www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 6 Issue 5 May 2017, Page No. 21534-21537
Index Copernicus value (2015): 58.10 DOI: 10.18535/ijecs/v6i5.57

Biomedical Named Entity Recognition - a swift review

S.Vijaya[1], Dr.R.Radha[2]
[1]Research Scholar, Dept. of Computer Science, S.D.N.B.Vaishnav College for Women, Chromepet, Chennai
vijeshnisna@gmail.com
[2]Associate Professor, Dept. of Computer Science, S.D.N.B.Vaishnav College for Women, Chromepet,Chennai
radhasundar1993@gmail.com

Abstract: ***The main focus of this paper is taking a swift review on the Biomedical Named Entity Recognition which is the most complex task in Information Extraction. This paper analyses various methods used for NER particularly in the field of Biomedical domain. The aim of this study is to discuss about the methods used to recognize Biological entities like genes, proteins and diseases etc., and propose an effective method to recognize heterogeneous entities.***

Keywords: Biomedical Named Entity Recognition, Information Extraction, Heterogeneous entities

1. Introduction

Named Entity identifies an item from a set of other items which has similar types of attributes. The word 'named' refers proper names but it is domain dependent. Named Entity Recognition refers extracting proper nouns like name of a person, organization, date, time, etc. In biomedical domain NER recognizes genes, proteins etc.[23] Recognizing proper names from textual information, classifying them into predefined set of entities are the major tasks involved in Named Entity Recognition. While comparing general domain and Biomedical domain the task of NER is most difficult in the field of Biomedical and this problem taken by many researchers. This motivated us to review various methods used for NER in Biomedical domain, analyse the performance and propose an efficient method to extract biomedical entities effectively.

2. Proposed Method

Objectives:

i) Recognition of Biomedical Entity

ii)Assigning the named entity to a predefined class

iii)Finding the most suitable name for the entity

To achieve better results than existing methods we propose Hybrid approach to extract named entities effectively and a classifier to filter the false negatives.

Figure1.Describes our proposed method.

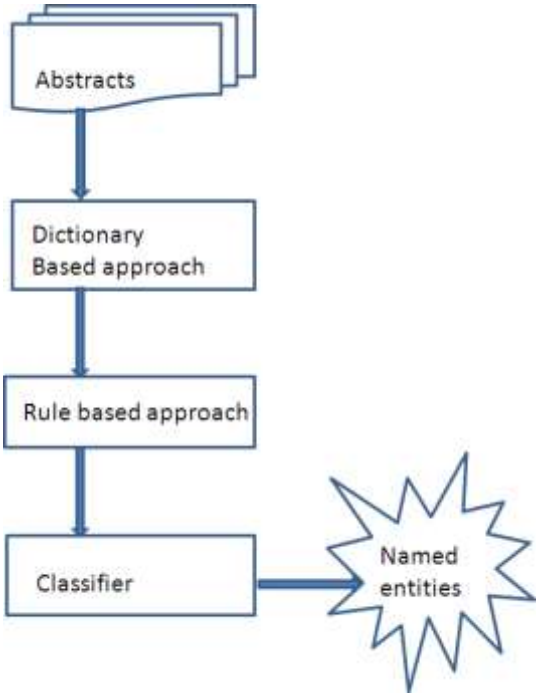

Figure1. Proposed system

3. Methods used by other researchers

3.1. Dictionary based Approaches

Dictionary based protein name recognition is used in extracting information from biomedical documents as it can provide ID information on recognized terms. It is very simple and efficient method. This method identifies Named Entities by

matching terms. The major problem in this approach is its performance as the accuracy is not satisfied because of limited coverage, spelling variations, homonymy, synonymy [1] and false recognition.

To overcome false recognition that is mainly caused by short names Yoshimasa Tsuruoka et al[29]have used machine learning to filter out false positives , to improve recall rate they have used appropriate string searching techniques and expanded the dictionary in advance with a probabilistic variant generator.

As Dictionary based approaches have limitations such as false positive recognition and lack of a unified resource that covers newly published names [28] addressed an approach with two-phase method where the first method scans text for protein name candidates and the second phase method filters irrelevant candidates by utilizing a Naïve Bayes Classifier.

3.2 Rule Based Approaches

In Rule based approach the named entities are recognized by predefined rules that describe typical naming structure. For example alphanumeric words, words with special symbol, capitalization, etc.,. Wide knowledge in linguistics, biomedical domain and programming skills are required to frame these rules. Designing a rule to deal with specific domain is its advantage. The difficulty in applying rules defined for a particular domain to other domain , requirement of wide domain knowledge to define rules, time consumption are major drawbacks in rule based approach.[1]

3.3 Machine Learning Model

To train data set supervised Machine Learning methods are used widely. Machine learning methods give better performance and it can be easily adopted on other domain. The limitations are these methods require reliable training resources and the number of features increases depends on problem size. Among many approaches used in machine learning approaches the Maximum Entropy, SVM and HMM methods are being used by many researchers because of their results shown outperform than other approaches.

3.3.1 Maximum Entropy

S.Raychaudhuri et al[24] used Maximum Entropy method on the genes appeared in Biomedical Literature and assigned Gene Ontology tags to that genes. They have used Naïve Bayes method and Nearest neighbour method with Maximum Entropy. Jon Patrick et al. [10] proposed Machine learning approach using Maximum Entropy model.

3.3.2Support Vector Machine

SVM approaches are most successful in classifying text automatically. S.Pakhomov et al [21] reported that Support Vector Machine (SVM) outperforms Maximum Entropy for Biological Named Entity Recognition. Zhenfei Ju et al[31] used SVM for biomedical NER , used Training data and Testing data from GENIA Corpus and got 84.24% Precision and 80.76% Recall rate in GENIA Corpus.

3.3.3Hidden Markov Model

The great success in recognizing biomedical named attracted many researchers to use this model. This model consists of states and observations. J.D.Kim et al[13] used their original machine learning method named self –organizing Hidden Markov Model(SOHMM) with a simple feature set.

4. NER using Hybrid methods

Hakenberg et al[9] proposed a machine learning system for extraction of gene and protein names form literature. They found a 10 percentage point difference in their strict and loose evaluation results. Hang-woo et al[8] developed a method with machine learning based Named Entity Recognition to filter out false recognitions of disease and gene names.

Yi-Feng et al.[30] have proposed hybrid method which used Maximum Entropy , dictionary based and rule based methods. Dictionary based and rule based methods used for post processing to overcome the inaccurate boundary detection that might occur while using Maximum Entropy method for Named Entity Recognition. They have used POS(Parts Of Speech)features annotated in the GENIA Corpus.

Haochang Wang et al.[7] have conducted experiments with Generalized Winnow, Conditional Random Fields(CRF), SVM and Maximum Entropy and explored local features for biomedical NER. They have used Ensemble approached for classification to improve recognition accuracy.

Shaodian Zhang et al[26] proposed unsupervised approach to overcome the challenges of entity boundary detection and entity type classification. Their method identified entities from raw text , leveraged existing terminology in lieu of task specific user defined rules or online information retrieval and added internal words using TF-IDF weights. Their work included a seed term extractor , an NP Chunker, an IDF filter , a classifier based on distributional semantics to provide a solution to Biomedical NER.

J-D Kim et al[12] used supervised learning approaches which require the annotated
corpora for the development , evaluation of Relation Extraction and shown good
performance. Manabu Torii et al[18] developed BioTagger-TM using i)rule/pattern based recognition methods characterized by handcrafted name/context patterns and associated rules ii) dictionary look up methods requiring a list of entity names and iii)Machine learning methods utilizing named entity tagged corpora . In their work on large entity corpus, machine learning methods had given promising performance.

5. Performance analysis

This section gives performance analysis based on the results obtained by various methods and researchers.

Table 1: Performance analysis table

S.No.	*Author*	*Methods used*	*Preci -sion*	*Recal l*	*F-Score*
1.	Yi-Feng et al[30]	ME,Dictio -nary based and rule based	77	80	78.5
			65.3	74.8	70
			71.6	78.8	55.6
2.	Y.Tsuruoka et al[28]	Support Vector Machine	49	66.4	56.5
3.	D.Hanisch et al[6]	Maximum Entropy	49.1	62.1	54.8
4.	Zhenfei Ju et al[31]	Support Vector Machine	84.24	80.76	-
5.	Haochang Wang et al[7]	General Winnow Algorithm	67.99	72.48	70.16
		CRF	70.02	72.35	71.17
		SVM	64.04	62.32	63.17
		ME	65.12	71.19	68.02
6.	Yoshimasa Tsuruoka et al[29]	Dictionary -based (without filtering)	46.5	65.4	54.3
		Dictionary -based (with filtering)	60.1	58.0	59.0
			68.2	59.8	63.7
			71.2	60.1	65.1
7.	Kazamat et al[11]	SVM	-	-	56.5
8.	Collier et al[2]	Hidden Markov Model	-	-	75.9
9.	Tanabe and Wilbus[27]	Statistical & Knowledg e- based Strategies	85.7	66.7	-
10.	Krauthamm er et al[15][24]	Dictionary based	78.8	71.1	-
11.	Fukuda et al.[4]	Rule based	94.7	98.8	96.7
12.	Proux et al.[22]	Rule based	91.4	94.4	92.9
13.	Gaizauskas et al[5]	Rule based	96 97	98 87	97 91.7

6.Conclusion

There are several common issues in recognizing biomedical entities such as no specific dictionary, same word referring different meaning, different phrase to a common entity, abbreviation, etc. To overcome these issues and recognizing entities effectively we propose Hybrid approaches as the results showed are greater than using the approaches alone. Using classifiers with Hybrid approaches can be used to improve the precision and recall rate.

References

[1] Baohua, "Recognizing Named Entities In Biomedical Texts",Thesis,Simon Fraser University 2008.

[2] N.Coolier,C.Nobata,J.TSujii,"Automatic acquisition and classification of molecular biology terminology using a tagged corpus", Journal of Terminology 7(2),(2001),239-258.

[3] Dietterich TG, "Ensemble methods in machine learning ", In: Proceedings of the First International Workshop on Multiple Classifier Systems, Lecture Notes in Computer Science, New York : Springer-Verlag 2000:1857.1-15.

[4] Fukuda K, et al., "Toward Information Extraction: Identifying Protein Names in Biomedical Papers",Proc. Pacific Symp. On Biocomputing, 1998,pp.707-718.

[5] Gaizauskas R , G.Demetriou and K.Humphreys,"Term Recognition and Classification in Biological Science Journal Articles", Proc. Workshop on Computational Terminology for Medical and Biological Applications, Patras, Greece,2000,pp.37-44.

[6] D.Hanisch, J.Fluck,H.Mevissen and R.Zimmer, "Playing biology's name game: identifying protein names in scientific text" , In PSB '03, 2003.

[7] Haochang Wang, Tiejun Zhao, Hong Ye Tan, Shuzhang, "Biomedical Named Entity Recognition Based on Classifiers Ensemble", International Journal of Computer Science and Applications, Vol.5,No-2,pp1-11.

[8] Hong-woo C, Y Tsoruoka, J D Kim,Rieshiba,N Nagata, T Hishiki, J.Tsujii, "Extraction of Gene-Disease relations from medline using domain dictionaries and machine learning",Pacific Symposium on Biocomputing 11: 4-15(2006) October 13, 2005, 15:5 Proceedings.

[9] Jorg Hakenberg,Steffen Bickel , Conrad Plake, Ulf Brefeld, Hagen Zahn, Lukes Faulstich, Ulf Leser and Tobians Scheffer,"Systematic feature evaluation for gene name recognition ", BMC, Bioinformatics,6 Suppl.1(2005)

[10] Jon Patrick and Yefend Wang," Biomedical Named Entity Recognition System", Proceedings of the 10th Australian Document Computing Symposium, Sydney, Australia, December-12,2005.

[11] J.Kazama, T.Makino, Y.Oh ta,J.Tsujii, "Tuning support vector machines for biomedical named entity recognition", in: Proceedings of the ACL-02, Workshop on Natural Language Processing in the Biomedical Domain, 2002,pp.1-8.

[12] J.D.Kim, J.Tsujii, "Corpus-based approach to biological entity recognition", in:Text Data Mining SIG(ISMB2002),2002.

[13] Kim J-D, Ohta t, Pyysalo S, Kano Y, Tsujii J ,"Overview of BioNLP'09 shared task on event extraction in BioNLP'09 Proc Work Curr Trends Biomed Nat Lang Process Shar Task. Association for Computational Linguistics;2009:1-9.

[14] Koike A, Niwa Y, Takagi T,"Automatic extraction of gene/protein biological functions form biomedical text", Bioinformatics 2005: 21-1227-36.

[15] M.Krauthammer, A.Rzhetsky, P.Morozov, C.Friedman, "Using BLAST for identifying gene and protein names in journal articles", Gene 259(2000) 245-252.

[16] Lafferty J,McCallum A, Pereira F, "Conditional random fields:probabilistic models for segmenting and labelling sequence data", In : Proceedings of the Seventeenth International Conference on Machine learning, San Francisco, CA:Morgan Kaufmann,2000:591-8.

[17] Liu H, Hu ZZ, Zhang J, Wu C, "BioThesaurus: a web-based thesaurus of protein and gene names",Bioinformatics 2006:22:103-5.

[18] Manabu Torii, Zhangzhi Hu, Cathy h, Wu, Hongfang Liu , "BioTagger-GM: A Gene/Protein Name Recognition System", Journal of the American Medical Informatics Association, Volume 16, Number 2, March / April 2009.

[19] McCallum A, Freitag D,Pereira F, "Maximum Entropy Markov Models for Information Extraction and Segmentation", In : Proceedings of the Seventeenth International Conference on Machine learning, San Francisco, CA:Morgan Kaufmann,2000:591-8.

[20] McDonald R, Pereira F, "Identifying gene and protein mentions in text using conditional random fields", BMC Bioinformatics 2005:6, Suppl 1:56.

[21] S.Pakhomov ," Semi-supervised maximum entropy based approach to acronym and abbreviation normalization in medical text", In ACL 2002.

[22] Proux F., et al., "Detecting Gene Symbols and Names in Biomedical Texts: A First Step Toward Pertinent Information", Proc. 9th Workshop on Genome Informatics, 1998,pp.72-80.

[23] Rahul Sharnagat, "Named Entity Recognition : A Literature Survey", June 30, 2014.

[24] S.Raychaudhuri, J.Chang,P.Sutphin and R.Altman , "Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature", Genome Research ,12,2002.

[25] Sha F, Pereira, "Shallow Parsing with Conditional Random Fields", Conference of Human Language Technology and North American Chapter of Association of Computational Linguistics, 2003.

[26] Shaodian Zhang, Noemie Elhadad, "Unsupervised biomedical named entity recognition: Experiments with Clinical and biological texts", Journal of Biomedical Informatics 46(2013) 1088-1098.

[27] L.Tanabe, W.J.Wilbur, "Tagging gene and protein names in biomedical text", BIOINFORMATICS 18(8) (200) 1124-1132.

[28] Y.Tsuruoka and J.Tsujii , "Boosting precision and recall of dictionary-based protein name recognition", In ACL 2003, 2003.

[29] Tsuruoka Y, Tsujii J, "Improving the performance of dictionary based approaches in protein name recognition", J Biomed Inform 2004,37:461-70.

[30] Yi-Feng Lin, Tzong-Han Tsai et al., "A Maximum Entropy Approach to Biomedical Named Entity Recognition ", BIOKDD04: 4th Workshop on Data Mining in Bioinformatics(With SIGKDD Conference)

[31] Zheinfei Ju, Jian Wang , Feizhu, "Named Entity Recognition from BioMedical Text using SVM", 2011 , IEEE.