Information Extraction of Diseases and its Application

[1]Mashmuma Qurban, [2]Dr. Syed Saif Ur Rehman

[1,2]Shaheed Zulfiqar Ali Bhutto Institute of Science and Technology, Karachi Pakistan

[1]mashmuma.qurban@yahoo.com
[2]saif.rahman@sse.habib.edu.pk

Abstract—Named Entity Recognition is an essential module of Information Extraction in the field of bio-medical and diseases are one of the most important sector to study in the medical field, but since the amount of incessantly updated information on diseases is huge and is merely accessible in the form of published journals or articles. An efficient Named Entity Recognizer is needed to extract diseases directly from the input given in the form of articles and to annotate the extracted terms with the knowledge base. The Named Entity Recognizer techniques must first identify the targeted terms. Though biomedical articles often consist of proper nouns recently prepared by the authors, and dictionaries which are conventional methods based on domain specific cannot identify such unidentified words. This study will identify a better and efficient Information extraction system which will extract diseases from the given biomedical text using techniques such as dictionary based and machine learning based (K-nearest neighbor and Naïve Bayes) techniques. The efficiency of both techniques and both algorithms have been measured through confusion matrix and machine learning approach more specifically K-nearest neighbor has been found more proficient as compared to other techniques.

***Keywords*—**Information Extraction, Disease, Text mining, Natural language processing, K-Nearest neighbor, Naïve Bayes, Machine learning, Dictionary based technique.

I. INTRODUCTION

Today we have a long list of diseases which cause harm to human beings. Data based on these diseases are available online in different structure, in different formats and on different sources. This makes those online published data, unfeasible in most of the time as well as in most situations. In the medical field, massive disease-related data are available. If these data can be extracted from different location and from different structure to one place with same structure and with one specified subject in focus, the data itself will be converted into a form which would be easy to understand and easy to refer. The more information about diseases stored in digital form, the more enhanced understanding would be created for disease, its causes, its environment, its symptoms and so forth. Discoverers, researchers, and scientist add content related to diseases to the web that is of an enormously varied in nature. Online disease information is developing closer to a real world knowledge base, but this online available information has one major issue/problem, which is related to its true interpretation and analysis of its context, which clearly indicates the need of logically accumulated information of disease [1].

Text mining focus on the recognition or discovery of patterns in natural language texts, same as data mining focus on the identification of patterns in databases. Applications of information processing can assist to access both types of information, unstructured information, like information that exists in documents or within databases as unstructured text fields and structured data which can be found in databases [2]. When retrieving this text format information, applications can also benefit from a more comprehensive linguistic analysis of the text, as compared to a narrower analysis based on words. There are extensive varieties of methods and techniques that can be used to study these natural language texts, as revealed in the substantial amount of researches that are done in the area of natural language processing [3], some of these techniques are categorizing content, extracting entities, clustering content, relationship extraction and fact extraction [4]. In this paper, we are using named entity recognition method of text mining to extract diseases which are our desired entity to extract from medical data in textual format.

NER is an abbreviation of Named Entity Recognition, it alludes all of the computational strategies that automatically identify named entities in natural language documents, for example, to relate it will be named entity in the desired domain. For biomedical field, a named entity is well-defined as a term consists of a single word or a phrase of multiple words that indicates a biomedical entity, for example, a disease, gene, protein or drug with which a semantic hierarchy is linked [5]. Named entity recognition of text mining in the biomedical field is most challenging. It is

evinced by the fact that several aliases, code-named, different naming conventions, acronyms, multiple naming conventions may discuss the same disease with diverse and changed terms, or a word may point to diverse, biologically changed entities. To deal with those problems different approaches and tactics have been applied on named entity recognizer using machine learning based techniques, rule-based and dictionary based on matching strings. With the speedy growth of biomedical texts which are being issued in thousands of journals, many spelling deviations of prevailing entity and new terms have been developed.

For such entities the dictionary based and the rule-based approaches lacks analysis and prediction power. Machine learning based methods, on the other side, have been proven as the most vigorous method for biomedical named entity recognizer due to its competence in dealing with high-dimensional discriminative vector structures in the text processing field and forecast of new entities or differences based on learned patterns [6]. To train a more reliable, a high performance and more accurate named entity recognizer model, it is essential to completely imprisonment structures surrounding the term in the background. In the past few years in the area of biomedical named entity recognition, many methods have been established that use semantic characteristics of the term like word lemmatization and word stemming, formation of the term such as is term in upper case or any symbols or any digits are used in the term which is also known as orthographic features, the suffixes and prefixes of words, char n-grams, and term shape which comes under morphological features, and local framework features [5]. Some methods are also assimilated with rigorous dictionary matching to identify named entities in a field specific wordlist. The binary encryption of the feature set is used as a response for the machine learning algorithm to train the named entity recognition model, along with the human explanation of named entity indications in the training dataset. For the past few years, much consideration has been given on named entity recognition of proteins and genes, while less work has been done on disease named entity recognition. In this paper, we compare a named entity recognition methods and machine learning algorithms to extract diseases and related information.

This paper is organized as follows. Section I describes the related concepts which are essential to execute text mining procedure in the medical domain and a concise introduction to the named entity recognizer. Section II describes the related work done in the same domain (Biomedical). Section III introduces the experiment performed to extract diseases and its application. In the end, in section VI and V experiment, the result and the conclusion are given.

II. LITERATURE REVIEW

To extract information on dealings with each article directly, the method must primarily recognize material names, for example, disease name, protein names, gene name and desire targeted word. Classifying desire entities from numbers of unrestricted and unstructured texts is a perplexing task for natural language processing, particularly in biological and medical articles. One will come across the following technical hitches: unfamiliar term processing, long, complex word identification, and the desire of robustness contrary to vague terminologies that are used only among in the expert's region. As far as it is identified, that there is no system which can extract technical terms with handling these issues at once [7].

A typical technique to identify technical words and proper nouns in textual data is to compare each term in the heading terms on prepared word lists. Though, biological and medical data usually have proper nouns recently prepared by the writers. Hence, one cannot ignore the chances to come across unidentified words, and adding such newly created and unidentified words to the vocabulary for future opportuneness is tremendously time-consuming and may cause many mistakes. Additionally, when the term expression in the text is changed from the glossary words, this conformist technique is ineffective and doesn’t provide desired results [5].

A. Information Extraction

Information Extraction (IE), is the technique which automatically extracts meaningful data related to the targeted category from natural language text. According to Russell and Norvig, it targets to process text and retrieve existences of a specific class of items or actions and existences of associations between them. Riloff gives the related opinion and states that information extraction is a method of processing natural language in which specific sort of information/ data must be predictable and dig out from the text [8]. Following can be considered as an example of an information extraction system, consider a method that develops a set of web pages and abstract information about countries and their administrative, financial and community pointers can be given as its related information. Some type of model that identifies what to search, for example population, cities, country, capital or any other term, this search is needed to director this procedure. The system will try to recall that information which will be matched according to this model and will ignore other types of information.

Russell and Norvig more stated that information extraction comes in midway among text understanding methods that are often referred to as text parsers that try to examine the text and dig out their semantic substances and information retrieval methods, information retrieval methods

simply find papers that are associated with the user's desires [8]. Many productive structures have been produced using studies made in information retrieval, for example, web-based search engines, although text understanding systems have not played any important role to contribute something very useful and successful. Meanwhile, the effort linked with information extraction methods comes under these two groupings; their accomplishment has also been somewhere between the heights accomplished by information retrieval and text understanding methods.

B. Ontology-Based Information Extraction

Ontology-based Information Extraction has just appeared as a sub-field of information extraction. At this point, the information extraction process uses ontologies as an input and the result is usually presented through ontology. It must be noted that ontology is well-defined as a formal and clearly specification of a public conceptualization. Generally, ontology is specified for a particular domain. Since information extraction is principally concerned with the task of repossessing information for a specific area, properly and unambiguously identifying the notions of this area through ontology can be supportive of this process. For example, a Diseases ontology that describes the concepts like Diseases, symptoms, and related genes can be used to monitor the information extraction method that was defined previously. This is the over-all idea behind information extraction base on ontology. It gives the impression that the term information extraction based on ontology has been considered only a few years ago. Nevertheless, there has been more or less working associated with this area before that, for example, Hwang presented a work which was published in the year 1999, he constructed ontologies using text. In recent times, there have been many publications that define Ontology-based information extraction system and even on this topic, many workshops have been arranged.

C. Natural Language Processing

Natural language processing contracts with the program processing and study of unstructured written information. One way of natural language processing investigation depends on numerical methods, usually connecting the dealing out of words set up in texts. Another methodology makes use of rule-based methods, leveraging data assets such as linguistic, ontologies and taxonomies rule bases [5]. Numerical human language processing methods call for groups of preparation material which demonstrate the necessary or unwanted associations and dependencies. Consequent alteration of the method, then needs some degree of reinstruction of the system.

As an alternative to reinstruction training material, rule-based methods wants information in the form of online wordlists, recognized linguistic concepts, and they are able to control present grouping methods or taxonomic contexts. Natural language processing could create use of either or both of these methods, and the choice of which method to use is often reliant on the accessibility of training materials, external properties, and the actual text analysis tasks required in the resulting application [3].

Numerous text mining methods have been carried out for biomedical named entity recognizer tasks using diverse methodologies. Those methodologies, in the precipitate, can be classified into following three categories.

1) Dictionary based method: Outdated information mining methods were initially built as a pattern, as a dictionary, and then utilize dictionary to mine required data or information from the original untagged writing. These mining methods are known as dictionary based method as well as called pattern based methods. The main point in this method is how to acquire the wordlist of patterns that can be used to classify the significant and relevant information from a text. AutoSlog was the first method to study text mining using wordlist from training examples [9]. It is the most direct and straightforward method that attempts to discover all named entities from textual data by looking up and matching words from the word list. Some terminologies have been comprehensively applied in the field of medical text mining. Contrasting to machine learning based method, one main benefit of the dictionary based method is that it has a peripheral database identifier which is also known as ID incorporated for each record, thus offers an external metadata explanation to the mined items [5]. Though, it agonizes from numerous restrictions together with, name abstruseness creates false positive, spelling deviations, and alternative word create false negative, and incapability to concealment newly generated items. In addition, it profoundly is influenced by the formation and interval of wordlist for the specific field, which may comprise of masses of records and is a very manual labor exhaustive. To deal with aforesaid spelling deviation concern, Tsuruoka et.al used estimated string examining and alternative originator techniques to accomplish a noteworthy improvement of F-measure which was 10.8% on GENIA corpora assessment as compared with rigorous or strict equivalent procedures [10].

2) Rule-based approach: it can deal in a better way with word morphological and orthographic configurations, as compare to vocabulary/ dictionary based method. In this technique by means of exterior evidence of character strings was presented to classify essential terms trailed by

handcrafted designs and rules to concatenate end-to-end words as named entity. The rule-based method principally is influenced by on the field precise named entities with public morphological or orthographic features. Thus creates it problematic to spread to other fields, meanwhile the handcrafted instructions are often field explicit and cannot be applied to a new field due to diverse naming resolutions [5].

3) Machine learning methods: the use of machine learning approaches in Information Extraction is primarily dedicated to the automatic attainment of the extraction patterns. These patterns are used to mine the information related to a particular task from every single text of a given collection [6]. Machine learning is a most frequently used and has accomplished the best performance in Bio-Creative II gene and protein named entity recognizer tasks. Diverse supervised machine learning approaches have been utilized in named entity recognizer methods. Furthermore to supervise approaches that consume only the marked text corpora, permissible to resolve data scantiness issues which often stumbles upon when using large feature set on a comparatively small training dataset, some semi-supervised methods are also accessible recently to take benefit of the large size of un-annotated text corpora. One precarious phase of machine learning method is to select the utmost discriminative feature established that signify the named entity. Frequently castoff features comprise of part of speech tagging, orthographically term creation designs, lemmatization, token window, morphological patterns and combination of relative features [5]. In this paper, we have chosen Naïve Bayes and KNN algorithms for disease classification because these algorithms perform well in the medical domain for text classification [11].

III. EXPERIMENT

A. *Data*

The data are in the text format which has an overall size of 3GB which consist of 90,643 PubMed articles. Along with biomedical domain articles (published in PubMed), Human Disease Ontology is also provided as input to perform this experiment which serves as a knowledge source and using this ontology I have mapped URI after extracting diseases from an article in dictionary based technique. To train machine learning algorithms for machine learning NER approach I have trained the model by splitting set into 70 percent for training and 30 percent for testing.

B. *Text Processing*

Data Pre-Processing Stage

- Read Articles
- Parsed those articles taken as input.
- Remove punctuation from the input.
- Remove N chars from input (I have taken 3 N char).
- Remove Numbers from the input.
- Remove Stop words from the articles taken as input and apply Porter Stemmer on them.
- Break Articles into sentences than further into words

C. *Applying Dictionary based approach*

I have used KNIME for performing this approach. First of all, I have created a dictionary based on Human Diseases ontology with the help of SPARQL query shown in figure 1 which produces a dataset shown in figure 2, the main advantage of using an ontology is that we don't need to update our dictionary manually whenever a new term has been introduced. For implementing this approach first we need to read the dataset and then parse it, I have perform 4 basic operations on input which are:

- Parsing
- Enrichment
- Transformation
- Preprocessing

First, I have parsed all the articles taken as an input applied pre-processing step (filter out punctuation, N chars (3 N char), numbers, stop words and porter stemmer. Then further processed set enriched by tagging desire terms which are diseases on them after this transformation process was applied extracted sentences from articles converted them into a bag of words , converted tag into a string and extracted terms into words. After applying all these processes the extracted terms are mapped on to the knowledge base which was extracted using ontology as shown in figure 3.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX void: <http://rdfs.org/ns/void#>
PREFIX sio: <http://semanticscience.org/resource/>
PREFIX ncit: <http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#>
PREFIX up: <http://purl.uniprot.org/core/>
PREFIX dcat: <http://www.w3.org/ns/dcat#>
PREFIX dctypes: <http://purl.org/dc/dcmitype/>
PREFIX wi: <http://http://purl.org/ontology/wi/core#>
PREFIX eco: <http://http://purl.obolibrary.org/obo/eco.owl#>
PREFIX prov: <http://http://http://www.w3.org/ns/prov#>
PREFIX pav: <http://http://http://purl.org/pav/>
PREFIX obo: <http://purl.obolibrary.org/obo/>

SELECT DISTINCT ?gda
    ?disease
    ?label
    ?title
    ?id
FROM <http://rdf.disgenet.org>
WHERE {
        ?gda sio:SIO_000628 ?disease.
        ?disease rdf:type ncit:C7057.
        ?disease rdfs:label ?label.
        ?disease dcterms:title ?title.
        ?disease dcterms:identifier ?id
filter regex(?disease, "umls/id").
}
```

Fig. (1). SPARQL Query to extract Diseases Information [12]

Fig. (2). Results from SPARQL Query

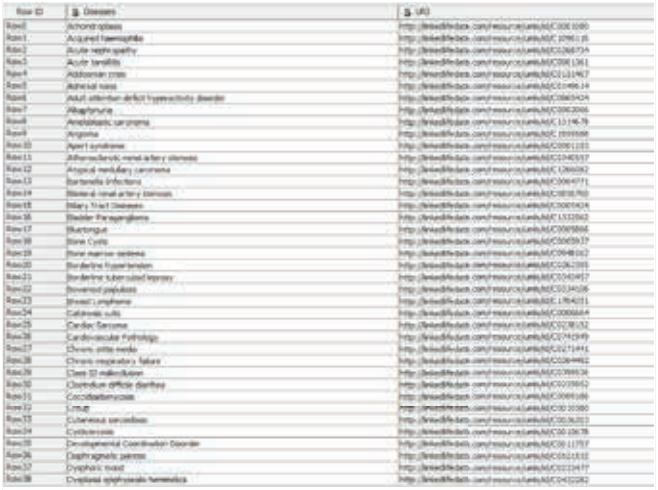

Fig. (3). List of extracted diseases and their respective URI

D. Machine Learning Approach

1) Applying Naïve Bayes Model

Naïve Bayes model is one of the most proficient and effective classification models, it works on the basis of probability and Bayes algorithm with strong independence assumptions [11]. It computes the number of rows for each feature value for each group for minimal features and the Gaussian distribution for statistics features. The generated model is used in the naive Bayes forecaster to predict the group association of uncategorized data. The model is used to categorize disease data with unidentified target feature, then this model is connected to its predictor. The predictor forecasts the group for each row grounded on the educated model. The group possibility is the output of the probability for each feature and the probability of the group feature itself. The probability for insignificant values is the number of incidences of the group / class value with the assumed value divided by the number of total incidences of the group / class value. The probability of statistical values is calculated by supposing an ordinary distribution for each feature and creates output/ result as the input table given with the addition of one column comprising the grouping, classification and the possibilities provisional on the decisions.

2) Applying K-Nearest Neighbor Model

Categorizes a set of test records grounded on the k nearest neighbor algorithm using the training data. The fundamental algorithm uses a KD tree and ought to, consequently, demonstrate rational performance. K nearest neighbors is an algorithm is a simple algorithm which saves entirely accessible cases and categorizes new and different cases built on a similarity measure for example distance functions. K-nearest neighbor has been used in numerical approximation and pattern recognition previously in the starting of 1970's as a non-parametric system.

E. Applications

There are many applications related to the bio-medical field in the domain of text mining. I have highlighted and implement few applications that are related to this paper, i.e. related to extracting diseases terms.

1) Tag Cloud

A tag cloud is an illustration of terms representing the significance of these terms by controlling the graphical properties. Here we use the distribution of transparency , font size, and boldness of the font. To improve the variation of the tag cloud the colors of the words are organized. I have used Blue and purple color for diseases and red for other terms. The result of this application is shown in figure 4.

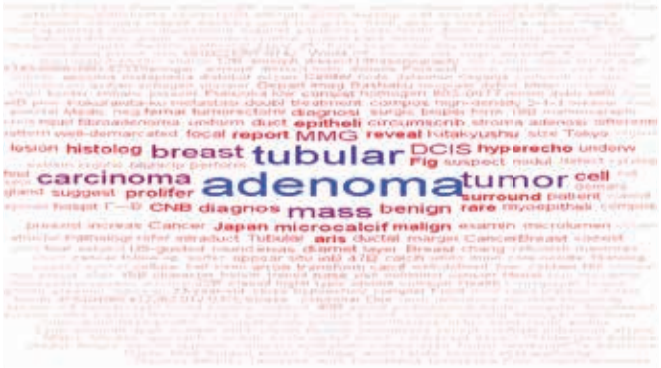

Fig. (4). Diseases Tag Cloud

2) Extract Most Used Diseases words from the List of Articles

By using Term Frequency technique we can extract most used diseases from the list of articles given as input. Analyze the term frequency of each term conferring to each article and adds a column enclosing the value of term frequency. The value is calculated by dividing the absolute frequency of a term rendering to a text by the number of all terms of that document.

3) Disease Gene Association

In this application, I had built up the association between the extracted diseases and genes from the biomedical articles. First, I have extracted all the diseases in the article along with this I have also extracted all the genes presented in the given input and have tagged them with the help of ontology after

extracting desired terms an associated URI have been given in the output with those terms which describe the relation between these terms in detail.

IV.RESULTS

Dictionary based and machine learning techniques were chosen for classifying diseases task , for Machine learning technique numerous algorithms can be applied, but in this study, I have chosen two of them which are K-nearest neighbor and Naïve Bayes because these two algorithms result in a pretty much better efficiency which is well-defined under a confusion matrix.

The confusion matrix of Dictionary based technique in table 1, tells us that it correctly identify 70 numbers of diseases and can identify 100 words as non-diseases words. The model incorrectly identifies 0 words of diseases and 34 words of non-diseases words, the model outputted 74 % of accuracy.

Table 1. Confusion matrix of Dictionary based technique

	True Positive	**False Positive**	**True Negative**	**False Negative**
Other	100	34	70	0
Diseases	70	0	100	34

Now I will discuss the results for machine learning approach and for this we will discuss the results of two algorithms which I have taken earlier.

The confusion matrix of Naïve Bayes in figure 5, tells us that it correctly identify 2985 numbers of diseases and can identifies 159 words as non-diseases words. The model incorrectly identify 341 words of diseases and words of non-diseases words, the model outputted 89% of accuracy, it have F-measure value.

Row ID	True Positive	False Positive	True Negative	False Negative	Recall	Precision	Sensitivity	Specifity	F-measure	Accuracy
Other	159	15	2985	341	0.318	0.914	0.318	0.995	0.472	
Diseases	2985	341	159	15	0.995	0.897	0.995	0.318	0.944	
Overall										0.898

Fig. (5). Confusion matrix of Naïve Bayes

The confusion matrix of k-nearest neighbour in figure 6, tells us that it correctly identify 3000 numbers of diseases and can identify 189 words as non-diseases words. The model incorrectly identifies 311 words of diseases and 0 words of non-diseases words, the model outputted 91 % of accuracy, and it have F-measure value

Row ID	True Positive	False Positive	True Negative	False Negative	Recall	Precision	Sensitivity	Specifity	F-measure	Accuracy
Other	189	0	3000	311	0.378	1	0.378	1	0.549	
Diseases	3000	311	189	0	1	0.906	1	0.378	0.951	
Overall										0.911

Fig. (6). Confusion matrix of K-Nearest Neighbor

Using the above results, it can be said that machine learning approach and more specifically K-nearest neighbour technique is more efficient to extract information related to diseases.

V. CONCLUSION

From the above results, we can deduce that the machine learning technique is efficient as compared to the Dictionary based technique and in machine learning specifically K-nearest neighbour. However, we can further increase the accuracy and performance of dictionary based by making our dictionary more comprehensive by including all synonyms related to diseases; add all the spellings available for diseases. Moreover, the model takes lots of time to be trained and RAM requirements would be greater than 8 GB. By increasing the training and testing data (input data) the error rate would be decreased subsequently but the time constraint will increase, i.e. it will take more time to train the algorithms and to pre-process dataset. In the proposed approach for machine learning technique the 70 - 30 split, 70 for training and 30 for testing have been applied.

VI. FUTURE WORK

In future different classification algorithm for machine learning technique such as Conditional Random Field, Maximum Entropy Markov, and Hidden Markov can be used. Moreover, in the future more focus can be placed on its application like in the temporal-spatial field, Linking Drugs or diseases synonyms to diseases and more.

REFERENCES

[1] P. C. Abeysiriwardana and S. R. Kodituwakku, "Ontology Based Information Extraction for Disease Intelligence," *International Journal of Research in Computer Science*, vol. 2, no. 6, pp: 7-19, 2012. DOI: 10.7815/ijorcs.26.2012.051

[2] I. Spasić, J. Livsey, J. A. Keane and G. Nenadić, "Text mining of cancer-related information: review of current status and future directions," *International Journal of Medical Informatics*, vol. 83, no. 9, pp: 605-623, 2014. DOI: 10.1016/j.ijmedinf.2014.06.009

[3] F. Popowich, "Using text mining and natural language processing for health care claims processing". *ACM SIGKDD Explorations Newsletter-Natural language processing and text mining*, vol. 7 no.1, pp: 59-66, 2005. DOI: 10.1145/1089815.1089824

[4] *Search Technologies Part of Accenture* [Online]. Available:https://www.searchtechnologies.com/blog/natural-language-processing-techniques

[5] Z. Huang and X. Hu, "Disease Named Entity Recognition by Machine Learning Using Semantic Type of Metathesaurus," *International Journal of Machine Learning and Computing*, vol. 3, no. 6, pp: 494, 2013. DOI: 10.7763/IJMLC.2013.V3.367

[6] A. Téllez-Valero, M. Montes-y-Gómez and L. Villasenor-Pineda. "A Machine Learning Approach to Information Extraction," in *Computational Linguistics and Intelligent Text Processing. CICLing* (Lec. N. Com. Sci.), A. Gelbukh, Eds. Springer, Berlin, Heidelberg, 2005, vol. 3406, pp. 539-547. DOI: 10.1007/978-3-540-30586-6_58

[7] K. Fukuda, T. Tsunoda, A. Tamura and T. Takagi, "Toward Information Extraction: Identifying protein names from biological papers," *International Journal of Computational Intelligence*, vol. 4, 2008.

[8] C. V. Monllaó, "Ontology-Based Information Extraction." PhD dissertation Thesis, Polytechnic University of Catalunya, 2011.

[9] J. Tang, M. Hong, D. Zhang, B. Liang, and J. Li, "Information extraction: Methodologies and applications." *Emerging Technologies of Text Mining: Techniques and Applications*, 2007.

[10] Y. Tsuruoka and J. Tsujii, "Improving the performance of dictionary-based approaches in protein name recognition." *Journal of biomedical informatics*, vol. 37, no. 6, pp: 461-470, 2004. DOI: 10.1016/j.jbi.2004.08.003

[11] K. M. Al-Aidaroos, A. A. Bakar and Z. Othman, "Medical data classification with Naive Bayes approach." *Information Technology Journal*, vol. 11, no. 9 p: 1166-1174, 2012. DOI: 10.3923/itj.2012.1166.1174

[12] *Disgenet* [Online]. Available: http://rdf.disgenet.org