ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

Constructing Synonymous Gene Database for Medline Abstracts Using SSFPOA

S.Jayaprada[1] , P.Bala Krishna Prasad[2] , I.Ramesh Babu[3]

[1] Sr. Assistant Professor, Dept. of CSE, VR Siddhartha Engineering College, (Autonomous) Vijayawada,India

[2]Principal, Dept. of CSE, Eluru College of Engineering and Technology, Eluru, India

[3]Professor, Dept. of CSE, Acharya Nagarjuna University, Guntur, India

ABSTRACT: Authors frequently use different names to refer to the same gene or protein names across Bio-medical articles. Identifying the alternate names for the same gene/protein would help biologists in the process of gene-protein interactions and protein-protein interactions. Biomedical databases such as SWISSPROT, GenBank, GOLD, UniGene and Karyn's Genome include synonyms, but these databases may not be always up-to-date. Therefore, it is necessary to automate this process, because of the increasing number of discovered genes and proteins. In this paper we considered this problem as Natural Language processing (NLP) problem and solved using SSFPOA semantic measure. Experiments were conducted on Medline abstracts and results are compared with existing methods. Machine learning algorithms are used in our work to analyze the performance of our method. Results are evaluated with the help of performance measures and results showed high percentage of accuracy when compared with existing works.

KEYWORDS: Information Extraction (IE), Gene name, Protein name, Medline abstracts, Synonym Identification

I. INTRODUCTION

The existing MEDLINE database includes over 12 million computer-readable records within the biomedical domain and is expanding rapidly. Automatically extracting synonymous gene/protein names from these biomedical text documents requires the knowledge of IE techniques such as recognizing the gene/proteins names in the text because human genes/proteins may be named with standard English words, Names may be alphanumeric, may include Greek or Latin letters, can be case sensitive, and may be composed of multiple words. Hence there is a need for an information extraction algorithm that can extract gene/ protein names accurately. Also this task requires the knowledge of NLP techniques such as stop words and duplicates words removal in order to make the identification task simpler. Next task is named entity recognition task by ABNER tagger. Finally similarity measure SSFPOA [1] is used to calculate similarity between any two gene/protein names present in Medline abstracts so as to update the database. This paper is organized as follows: Section II discusses related work, Section III discusses about our proposed system. Section IV discusses about Evaluation of Experimental results and Section V describes Conclusions and Future work.

II. RELATED WORK

Many approaches were proposed for constructing synonymous database such as rule based, machine learning and statistical techniques. Dictionary-based methods uses predefined terminological resources and various string matching approaches to locate gene within the text. In rule-based approach several rules are defined to extract gene name. Statistical approach requires bulk training corpora for doing the same job. Rule based approach such as [2] uses combination of matchers such as exact, exact-like and token-based approximate matching during the pre-processing phase. A number of token based transformation rules are used iteratively to map semantically equivalent or related tokens until no new forms are generated. Problem with this method is limited usage of approximation matchers. Also it does not use synonym dictionary for resolving ambiguous gene/protein names. Our proposed approach uses SSFPOA which is a compound similarity measure that uses 12 matchers varying from exact to approximation and domain dependent to domain independent with various [3] proposed a Named Entity Recognition technique, TaxonGrab, which is based upon some nomenclature rules (comprising linguistic and syntactic properties of taxonomic names) that are used for taxonomic nomenclature in scientific publications. Their work is based on the fact that "organism nomenclature conforms closely to prescribed rules". This work cannot be applied to all types of biomedical resources

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

as their rules extracts taxonomic names that follow a set of prescribed syntax and linguistic rules. In our proposed work, we considered Medline abstracts. [4] assess the performance of text mining systems applied to biomedical texts including tools which recognize named entities such as genes and proteins, and tools which automatically extract protein annotations [5] The ProMiner system is a rule based system that uses a pre-processed synonym dictionary (case –sensitive) to identify potential name occurrences in the biomedical text and associate protein and gene database identifiers with the detected matches. Based on all detected synonyms for one abstract, the most plausible database identifiers are associated with the text. Porter Stemmer algorithm is used for stemming in pre processing phase. Gene Ontology is used to resolve disambiguation. In addition to this the system uses an abbreviation dictionary containing abbreviations and their long forms, which do not correspond to protein or gene names. If match overlaps, the match with higher acceptance score is considered. Even though Pro Miner system could get F-Measure of 0.83, but detection of synonyms can be evaluated using machine learning methods. Identifying gene/protein names (Named entity Recognition- NER) for the mouse and yeast organisms is done using simple matching procedure. This approach may not work for organisms with high prevalence of unspecific names, such as fly. In our proposed work, we considered ABNER for tagging the gene/protein names. [6] GAPSCORE identifies protein and gene names from text. It uses a word-based approach and scores the confidence that a word may be a gene based on appearance, morphology and context criteria that includes information from all of MEDLINE. To identify the boundaries of multi-word gene names, GAPSCORE extends the name using heuristics on part of speech tags.

The algorithm consists of five steps: (1) TOKENIZE to split the document into sentences and words; (2) FILTER to remove from consideration any word that is clearly not a gene name;(3) SCORE to score words using a machine learning classifier;(4) EXTEND to extend each word to the full gene name; and (5) MATCH ABBREVIATION: to score abbreviations of the gene names identified. But this work does not handle ambiguous names. Another rule based system is proposed by [7] that combines morphological cues, functional keywords, and position functional keywords to filter non-gene/protein terms. Extraction is specified with rules to abbreviations and full names. When a string is mapped to several terms, then the rule is to prefer longer term mapping. One limitation of this work is that if gene/ protein symbols and full names are defined in the abstract only then its performance increases otherwise it cannot capture gene/protein symbols and full names. [8] proposed an algorithm for extraction of abbreviations from biomedical Medline abstracts. First it extracts text of the form pair candidates. Then it identifies the correct long form from among the candidates in the sentence that surrounds the short form such as adjacency to parentheses (. This algorithm can be improved by using syntactic information during pre processing step. One more drawback of this algorithm is it identifies abbreviation only when the definition is enclosed in parentheses. In our proposed approach we used certain words such as "known as", "also called", "also known as" and parentheses to recognize synonymous gene/protein name. [9] proposed a method for tagging gene and protein names in biomedical text using a combination of statistical and knowledge based strategies. Rules of this method use Part-ofspeech (POS) tagger, morphological clues, trigrams, suffixes. Errors were introduced in this method because of discovery of gene/protein names. [10] Developed SGPE (for synonym extraction of gene and protein names), a software program that recognizes patterns and extracts synonymous terms from MEDLINE abstracts. SGPE then applies a sequence of filters to remove unwanted gene and protein names with the help of pre-fetched synonymous gene/protein names from the SWISSPROT databank. Our method is different from SGPE, where we consider synonyms basing on NLP. Problem with SGPE is that it relies on authors to list synonymous gene and protein names in the literature. But all the authors may not list synonymous gene and protein names and hence the extraction method is not complete. [11] this method uses pattern-based abbreviation rules in addition to text markers and cue words for finding abbreviations. The pattern-based rules describe how abbreviations are formed from definitions. Rules can be generated automatically and/or manually and can be augmented when the system processes new documents. [12] proposed a rule based approach for development of gene and protein names dictionary. In the first phase, the Medline abstracts are processed by set of functions such as tokenization, filtering and stemming to make the extraction process simpler. In the second phase, the set of rules are used to identify and extract gene and protein names from preprocessed Medline abstracts and subsequently updates the created dictionary. The third phase verifies and validates the performance and efficiency of the created dictionary by using precision, recall and F-measure metrics. Our proposed approach is a hybrid approach that uses both dictionary and rule based approaches. In the first phase, preprocessing is carried out to remove the inconsistencies from the dataset. In the second phase, the Gene and Protein names are extracted from Medline abstracts using regular

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

expressions and added to dictionary and in third phase the extracted gene and protein names are validated and verified using precision, recall and F-measure.

III. PROPOSED METHOD

Figure 1 presents our proposed system for identifying Synonymous Gene. The framework consists of three phases:

- Pre-processing
- Synonyms identification using semantic similarity measure SSFPOA
- Constructing/ Updating Gene database

A. Preprocessing: Pre-processing is a technique to retrieve gene or protein terms from Biomedical documents. The following are the Natural language techniques used in preprocessing:

1. Tokenization Tokenization is the process of identifying various elements (tokens) from the given text (Medline abstracts). These tokens are given as input for the next phases. In our approach we used the heuristic rules proposed by [14] to remove nonfunctional characters and the following heuristic rules during tokenization. do not split on hyphens, do not split on single quotation marks, do not split on commas, and do not split on parentheses and brackets. Example: Lymphocyte associated receptor of death. |Lymphocyte| |associated| |receptor| |of| |death| |.|

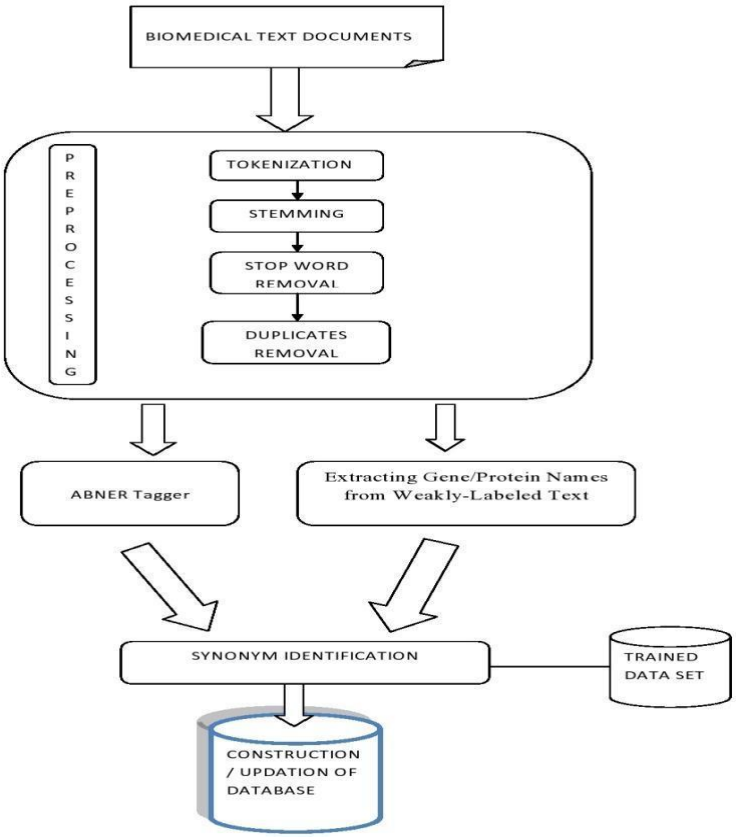

Fig 1: Proposed System for Identifying Synonymous Gene

2. Stop word Removal A word which has more frequency is termed as stopwords. Words such as pronouns, prepositions, conjunctions etc are normally used in English language to bridge the words and carry no information. . Removal of stop-words improves information extraction process. In our systems we used 371 stop words listed in PubMed [15] and 318 words list by referring to Cambridge University, 137 words list given in [16] ,429 words given in [17]

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

3. Stemming Next step is to remove suffixes by using our proposed improved stemmer algorithm. We studied improved porter stemmer algorithm [18] and identified the following various errors for which solutions are described in the form of rules in [19]

4. Duplicate Removal In order to reduce corpus we identify duplicate words and filter them. Reducing the corpus also increases the efficiency of preprocessing phase.

5. Named entity recognition: From the output corpus we should next extract gene/protein tokens. The ABNER (A Biomedical Named Entity Recognizer) tagger [20], which is an open source tool for is an open source software tool for automatically tagging genes, proteins and other entity names in text is used in our proposed approach.The latest version is 1.5, which has an intuitive graphical interface and includes two modules for tagging entities (e.g. protein and cell line) trained on standard corpora, for which performance is roughly state of the art. Difficulties for Biomedical NER.

B. Identifying Synonyms There are two types of synonyms for gene and protein names, Type I and Type II. We distinguish between Type I and Type II, Type I consist of the correspondence between the short and long forms of gene and protein names (e.g., LARD and lymphocyte associated receptor of death). Type II consists of the correspondence between all short forms (e.g., Apo3, DR3, TRAMP, LARD, and wsl). We identify synonymous gene/protein by using SSFPOA measure (only 11 matchers are used in this paper) which is compound similarity measure consisting of 12 matchers as listed below: Levenstein Distance, Smith- waterman, Needleman-Wunsch, Monge-Elkan measure, Stoilos Similarity, Boyer-Moore, synonymn dictionary, Soundex, Tri-gram , exact matcher, Jarowinkler.

C. Construction/Updation of Gene Database The Extracted gene/protein synonyms are stored in a database. If the database already consists of these gene/protein names, they are discarded; otherwise add new gene/protein names to the database (possible after filtering by a human biologist) that will enhance the current database. To Extract Gene-Protein Names from Weakly-Labeled Text we can use semantic similarity such as

eq. (1)

Where LD(s,s') is Levenshtein Distance between strings s and s', and length(s) is the number of characters in s.

IV. EVALUATION RESULTS

We took 50 MEDLINE Abstracts from [13] http://www.biomedcentral.com/ to construct synonymous gene/protein database. The following table 1 shows the statistics of extracted gene/protein names and their cluster similarities. Weka tool 3.7.9, 4GB RAM, Intel core i5 with 2.4 GHz processor is used for finding clusters. .arff file is created with 11 matcher values and is given to various classification techniques like Bayesnet, LibSVM and Multiplayer Perceptron. The output which represents similarity is not just 0 or 1 but in the range of 0-1 to represent different levels of similarity.

Table 1: Semantic Similarity measures applied on Gene Synonyms

Doc	Num of Gene names in the doc	Num of genes referring synonyms within the doc	No of gene names belonging to cluster				
			C1	C2	C3	C4	C5
1	80	40	15	14	2	9	0
2	66	33	11	17	5	0	0
3	74	37	6	17	11	3	0
4	90	45	19	11	15	0	0
5	26	13	2	8	3	0	0
6	46	23	1	17	5	0	0
7	112	56	6	28	21	1	0
8	244	122	26	72	22	2	0
9	172	86	7	58	20	1	0
10	60	30	7	20	3	0	0
11	154	77	11	38	26	2	0

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

Various experimental studies are performed by considering only 1 matcher, 3 matchers,5 matchers, 6 matchers,11 matchers, so as to understand the importance of each matcher in the semantic measure SSFPOA. The following Table-2,3,4 represents the comparison of accuracy in results while using a combination of matchers for various classifiers.

Table 2: Comparison of accuracy when using different matchers (Bayesnet)

Sno	Number of matchers used	Precision	Recall	F-Measure
1	1-matcher (Exact matcher)	.267	.467	.326
2	3-matchers (Exact, soundex, synonymn matchers)	.258	.478	.341
3	5-matchers matchers (Exact, soundex, synonymn, trigram, boyer-moore matchers)	.587	.661	.601
4	6-matchers(Levenstein Distance, Smith-waterman, Needleman-Wunsch,Monge-Elkan Distance , Stoilos Similarity,Jaro-winkler)	.559	.704	.614
5	11- matchers (Levenstein Distance, Smith-waterman,Needleman-Wunsch, Monge-Elkan measure, Stoilos Similarity,Boyer-Moore, synonymn, Soundex, Tri-gram , exact matcher,Jaro-winkler)	0.757	0.8	0.757

The accuracy is calculated according to Precision,Recall, and Fmeasure .Then compare the results using different matchers in SVM classifier is shown in Table-3

Table 3: Comparison of accuracy when using different matchers (SVM)

Sno	Number of matchers used	Precision	Recall	F-Measure
1	1-matcher (Exact matcher)	.285	.488	.347
2	3-matchers (Exact, soundex, synonymn matchers)	.195	.424	.284
3	5-matchers matchers (Exact, soundex, synonymn, trigram, boyer-moore matchers)	.721	.661	.653
4	6-matchers(Levenstein Distance, Smith-waterman, Needleman-Wunsch,Monge-Elkan Distance , Stoilos Similarity,Jaro-winkler)	.683	.768	.707
5	11- matchers (Levenstein Distance, Smith-waterman,Needleman-Wunsch, Monge-Elkan measure, Stoilos Similarity,Boyer-Moore, synonymn, Soundex, Tri-gram , exact matcher,Jaro-winkler)	.646	.774	.694

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

The accuracy is calculated according to Precision,Recall, and Fmeasure .Then compare the results using different matchers in MultiLayer classifier is shown in Table-4

Table 4: Comparison of accuracy when using different matchers (MultiLayer)

Sno	Number of matchers used	Precision	Recall	F-Measure
1	1-matcher (Exact matcher)	.269	.473	.331
2	3-matchers (Exact, soundex, synonymn matchers)	.182	.409	.270
3	5-matchers matchers (Exact, soundex, synonymn, trigram, boyer-moore matchers)	.662	.625	.601
4	6-matchers(Levenstein Distance, Smith-waterman, Needleman-Wunsch,Monge-Elkan Distance , Stoilos Similarity,Jaro-winkler)	.376	.558	.433
5	11- matchers (Levenstein Distance, Smith-waterman,Needleman-Wunsch, Monge-Elkan measure, Stoilos Similarity,Boyer-Moore, synonymn, Soundex, Tri-gram , exact matcher,Jaro-winkler)	.825	0.9	0.857

One problem noticed while evaluating SSFPOA is some of the Medline abstracts taken by us involve more number of cryptic.Pre-processing phase does not replace these cryptic with corresponding full names such as (LARD - lymphocyte associated receptor of death). This was reflected while evaluating with 1 matcher (exact matcher) as shown in table 2. Mapping cardinality of 1:1 (simple) is used in this paper and have not resolved 1:n (complex) mappings.
The following Fig.2 shows a graph of overall performance of SSFPOA on the 10 Medline Abstracts when compared to [12]

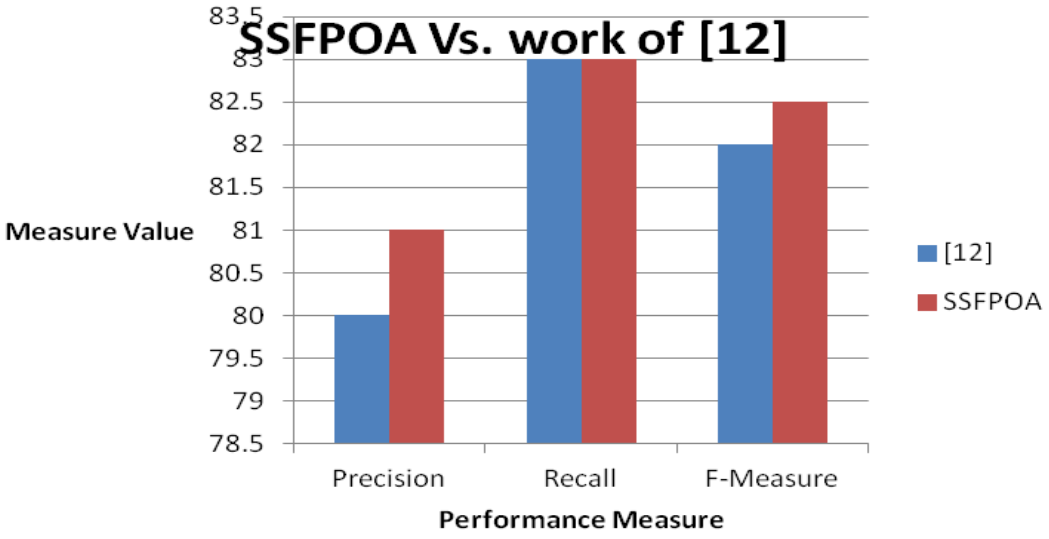

Fig 2: Overall performance of SSFPOA

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

V. CONCLUSION AND FUTURE WORK

This paper proposed a hybrid approach for constructing synonymous gene/protein names database dictionary from Medline abstracts that consists of three phases. In the first phase, pre-processing is carried out to reduce the corpus of the given abstracts. In the second and third phase, synonymous gene names are identified and added to the database. Experimental result shows that the proposed work provides 82% accuracy in identifying Gene and Protein names, which is evaluated and verified using the Precision, Recall and F-Measure. The Proposed architecture improved the performance of stemmer algorithm by applying new rules to the original stemmer algorithm. As there is no universally accepted tokenization method for processing text documents, our future work concentrates on improving biomedical tokenization process. Our work should also consider abstracts from other biomedical journals such as Pubmed etc. Also we need to compare Gene database constructed by us with one of the existing databases such as GenBank so as to analyse whether our approach can put update-to-date information or not.

REFERENCES.

[1]. S. Vasavi, S. Jayaprada, V. Srinivasa Rao, "Extracting Semantically Similar Frequent Patterns Using Ontologies", SEMCCO'11 Proceedings of the Second international conference on Swarm, Evolutionary, and Memetic Computing - Volume Part II, Pages 157-165.
[2]. Hui Yang, Goran Nenadic, John A. Keane, "A cascaded approach to normalising gene mentions in biomedical literature", Biomedical Informatics Publishing Group, Bioinformation 2(5): 197-206, 2007.
[3]. Koning D, Sarkar I, Moritz T: " TaxonGrab, Extracting taxonomic names from text ", Biodiversity Informatics, 2, 2005, pp. 79-82
[4]. Martin Krallinger, Maria Padron and Alfonso Valencia, "A sentence sliding window approach to extract protein annotations from biomedical articles", BMC Bioinformatics 2005, 6(Suppl 1): S19 doi: 10.1186/1471-2105-6-S1-S19.
[5]. Hanisch D, Fundel K, Mevissen HT, Zimmer R, Fluck J , ProMiner: rule-based protein and gene entity recognition, BMC Bioinformatics, 6: S14 2005.
[6]. Chang JT, Schutze H, Altman RB, "GAPSCORE: finding gene and protein names one word at a time". Bioinformatics Vol. 20 no. 2 2004, pages 216–225 DOI: 10.1093/bioinformatics/btg393
[7]. Hong Yu, Vasileios Hatzivassiloglou, Andrey Rzhetsky, and W. John Wilbur, "Automatically identifying gene/protein terms in MEDLINE abstracts". Biomedical Informatics 2003. doi:10.1016/S1532-0464(03)00032-7
[8]. Schwartz, A.S ,Hearst, M.A,"A simple algorithm for identifying abbreviation definitions in biomedical text", In Proceedings of the Pacific Symposium on Biocomputing 8:451-462,2003.
[9]. Tanabe L, Wilbur WJ , "Tagging gene and protein names in biomedical text". Bioinformatics 2002, 18(8):1124-1132.
[10]. Hong Yu,Vasileios Hatzivassiloglou, Carol Friedman, Andrey Rzhetsky, W. John Wilbur, "Automatic extraction of gene and protein synonyms from MEDLINE and journal articles". Proceedings of AMIA Symposium 2002:919-923.
[11]. Youngja Park, Roy J. Byrd, "Hybrid Text Mining for Finding Abbreviations and their Definitions", Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing, Pittsburgh, PA
[12]. R.Porkodi, B.LShivakumar, "Rule based approach for constructing Gene/Protein names Dictionary from Medline abstract", International journal of advances in computing and information technology pageno: 457-468 June 2012.
[13]. [13] http://www.biomedcentral.com/
[14]. Jing Jiang,ChengXiang Zhai ,"An Empirical Study of Tokenization Strategies for Biomedical Information Retrieval", *Inf. Retr*. 2007, **10**(4-5)**:**341-363
[15]. http://www.oocities.org/athens/sparta/7124/physicians/advanced/stopwords.pdf
[16]. http://www.ncbi.nlm.nih.gov/books/NBK3827/table/pubmedhelp.T43/
[17]. http://www.lextek.com/manuals/onix/stopwords1.html
[18]. Fadi Yamout, Rana Demachkieh, Ghalia Hamdan and Reem Sabra, Further Enhancement to the Porter's Stemming Algorithm CandE American University, 2004.
[19]. B.Jayanag, S.Vasavi, "Dynamic feature subsumption based multiclass sentiment analyzer using machine learning techniques", Communicated to IEEE2014.
[20]. http://pages.cs.wisc.edu/~bsettles/abner/