Australian Journal of Basic and Applied Sciences, 9(7) April 2015, Pages: 393-398

ISSN:1991-8178

Australian Journal of Basic and Applied Sciences

Journal home page: www.ajbasweb.com

Named Entity Recognition from Biomedical Abstracts–An Information Extraction Task

[1]N. Kanya and [2]Dr. T. Ravi

[1]Research Scholar, Manonmanium Sundaranar University, Department of Computer Science and Engineering, Tirunelveli, India.
[1]Asst Professor, Dr.M.G.R Educational and Research Institute-University, Department of Computer Science and Engineering,Chennai, India.
[2]Principal, Srinivasa Institute of Engineering and Technology, Tamil Nadu, India.

ARTICLE INFO

Article history:
Received 12 October 2014
Received in revised form 26 December 2014
Accepted 1 January 2015
Available online 27 February 2015

Keywords:
Information Extraction, Text Mining, Named Entity Recognition, Machine Learning, Conditional Random Field s

ABSTRACT

Background: The task of Information Extraction(IE) is to recognize a predefined set of concept in a particular domain, ignoring other irrelevant information's. Where a domain consist of a corpus of text together with a clearly specified information need. Information extraction includes three basic tasks such as Named entity Recognition , Co-reference Resolution and Relation extraction. On that Named entity Recognition is primary Information Extraction task to identify the named entities in the given text. It involves in identifying textual mentions of named entities that fit in to a predefined set of categories. **Objective:** Named Entity Recognition(NER) is a crucial initial step in information extraction. In the biomedical Information Extraction It is an essential task in identifying various kinds of key terms mentions such as genes, proteins, cell trace etc. The biomedical society makes wide use of text mining technology. NER is one of the most primary and significant tasks in biomedical information extraction of text mining technology. NER involves in processing structured and unstructured documents to recognize the definite kinds of entities and categorization of them into some predefined classes. **Result:** In this article we are presenting a framework for biomedical Information Extraction pipeline, Which includes various domain independent and domain specific tasks such as sentence splitter, tokenizer, POS Tagger, morphological analyzer, chanker, NER etc. Several NER systems have been developed based on the techniques of Rule-Based, Dictionary based and Machine Learning based for Biomedical Domain. Machine learning based approaches have many advantages than other approaches. In this paper we are proposing an Machine learning based framework for recognizing named entities from biomedical abstracts. For this study we used benchmarked datasets such as GENETAG and JNLPBA. **Conclusion:** Considering the framework and the underlying characteristics, we believe that this framework is an important contribution to the biomedical community in the development of Named entity Recognition.

To Cite This Article: N. Kanya and Dr. T. Ravi, Named Entity Recognition from Biomedical Abstracts –An Information Extraction Task. ***Aust. J. Basic & Appl. Sci.**, 9(7): 393-398, 2015*

INTRODUCTION

The rapid growth of internet has resulted in huge amount of information generated and available in the form of textual data, image, video and sounds. Text mining is the discovery of interested , non trivial knowledge from the unstructured text. Information Extraction is a essential text mining task. Information Extraction is about deriving structured factual information from unstructured text. The IE pipeline framework is to identify instances of a particular pre specified class of entities, Co-reference Resolution, relationships and events from natural language text.

In the biomedical domain vast amount of data is available, in the form of articles, research finding and literature abstract. The MEDLINE literature database contains over 20 million references to journal papers, covering a wide range of biomedical field with a growth rate of about 400,000 articles per year. Named entity recognition in the domain of biomedical aims to identify Bio-entities corresponding to the instances of concepts that are of interest to biologists.

This primary task of Named Entity Recognition (NER) includes various preprocessing stages. In support of the correct identification and normalization of Named Entities.

Named Entity Recognition is the most important and fundamental task in biomedical information extraction. The Named Entity Recognition involves in two different stages, The first step is Identification of specific kinds of entities and the second step is to classify them into some predefined classes. This

Corresponding Author: N. Kanya, Research Scholar, Manonmanium Sundaranar University, Department of Computer Science and Engineering, Tirunelveli, Tamil Nadu, India-627012,
Tel: 09884499949 E-mail: kanyamtech@yahoo.co.in,

entire task is termed as Named Entity Recognition and Classification (NERC) McCallum (2003). Biomedical named entities (NEs) includes mentions of proteins, genes, DNA, RNA etc. which, typically, have compound structures and not easy to recognize.

For the Biomedical domain several Named Entity Recognition and Classification (NERC) systems have been developed using different techniques and approaches. It can be categorized as

Rule-Based: It focuses on extracting names using lots of human-made rules set. Result will be good if the domain is restricted.

Dictionary Based: Depended on predefined terms of names.

Machine Learning Based : This approach will have advantage over other two approaches. It looks for patterns and relationships, using statistical models and various ML Algorithms.1, Settles B (2004)
There are two types of machine learning model for NER.

- Supervised machine learning model.
- Unsupervised machine learning model.

Supervised machine learning process takes a known set of input data and known responses to the data, and seeks to build a predictor model that generates reasonable predictions for the response to new data.

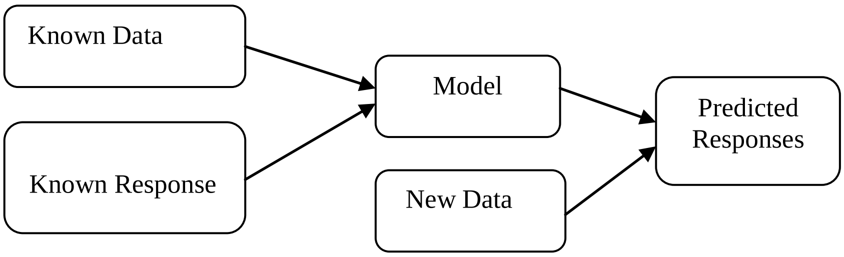

Fig. 1: Supervised Learning.

The disadvantage of this approach is it requires preparing labeled training data to construct a statistical model, but it cannot achieve a good performance without a large amount of training data, because of data sparseness problem.

The progress of Machine Learning based solutions has two essential process. In the first process the Machine Learning Model must be trained using the annotations present on the annotated documents. In the second process after storing the model in a physical resource, raw documents can be annotated, as long as entity names based on the past experience inferred from the annotated document Dai H *et al* (2010), 13.

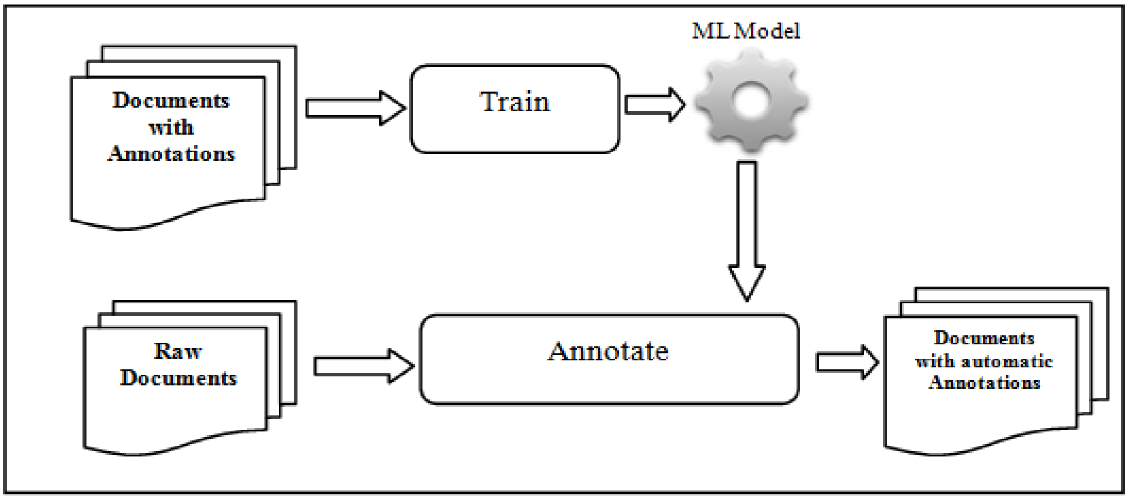
ML Model
Documents with Annotations
Train
Raw Documents
Annotate
Documents with automatic Annotations

Fig. 2: The progress of Machine Learning based solutions.

Methodology:

The Machine learning based solution such as training and annotation tasks depends on range of processing steps and resources. These set of processes have grouped in to an integrated framework for machine learning based Named Entity Recognition task. Which consist of the following modules. Zhao S(2004), Sun C *et al* (2007)

- ***Documents with Annotations:***

It is collection of texts and the annotations related with the target domain.

- ***Preprocessing:***

It process the input data from the corpora in order to simplify the target domain.

- ***Feature Selection:***

This process involves in extracting, selecting and features from the preprocessed input data.

- ***Machine Learning Model:***

This process involves in identifying the entity names from the generated features using the various Machine Learning Algorithms 1, Finkel .J, S. Dingare, (2004).

- **Post-Processing:**

This process will extend the task by refining the generated annotations. solving the conflicts ion the recognition process.

The input to this frame work will be Corpora of annotated documents. And the Output of the framework will be the annotated documents and Extracted information's in the structured form. Fig.3. illustrates the processes of the framework. Tanabe L(2002)

2.1 Corpora / Documents with Annotations:

Documents with Annotations is a collection of text documents that usually holds the annotations of one or more entity types. Zhou, G (2004). These annotations are used to train the Machine Learning models. The quality of the annotations present in the corpora makes the high impact on the training model.

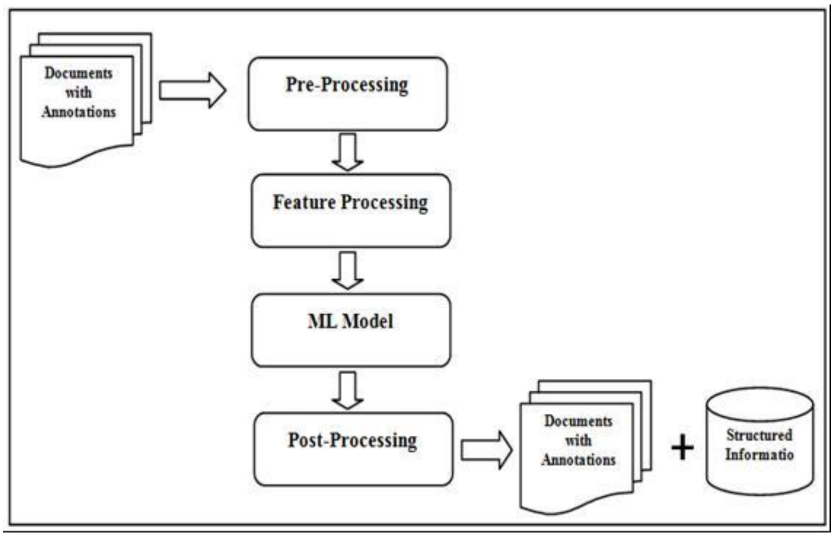

Fig. 3: Framework for Machine Learning based Named Entity Recognition System.

Table 1: Presents Gold Standard Corpora available for various gene and protein entity types.

Entity	Corpus	Type	Size (Sentences)
Gene and Protein	PennBioIE	Abstracts	≈28977
	JNLPBA	Abstracts	≈25402
	FSUPRGE	Abstracts	≈29447
	GENETAG	Sentence	≈23120

Table 3: Features of NER Systems.

Features		AIIAGMT	Gimli	BIOEnEx	NERSuite
Linguistic	Normalization	❋	❋	❋	❋
	POS	❋	❋	❋	❋
	Chunking	❋	-	-	❋
	Dependency	❋	-	❋	-
Orthographic	Capitalization	❋	❋	❋	❋
	Counting	❋	❋	-	❋
	Symbols	❋	❋	❋	❋
Morphological	suffix and prefix	❋	-	❋	-
	n-gram	❋	❋	❋	❋
	Word shape	❋	❋	-	❋
Lexicons	Target names	❋	-	❋	❋
	Trigger names	❋	❋	-	-
ML Model	Supervised	CRF	CRF	CRF	CRF

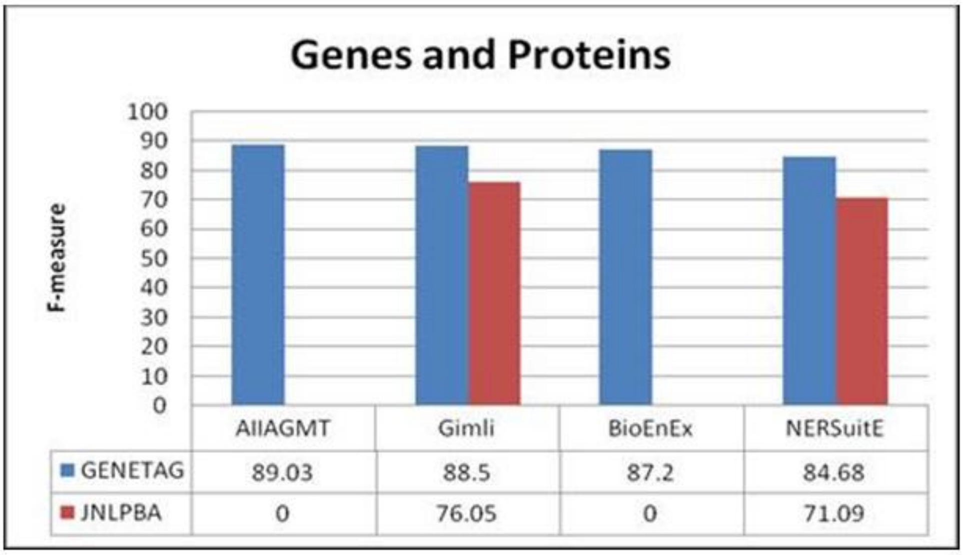

Fig. 4: Performance Comparison of various Machine Learning based NER

The two standards of annotated corpora are

- Gold Standard Corpora. (GSC)
- Silver Standard Corpora (SSC)

In the Gold Standard Corpora (GSC), the annotations are created manually by domain expert annotators. In the Silver Standard Corpora, annotations are generated automatically by the computerized Systems.

In most of the Biomedical Named Entity System research the efforts have been on the gene and protein names. The two different factors for this are:

- The importance of gene and protein names on the biomedical domain.
- High inconsistency and Non-standardization of names.

Pre-Processing:

Natural Language Processing (NLP) solutions look forward to their input to be segmented into sentences and each sentences into tokens. As the biomedical documents lacks in the well defined structure. We need to implement the following process.

- Sentence Splitting
- Tokenization
- Annotation encoding

Sentence Splitting:

The process of breaking a text document in to sentences is known as Sentence Splitting. Finally the tasks provide a detailed logical and meaningful context for further process. For Biomedical abstracts and documents various tools were developed to achieve the best result (Dai, 2010). The best fit tools with high accuracy are JSBD, OPENNLP and SPECIALIST NLP Geetha S, GS Mala (2014).

Table 1 presents Gold Standard Corpora available for various gene and protein entity types.

Pre-Processing:

Natural Language Processing (NLP) solutions look forward to their input to be segmented into sentences and each sentences into tokens. As the biomedical documents lacks in the well defined structure. We need to implement the following process.

- Sentence Splitting
- Tokenization
- Annotation encoding

Sentence Splitting:

The process of breaking a text document in to sentences is known as Sentence Splitting. Finally the tasks provide a detailed logical and meaningful context for further process. For Biomedical abstracts and documents various tools were developed to achieve the best result (Dai, 2010). The best fit tools with high accuracy are JSBD, OPENNLP and SPECIALIST NLP He, Y., & Kayaalp, M. (2006).

Tokenization:

The process of breaking the sentences into meaningful units, called tokens. Tokenization is a important process of the Information Extraction task. Zhou, G(2004)All the Named Entity Recognition processes depends on the tokens generated by this task. For Tokenization processes in the biomedical domain a range of tools such as GENIA Tagger, JTBD and SPECIALIST NLP were developed with high accuracy around 96%.

Annotation encoding:

Encoding schemes are necessary to give a tag the each annotated tokens of the text (McCallum 2003). The IO encoding tag is the simplest technique for annotation encoding ,but it cannot annotate two entities next to each other. This issue related to boundary problem can be overcome by “Bio encoding". In this encoding technique Hanisch D (2003)

- Tag "B" represents the First token of or beginning of entity name.
- Tag "I" represents the remaining.

The extension of "Bio encoding" is "BMEWO". This includes:

- Tag "E" represents the end of the entity.
- Tag "M" represents the tokens from the middle of the entity.
- Tag "W" represents the new tag for the entities with only one tokens.

Feature Selection:

Feature Selection is a crucial Named Entity Recognition task. To represent the target entity names properly it is necessary to have the rich set of features.

Linguistic:

The tokens are most essential basic inner features. Most of the tokens have the morphological variants. That is the tokens have similar semantic interpretations (He, 2006). This tokens can be considered as equivalent. The Stemming or lemmatization can be used to group all inflected tokens. So that they can identified as a similar token. The stemming used to identify the tokens with similar prefix. Whereas the Lemmatization is used to identify the root term of the variant tokens. Using the Part-of-Speech (POS) tagging tokens can be classified based on grammatical category of its context. Further dividing the text in to syntactically correlated words can be done by the process chunking. To identify the tokens as the sentence these linguistic features were used. Parsing tools can be used to collect the relations between the ranges of tokens in the sentences.

Orthographic:

The orthographic features are used to get the knowledge about word formation. The features such as words starts and presence of Upper and lower case letters , presence of symbols, or counting the number of digits , upper and lower case characters in the tokens can be used to analyze the orthographic features.

Morphological:

To identify the similarity between different tokens, three kinds of morphological features can be considered:

1. Suffix and Prefix can be used to differentiate between entity names.
2. The n-grams techniques
3. Word shape patterns.

Lexicons:

To further optimize the Named Entity Recognition system can be done by adding domain knowledge of biomedical field.

Two different dictionaries been used to match the tokens with the annotated tags.

- Dictionary of Target Entity Names
- Dictionary of Trigger Name

Dictionary of Target entity name compares the tokens with dictionary content with the set of name of the target entity names.

Dictionary of Trigger names indicate the presence of the surrounding tokens with the biomedical names.

Machine Learning Model:

Machine Learning Model can be classified as supervised and Unsupervised/ Semi supervised model.

Supervised learning uses annotated data, has received most research interest in recent years. Consequently, different supervised models have been used on NER systems, namely (Adafre 2005)

- Conditional Random Fields (CRF)
- Support Vector Machines (SVM)
- Hidden Markov Models (HMMs)
- Maximum Entropy Markov Models (MEMMs)

CRFs have several advantages over other methods. CRFs overcome the label bias problem, which is the main weakness of MEMMs. Also, CRFs have advantages over HMMs, a consequence of their conditional nature that results in relaxation of the independence assumptions. Tsuruoka. Y (2005) Finally, although SVMs can provide equivalent results, but to train complex models more time is required . Semi-supervised solutions use both annotated and unannotated data, in order to solve the data sparseness problem.

Thus, the main goal is to collect features of the unannotated data that are not present in the annotated data, which may contribute to a better identification of the entity names boundaries.

There are various approaches to implement semi-supervised solutions, such as Semi- CRFs Wallach H (2014), Semi-SVMs , ASO and FCG.

Post-Processing:

Post-processing techniques are used to solve some recognition issues, It can be easily corrected through simple rules or methods: Adafre SF (2005)

- Remove or correct recognition mistakes: Annotations with an odd number of brackets may be removed or corrected.
- Extend or make annotations more precise: Abbreviation resolution methods can be used to extend detected annotations. Moreover, curated dictionaries can be also used to correct generated annotations.
- Remove uninformative terms:

Some annotations may be known for being non-informative or unwanted terms, and consequently must be removed.

Evaluation:

In order to understand the behavior of the system, it is important to measure the

The performance results are obtained using F-measure. It is the harmonic mean of precision and recall.

$$\text{F-Measure} = 2\ \frac{\text{Precision x Recall}}{\text{Precision} + \text{Recall}}$$

Tools:

Since many tool are available for recognition of a specific entity types such as gene and protein, we decided to experiment the systems that better reflect the overall process of the domain. The applications main target will be gene and protein names. In this study we are having four systems for that process. The systems are AIIAGMT, Gimli, BIOEnEx, NERSuite. The features of the above systems mentioned in the Table.3. The two Biomedical corpora GENETAG and JNLPBA were used., Sun C *et al* (2007)

Conclusion:

In this paper we presented a study on various machine learning based named Entity Recognition tools for Biomedical data. Initially we described the fundamental processes involved in the progress of the tools. Then we analyzed the features of the tools for NER. Such analysis allowed us to expose the current trends of Machine Learning based recognition of Named Entities for Biomedical Documents, and we compared the performance of the four systems for NER. In future we are planning to enhance the work. Kim JD(2004)

Entities extracted from biomedical documents can be used to identify the multiple (coreferring) mentions of the same entity in the text identified. And the associations between the entities and Event Extraction.

Future Enhancement:

The IE framework consist of Named Entity Recognition, Co reference resolution, Relationship Extraction and Event identification. This paper focused on Named Entity Recognition Our future research work will focus on Relationship extraction and identifying the Events from that findings.

REFERENCES

Adafre, SF., 2005. Rijke Md: Feature Engineering and Post-Processing for Temporal Expression Recognition Using Conditional Random Fields. In ACL-05 Workshop on Feature Engineering Ann Arbor, pp: 9-16.

- True Negative (TN): the non existence of an annotation is correct according to the curated corpus;
- False Positive (FP): the system provides an annotation that does not exist in the curated corpus;
- False Negative (FN): the system does not provide an annotation that is present in the curated corpus.

Dai, H., 2010. New challenges for biological text-mining in the next decade. J Comput Sci Technol, 25(1): 169–179.

Finkel, J., S. Dingare and H. Nguyen, 2004. Exploiting context for biomedical entity recognition: from syntax to the Web. Proc. of the Joint Workshop on NaturalLanguage Processing in Biomedicine and Its Applications,Geneva, Switzerland, August (28–29): 88–91.

Geetha, S., GS. Mala, 2014. Automatic database construction from natural language requirements specification text. journal of engineering & applied sciences,

Hanisch, D., J. Fluck, H. Mevissen, R. Zimmer, 2003. Playing biology's name game: identifying protein names in scientific text. Pacific Symposium on Biocomputing '03 2003.

He, Y., M. Kayaalp, 2006. A Comparison of 13 Tokenizers on MEDLINE. Bethesda, MD: The Lister Hill National Center for Biomedical Communications

Kim, JD., T. Ohta, Y. Tsuruoka, Y. Tateisi, N. Collier, 2004. Introduction to the Bio-Entity Task at JNLPBA. 2004.

McCallum, A., 2003. Efficiently Inducing Features of Conditional Random Fields. UAI-03 2003.

Settles, B., 2004. Biomedical Named Entity Recognition Using Conditional Random Fields and Rich Feature Sets. COLING 2004 International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA).

Sun, C., 2007. Rich features based conditional random fields for biological named entities recognition. Comput Biol Med, 37(9): 1327–1333.

Tanabe, L., WJ. Wilbur, 2002. Tagging gene and protein names in biomedical text. Bioinformatics, 18(8).

Tsuruoka, Y., Y. Tateishi and J.D. Kim, 2005. . De veloping a robust part-of-speech tagger for biomedical text. Panhellenic Conference on Informatics, Volos, Greece, November (11–13): 382–392.

Wallach, H., 2004. Conditional Random Fields: An Introduction. CIS, U of Pennsylvania.

Zhao, S., 2004. Named Entity Recognition in Biomedical Texts using an HMM Model. COLING 2004 International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA).

Zhou, G., J. Zhang, J. Su, D. Shen, C. Tan, 2004. Recognizing names in biomedical texts: a machine learning approach. Bioinformatics, 20: 1178-1190.