998) Toward information extraction: identifying protein names from biological papers. In: Proceedings of the Pacific Symposium on Biocomputing-98 (PSB_98); pp. 707–18.

[3] Gaizauskas R, Demetriou G, Humphreys K. (2000) Term recognition and classification in biological science journal articles. In: Proceedings of the computational terminology for medical and biological applications workshop of the 2nd international conference on NLP; pp. 37–44.

[4] Proux D, Rechenmann F, Julliard L, Pillet V, Jacq B.(1998) Detecting gene symbols and names in biological texts: a first step toward pertinent information extraction. In: Proceedings of genome inform ser workshop genome inform; pp. 72–80.

[5] Z. GuoDong, S. Jian, N. Collier, P. Ruch, and A. Nazarenko, "Exploring Deep Knowledge Resources in Biomedical Name Recognition," COLING 2004 International Joint workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP), 2004, pp.99-102.

[6] Pereira, F., Tishby, N., Lee, L. 1993. *Distributional Clustering of English Words*. In *Proceedings of ACL-93*. pp. 183-190. Columbus, Ohio.

[7] Nobata C, Collier N, Tsujii J. (1999) Automatic term identification and classification in biology texts. In: Proceedings of the 5th NLPRS; pp. 369–74.

[8] Zhou G, Zhang J, Su J, Shen D, Tan C: Recognizing names in biomedical texts: a machine learning approach. Bioinformatics 2004, 20(7):1178–90.

[9] S. Zhao, "Named Entity Recognition in Biomedical Texts using an HMM Model," The Proc. of the COLING 2004 International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA), 2004.

[10] Z. Ju, J. Wang, and F. Zhu, "Named Entity Recognition From Biomedical Text Using SVM," Bioinformatics and Biomedical Engineering (iCBBE 2011), 2011.

[11] J. Finkel, S. Dingare, H. Nguyen, M. Nissim, C. Manning, and G. Sinclair, "Exploiting Context for Biomedical Entity Recognition: From Syntax to the Web," Joint Workshop on Natural Language Processing in Biomedicine and Its Applications at Coling 2004, 2004.

[12] B. Settles, "Biomedical Named Entity Recognition Using Conditional Random Fields and Rich Feature Sets," The proc. of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-2004), 2004.

[13] L. Yang, and Y. Zhou, "Two-phase Biomedical Named Entity Recognition based on Semi-CRFs," Bio-inspired Computing: Theories and Applications (BIC-TA), 2010.