ISSN 1996-3343

Asian Journal of Applied Sciences

http://knowledgiascientific.com

Knowledgia SCIENTIFIC
A Place to Publish Outstanding Research

Asian Journal of Applied Sciences 8 (3): 210-216, 2015
ISSN 1996-3343 / DOI: 10.3923/ajaps.2015.210.216

A Hybrid Method of Linguistic Features and Clustering Approach for Identifying Biomedical Named Entities

E. Alharbi and S. Tiun
Center for Artificial Intelligence Technology (CAIT), Faculty of Information Science and Technology, University Kebangsaan Malaysia, Bangi, 43600, Selangor Darul Ehsan, Malaysia

Corresponding Author: E. Alharbi, Center for Artificial Intelligence Technology (CAIT), Faculty of Information Science and Technology, University Kebangsaan Malaysia, Bangi, 43600, Selangor Darul Ehsan, Malaysia

ABSTRACT

Named entity is a term that has been widely used in the field of Natural Language Processing (NLP). It contains the names of persons, organizations, locations, dates and currencies. The process of extracting such names called Named Entity Recognition (NER). Biomedical Named Entity Recognition (BNER) is one of the fields that contains variety of named entities such as genes, DNA, RNA, chemical compounds. The key characteristic behind BNER lies on selecting an appropriate method that has the ability to identify the named entities effectively. Each entity (e.g., DNA, RNA, drugs) has its own features which are different from the others. Recently, identifying chemical compounds have caught researchers' attentions due to the various types of entities that included. Many approaches have been proposed in terms of extracting chemical compounds however, most of these approaches depend on a supervised learning techniques where the class label is predefined. In fact, chemical compounds have tremendous types of entities which requires more analytical categorization. For instance, Tramadol and Aspirin are drugs but each of them belongs to different classes of drugs. Hence, investing the unsupervised learning techniques may enrich these classifications. This study attempts to address the role of unsupervised learning specifically clustering K-means approach combining with feature extraction method including Part-of-Speech (POS) tagging and affixes (suffixes and prefixes). The experimental results of the proposed method have demonstrated an enhancement by obtaining 90% of F-measure. It is concluded that future efforts may concentrate on utilizing more linguistic features which could improve effectiveness.

Key words: NER, BNER, POS tagging

INTRODUCTION

Named entity is a term that have been widely used in the field of Natural Language Processing (NLP) (Nadeau and Sekine, 2007). It contains the names of persons, organizations, locations, dates and currencies. The process of extracting such names called Named Entity Recognition (NER). Biomedical information has been expanded due to the huge number of resources such as articles, books and publications that generated every year. One of these resources is MEDLINE literature database which consists of 20 million references to journal papers that related to biomedical areas (Campos *et al*., 2012). The problem of extracting named entity is challenging nowadays due to its significant role within the text. In fact, it occupies most of the text domains by including names, locations, organizations, currencies and dates. This significance is increasing when dealing with domain like biological which contains necessary entities such as genes, proteins and DNA. Several

approaches have been proposed in terms of extracting chemical compounds however, most of these approaches depend on a supervised learning techniques where the class label is predefined. In fact, chemical compounds have tremendous types of entities which requires more analytical categorization. For example, drugs yield several information such as name of the drug, Code Company, the hierarchy of family and molecular formula (Campos *et al*., 2012). This could bring many challenges in terms of the variety and differences between those chemical entities. Therefore, selecting an appropriate approach for extracting named entities differs between different domains which can also be a challenging issue.

The earliest effort has been introduced toward drug names was performed by Rindflesch *et al*. (2000) which aimed to propose an approach called EDGAR (Extraction of Drugs, Genes and Relations). This approach extracts automatically drugs and genes that related to cancer using a biomedical database called MEDLINE (containing a huge number of publications and journal papers related to biomedical). In fact, authors have tend to figure out some relations between genes and drugs. This relation is represented by the effect of gene expression on the drug sensitivity of a cell. On other hand, drug treatment usually causes a variations in the cell's gene expression. Hence, the author have used a part-of-speech tagging in order to determine the actual tag for each word (e.g., verb, noun, adjective, etc.) and a semantic approach using Unified Medical Language System lexicon which provides syntactic and semantic information about the biomedical terms.

However, recent researchers tend to utilize machine learning techniques (most of them used supervised learning technique). For instance, Friedrich *et al*. (2006) have proposed several machine learning techniques with a dictionary-based approach in terms of identifying biomedical entities using the benchmark settings of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-2004) on the GENIA corpus. They have used Maximum Entropy Markov Model (MEMM), Support Vector Machine (SVM), Naïve Bayes (NB) and Conditional Random Fields (CRF) classifiers. On other hand, they have utilized multiple features such as capitalization, containing digits, containing delimiters and containing Roman or Greek letters. Furthermore, they have stemmed the words and used a part-of-speech tagger in order to assign each word with its exact tag. As a result, conditional random field outperforms the other classifiers in terms of the accuracy. Corbett *et al*. (2007) have prepared a set of guidelines and suggestion for chemical compounds and related entities recognition. Similarly, Degtyarenko *et al*. (2008) have built an ontology called ChEBI (Chemical Entities of Biological Interest) which contains terminologies, definition and hierarchy of chemical compounds. The use of this ontology has demonstrated a valuable rewards regarding to chemical compounds recognition.

Furthermore, Kolarik *et al*. (2008) have developed a system to use character-based n-grams, Maximum Entropy Markov Models (MEMM) and rescoring to recognize chemical names and other such entities. The authors have produced a set of annotation guidelines for chemical named entities and used them to annotate a set of 42 chemistry papers. Three groups of classifiers have been applied in order to recognize chemical names. The first classifier uses character-level n-grams to estimate the probabilities of whether tokens are chemical or not. The output of this classification is combined with information from the suffix of the word and is used to provide features for the MEMM. The second group of classifiers constitute the MEMM proper named entities by using a POS tagger which aims to provide tags for each word. The third group of classifiers aims to utilize a set of features for each entity. These features are derived from the probabilities of other entities that share the same text string as the entity, from probabilities of potential synonyms found via acronym matching and other processes and most importantly, from the pre-rescoring probability of the entities themselves.

Moreover, De Matos *et al*. (2009) have attempted to extract new feature from ChEBI corpus by making substantial improvements to the data quality of ChEBI via additional data annotations as well as extensions of the ChEBI ontology.

On other hand, Rocktaschel *et al*. (2012) have proposed a supervised machine learning technique using a hybrid system called ChemSpot for recognizing chemical and drugs named entities. The system is a combination of Conditional Random Field (CRF) with a dictionary in order to recognize chemical or drug names that formulated according to International Union of Pure and Applied Chemistry (IUPAC) names. The IUPAC contains information about the family of drug (e.g., Alcohol), company code (e.g., ICI304428) and molecular formulas (e.g., COOH). The author have used a training set of annotated entities with the use of a dictionary for the entities that are not annotated. Similarly, Lamurias *et al*. (2013) have proposed a supervised learning technique which is Conditional Random Filed (CRF) in terms of recognizing chemical compounds and drug names by using semantic similarity. The author have used ChEBI ontology which provides information about family name, chemical name and company code of the drug. Such information has been used as training examples for CRF classifiers.

MATERIALS AND METHODS

Figure 1 illustrates the research framework of this study. The corpus that have been used in this study is Scientific Computing and Algorithms Institute (SCAI) (Kolarik *et al*., 2008) which is specific sub-entity type corpus, containing only annotations of chemicals that follow; that

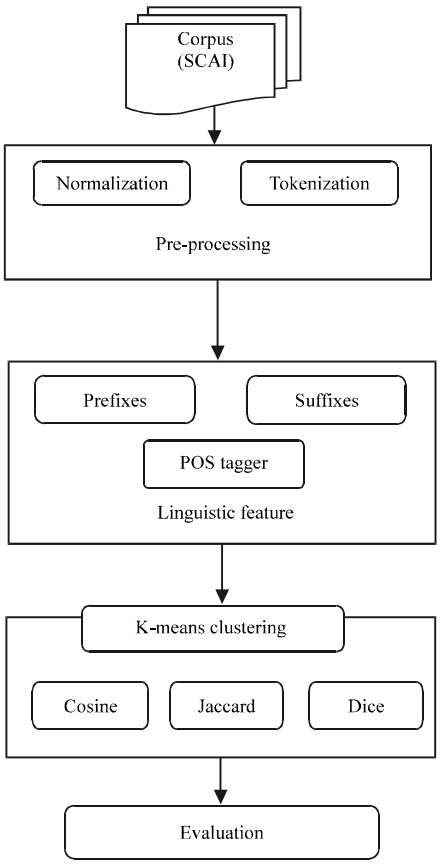

Fig. 1: Framework of study

formulated according to International Union of Pure and Applied Chemistry (IUPAC) nomenclature. Basically, several steps of pre-processing phase have been taken in order to turn the data into a suitable form that can be processed including normalization, tokenization and stemming. Normalization step performs an elimination for all the unnecessary data such as special characters, numbers and stop-words. Tokenization step performs a dividing task for the text into tokens of string. Stemming step perform a filtering task which aims to eliminate all the inflectional derivations. Porter stemmer (Porter, 2001) has been used in this study in order to retrieve the stem for each word.

After that, a linguistic features extraction phase is taking including Part-of-Speech (POS) tagger and affixes (prefixes and suffixes). The POS tagging is one of word sense disambiguation methods that aims to assign each word in certain text with a fixed set of parts of speech such as, noun, verb, adjective or adverb (Navigli, 2009). There are tremendous words that have several potential tags thus, POS have been came up in order to disambiguate these words. Therefore, the main role of POS is to determine the exact tagging for each words in the corpus. Whereas, affixes are consisting of prefixes and suffixes. Since the drugs' names usually contain a suffix or prefix (Table 1 and 2), the proposed method utilized affixes feature in order to recognize the drug and its family.

Hence, the results of previous phases (POS tagging and affixes) have been used in order to determine the initial centroids for the clustering approach. This process is to make sure that the chosen seeds are drugs' names. After that, clustering approach will be taken place by carrying out K-means clustering using three similarity measures in order to classify each drug with its similar clusters. K-means clustering technique requires a prior user defined for the number of clusters, then it selects a random centroids based on the defined number of cluster and then measure each point of the data with the centroids, the nearest will be joined (Agarwal *et al*., 2012). The proposed clustering technique relies on three parameters which are similarity measures (Dice, Cosine and Jaccard), number of clusters (k) and threshold (2). Such parameters are explained as follows.

Similarity measures: According to Cha (2007), the similarity measures are the token-based distance functions and the string representation as an m-dimensional vector of tokens with weights for instance, the string $s = \{"_1, "_2,..,"_m\}$ as a vector of tokens will be represented as a vector $s' = \{W1("_1), W_2("_2),..,W_m("_m)\}$ where $W_i("_i)$ is the number of occurrences of the token $("_i)$ in the string s. Hence, the cosine similarity between two strings s and r is defined as:

Table 1: Sample of prefixes in drugs

BioNEs	Prefixes
Cefepime	Cef
Cefluprenam	Cef
Cefrom	Cef
Cephems	Cef

Table 2: Sample of suffixes in drugs

BioNEs	Suffixes
Lymecycline	Cycline
Methacycline	Cycline
Minocycline	Cycline
Tetracycline	Cycline

$$\text{Cosine}(s, r) = \frac{s.r}{|s|.|r|} \quad (1)$$

While Dice is defined as:

$$\text{Dice}(s, r) = \frac{2 \times |r \cap s|}{|r| + |s|} \quad (2)$$

Eventually, Jaccard is defined as:

$$\text{Jaccard}(s, r) = \frac{|r \cap s|}{|r \cap s|} \quad (3)$$

Number of clusters (k): The number of clusters plays an essential role in terms of the effectiveness of the clustering method where this number has to cover all the data. Therefore, three number of clusters have been used in the experiments which are 3 clusters, 4 clusters and 5 clusters. The reason behind selecting such numbers is that the observation of the biomedical named entities that lie on the dataset has shown that 3, 4 or 5 kinds of suffixes or prefixes could fairly include all these named entities. Therefore, these number have been used in the experiments in terms of seeking the most accurate results.

Threshold (2): In fact, threshold referred as the required value of similarity between words to be included into specific cluster. Three numbers of threshold value have been selected which are 0.3, 0.4 and 0.5. The reason behind selecting such numbers is to address the performance of each similarity within these parameters.

RESULTS

As mention earlier, the clustering technique is depending on three parameters; similarity measures (Dice, Cosine and Jaccard), number of clusters (k) and threshold which are stated in Table 3, the results of the three similarity measures have been stated when k = 3, k = 4

Table 3: Results of clustering technique

Similarity measure	2 = 0.3	2 = 0.4	2 = 0.5
K = 3			
Cosine	0.84	0.86	0.88
Dice	0.79	0.82	0.83
Jaccard	0.77	0.83	0.84
K = 4			
Cosine	0.81	0.87	0.90
Dice	0.78	0.81	0.87
Jaccard	0.79	0.82	0.84
K = 5			
Cosine	0.82	0.85	0.86
Dice	0.81	0.85	0.89
Jaccard	0.80	0.84	0.86

Table 4: Comparison with other approaches

Study	Pattern	Method	Corpus	Accuracy (%)
Klinger *et al*. (2008)	Chemical names	Conditional Random Fields (CRF)	SCAI-IUPAC	85.6
Friedrich *et al*. (2006)	Chemical names	A hybrid method of feature extraction and CRF	SCAI-IUPAC	71.5
Usie *et al*. (2013)	Chemical names	Combination of CRF and regular expression trigger	SCAI-IUPAC	73.6
Rocktaschel *et al*. (2012)	Chemical names	CRF and dictionary-based	SCAI-IUPAC	69.0
Proposed method of this study	Chemical names	Hybrid of feature extraction and K-means clustering	SCAI-IUPAC	90.0

and k = 5 with β = 0.3, β = 0.4 and β = 0.5. These values are based on the common information retrieval metric F-measure. Basically, Dice and Jaccard have obtained their greatest values of F-measure when k = 5 and β = 0.5 by achieving 0.89 and 0.86, respectively. Unlikely, Cosine has achieved the highest values of F-measure when k = 4 and β = 0.5 by gaining 0.90 of F-measure. Obviously, Cosine has outperformed the other similarity measures in terms of the effectiveness.

DISCUSSION

In order to evaluate the proposed method of this study, several approaches have been brought for the purpose of comparisons. These approaches have used the same corpus that used in this study which is SCAI (IUPAC) with different method which is illustrated in Table 4.

As shown in Table 4, there are several approaches that have been proposed in terms of identifying chemical names. However, most of these approaches have been implemented via supervised machine learning. For instance, Klinger *et al*. (2008) have proposed a Conditional Random Fields (CRF) classifier in order to extract chemical names. The authors have obtained a good accuracy of approximately 86%. Whereas, the other approaches have combined CRF with different techniques such as dictionary-based, suffixes and prefixes, regular expressions and POS tagger. Since, classification has a predefined class label so that, the class label will be ranged between yes or no. Unlikely unsupervised learning has the ability to provide more categorizations. This is the main reason behind obtaining the highest accuracy by the proposed clustering technique.

CONCLUSION

This study aimed to propose a hybrid method of linguistic features and K-means clustering approach toward enhancing the effectiveness of identifying Biomedical named entities. The experimental results have reported a 90% of F-measure in terms of retrieving candidates. The future efforts may concentrate on utilizing more linguistic features which could improve the effectiveness.

REFERENCES

Agarwal, S., S. Yadav and K. Singh, 2012. K-means versus k-means++ clustering technique. Proceedings of the Students Conference on Engineering and Systems, March 16-18, 2012, Allahabad, Uttar Pradesh, pp: 1-6.

Campos, D., S. Matos and J.L. Oliveira, 2012. Biomedical Named Entity Recognition: A Survey of Machine-Learning Tools. In: Theory and Applications for Advanced Text Mining, Sakurai, S. (Ed.). Chapter 8, InTech Publisher, Reijika, Croatia, ISBN: 9789535108528, pp: 175-195.

Cha, S.H., 2007. Comprehensive survey on distance/similarity measures between probability density functions. Int. J. Math. Models Meth. Applied Sci., 1: 300-307.

Corbett, P., C. Batchelor and S. Teufel, 2007. Annotation of chemical named entities. Proceedings of the Workshop on BioNLP 2007: Biological, Translational and Clinical Language Processing, June 29, 2007, Association for Computational Linguistics, pp: 57-64.

De Matos, P., R. Alcantara, A. Dekker, M. Ennis and J. Hastings *et al*., 2009. Chemical entities of biological interest: An update. Nucleic Acids Res., 10.1093/nar/gkp886

Degtyarenko, K., P. de Matos, M. Ennis, J. Hastings and M. Zbinden *et al*., 2008. ChEBI: A database and ontology for chemical entities of biological interest. Nucleic Acids Res., 36: D344-D350.

Friedrich, C.M., T. Revillion, M. Hofmann and J. Fluck, 2006. Biomedical and chemical named entity recognition with conditional random fields: The advantage of dictionary features. Proceedings of the 2nd International Symposium on Semantic Mining in Biomedicine, Volume 7, April 9-12, 2006, Jena, Germany, pp: 85-89.

Klinger, R., C. Kolarik, J. Fluck, M. Hofmann-Apitius and C.M. Friedrich, 2008. Detection of IUPAC and IUPAC-like chemical names. Bioinformatics, 24: i268-i276.

Kolarik, C., R. Klinger, C.M. Friedrich, M. Hofmann-Apitius and J. Fluck, 2008. Chemical names: Terminological resources and corpora annotation. Proceedings of the 6th Edition of the Language Resources and Evaluation Conference, May 28-30, 2008, Marrakech.

Lamurias, A., T. Grego and F.M. Couto, 2013. Chemical compound and drug name recognition using CRFs and semantic similarity based on ChEBI. Proceedings of the 4th BioCreative Challenge Evaluation Workshop, October 8, 2013, Washington, DC., USA., pp: 75.

Nadeau, D. and S. Sekine, 2007. A survey of named entity recognition and classification. Lingvisticae Invest., 30: 3-26.

Navigli, R., 2009. Word sense disambiguation: A survey. ACM Comput. Surv., 41: 1-69.

Porter, M.F., 2001. Snowball: A language for stemming algorithms.http://www.snowball.tartarus .org/texts/introduction.html.

Rindflesch, T.C., L. Tanabe, J.N. Weinstein and L. Hunter, 2000. EDGAR: Extraction of drugs, genes and relations from the biomedical literature. Proceedings of the Pacific Symposium on Biocomputing, July 13-15, 2000, NIH Public Access, pp: 517-528.

Rocktaschel, T., M. Weidlich and U. Leser, 2012. ChemSpot: A hybrid system for chemical named entity recognition. Bioinformatics, 28: 1633-1640.

Usie, A., J. Cruz, J. Comas, F. Solson and R. Alves, 2013. A tool for the identification of chemical entities (CheNER-BioC). Proceedings of the BioCreative Challenge Evaluation Workshop, Volume 2, October 7-9, 2013, Bethesda, pp: 66.