International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 9 (2015) pp. 23073-23085

http://www.ripublication.com

Context-Based Term Identification and Extraction For Ontology Construction In Indian Regional Language

Rajeshwari S B and Dr. S R Swamy
Department of Computer science and Engineering
RVCE, Bangalore
rajeshwari.sb@msrit.edu and swamysr@rvce.edu.in

Abstract

Ontology construction requires a domain specific corpus in conceptualizing the domain knowledge; it is an association of terms, relation between terms and related instances. In this paper, we proposed a new approach using ontology to improve precision of terminology extraction from documents. Firstly, a linguistic method was used to extract the terminological patterns from documents. Then, similarity measures within the framework of ontology were employed to rank the semantic dependency of the noun words in a pattern. In this project, the use of a context-based term identification and extraction methodology for ontology construction from text document in Indian regional language (Kannada) has been considered. The context-based term identification and extraction methodology is viable in defining topic concepts and its sub-concepts for constructing ontology and also to be applied in a small corpus / text size environment in supporting ontology construction for an Indian regional language.
Experiments on Retuers-21578 corpus has shown that *WordNet* ontology, that we adopted for the task of extracting terminologies from Kannada documents, can improve the precision of classical linguistic method on terminology extraction significantly.

Keywords: Wikipedia, automatic term identification, context-based term identification, extraction methodology, semantic dependency, WordNet.

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-learning ,to support emergency decision making and to resolve gathered information ambiguity. A good ontology depends on its successfulness in solving a given domain problem.

Recently, terminology and its related lexical units such as multi words, collocations, etc. have been widely studied in both linguistics and text mining field. From the cognitive point of view, human beings recognize, learn and understand the entities and concepts in texts for a complete natural language comprehension. It is commonly accepted by researchers in natural language processing (NLP) that terminologies can better capture the topics of texts and describe the contents of texts more accurate than individual words, because their distinctive entities in a domain and their referents are more specific and unambiguous than their constituents as individual words where *polysemy* may usually occur.

There are mainly two types of methods developed for terminology extraction: linguistic method and statistical method. Linguistic methods utilize structural properties of phrases and sentence grammar of a special language to extract terminologies from documents (Chen & Chen, 1994; Choueka, 1983; Church, 1989; Justeson& Katz, 1995). Statistical methods utilize corpus learning with statistical indicators to measure the words' association for co-occurrence pattern discovery.

In the linguistic aspect, Choueka's methodology for handling the large corpora can be considered as a first step toward computer aided lexicography. In his method, the consecutive sequences with two to six words were retrieved as collocations (Choueka, 1983). Justeson and Katz proposed a regular expression for individual words in a consecutive sequence to retrieve terminology (Justeson& Katz, 1995). They reported that their method can produce a recall as 95% and precision as 96% at average on all the categories of SUSANNE corpus (online: http://www.grsampson.net/Resources.html) (Chen & Chen, 1994).

In the statistical aspect, language is modeled as a stochastic process and the corpus is used to estimate that whether or not a given sequence occurs in the corpus by chance. Church and Hanks proposed the association ratio for measuring word association basedon the information theoretic concept of mutual information to retrieve the pairs of words which occurred frequently in corpus together (Church & Hanks, 1990). Smadja developed *Xtract* to extract collocations from documents using the relative positions of two words in a corpus (Smadja, 1993). In *Xtract*, four parameters were used: strength, spread, peak z-score and percentage frequency.

The contribution of this paper is threefold. Firstly, ontology, WordNet and the relationships of words within ontology is introduced. Secondly, a method is proposed to rank words' semantic dependency in a sequence based on word similarities within ontology. Thirdly, a classical linguistic method is employed to extract word patterns from an English corpus and the proposed method is examined.

Related Work

A classic linguistic method for terminology extraction was presented and the semantic relationship of noun words within ontology was discussed in this section. Termrecognition 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.

A Linguistic Approach For Terminology Extraction

From Justeson and Katz's point of view, Noun Phrases (NP) can be divided into two groups: lexical NPs and non-lexical NPs. Lexical NPs are subject to a much more restricted range and extent of modifier variation, on repeated references to the entities they designate, than non-lexical NPs. And the terminological NPs differ from other NPs because they are lexical (Justeson& Katz, 1995).

When a terminological NP is a topic of significant discussion within a text, they tend to be repeated intact on repeated references to the entities they designate. The non-lexical NPs usually do not repeat many times within a text because they can simply be represented by the head noun and their modifiers often vary. For this reason, one effective criterion for terminology identification is simple repetition: a noun phrase having a frequency of two or more can be entertained as a likely terminological unit, i.e., as a candidate for inclusion in a list of technical terms from documents.

Ontology

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. In philosophy, ontology is a study of being or existence and forms the basic subject matter of metaphysics. It seeks to describe the basic categories and relationships of being or existence to define entities and types of entities within its framework (Ontology). In the area of knowledge management, ontology refers to using hierarchical trees to represent the background knowledge, for example, MESH ontology and Word Net .Although no formal definition of ontology is generally recognized and how it should be implemented is controversial. That is, ontology is constructed based on a controlled vocabulary and the relationships of the concepts in the controlled vocabulary.

Definition: Controlled vocabulary CV = name set of concepts c, where c = (name, definition, identifier, synonyms). In ontology the concepts are linked by directed edges, then form a graph. The edges of an ontology specify in which way concepts are related to each other, e.g., ‘‘is-a” or ‘‘part-of”.

Wikipedia

Wikipedia is the world largest online encyclopedia lies in its size and coverage. It reaches approximately 3 million articles in English.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. In our work, Wikipedia is used to provide context and background knowledge to topic in domain taxonomy. An assumption is formed where its specific keywords in providing a context and background knowledge.

Semantic Relationship of Words In Word Net

WordNet try to make the semantic relations between word senses more explicit and easier to use. Because terminologies are usually nouns, in this paper, we concentrate on using noun words in WordNet to improve the performance of terminology extraction. WordNet contains 80,000 noun word forms organized into some 60,000 lexicalized concepts. Many of these nouns are collocations; a few are artificial collocations invented for the convenience of categorization. WordNet divided the nouns into 11 hierarchies, each with a different unique beginner which corresponds to a primitive semantic component in a compositional theory of lexical semantics (Miller, 1998). The basic relationship in WordNet is synonymy. A set of synonym is called a *synset*. And the relationship between noun *synsets* in WordNet is either *hypernym* or *hyponym*. For instance, the *synset* ‘‘person, human” is a *hypernym* of the concept as ‘‘adult, grownup” and the relationship is hyponym in reverse. A synset has only one *hypernym* but it may have more than one hyponyms. This design for concepts in WordNet is very similar to the concept organization in human natural

language. The distinctiveness of WordNet from conventional dictionary is that we can use the semantic relationships between *sysnets* for inferences besides it is readable by computer. For instance, if we have a concept as ‘‘human”, then we can infer that perhaps this ‘‘human” is ‘‘male” in gender and a ‘‘teacher” in vocation.

Using Ontology To Improve Terminology Extraction

In this section, the motivation of adopting ontology into terminology extraction is specified. Two methods for similarity measure of words within ontology are described. The new method of combing ontology into terminology extraction is proposed.

The Motivation

The main motivation to adopt ontology for terminology extraction is that, we want to make use of the background knowledge and words’ semantic relationships compiled in ontology to capturethe semantic features of terminologies in documents. We conjecture that ontology will take a positive effect on terminology extraction based on the following reasons:

1. Terminology is expression of a specific concept in a domain and ontology is also constructed on different domains. For this reason, we can use the concepts in ontology to match the terminologies in documents directly.
2. The constituents as individual words of a terminology are highly semantically correlated with each other and most lexical noun phrases have the property as compositional meaning in its sense. For this reason, we can deduce that the individual words in a terminology are of closer semantic relationships than those individual words not in a terminology but co-occurred together.
3. In document writing, terminologies are fixed phrases and its constituents co-occur together to express a complete concept. In WordNet ontology, all senses of a word are listed to relate its subordinates and super ordinates. Although some words co-occur infrequently in a corpus, we also can extract it by matching their senses. That is, ontology can compensate the loss of the statistical methods in terminology extraction.

Similarity of Words Within Ontology

In order to gauge the semantic dependency of individual words in a string sequence, similarity measures within framework of ontology was employed. Although many methods are proposed to measure semantic similarity between words (Rada, 1989; Richardson et al., 1994; Li, 2003), here a traditional method and a newly developed method are attempted in our study. That is, Rada et al.’s method (Rada, 1989) (referred as Rada method hereafter) and Li et al.’s method (Li, 2003) (referred as Li method hereafter).

In Rada method, similarity of two words is measured by the length of the shortest path between them in the hierarchical tree. The basic idea behind this method is very intuitively: words are associated with concepts in the ‘‘is a” (ISA) hierarchy, therefore, we can find the first concept in the hierarchical semantic network that

subsumes the concepts containing the compared words and then a path that can connect these two words is found.

In this paper, the similarity between words using Rada method is calculated as with the following formula.

$$sim(w1, w2) = e^{-\alpha l}$$

Whereα is a predefined constant and l is the length of the shortest path of word $w1\ andw2$ inthe hierarchical tree. The exponential form in similarity calculation is adopted because of Shepard's law which claims that exponential-decay functions are a universal law of stimulus generalization for psychological science.

The difference of Li method from Rada method is, in that not only the shortest path between compared words, but also the depth of their *subsumer* in the ontology hierarchy, and the *subsumer's* local semantic density are considered to calculate the similarity in Li method. The basic idea behind this method is to overcome the weaknesses in Rada method.

In this paper, the similarity between words using Li method is calculated with formulas.

$$\mathrm{sim}(\mathrm{w1}, \mathrm{w2}) = f1(l) + f2(d) + f3(fr)$$

$f1(l)$, $f2(d)$and $f3(fr)$are defined as follows:

$$f1(l) = e^{-\beta l}$$

Whereβ is a predefined constant and $l$is the length of the shortest path of $w1\ andw2$ in the hierarchical tree.

$$f2(d) = \frac{\mathrm{e^{cd}} - e^{-cd}}{\mathrm{e^{cd}} + \mathrm{e^{-cd}}}$$

Where c is a predefined constant and dis the depth of the *subsumer* of $w1\ andw2$ in the hierarchical tree.

$$f3(fr) = e^{-\gamma/fr}$$

Wherefr is the frequency of the extracted terminology candidate which contains $w1\ andw2$ using JK method.

The Proposed Approach For Terminology Extraction Using Ontology

Based on JK method and the similarity of words within ontology, a new approach is proposed to use ontology to improve terminology extraction from documents. Here are the main steps of ourapproach and the details will be discussed later.

1. Extract the repetitions from documents.
2. Conduct POS (part of speech) processing for repetitions and extracted patterns from repetitions using JK regular expression.
3. If an extracted pattern is a collocation already included in ontology hierarchy such as ''professional person", then it will be accepted as a terminology. Otherwise, similarity dependency will be given for this pattern.
4. Accept the patterns whose semantic dependencies are greater than the critical semantic dependency on the point of Retaining Proportion (RP) as

terminologies. RP is a predefined threshold for patterns' proportion with highest semantic dependency at a ratio.

Input:
s_1, the first sentence; s_2, the second sentence
Output:
Multi-word extracted from s_1 and s_2.
Procedure:

$s_1 = \{w_1, w_2, \ldots, w_n\}$, $s_2 = \{w_{1'}, w_{2'}, \ldots, w_{m'}\}$, k=0
For each word w_i in s_1
For each word $w_{j'}$ in s_2
While(w_i equal to $w_{j'}$)
k++
End while
If k>1
extract the words from w_i to w_{i+k}
toform a repetition
k = 0
End if
End for
End for

Figure 1: The Algorithm Used For Repetition Extraction From Sentences

In Step 1, the repetitions are extracted by matching the same sequences between two sentences. For example, if we have the following two sentences:

- ರಾಮನತಲೆಗೆಬಲವಾದಗಾಯವಾಯೀತ್ತು,
- ಬಂಗಾರದಕಿರೀಟವನ್ನುರಾಮನತಲೆಗೆತೊಡಿಸಿದರು .

From the above two sentences, ''ರಾಮನತಲೆಗೆ" will be extracted as a repetition. The algorithm we used for extracting repetitions from sentences is shown in Fig. 1.

In Step 2, we conduct the POS tagging for repetitions using QTAG which is a probabilistic POStagger and can be downloaded freely.

In Step 3, semantic dependency of a pattern is produced as the maximum similarity, which is measured by either Rada method or Li method, of two nouns from the pattern. That is, assuming p is a pattern, and (w_1, w_2, . . .,w_n) is the noun words from p, the semantic dependency of p is defined as follows:

$$sd(p) = \max [sim(wi, wj)] \text{ i} \neq j, 1 \leq i, \text{j} \leq n$$

In Step 4, RP is set by trial and error practice.

Encoding Standards for Indian Languages
The two main standards in character representation of Indian languages are ISCII and Unicode:

1. Indian Script Code for Information Interchange (ISCII) is an 8-bit code and covers10 Indic scripts (Devanagari, Gujarati, Punjabi, Bengali, Assamese, Oriya, Telugu, Tamil, Malayalam and Kannada), ISCII uses extended ASCII in an 8-bit environment.
2. The Unicode consortium was initiated in January 1991,under the name Unicode Incto promote the Unicode Standard as an international encoding system for information interchange, to aid in its implementation and to maintain quality control over future revisions.Currently, Unicode is in version 4.1.0. The Unicode Standard provides with three encoding formats:UTF-8,UTF-16andUTF-32. Any one of these forms can be used to represent the Unicode characters. Each of these is used in different environments. The default encoding form of Unicode is UTF-16.

System Framework

The prototypical implementation of context-based term identification and extraction of our approach is illustrated in Figure 2 and it details algorithm is formulated as 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 thesource documents for term recognition and extraction.

Step 2: Discovery of domain taxonomy

Given the identified domain, related domain taxonomy is defined. In the experiment, a domain taxonomy corresponding to the domain is adopted from Indian Culture website.

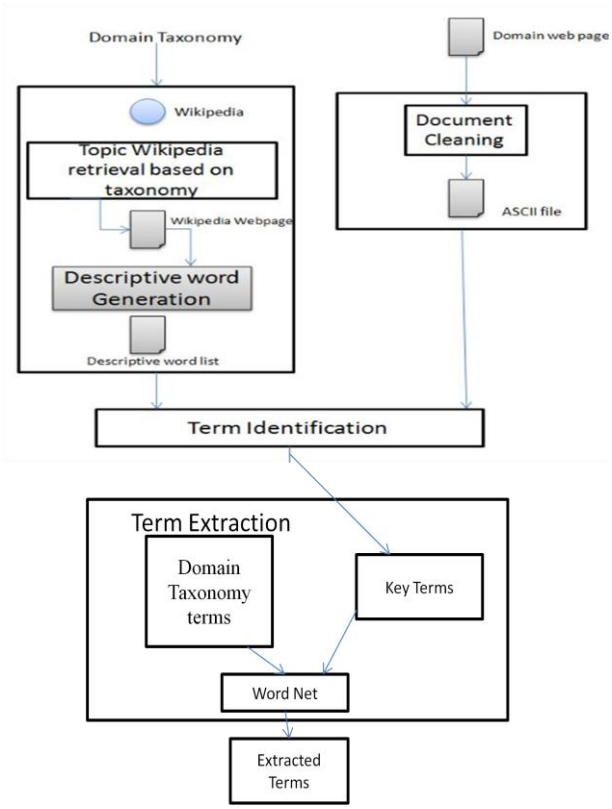

Figure 2: System framework

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.

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:

$DW = \{dw_1, dw_2, ...dw_n\}$

Step 5: Document purging

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 sentenceswhere 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.

Experiment and Evaluation

In this section, a series of experiments were carried out to evaluate our method on terminology extraction from Kannada documents. Both Rada method and Li method were used to rank the semantic dependencies of the noun words in patterns.

Corpus

Reuters-21578 text collection was applied as our experimental data. It appeared as Reuters-22173 in 1991 and was indexed with 135 categories by personnel from Reuters Ltd. in 1996. By our statistics, it contains in total 19403 valid texts with average 5.4 sentences for each text. Because these documents are mostly short passages and there are not enough sentences in each one of them to extract the repetitions, we only fetched out 196 documents whose sizes are larger than 4 K from the corpus.

Experimental Design

For convenience of evaluation, a standard terminology base for 30 documents randomly selected from the target 196 documents is constructed manually. In order to extract repetitions from documents, individual sentences are aligned using the sentence boundary determination method described in Weiss (2004). Thus, 8694 sentences with 139,836 words are aligned for the 196 target documents. Then the repetition extraction method depicted in Fig. 1 was utilized and 7945 repetitions were produced. Next, QTAG is used to conduct the POS tagging for repetitions and the

regular expression in JK method employed to extract the patterns, i.e. the final terminologies in JK method,

Evaluation Metrics

The most challenging activity in term extraction lies in its evaluation method as there is no formal way to evaluate terms. 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 identification and extraction were evaluated using two metrics:Precision and Recall as below:

$$Precision = \frac{\text{No of correctly extracted term}}{\text{Total term extracted by the system in Investigated topic}}$$

$$Recall = \frac{\text{No of correctly extracted term}}{\text{Total terms exist in the Investigated topic}}$$

Conclusion and Future Work

The work proposed in this paper is meant to provide a better way for term identification and extraction by taking into consideration of different topics might occur in a domain corpus for supporting ontology construction. The multi-topic is represented in document as a multi-level tree representation.

In this paper, we adopt ontology to improve the performance of terminology extraction from documents. Firstly, we present a review of current trends on terminology extraction. Secondly, ontology and two popular methods of words similarity measure within the framework of ontology were introduced. Then, JK method was adopted to extract the terminological candidates from documents and ontology methods were adopted to rank the semantic dependencies of the terminological candidates with the goal to improve extraction precision. Finally, we carried out a series experiments on terminology extraction from the Reuters-21578 corpus.

As far as future work is concerned, terminology extraction is still of our interest. We will combine the statistical and linguistic methods based on their superiorities for solving this problem. On the other hand, ensemble ontology method such as combing Mesh ontology and WordNet ontology will be attempted for text clustering and information retrieval, so that the contextual and background knowledge can be integrated into practical intelligent information processing applications.

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 Informatics IEEE/SOLI, Beijing, China, pp. 801-805, October 2008.

[3] Fonseca. F. T., and Egenhofer. M. J., 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] hang. 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 Workshop 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. O., and Zielinski. D., and Ties. I., and Voltmer, L., "Term Extraction for Ladin: an Example Approach", TALN: TraitementAutomatique des LanguesNaturelles, 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 Franzén. 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. I., 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 Biomedicine", Text Mining for Biology and Biomedicine, S. Ananiadou and J. McNaught (eds), Artech House, London, Ch.4, pp. 67-98, 2006.