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
DOI: 10.47893/IJCSI.2012.1042  
Available at: https://www.interscience.in/ijcsi/vol1/iss3/16

This Article is brought to you for free and open access by the Interscience Journals at 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

Rochie & 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 tongue of “Hindi Belt” of north and central India, and it is the world’s fourth major spoken language.

To write Hindi (हिंदी), Devanagari (देवनागरी) is a widely used script (लिपि). Each devanagari script character represents a syllable, not a alphabet. It is written from left to right. Alphabet of Devanagari called ‘var/Namaala’ (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 समानाथर्क शब्द and विपरीताथर्क शब्द respectively. Thus Hindi thesaurus is defined as a collection of words of समानाथर्क शब्द and विपरीताथर्क 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), ग (sha), ग (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”**

<table>
<thead>
<tr>
<th>Synonyms</th>
<th>Antonyms</th>
</tr>
</thead>
<tbody>
<tr>
<td>पीयुष</td>
<td>दिव</td>
</tr>
<tr>
<td>सुधा</td>
<td>सोम</td>
</tr>
<tr>
<td>सोम</td>
<td>सोम</td>
</tr>
</tbody>
</table>

**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 “अंधकार” is the word having two contexts. One represents as meaning “अंधकार” (i.e. Darkness) and second represents it as “अंधकार”.

- **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 ख ग च घ ण त द ध न 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 = बङ्द, लम्बा = लङ्का and ख़ंड = खङ्क.

- Different way to write same word. Some words with same meaning can be written differently example: पंजाबी = पंज़ाबी, डॉक्टर = डॉक्टर=डॉक्टर
- There are many words which contain “-” special symbol for their continuation.
  Example: छियर-मिश्र, टूटा-फूटा, अल-अस्त, तितर-तिकर, डूक-डूक

- **Nukta letters**: There are 11 characters in Hindi that are carrying nukta (Diacritic Mark/ तूक, ◌) क ख ग ज झ ञ ढ ण त थ ध न च छ जौ जं झं ञं ढं णं तं थं धं नं चं छं जू जे ज़ ज़ं झ़ं ञ़ं ढ़ं ण़ं त़ं थ़ं ध़ं ऩं च़ं छ़ं ज़ू ज़े ज़़ं झ़़ं ञ़़ं ढ़़ं ण़़ं त़़ं थ़़ं ध़़ं ऩ़ं च़़ं छ़़ं ज़ूं ज़़ें ज़़़ं झ़़़ं ञ़़़ं ढ़़़ं ण़़़ं त़़़ं थ़़़ं ध़़़ं ऩ़़ं च़़़ं छ़़़ं ज़ूंं ज़़ेंं ज़़़ंं झ़़़ंं ञ़़़ंं ढ़़़ंं ण़़़ंं त़़़ंं थ़़़ंं ध़़़ंं ऩ़़ंं च़़़ंं छ़़़ंं ज़ूंंं ज़़ेंंं ज़़़ंंं झ़़़ंंं ञ़़़ंंं ढ़़़ंंं ण़़़ंंं त़़़ंंं थ़़़ंंं ध़़़ंंं ऩ़़ंंं च़़़ंंं छ़़़ंंं ज़ूंंंं ज़़ेंंंं ज़़़ंंंं झ़़़ंंंं ञ़़़ंंंं ढ़़़ंंंं ण़़़ंंंं त़़़ंंंं थ़़़ंंंं ध़़़ंंंं ऩ़़ंंंं च़़़ंंंं छ़़़ंंंं।

For example: Single stroke लड़की = ल+ड+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](image1.png)

  **Fig. 1**: Select all characters of the word.

  ![Fig.2](image2.png)

  **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 ~ ! @ # $ % ^ * ( ) _ + { } [] ; : , / ? 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** 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


[2] O. Gerb’e, B. Kerherv’e, “A Model-driven Approach to SKOS Implementation”, ICIW-
Identification of the Problems during the Implementation of Thesaurus for Hindi Language

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


