ISSN No. 0976-5697

Volume 6, No. 6, July - August 2015

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

Feature Detection & Classification Methods in Biomedical Named Entity Recognition Systems: A Review

Nishant Dubey
C-DAC, Mohali,
India

Iqbal Singh
C-DAC, Mohali,
India

Abstract: The Named Entity Recognition refers to identification of text from given samples and is most important & fundamental task in biomedical terms extraction. This field is very challenging in recent years. Its aim is to extract and classify the biomedical text terms like proteins, genes, DNA, RNA etc. which, in general, have complex structures and are difficult to recognize. This paper briefly defines Biomedical Named Entity Recognition. In this, the various methods for feature detection & classification like SVM, Neural Networks & K-nearest neighbour and along with various previous works has been discussed. Different NER features in context to identification and classification of named entities have also been reviewed. In which SVM function is used to increase efficiency of biomedical terms extraction process.

Keywords: BioNER, SVM,K-Nearest Neigbour, Neural Network, F-Score.

I. INTRODUCTION

With the increase in information in biomedical domain, there is a great demand for biomedical information extraction techniques. Recognising the entities such as RNAs, cells, DNAs etc. has become in biomedical knowledge discovery one of the important task. Though a lot of algorithms have been given for this purpose but NER [BIOMEDICAL NAMED ENTITY RECOGNITION] still remains a challenge and an area of active research[1], as still there is huge difference in F-score of 10 points between general newswire named entity recognition and biomedical named entity recognition. For biomedical NER it is more difficult in following ways:

a. ***Biomedical NEs –*** most types do not have a complete dictionary and new NEs are being created continuously.

b. Same phrase or word can point to different entities relying on their contexts. Biological NEs conversely have many spelling reform

c. Quite often before NEs modifiers are used and biomedical NEs are sometimes very long. These points marks the difficulties for NEs boundary identification.

d. NEs can be cascaded. Embedment of one NE can be done in another. For identification of these kinds of NEs more efforts must be made.

In biomedical domain abbreviations are used quite often. As there are not many evidences in abbreviation for some NE class, it becomes difficult to classify them rightly to face these problems , it is required to explore rich features and effective methods. In biomedical literature there has been many trials to develop techniques to identify NE. They roughly categorise into three approaches- dictionary based approach, statistical machine learning based approach and heuristic rule based approach. However techniques for biomedical NER don't gain satisfactory results. Problems propose that individual biomedical NER system might not involve entity representations with a lot of rich features and no algorithm of single type is practical to gain best performance.

To judge output quality of NER system, many measures have been given. One possibility is accuracy on token level [2] . it is facing two problems, the huge majority of tokens in real world text are not included in entity names as generally defined, leading to baseline [always predict not an entity] accuracy which is extravagantly high more than 90%; and wrongly predicting full span of entity name is not correctly penalized [finding a person's first name when last name follows is marked as ½ accuracy].

A variant of F1 in academic conferences has been defined below:

a. Firstly, Precision is the number of predicted entity name spans that line up exactly with spans in the gold standard evaluation data.

b. Similarly Recall is the number of names in the gold standard that appear at exactly the same location in the predictions.

c. F1 is combined meaning of above two.

It can be derived from above that any prediction has a wrong class if it misses a single token that is do not contribute to either recall or precision.

II. NAMED ENTITY RECOGNITION

NER [Named entity recognition] also called as entity chunking ,entity extraction and entity identification is a subtask of information extraction that classify and locate elements in texts in predefined entities like –names of organisations, locations, persons quantities expressions of time , percentages , monetary values etc [3]. Two modules of recognition are Named Entity Detection & classification.

a. ***Named Entity Detection:*** Named entity detection consists of two phases; Feature extraction and learning classification. Features are characteristic or descriptors attributes of words designed for algorithmic consumption**.** Feature selection is of utmost importance for the applications of statistical machine learning models. The main aim of selecting features is to find textual attributes that contribute to improving the recognition accuracy. In order to deal with the special phenomena in the biomedical texts, we make extensive use of a diverse set of features**,** including local features, external resources features and full text features**.** Here these types of features are described in detail and their effectiveness to the NER system is also discussed [4]. Local features are the immediate context of each word. Different types of features are there used for detection. They are:

b. ***Word Level Features:*** The character makeup of words is related to Word-level features. They specifically describe special characters. Numerical value, word case and punctuation.

a) ***Digit pattern:***

(a). Common word ending
(b). Functions over words
(c). Patterns and summarized patterns

c. ***List lookup features*****:** In NERC Lists are the privileged features. The term "list" Are quite often used interchangeably with terms like "lexicon" and "gazetteer". List inclusion is a route to show the relation "is a" (e.g., Paris is a city). It may appear to be quite obvious that if a word like (Paris) is a part of a list of cities then the probability of this word in beinga city in a given text, becomes high. However, the probability is almost never 1 (e.g., the probability of "Fast" to represent a company is low because of the common adjective "fast" that is much more frequent).

a) ***General dictionary:***

(a). Words that are typical of organization names
(b). On the list lookup techniques

d. ***Document and corpus features*****:** Document features are defined over both document structure and document content. Documents (corpora) large collections are also good sources of features.

(a). Multiple occurrences and multiple casing
(b). Entity co reference and alias
(c). Document meta-information
(d). Statistics for Multiword units

Feature extraction Supervised machine learning systems cannot be directly trained on a corpus annotated with named entities. The corpus is to be transformed into a collection of instances. Usually instances are generated for consecutive tokens excluding punctuation marks, sometimes punctuation marks are stored as part of tokens [5]. Since punctuation marks carry a lot of information about named entities in Lithuanian language and their loss would be harmful. Language independent features are very general based on the orthographic information directly available in the corpus, language dependent features resort to external resources such as special purpose grammatical tools (part-of-speech tagger, lemmatizer, and stemmer) or gazetteers.

e. **Classification in NER:** As the task of NERC has developed over the years and likewise has the applied methods. One major goal of the classification is to make training data available to machine learning systems. Named entities are unknown words because they cannot be looked up in any ordinary lexicon. To identify them in a machine learning scenario, a set of distinct features is needed to tell positive and negative examples apart.

Over the years, new methods from the machine learning field became more and more popular, leaving behind systems which use handcrafted rules. Machine learning techniques allow the automatic induction of rule-based systems or sequence labeling algorithms from allocated training data. This is achieved by analyzing the discriminative features of positive and negative examples. Similar cases and repetitions occurring in the data are merged into rules and hence gain abstraction over concrete examples. Three different types of learning methods can be distinguished by their requirements for the training data:

(a). Supervised learning
(b). Semi-supervised learning
(c). Unsupervised learning

Various classifiers do exist following are few of them:

f. ***SVM:*** Kernal function with SVM is a model which works well with a large range of problem sets. It's a binary classifier that can be extended to classification of multi-class by training a group of binary classifiers and using 'one vs. one' or 'one vs. all' to predict. This technique is quite powerful and performs best in non linear classification problems of wide range. In input features of small set it works well because it expand the features into higher dimension space., provided one also have training data of good size[or else over fit can happen] . SVM in dealing with huge number of training data is not scalable, so therefore logistic regression along with manually expanded feature set will be more pragmatic.

g. ***K Nearest Neighbor***: It is also called as instance based learning and not model based learning as it is not related to learning any model. Training process is only memorizing all training data. For predicting new data point, we find closest K [parameter which is tunable] neighbors from the array of training set and allowing them to vote for final prediction. To identify' nearest neighbors' it is necessary to define a distance function [e.g. Euclidean distance – common for numeric input variables]. Weighting of voting can also be done among K neighbors depending on its distance from new data point.

h. ***Neural Networks*****:** Apart from learning **multiple** outputs at same time, in learning non linear function it is also very good. Time of training is comparatively long and is also vulnerable to local minimum traps. This problem can be eliminated by picking best learned model and doing multiple rounds.

III. RELATED WORK

Lishuang et al. (2013) in the paper, "A Two-Phase Bio-NER System based on Integrated Classifier & Multi agent Strategy" proposed a two phase Bio-NER model. Their two-phase method separated the task into two subtasks: named entity classification (NEC) and named entity detection (NED). The NED subtask is achieved through two-layer stacking method in the first phase, where non named entities (NEs) are differentiated from named-entities (NNEs) in biomedical literatures. Six classifiers are made through four toolkits [Yam Cha, CRF++, Mallet, maximum entropy,) with separate training methods and are integrated based on the two-layer stacking method. In second phase for the NEC subtask a multi agent named strategy is given to produce the right entity type for entities matched in the first phase. Their experimental results tell that their method can gain 76.06 percent F-score that outperforms most of the state-of-the-art systems**.**

Jongwoo Kim (2013) in the paper, "Identification of Investigator Name Zones using SVM Classifiers and

Heuristic Rules" The research presented in biomedical articles quite often involves a lot of investigators at various institutions so that their names are brought up in the article. These Investigator Names (IN) now presents a needed field in the MEDLINE® citation for the given article. these names automated extraction is executed in a system produced by a research group at the U.S. National Library of Medicine, containing three modules dependant on Support Vector Machine (SVM) classifiers and heuristic rules. The SVM classifiers label text blocks ("zones") which perhaps contain Investigator Names and the heuristic rules identify the actual zones. They have gathered eleven sets list of words to test and train each set of classifiers containing 100 to 56,000 words. Experimental results done on online biomedical articles show a Recall, , F-measure, Precision ,Accuracy of, 0.95, 0.99 , 0.90 and 0.92 respectively.

Zhihua et al. (2012) in the paper, "Biomedical Named Entity Recognition Based on Skip-Chain CRFS" show a skip-chain conditional random fields (CRFs) model for BioNER. The model considers to the long-range dependencies about biomedical information. Such distant dependencies are powerful to identify some frequent appearing named entities and to classify them specifically for both classes protein and cell type. When they test the GENIA corpus, their approach obtains significant improvement over other methods, which achieves precision 72.8%, recall 73.6% and F-score 73.2%.

Chowdary et al. (2012) in the paper "Decision Tree Induction Approach for Data Classification Using Peano Count Trees" produced a new method for decision tree for classifying data using a data structure called Peano Count Tree (P-tree) that increases the scalability and efficiency. They apply Attribute Relevance techniques and data smoothing together with a classifier. Experimental results display that the P-tree method is quite faster than already existing classification methods and the preferred method for mining on data to be classified.

Kishana et al. (2012) in the paper "Performance Analysis for Visual Data Mining Classification Techniques of Decision Tree Ensemble and SOM" concerns on visual data mining applications in order to increase business decisions. The software based system is used as an intelligent and fully automated system that takes into effect every sales transaction. It modifies and updates forecasting statistics by receiving input through sales data directly and from sales counter by a networked connectivity. It may be wireless or wired. Three artificial intelligence tools: ensemble classifier, Self Organizing Maps (SOM) and efficiency decision tree are used for data analysis and data processing. The visual data mining concept is put forth by producing results in the way of visual interpretation in simple and possible way to understand complex statistics.

The present research results are matched with interactive visualization through multi- bar charts, multi level pie charts, multi, histograms,, tree maps dataflow diagrams and scatter plots. The separate visualization techniques are used in understanding various levels of information concealed in huge data sets. The results analysis depict that the predictions using SOM has accuracy of 90.0 %, up to a 86.0 % decision tree has classified data correctly and ensemble techniques produced an average of 88.0 % . The survey done after putting forth and use of the system depicts that the system is quite easy to comprehend and can be quickly interpreted with least efforts.

IV. CONCLUSION

In this paper, different BioNER System has been discussed where different feature detection techniques, classifiers and various previous works were under consideration. In recognition there are two phases named as detection and classification. A no. Of techniques were already used for feature detection where word, list or patterns types feature are used for further process. The best feature is to consider for better recognition are words & patterns concluded from the above studies. Also there are different classification methods for recognition phase but it concludes that SVM classifier works better on the above selected features. So, this paper concludes that biomedical named entity recognition system will achieve better results by selection of best features with classifiers.

V. REFERENCES

[1]. Lishuang Li, Wenting Fan, and Degen Huang "A Two-Phase Bio-NER System Based on Integrated Classifiers and Multiagent Strategy" IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol.10, No.4, pp. 897-904, 2013.

[2]. Jongwoo Kim*, Daniel X. Le, George R. Thoma "Identification of Investigator Name Zones using SVM Classifiers and Heuristic Rules" 12th International Conference on Document Analysis and Recognition (ICDAR), 2013, pp. 140-144.

[3]. Zhihua Liao "Biomedical Named Entity Recognition Based on Skip-Chain CRFS" International Conference on Industrial Control and Electronics Engineering (ICICEE), 2012, pp. 1495-1498.

[4]. B V Chowdary, Annapurna Gummadi, UNPG Raju,B Anuradha and Ravindra Changala "Decision Tree Induction Approach for Data Classification Using Peano Count Trees" International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 2, No. 4, pp. 475-479, April 2012

[5]. C. M. Velu, Kishana R. Kashwan "Performance Analysis for Visual Data Mining Classification Techniques of Decision Tree, Ensemble and SOM" International Journal of Computer Applications (0975 – 8887), Volume 5, No.22, pp. 65-71, November 2012.

[6]. Krishnalal G, S Babu Rengarajan and K G Srinivasagan "A New Text Mining Approach Based on HMM-SVM for Web News Classification" International Journal of Computer Applications (0975 - 8887) Volume 1, No. 19, pp. 98-104, 2010.

[7]. Nadeau, David, and Satoshi Sekine. "A survey of named entity recognition and classification." Lingvisticae Investigationes 30, no. 1, pp. 1-20, (2007).

[8]. Mónica Marrero, Julián Urbano, Sonia Sánchez-Cuadrado, Jorge Morato, Juan Miguel Gómez-Berbís "Named Entity Recognition: Fallacies, Challenges and Opportunities" Computer Standards & Interfaces Volume 35, Issue 5, September 2013, pp. 482–489.

[9]. Support Vector Machine http://in.mathworks.com/help/stats/support-vector-machines-svm.html