Term Recognition and Extraction based on Semantics for Ontology Construction

Akalya.B and Nirmala Sherine.F

**Dept. Information Technology, Periyar Maniammai University,
Vallam,Thanjavur-613403,Tamil Nadu,India.**

**Dept. Information Technology, Periyar Maniammai University,
Vallam,Thanjavur-613403,Tamil Nadu,India.**

Abstract

In recent years the development of Ontology's, leads to explicit formal specifications of the terms in the domain and relations among them. The construction of Ontology often requires a domain specific corpus in conceptualizing the domain knowledge. It is an indispensible task to identify a list of significant terms for constructing a Structured Ontology. In this paper, we investigate the use of Semantic Similarity-based metrics for term recognition and extraction, for ontology construction from the text document. The methodology uses Taxonomy and Wikipedia to reinforce the automatic term recognition and extraction from structured documents. It is done with the assumption of candidate terms for a topic are often associated with its topic-specific keywords through Semantic Similarity-based metrics by making use of WordNet. A hierarchical relationship of super-topics and sub-topics is defined by Taxonomy, meanwhile, Wikipedia is used to provide a semantic relationship and background knowledge for topics that are defined in the Taxonomy to surpervise the term recognition and extraction. Experimental results show that the proposed methodology is viable to be applied in a small corpus supporting Ontology construction, which renders the foundation for higher recall and precision, when compared with existing methodologies.

Keywords: *Ontology, Taxonomy, Noun phrase, POS tagger, Precision, Recall, Semantics.*

1. Introduction

The field of ontology construction always received attention due to the increasing needs in conceptualizing the domain knowledge in resolving various jobs' demand. Ontology has been used to support personalized e-Iearning [1], to support emergency decision making [2] and to resolve gathered information ambiguity [3]. A good ontology depends on its successfulness in solving a given domain problem. Ontology is basically comprised of terms, relation between terms and related instances. Term represented in ontology denotes a set of words (single word and/or complex words) that is significant to explicate the domain investigated. It is usually found explicitly on the surface of the investigated domain text. The basic requirement in constructing ontology is identifYing an appropriate corpus. Ontology construction requires domain-specific corpus for acquiring concepts and building corresponding hierarchy of one domain [4].

Ontologies have become common on the World-Wide Web. The ontologies on the Web range from large taxonomies categorizing Web sites (such as on Yahoo!) to categorizations of products for sale and their features (such as on Amazon.com). The WWW Consortium (W3C) is developing the Resource Description Framework (Brickley and Guha 1999), a language for encoding knowledge on Web pages to make it understandable to electronic agents searching for information. The Defense Advanced Research Projects Agency (DARPA), in conjunction with the W3C, is developing DARPA Agent Markup Language (DAML) by extending RDF with more expressive constructs aimed at facilitating agent interaction on the Web (Hendler and McGuinness 2000). Many disciplines now develop standardized ontologies that domain experts can use to share and annotate information in their fields. Medicine, for example, has produced large, standardized, structured vocabularies such as SNOMED (Price and Spackman 2000) and the semantic network of the Unified Medical Language System (Humphreys and Lindberg 1993). Broad general-purpose ontologies are emerging as well. For example, the United Nations Development Program and Dun & Bradstreet combined their efforts to develop the UNSPSC ontology which provides terminology for products and services (www.unspsc.org).

Some of the reasons to develop Ontology:

- To share common understanding of the structure of information among people or software agents
- To enable reuse of domain knowledge

- To make domain assumptions explicit
- To separate domain knowledge from the operational knowledge
- To analyze domain knowledge

2. Related Work

To date, various approaches have been explored to improve the outcome of Term Recognition and extraction. They can be categorized into linguistic-based approach, statistical approach, machine learning approach, ontology-based approach and hybrid approach. Linguistic-based approach [5], [6], [7] uses dictionary, rules or patterns to extract desired terms; high workload is imposed on linguists as all rules and patterns have to be coded manually but it produces better quality outcome. Meanwhile, statistical approach [8], [9], [10], [11], [12], [13], [14], [15] uses solely mathematical computation to identify the co-occurrence of lexical(s) in one or more specified set of documents. Generally, it performs solely on the term counting without any semantic understanding. On the other hand, the advantage of machine learning approach [16], [17], [18] is that it is relatively easy to tune to new domains, provided that tagged training data is existed. [19], [20] have adopted ontology in extracting terms, where it can achieve higher performance results; however, it requires the availability of the domain specific ontology prior the extraction. Lastly, hybrid approach is the combination of various approaches in performing Term Recognition and extraction. For example, CINC value [21], a hybrid approach that use linguistic-based to form terms' patterns for identification and extraction before statistically counting its contextual information. Nevertheless, term play a vital role in ontology construction as it represents the summary of the domain investigated. However, all above mentioned term recognition and extraction approaches are mainly constrained by three major issues as discussed below.

A. Domain Level: Most research works required the selection of corpus resources. Some researchers prefer to use manually compiled corpus by linguistic expert such as Brown Corpus, web pages crawled from the Internet and existing business documents or manuals. Often, the selection of corpus is always domain specific depending on the intended problem to be solved and a domain corpus might consist of various distinct topics. The above mentioned approaches are solely focused on one level (domain level) Term Recognition and extraction, where it often accommodates to the nature of the investigated domain as a whole context. This might constraint its identification and extraction outcome/result when applying it to other domains or various topics as they might be different in context and background knowledge.

B. Nature of the terms: Research works done in term identification and extraction are mainly falling into two areas, which are technical and non-technical. In the technical area, it implies the use of specialized knowledge of applied sciences such as for medicine and biology domain. Meanwhile, the non-technical area denotes the use of general knowledge such as for tourism and educational domain. Both areas exhibit the increasing usage of diversity in term of morphology and collocation. Despite to the distinct nature of terms between the technical area and the non-technical area, the technical areas often expose certain pattern in its terminological presentation, for instance, biological terms often contain prefixes and suffixes that give an indication of their class. On the other hand, the non-technical areas are always clueless to accommodate precisely the likelihood of potential terms.

C. Text / corpus size: Research works done in Term Recognition and extraction often involve multi-documents with the aim to conciliate the relevancy of extracted terms to the domain investigated. Various statistical metrics are then used to validate the extracted terms relevancy to the domain chosen. Frequency-based counting and Term Frequency - Inverse Document Frequency (TF-IDF) are the two most commonly used metrics in validating true terms. In this case, corpus size does impact the Term Recognition and extraction. However, it will be a problematic issue for domains which have less resources and small corpus size in
term extraction.

In this paper, we propose a Semantic Similarity-based term recognition and extraction using taxonomy and Wikipedia to overcome the above mentioned issues. A hierarchical relationship of super-topics and sub-topics is defined by a taxonomy, meanwhile, Wikipedia is used to provide context and background knowledge for topics that defined in the taxonomy. Further with the help of WordNet we identify the terms similar to topics semantically, to perform the term recognition and extraction.

3. Proposed System

3.1 Taxonomy

To date, various research works have been carried out on term extraction, however their works are mainly focus at domain level extraction. As defined, a domain might consist of various topics. For example, Tourism domain comprises of culture and heritage, hotel, transportation and places to visit as its topics., In our work, taxonomy is proposed to be used as it provides a structure for various

topics in a domain. It defines a hierarchical relationship of super-topics and sub-topics. Hence, the domain level Term Recognition
and extraction is performed by taking into the consideration of different topics might appear in the investigated domain in which topic related documents will be handled specifically according to its context and background during Term Recognition and extraction.

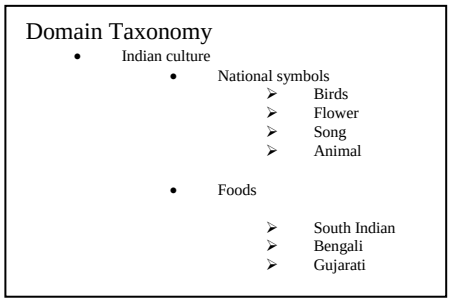

Fig. 1 Shows a portion taxonomy for Indian Culture as an example.

As illustrated in the Fig1, the Indian Culture domain webpage consists of various topics; each topic is distinct in its contents. For example, the parent concept " Indian Culture " consists of "National Symbols" and "Foods" as the first level sub-topic in domain taxonomy; meanwhile "Birds", "Flower", "Song" and "Animal" and are represented as the second level sub-topics of " National Symbols " whereas "South Indian", “Bengali” and "Gujarati" are represented as the second level sub-topic of "Foods" in the domain taxonomy.

3.2 Wikipedia

Wikipedia is the world largest online encyclopedia lies in its size and coverage. It reaches approximately 3 million articles in English as dated on May 2010 since its establishment in year 2001. It covers a rich resource of general knowledge as well as in depth clarification of many specialized knowledge which might be potentially contribute in various aspects to knowledge extraction.

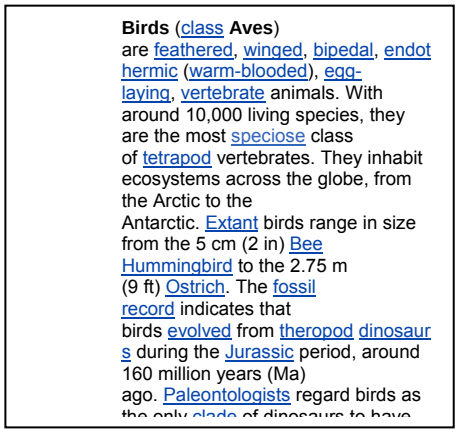
Birds (class **Aves**) are feathered, winged, bipedal, endothermic (warm-blooded), egg-laying, vertebrate animals. With around 10,000 living species, they are the most speciose class of tetrapod vertebrates. They inhabit ecosystems across the globe, from the Arctic to the Antarctic. Extant birds range in size from the 5 cm (2 in) Bee Hummingbird to the 2.75 m (9 ft) Ostrich. The fossil record indicates that birds evolved from theropod dinosaurs during the Jurassic period, around 160 million years (Ma) ago. Paleontologists regard birds as

Fig .2 Wikipedia content on topic "Birds"

In our work, Wikipedia is used to provide context and background knowledge to topics in a domain taxonomy. An assumption is formed where candidate terms of a topic are often associated with its topic-specific keywords in providing a context and background knowledge.

Fig .3 Extracted dw on topic "Birds"

Wikipedia article corresponding to the topics represented in domain taxonomy is elicited. Hyperlinks exist in each Wikipedia article symbolizes descriptive words (dw); it is the description of context and background knowledge to the investigated topic. Fig 2 shows the introduction part of the topic "Birds" described in Wikipedia. The underline words (hyperlinks) are example of its dw. Fig 3 is the list of extracted dw for topic " Birds " and it symbolized the topic "Birds" context and background knowledge.

3.3 Frame Work Description

The core theme of this paper is to perform Term recognition and Term Extraction, for ontology construction from the text document. The methodology uses Taxonomy and Wikipedia to reinforce the automatic term recognition and extraction from structured documents. It is done with the assumption of candidate terms for a topic are often associated with its topic-specific keywords through Semantic Similarity-based metrics by making use of WordNet. A hierarchical relationship of super-topics and sub-topics is defined by Taxonomy, meanwhile, Wikipedia is used to provide a semantic relationship and background knowledge for topics that are defined in the Taxonomy to surpervise the term recognition and extraction.This renders a foundation of higher recall and precision.

The prototypical implementation of Semantic Similarity-based term recognition and extraction of our approach is illustrated in Figure 4 and the details of algorithm is formulated below:

Step 1: Discovery of domain source

A domain is identified as input source for term recognition and extraction. For the experimental purpose, tourism domain is chosen to test the proposed framework. Domain web pages are taken from Indian Culture website as the

source documents for term recognition and extraction. "Birds", "Flower", "Song" and "Animal" are the selected topics with their corresponding web page in the selected domain web page.

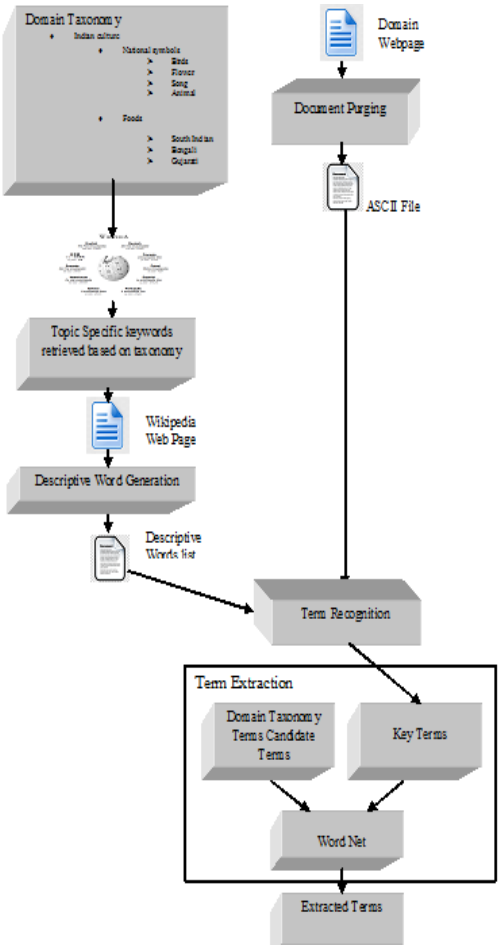

Fig.4 Semantic Frame Work for Term Recognition and Extraction

Step 2 : Discovery of domain taxonomy

Given the identified domain, a related domain taxonomy is defined. In the experiment, a domain taxonomy corresponding to the domain is adopted from Indian Culture website. Its hierarchical relationship of super-topics and sub-topics defined by the taxonomy as illustrated in Fig 1. is to provide structure for Term Recognition and extraction from structured documents.

Step 3 : Retrieve related Wikipedia articles of the topics defined in the domain taxonomy

A freely available Wikipedia API3 is used to provide automatic access to Wikipedia articles. For example, topic's name, "Birds" in the domain taxonomy is served as a keyword for retrieving a related Wikipedia article. The Wikipedia API implements OpenSearch protocol to retrieve Wikipedia article using the provided keyword. The related Wikipedia article(s) corresponding to the provided keyword is returned.

Step 4 : Generation of the list of descriptive words

The obtained Wikipedia article (webpage) is rendered into HTML format automatically using info.bliki.api.creator, a package in the Wikipedia API (Bliki engine). The objective of this step is to ease the generation of the list of descriptive words. A set of a word list is defined as follow:

$$DW = \{dw1, dw2, \ldots dwn\}$$

where DW denotes a list of descriptive word (dw) as described in Figure 3. All hyperlinks (descriptive words) will be extracted from the converted HTML document. The below HTML syntax is the sample hyperlink of a related descriptive word.

<ahref= "/wiki/peacock" title = "Peacock ">peacock

<a> .... is the anchor name used to display information within a document, href= "1wiki/Peacock denotes the URL of the descriptive word, and title = "Peacock" indicates title/description of the descriptive word. "peacock" is a descriptive word as explained in Section Wikipedia. In this case, "peacock" is extracted.

Step 5 Document purging

"Birds", "Flower", "Song" and "Animal" are the topics available in the selected domain web pages as stated in step 1. Each topic webpage is retrieved from the domain web page and cleaned up using HTML Content Extractor to eliminate non-text contents such as ads, banners, videos, audios, navigations links and menus. The cleaning task is performed automatically and does not require any user interaction during the cleaning process. At the end of the process, a pure text file of the topic is produced.

Step 6 Term recognition

Given an assumption that the candidate terms of a topic are often associated with its topic-specific keywords, each dw in the list of descriptive word is examined against each sentence in pure text file of the topic. Sentence which contains dw is extracted. Finally, a file with sentences where each sentence contains at least one dw is being generated.

Step 7 Term extraction

This step consists of three processes which are tagging, Stemming and Semantic Analysis.

a. Tagging

Treetagger (a multi-lingua tool for annotating text with part-of-speech and lemma information) is used to shallowly tag all the extracted sentences in the step 5 and to elicit terms which are tagged with "NP" (Noun Phrase). Generally, most candidate terms possess to be tagged with Noun Phrase. A list of terms is generated as an input for next process.

b. Stemming

Extracted list of terms might contain redundancy. For example, the word "malays" and "malay" are in fact referring the same item. Hence, Porter Stemmer is used to reduce all terms into their stem, base or root form. In this case, after the stemming process, the base form of "malays" and "malay" is "maim"'. The terms with the same base form will be considered as one term.

c. Semantic Similarity-based term extraction

Domain Taxonomy Terms and Key terms are analyzed for identifying the extracted terms with the mode of WordNet. The term which has the higher percentage of Semantic Similarities are extracted.

4. Results And Discussion

4.1 Experimental Results

The most challenging activity in term extraction lies in its evaluation method as there is no formal way to evaluate terms. Same resource corpus might produce different terms, all are depends on their usage in the developed application. Hence, it is difficult to obtain a suitable "gold standard" that can be used to evaluate extracted terms. Having all the known evaluation difficulties in mind, we manually evaluated the result with the help of human expert. The Term Recognition and extraction were evaluated using two metrics: recall and precision.

$$recall = \frac{|\{RelevantDocuments\} \cap \{RetrievedDoaments\}|}{|\{RelevantDocuments\}|}$$

$$precision = \frac{|\{RelevantDocuments\} \cap \{RetrievedDoaments\}|}{|\{RetrievedDoaments\}|}$$

Indian Culture website was used to examine our methodology. The available domain taxonomy on the designated Indian Culture website is adopted to guide the term recognition and extraction. Four distinguished topics in the domain taxonomy, "Birds", "Flower", "Song" and "Animal" as illustrated in Fig1 are chosen for testing our proposed approach. The content of each topic is context sensitive to its title. Thus, it is significantly novel to experiment the taxonomy utilization in term recognition and extraction.

Topic	WordCount
Birds	923
Flower	667
Song	234
Animal	814

Table.1 Displayed the word count for each topic in a pure text file.

The results of term extraction are presented in Table2 and Figure 5, they showed that the domain taxonomy and Wikipedia have contributed an impact in our work. The use of dw in topic web page hints the exact location of the topic candidate terms. For instance, "Birds", "peacock" and "Ostrich" are the most referred dw against topic "Birds".

Topic	Recall	Precision
Birds	100	57.4
Flower	63	39
Song	64.5	40.6
Animal	100	45

Table.2 Recall and Precision of Term Extraction

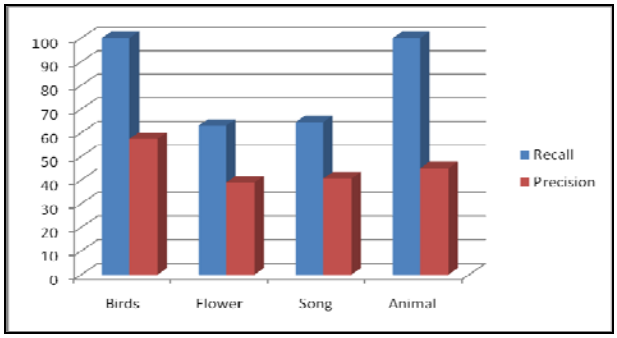

Fig5. Recall and Precision of Term Extraction

Whenever the descriptive word occurs in a sentence, the topic candidate term was existed before or after it. Out of 60% of the extracted sentences in the topic "Birds", it contains 100% of terms in it. This has proven that context-based of using domain taxonomy and Wikipedia worked well in handling term recognition and extraction by taking into the consideration of different topics might appear in the investigated domain. This experiment has also discovered that the page size of topic has not contributed to its performance metrics. The largest size of topic webpage, "Birds" with 923 word count gives 100% recall and 60% precision whereas the smallest topic webpage, "Animal" with only 196 word count gives a recall of 100% and precision of 44%. This gives an indication that our approach can be applied in any domain regardless of the text / corpus size.

Table.3 Recall and precision of term extraction tools

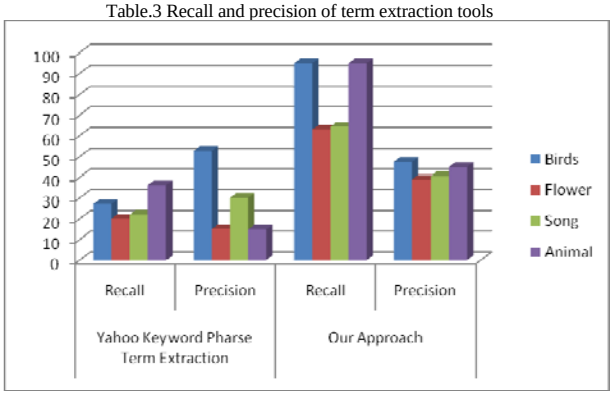

Fig. 6 Performance among term extraction tools

As indicated in Table 3 and Figure 6, we performed a comparison study between our approach with Yahoo Search Keyword Terms Extraction. The methodology of the context-based Term Recognition and extraction using taxonomy, and Wikipedia performed domain level term recognition and extraction by taking into the consideration of different topics might appear in the investigated domain compared to the both tools which have performed the term extraction on the domain level regardless of the context and background knowledge of its content.

A Part-of-Speech (POS) tagger and a noun phrase chunker [22] are used when extracting terms. When measuring the words Semantic Similarity based on WordNet, we choose Lin's method. The full text retrieval tool, lucene [23], is the basis of our experiments.

5. Conclusion

The work proposed in this paper is meant to provide a better way for term recognition and extraction by taking into consideration of various topics might occur in a domain corpus which supports Ontology construction. The multi-topics are represented in taxonomy as a multi-level tree representation and the Wikipedia is used to provide multi topics' context and background knowledge. We make use of WordNet to incorporate semantic similarities for extracting relevant terms for Ontology Construction. The hypothesis proven is that carefully taking care of the need of each domain topics can improve the performance metrics.

Future work is directed towards, considering many domains and continuous revamping of Ontology which provides an avenue for huge Ontology Construction.

	Yahoo Keyword Pharse Term Extraction		Our Approach	
Taxonomy Concept	Recall	Precision	Recall	Precision
Birds	27	52.8	95	47.4
Flower	20	15	63	39
Song	22	30	64.5	40.6
Animal	36.2	14.7	95	45

References

[1] Henze.N.,and Dolog.P., and Nejdl. W., "Reasoning and Ontologies for Personalized e-Learning in the Semantic Web", Educational Technology & Society, 7(4),pp. 82-97,2004.

[2] Yu. K., and Wang. Q. Q., and Rong. L. L., "Emergency Ontology Construction in Emergency Decision Support System", Proceedings of 2008 IEEE International Conference on Service Operations and Logistics and Info rmatics IEEE/SOLI, Beijing, China, pp. 801-805, October 2008.

[3] Fonseca. F. T., and Egenhofer. M. 1., and Agouris. P., and Camara, G., "Using Ontologies for Integrated Geographic Information Systems", Transactions in GIS 6(3), pp. 231-257, 2002.

[4] Cui. G. Y., and Lu. Q., and Li. W. J., and Chen. Y. R., "Corpus Exploitation from Wikipedia for Ontology Construction", Proceedings of the Sixth International Language Resources and Evaluation (LREC2008), Morocco, 2008.

[5] Bajwa. I. S., and Siddique. M. I., and Choudhary. M. A.:, "Automatic Domain Specific Terminology Extraction using a Decision Support System", In the Proceedings of 4th IEEE - International Conference on Information and Communication Technology-ICICT, pp. 651-659, Cairo, Egypt, 2006.

[6] Wermter. J., and Hahn, U., "Finding New Terminology in Very Large Corpora", Proceedings of the 3rd international conference on Knowledge capture Banff, pp. 137-144, Alberta, Canada, 2005.

[7] Mukherjea. S., and Subramaniam. L. V., and Chanda. G., and Sankararaman. S., and Kothari. R., and Batra. V., and Bhardwaj. D., and Srivastava. B., "Enhancing a Biomedical Information Extraction System with Dictionary Mining and Context Disambiguation", IBM Journal of Research and Development, Volume 48 , Issue 5/6, pp. 693 701, 2004.

[8] Chang. J.S., "Domain Specific Word Extraction from Hierarchical Web Documents: a First Step Toward Building Lexicon Trees from Web Corpora", Proceedings of the Fourth SIGHAN Workshop on Chinese Language Learning, Korea, pp.64-71, October 2005.

[9] Chen. Y. R., "The Research on Automatic Chinese Term Extraction Integrated with Unithood and Domain Feature", Master Thesis in Beijing, Peking University 2005.

[10] Kurz. D.F., and Xu. Y., "Text Mining for the Extraction of Domain Relevant Terms and Terms Collocations", Proceedings of the International on Computational Approaches to Collocations, Vienna, Austria, July 2002.

[11] Church. K. W., and Gale. W. A., "Concordances for Parallel Text", In Proceedings of the Seventh Annual Conference of the UW Center for the New OED and Text Research,

Association for Computational Linguistics, Oxford, UK, pp. 40□62, 1991.

[12] Streiter. 0., and Zielinski. D., and Ties. I., and Voltmer, L., "Term Extraction for Ladin: an Example Approach", TALN: Traitement Automatique des Langues Naturelles, VVF-Batz-sur-Mer(44), France, 2003.

[13] Dunning. T., "Accurate Methods for the Statistics of Surprise and Coincidence", Computational Linguistics, Vol. 19, no. 1, pp. 61□74, 1993.

[14] He. T.T., and Zhang X.P., and Ye X.H., "An Approach to Automatically Constructing Domain Ontology", PACLIC 2006, Wuhan, China, pp. 150-157, 1-3 November, 2006.

[15] Alexander G., and Grigori S., and Eduardo LV., and Liliana C.H., "Automatic Term Extraction using Log-likelihood based Comparison with General Reference Corpus", LNCS 6177, pp. 248-255, 2010.

[16] Eriksson. G., and Franzen. K., and Olsson. F., and Asker. L., and Liden, P., "Exploiting Syntax when Detecting Protein Names in Text", EFMI Workshop on Natural Language Processing in Biomedical Applications, Nicosia, Cyprus, March 2002.

[17] Zhang. Q.L., and Lu. Q., and Sui. Z.F., "Measuring Termhood in Automatic Terminology Extraction", International Conference on Natural Language Processing and Knowledge Engineering, Beijing, China, pp.328 □ 335, 2007.

[18] Zhou. G. D., and Suo J., "Named Entity Recognition using an HMM-based Chunk Tagger", Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, pp. 473-480, 2002.

[19] Zhang. W., and Yoshida. T., and Tang, X.J., "Using ontology to improve precision of terminology extraction from documents", Expert Systems with Applications, Vo. 36, Issue 5, pp. 9333-9339, 2009.

[20] Zhou. X.H., and Han. H., and Chankai. 1., and Prestrud. A., and Brooks. A., "Approaches to Text Mining for Clinical Medical Rrecords", Proceedings of the 2006 ACM symposium on Applied computing, Dijon, France, pp.235 239, 2006.

[21] Ananiadou. S., and Nenadic. G., "Automatic Terminology Management in Biomedicine", Text Mining for Biology and Biomedicine, S. Ananiadou and J. McNaught (eds), Artech House, London, ChA, pp. 67-98, 2006.

[22] Noun Phrase Chunker:- http://www.dcs.shef.ac.uk.

[23] Lucene Available: http:-//lucene.apache.org

[24] Hui-Ngo GOH , Ching-Chieh KJV ,"Context-Based Term Identification and Extraction for Ontology Construction " IEEE 2010.

[25] Xin Peng, Wenyun Zhao,"An Incremental and FCA-based Ontology Construction Method for Semantics-based Component Retrieval "Seventh International Conference on Quality Software IEEE 2007.

[26] Che-Yu Yang, Shih-Jung Wu "A WordNet Based Information Retrieval on the Semantic Web "pp 324-328.

[27] Jibran Mustafa, Sharifullah Khan, Khalid Latif,"Ontology Based Semantic Information Retrieval" 4th International IEEE Conference "Intelligent Systems"pp 14-19,2008