International Journal of Computer Science and Informatics

Volume 1 | Issue 3 Article 16

January 2012

Identification of the Problems during the Implementation of Thesaurus for Hindi Language

Roochie Sharma
Department of Computer Science, Punjabi University, Patiala, India, roochie66@gmail.com

Dharam Veer Sharma
Department of Computer Science, Punjabi University, Patiala, India, dveer72@hotmail.com

Follow this and additional works at: https://www.interscience.in/ijcsi

 Part of the Computer Engineering Commons, Information Security Commons, and the Systems and Communications Commons

Recommended Citation

Sharma, Roochie and Sharma, Dharam Veer (2012) "Identification of the Problems during the Implementation of Thesaurus for Hindi Language," *International Journal of Computer Science and Informatics*: Vol. 1 : Iss. 3 , Article 16.
Available at: https://www.interscience.in/ijcsi/vol1/iss3/16

This Article is brought to you for free and open access by Interscience Research Network. It has been accepted for inclusion in International Journal of Computer Science and Informatics by an authorized editor of Interscience Research Network. For more information, please contact sritampatnaik@gmail.com.

Identification of the Problems during the Implementation of Thesaurus for Hindi Language

Roochie & Dharam Veer Sharma

Department of Computer Science, Punjabi University, Patiala, India
E-mail : roochie66@gmail.com, dveer72@hotmail.com

Abstract **-** Well designed thesaurus reveals its results in the applications of browsing and information searching in the document. Thesaurus is an important tool, which is well suited to find more and/or better terms during writing and reading the documents. Such monolingual thesaurus for Hindi language has been presented in this paper. A thesaurus contains synonyms (words which have basically the same meaning) and antonyms, which is important for many other applications in NLP too. Implementation of thesaurus includes various challenges the foremost being designing the database for synonyms and antonyms for efficient retrieval of results. Therefore this paper describes the various problems that arise during the implementation of monolingual thesaurus for Hindi language.

Keywords **-** *Content-based data mining queries, organizing temporal patterns, mark-based indexing methods.*

I. INTRODUCTION

Natural language processing (NLP) is concerned with the interaction of computers and human languages. Thesaurus is one of the research works of NLP. Thesaurus is usually embedded in an application system such as document analysis system or a text retrieval system. Thesaurus lists words which are grouped together according to the similarity of meaning called synonyms and sometimes contain their antonyms. Synonym is a word or phrase that is perfectly substitutable in a context for another word or phrase. It may be possible that the substituted word is not the ideal synonym.

Monolingual means concerned with only one language as in this case it is Hindi. Thesaurus is a controlled vocabulary. The primary purpose of vocabulary control is to achieve consistency in the description of content objects and to facilitate retrieval [8]. We go to thesaurus when we have an idea, some concept or meaning in our mind but we are unable to get just the right word that fits our need or when we want to put more weight behind a concept by using some more appropriate word..

Thesaurus helps to understand the meaning of a term. For example: if user is not aware of word “tremor” then while going through the list of words of synonyms user may understand the word tremor i.e. earthquake or quake.

The paper is divided into seven sections: section two discusses the need of a thesaurus, section 3 covers the related work already done, section 4 covers the properties and nature of Hindi language, section 5 discuss the thesaurus for Hindi language section 6 details the problems faced during development of thesaurus for Hindi language and section 7 concludes the paper and references are given at the end.

II. NEED FOR THESAURUS

Tool like thesaurus is essential because all documents and queries are expressed in language. Language is complex and ambiguous. Ambiguity means same word having different meaning in different context i.e. Mercury (is a Planet) and Mercury (is a Metal) also. For example: we know “cycle” is an object which is named as “cycle” but we can also call “cycle” as “Bicycle” or “Push-ride”. Likewise “Beam balance” is also called “pair of scales” or “balance” or “scales”. It is difficult to decide that which one is to choose? Methods for solving the language issue are difficult. Even some systems don’t attempt to deal with such issues. Such problems of ambiguity are resolved by thesaurus, as thesaurus helps us to understand the meaning of term. It also helps to express and improve the sentence, paragraph and queries of the document in a better way. Thus, thesaurus is organized for us to help and find those words that we want, but cannot think of.

III. NATURE OF HINDI LANGUAGE

Hindi is an Indo-Aryan and a national language of India. Hindi share the title of India's constitutionally recognized national language with English. It is the mother tough of "Hindi Belt" of north and central India,and it is the world's fourth major spoken language.

To write Hindi (हिन्दी), Devanagari **(**देवनागरी**)** is a widely used script (लिपी/lipi). Each devanagari script characters represents a syllable, not a alphabet. It is written from left to right. Alphabet of Devanagari called 'varNNamaalaa' (वर्णमाला; varnamala). Varnamala is also called 'Aksharmala' (अक्षरमाला; AkShar-maalaa). Vowels and consonants together are called AkShars. The basic units of the writing system are referred to as *Aksharas*. The shape of an Akshara depends on its composition of consonants and the vowel, and sequence of the consonants.

India is a country with rich diversity in languages, culture, customs and religions. But, the language is making hindrances in the advantages of Information Technology revolution in India. So, there is the need of the adequate measures to perform natural language processing through computer processing so that computer based system can be interacted by users through natural language like Hindi and handled by users who have knowledge of regional language.

We require such a tool which can resolve all the Indian queries written in Hindi. With the use of thesaurus they can improve their vocabulary also. Yet many of the major languages of India have no thesaurus till date.

IV. HINDI THESAURUS

The idea of Hindi Thesaurus is inspired by the English Thesaurus. Hindi Thesaurus is such a tool which is important to the country like India where a very large fraction of the population is not conversant with English and consequently does not have access to the vast store of information that is available in English on the internet. In India, there are also many people who know English, but not fluent enough to be able to formulate their queries in it. Moreover Hindi is the official language of India.

In Hindi language, synonyms and antonyms are called समानार्थक शब्द/ पर्यायवाची शब्द and विपरीतार्थक शब्द/विलोम शब्द respectively. Thus Hindi thesaurus is defined as a collection of words of समानार्थक शब्द and विपरीतार्थक शब्द.

The biggest advantage to thesaurus is that once we find the correct term; all other relevant terms are grouped together in one place under all of the other synonyms for that term and antonyms, when sometimes user want to know the term with opposite in their meaning. Using a thesaurus routinely can help to expand a writer's vocabulary. Most of the Indian languages have letters that sounds mostly alike. Example श (sha), ष (sha), स (sa). This is the difficult part to recognized the words if pronounced.

In Microsoft Word you can look up a word quickly if you right-click anywhere in your document, and then to find a synonyms for a specific word, either type the word in the task pane search field or highlight it in your document. Then list of all possible synonyms appear in the context menu. Likewise Hindi Thesaurus is worked for you.

For example if user select and right clicked on word "अमृत" then resultant words are shown in the popup menu and synonyms as well as antonyms are listed as shown below in the table 1.

TABLE I : LIST OF ALL THE RESULTANT WORDS OF "अमृत"

Synonyms		
पीयूष	सुधा	सोम
Antonyms		
विष		

V. PROBLEMS FACED

There are several difficulties that are arise while building the Hindi thesaurus. Thesaurus builders should keep in mind all the problems mentioned below, while building the thesaurus.

- **Ambiguity :** Ambiguity arises when same word having different meaning in different context. Ambiguity which won't be resolved by computers is वह आम खा रहे। (आम as mango)

आम आम आदमी की परिधि से परे हो गया है । **(**आम as common person**).** Here in the previous sentence 'आम' can have two different meaning 'Mango' and 'common person'. Another example "काल" is the word having two contexts. One represents as meaning "अंधकार" (i.e. Darkness) and second represents it as "यमराज".

- **Use of half Consonants and Use of Dot with Vowels:** Half consonants mean half character, In Unicode half consonants are formed if characters make use of **(**हलंत**)** (्) like क् ख् ग् घ् च् छ् ण् त् द् ध् न् etc. Some words are written with half consonants and sometimes

same words are written with the help of small dot called Anusvaar/अनुस्वार /bindu, placed above the word (ं) which is nasalized that is a nasal quality is added to the vowel sound. For example बन्द = बंद, लम्बा = लंबा and खण्ड = खंड.

- Different way to write same word. Some words with same meaning can be written differently example: पंजाबी = पंञ्जाबी, डाक्टर = डॉक्टर= डौक्टर

- There are many words which contain "-" special symbol for their continuation.

Example: छिन्न-भिन्न, टूटा-फूटा, अस्त-व्यस्त, तितर-बितर, टूक-टूक

- **Nukta letters :** There are 11 characters in Hindi that are carrying nukta (Diacritic Mark/ नुक्ता; ़) क़ ख़ ग़ ज़ ड़ ढ़ फ़ य़ ऩ ऱ ऴ and represent new sounds created by adding a nukta to another letter shape i.e. क ख ग ज ड ढ फ य न र ळ. Character are written with single key stroke as well as with double key stroke i.e. double key stroke means one with nukta (़) as well as with these क ख ग ज ड ढ फ य न र ळ characters.

For example: Single stroke लड़की = ल+ड़+क+ी. If written with double key stroke लड़की = ल+ड+़+क+ी.

- In thesaurus, identifying the words that are semantically related to one another is the major difficult task.

- **Selection of the word :** There are two ways to select the words for look up Hindi thesaurus. Fig.1 and Fig.2 has shown these two ways as below

Fig. 1: Select all characters of the word.

Fig. 2 : Put the blinked cursor on that word.

Both the ways of selection will give the correct popup button of Hindi thesaurus in the context menu. But are some difficulties with second method of selection.

There are two main reasons explained below as:

First reason is if document is written in Unicode format then both ways of selection of the word is done except to the cases when words contained nukta letters as mentioned above in (v) options. This nutka is considered as delimiter which separates the word into two parts if written with double key stroke.

For example : With double key stroke लड़की = ल+ड+़+क+ी in which (़) is considered as delimiter or end of the word, which split the word into two parts as ल+ड in first part and क+ी in second part.

Second reason is if word document is in Non-Unicode format then also we have to select the whole word. Because In Non-Unicode format there are many characters which are generated with the key stroke of different special characters as well as with the delimiters also. When these ~ ! @ # $ % ^ and * ( ) _ - + { } [ ] \ : ; ‘ , . / ? characters are encountered then as mentioned above because of the same reason, user need to select the whole word.

- **Context Menu:** A **context menu** (also called **contextual**, **shortcut**, and **popup** or **pop-up menu**) is a menu in a graphical user interface (GUI) that appears upon user interaction, such as a right mouse click operation.

Text that is written in the word document use different properties of MS word like make use of bullets, tables, hyperlink etc. When user selects the word to check the thesaurus facility, there are different types of context menus which are popped-up. For example: for simple text, text written within the table, text with bullets and numbering, text with the hyperlink properties etc; all generate different context menus. It is difficult to recognize the context menu, because there are near about 180 context menu.

VI. CONCLUSION

This paper provides the knowledge about the experiments and their effects involving in the applications of thesaurus. This paper also presented the various difficulties that are occurred during the implementation of Hindi thesaurus and various major challenges. The biggest challenge in constructing a thesaurus, therefore, is to find out the context menu for Hindi thesaurus in which it is popped-up and to identifying the words that are semantically related to one another. To implement Hindi thesaurus one should keep in mind all these difficulties as mentioned in section 5.

REFERENCES

[1] C. Lu, K.H. Lee, and H.Y. Chen, "TheSys-A Comprehensive Thesaurus System For Intelligent Document Analysis And Text Retrieval", In the proceedings of Third International Conference on Document Analysis and Recognition (ICDAR'95), vol. 2. 1995, pp. 1169-1173.

[2] O. Gerb´e, B. Kerherv´e, "A Model-driven Approach to SKOS Implementation", ICIW-

10.79. In the proceedings of Conference on Internet and Web Applications and Services, 2010, pp. 484-488.

[3] M. Amirhosseini, J. Salim, "Quantitative Evaluation of Simplicity Invisible Domain in Islamic Knowledge Organizations", In the proceedings of International Conference on Information Retrieval and Knowledge Management, 2010, pp. 119-124.

[4] A. Hosseinizadeh, "The Problematiques of Thesaurus Construction in Iran from the point of view of thesaurus makers", In the proceedings of A-LIEP, 2009, pp. 563-569.

[5] E.A.Fox, J.T.Nutter, T.Ahlswede, M.Evens, and J.Markowitz, "Building a large thesaurus for information retrieval", In Proceedings of the Second Conference on Applied Natural Language Processing, pages 101108, Austin, TX, 1988. ACL.

[6] M.Santhosh Kumar and Kavi Narayana Murthy, "Automatic Construction of Telugu Thesaurus from available Lexical Resources", In the proceedings of Creation of lexical resources for Indian languages computing and processing LRIL-2007, C-DAC. Mumbai, March 2007.

[7] R. Nikolai, A. Traupe, and R. Kramer, "Thesaurus Federations- A Framework for the Flexible Integration of Heterogeneous, Autonomous Thesauri", In the proceedings of the Advances in Digital Libraries Conference, 1998, pp. 46-55.

[8] National Information Standards Organization: Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies - ANSI/NISO Z39.19-2005, 2005.

◈◈◈