Volume 5, Issue 5, May 2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Mining Textual Information via Linguistic Resources: Lexicon and Synset

Beenu Yadav[*]
Assistant Professor
College of Professional Education,
Meerut, U.P., India

Sunita Yadav
M. Tech (CS) Scholar
Translam Group of Institutions,
Meerut, U.P., India

Minu Yadav
MBA
CCS University,
Meerut, U.P., India

Abstract— Concepts are sequences of words that represent real or imaginary entities or ideas that users are interested in. As a first step towards building a web of concepts that will form the backbone of the next generation of search technology, we develop a novel technique to extract concepts. This paper proposes a technique to extract concepts from plain text. The extraction is based on existing linguistic resources like lexicon and synset. A Universal Decimal Classification is associated with each concept to classify the concepts. The noun/noun phrases are extracted from the preprocessed text which are input to the concept extractor which extracts the potential nouns as the concepts. It uses Statistically indexed table which is generated with the validation of the concept in text. Those concepts are extracted which are occurring most frequently in the text. This technique helps to extract the concepts from the plain text using linguistic resources. It can be applied in multilingual environment because of the use of linguistic resources for the concept extraction and hence expand the domain of input.

Keywords— Lexicon,, Sysnset, Universal Decimal Classification (UDC), Statistically Indexed Table.

I. Introduction

Mining textual information presents challenges over data mining of relational or transaction databases because text lacks any predefined fields, features or standard formats. However, mining information from unstructured textual resources such as email archives, news articles, and the World Wide Web has received much recent attention. One approach for effectively mining relevant information from raw text is based on finding common concepts in a given document. The field of Information Retrieval is mainly concerned with the specific problem, while we are primarily interested in concept identification. The concept extractor is designed for the determination of concepts of the ontology. We have developed a method to robustly extract and identify structures which closely correspond to concepts from raw text. The nouns and the noun phrases are the keys which form concepts [7, 8, 9].

In this paper we describe a method for extracting some semantic features from raw text which are then linked together in a structure which represents the text's thematic content. For this purpose we scale some existing linguistic resources according to our requirement and design new components using some existing resources. The used resources and the components to identify the concepts from noun and noun phrase are explained further.

II. Lexicon

A Lexicon is a repository of words and knowledge about those words. A lexicon is a list of words together with additional word-specific information. It is a list of corresponding terminology in different languages, usually locale, industry or project specific [4].

There are a number of special cases, which are usually researched and produced separately from general-purpose lexicons.

Lexicon incorporates-

1. Collection of Words
2. Unique Id(s) respective to each word: It is a Universal Decimal Classification (UDC) that uniquely identifies the concepts. The UDC(s) are determined from the SynSet table.
3. The category to which the word belongs based on classification of concepts is attached. The classification of concepts is given in the forthcoming section.

The word extracted from text/document for the identification of concept may or may not be matched with any word from the collection of words in Lexicon. When word does not match with any entry of Lexicon directly then morphology [1, 2, 3] is used.

For Example, words like Networks, Leaves etc., are not found in Lexicon. In these words morphemes are -

1. Network, -s
2. Leaf, -ves

To identify UDC(s) for these words, these words are analyzed as sequence of morphemes so that one of the word forms gets matched in Lexicon.

III. SynSet Table

The SynSet Table is a table developed for the identifications of words possessing the same meaning. It is the collection of synonymous words with the attribute set. The unique identification number is given to the set of words that have the identical meaning and such set identify the unique concept.

To each unique concept we give *UDC (Universal Decimal Classification)* identification as its unique identification number [5]. The UDC is the world's foremost multilingual classification scheme for all fields of knowledge. An advantage of this system is that it is infinitely extensible, and when new concepts are introduced, they need not disturb the allocation of numbers to the existing concepts [6].

In every language there are some words that express multiple meanings when used in different contexts. The exact meaning of such word is determined from the context of sentence in which the word is used. For this purpose we attach an attribute set with such words in the SynSet Table. In case when a word with different meaning in different contexts is encountered then the attribute set is exploited for the identification of exact word.

Each row in the SynSet table consists of three columns.

a. The first column of every row has UDC.
b. The second column has synonymous words having the same concept.
c. The third column has Attribute Set. The motivation for this is to provide a framework for finding semantically sensible concept of a multi-contextual word provided by the Lexicon.

For Example,

TABLE 1 SYNSET TABLE

UDC	Synonym Set	Attribute Set
5/6:523.31.12	Space, Area, Volume, Region	one, two, or three dimensional; bounded, occupied by objects
5/6:528.93	Space, Outer Atmosphere	Related to solar system, beyond the earth's atmosphere, boundless

IV. Statistically Indexed Concept Table

Extracting concepts requires a technique that can retrieve the appropriate concepts from documents of any subject domain. Statistical indexing technology is accurate enough to compute extraction of concepts [8].

The Vibhakti Parser extracts the units, such as noun phrases; they can be used to depict concepts by computing their frequency across the document. The indexing can be accomplished by computing the statistical frequency of extracted noun phrases within each document in a collection. The Statistically Indexed Concept Table is constructed by entering each noun phrase with its UDC. The UDC is determined from Lexicon and SynSet table. The noun/noun phrases, their UDC identification and their count altogether shape the Statistically Indexed Concept Table. Example:

TABLE II STATISTICALLY INDEXED CONCEPT TABLE

Row No.	Nouns/Noun Phrases	Frequency	UDC
1	TCP/IP, TCP and IP	7	681.324.003
2	Local Area Network, LAN, LAN operations	3	681.324.001
3	Computer Networks	5	681.324

The frequency index of each noun/noun phrase changes while the document is read. The frequency index of the table corresponding to each concept determines the validated concepts of the ontology.

V. Concept Extraction Method

This section outlines the methodology for figuring out the concepts for an ontology using above illustrated components and resources. Lexicon and SynSet Table are used to develop the Statistically Indexed Concept table, which is used to determine the concepts for the ontology. The step wise procedure is given as:

1. The word/phrase is extracted from the sentence to determine its concept.
2. This extracted word/phrase is mapped to the Lexicon. The Lexicon consists of UDC(s) relative to each word. These Unique Id(s) is used to find the concept(s) from SynSet table.
3. There may be more than one Unique Id corresponding to each word, which indicates that the word is used in different senses or contexts. The context of the extracted word is resolved using Attribute Set which is defined in SynSet Table.
4. The Unique Id found by the concept extractor is searched into the Statistically Indexed Concept Table. If it is found then the frequency corresponding to that Unique Id is increased by one and the extracted noun/noun phrase is appended to the Noun/Noun Phrase column.
5. For each extracted word/phrase

a) If the extracted word/phrase has one UDC in the Lexicon then this identification is fed into Statistically Indexed Concept Table.

b) Otherwise the complete sentence is read and the SynSet table is referred to determine its unique concept. With the help of Attribute Set and the sentence, the unique concept of the word/phrase is determined. Corresponding to the unique concept the UDC is identified and fed into the Statistically Indexed Concept Table.

c) Unique Id and the extracted noun/noun phrase are made as a new entry into the table with the frequency 1.

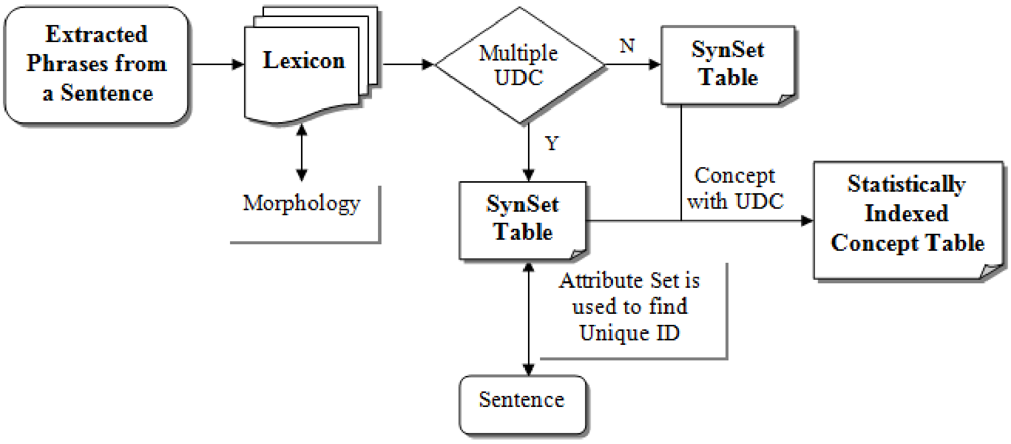
Extracted Phrases from a Sentence
Lexicon
Multiple UDC
N
SynSet Table
Y
Morphology
SynSet Table
Concept with UDC
Statistically Indexed Concept Table
Attribute Set is used to find Unique ID
Sentence

Fig. 1 Functioning of Concept Extractor

VI. CONCLUSION

Thus the Concept Extractor determines concepts from the input document. The Concept Extractor gives Statistically Indexed Concept Table as the output. The concepts are finalized according to the frequency column's entry of the table. Hence, Concept Extractor is an important tool for mining textual information.

REFERENCES

[1] Anusaaraka: Overcoming the Language Barrier In India. [Online]. Available: http://www.iiit.net/ ltrc/Publications/anuvad.html.

[2] Morphology (Linguistics) [Online]. Available: http://en.wikipedia.org/wiki/Morphology_%28linguistics %29.

[3] Ruslan Mitkov, 2004, "The Oxford Handbook of Computational Linguistics". pp. 25-40.

[4] Lexicon. [Online]. Available: http://en.wikipedia.org/wiki/Lexicon.

[5] Om Vikas, Siddhartha Kar, Anil Kumar, Arun Kumar, "An Integrated WordNet for English, Hindi and Bengali" 2004.

[6] UDC Consortium, http://www.udcc.org/.

[7] Nuala A. Bennett, Qin He, Conrad Chang, Bruce R. Schatz, "Concept Extraction in the Interspace Prototype". [Online]. Available: http://www.canis.uiuc.edu/archive/techreports/UIUCDCS-R-99-2095.pdf, Technical Report, Digital Library Initiative Project, University of Illinois at Urbana-Champaign, 1999.

[8] Bruce R. Schatz, "The Interspace: Concept Navigation Across Distributed Communities",. [Online]. Available: http://www.canis.uiuc.edu/archive/papers/interspace.computer.pdf, IEEE Computer, 2002.

[9] Spela Vintar, Paul Buitelaar Martin Volk, "Semantic Relations in Concept-Based Cross-Language Medical Information Retrieval". [Online]. Available: http://www.dcs.shef.ac.uk/~fabio/ ATEM03/vintar-ecml03-atem.pdf, 2003.