Importance of Hindi Thesaurus in NLP and Hindi Language Script Description

Roochie
Dept. of Computer Science and Engineering, Punjabi University, Patiala, Punjab, India

Abstract
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. The use of thesaurus in documents is considered to be the most important resource to writers. A thesaurus contains synonyms (words which have basically the same meaning) and antonyms, which is important for many other applications in NLP too. This paper describes the script of Hindi language and importance of Hindi thesaurus written in Hindi documents.

Keywords
Thesaurus; Natural Language Processing; Controlled Vocabulary; Synonyms and Antonyms.

I. Introduction

Natural language processing (NLP) is a field of computer science and linguistics concerned with the interactions between computers and human (natural) language. 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 is a reference tool which groups synonyms and antonyms together. In another words, Thesaurus lists words which are grouped together according to the similarity of meaning called synonyms and sometimes contain their antonyms i.e. ({rise, ascend} vs. {fall, descend}). And it is very useful when writer or reader wants to know or distinguish between the terms, because there are several possible terms designating a single concept.

People always relate the term thesaurus with the dictionary. But there is a difference between dictionary and thesaurus. A dictionary is typically organized in, alphabetical order so that you can quickly locate the word of interest and then you can get the correct spelling, pronunciation, meanings, usage, and other such pieces of information. A thesaurus, on the other hand, could be organized the words into groups that convey a specific meaning, in contrast to a dictionary, which contains definitions and pronunciations. Therefore difference between a dictionary and thesaurus is lying between, therefore, is more of structure and organization.

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 [1]. 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.

The paper is divided into seven sections: section II covers the properties and nature of Hindi language, section III discuss the thesaurus for Hindi language, section IV discusses the need of a thesaurus, and section V concludes the paper and references are given at the end.

II. Nature of Hindi Language Script

This section describes the nature of scripts of Hindi language, which is an Indian language. Such introduction is necessary as in Hindi language the alphabet set is very large. Most importantly this alphabet set and shape characteristics of Hindi language scripts are utilized in the development. Hindi is an Indo-Aryan and a national language of India, ranking 4th by majority spoken in the world. The number of languages listed for India is 418. Of those, 407 are living languages and 11 are extinct. Eighteen are constitutionally recognized languages written in a variety of scripts. But After "The Eighth Schedule to the Constitution", Indian Constitution contains a list of 22 scheduled languages or official languages [3]. Among these 22 languages, Hindi is one of the official languages of India and is the main language used in the northern states of Rajasthan, Delhi, Haryana, Uttarakhand, Uttar Pradesh, Madhya Pradesh, Chhattisgarh, Himachal Pradesh, Jharkhand and Bihar, and is spoken in much of north and central India.

To write Hindi (हिन्दी), Devanagari (देवनागरी), is a widely used script (लिपी/lipi) [2]. It is written from left to right. Alphabet of Devanagari called 'varNNamaalaa' (वर्णमाला; varnamala). Varnamala is also called 'Aksharmala' (अक्षरमाला; AkShar-maalaa). There are 13 vowels (Swars /स्वर), 33 consonants (vyaNjans /व्यंजन), 8 chihns (चिह्न/ Marks / chihnas) and 12 MaatraaEN (मात्राएँ / Vowel-Signs). Vowels and consonants together are called AkShars. The basic units of the writing system are referred to as Aksharas(अक्ष). The shape of an Akshara depends on its composition of consonants and the vowel, and sequence of the consonants.

A. Notable features

- Type of writing system: alphasyllabary / abugida.
- Direction of writing: left to right in horizontal lines.
- Consonant letters carry an inherent vowel which can be altered or muted by means of diacritics or matra.
- Vowels can be written as independent letters, or by using a variety of diacritical marks which are written above, below, before or after the consonant they belong to. This feature is common to most of the alphabets of South and South East Asia.
- When consonants occur together in clusters, special conjunct letters are used.
- The order of the letters is based on articulatory phonetics.

B. Marks

MARKS (चिह्न/ chihnas) A mark is called chihn (चिन्ह; chihna). Following chihn/s are as important as mukya-maatraa/s. Halant(Vowel Omission Sign/हलंत/◌्) Anusvaar (Nasal Consonant Sound Sign/अनुस्वार/bindu/◌ं) Chandrabindu (Nasalisation Sign/◌ँ)Visarg (Aspirate Sign?) Nuktaa (Diacritic Mark) (नुक्ता/◌़)

- **Halat.** In Unicode font, half consonants are formed if characters make use of (हलंत) (◌्) like क् ख् ग् घ् च् छ् ण् त् द् ध् न् etc. example खण्ड Half consonants mean half character.
- **Anusvaar.** 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. Example अंग
- **Chandrabindu** (meaning "moon-dot" in Sanskrit) it is a diacritic sign having the form of a dot inside the lower half

of a circle. It usually means that previous vowel is nasalized. In Hindi, it is replaced in writing by anusvaar when it is written above a consonant which carries a vowel symbol which extends above the top line [8]. It is to be written as 'Nn'.

- Visarg (Aspirate Sign?) The visarga has an "aspirate" sound. The visarga is denoted by "ḥ". For example, "aḥ" is pronounced as "aha."
- Nuktaa (Diacritic Mark) (नुक्ता/ ़) is a diacritic mark. Following two vyaNjan are very common in Hindi.

ड़ (Da+nukta) = Dda
ढ़ (Dha+nukta) = Ddha

India is 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. As all types of NLP tasks need a thesaurus [7].

III. Hindi Thesaurus

The idea of Hindi Thesaurus is inspired by the English Thesaurus. 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. Synonym is a word or phrase that is perfectly substitutable in a context for another word or phrase. For example: the word झंडा (flag) has synonyms झंडा, पताका, ध्वज, ध्वजा, केतु, केतन.

Likewise तारीफ़ (praise) has synonyms मोद, प्रसंनता, उल्लास, प्रमोद, प्रशंसा, तारीफ़, हर्ष, प्रसंन

Antonyms are the negative connotation of a particular word. Antonym is a lexical relation between word forms. For example: आग is the word having पानी as antonym.

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 "सुन्दर" then resultant words are shown in the popup menu and synonyms as well as antonyms are listed as shown below in the Fig. 1.

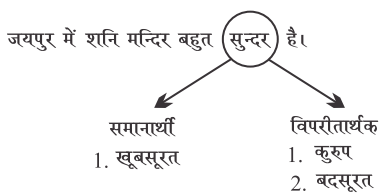

Fig. 1: illustration of Hindi thesaurus example "सुन्दर"

V. Importance of Hindi 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.
- **Synonyms help to express.** For example: we know "प्रभु" is an object which is named as "प्रभु" but we can also call "देवता" as "भगवान" or "परमात्मा". Likewise "मनुष्य" is also called "मानव" or "व्यक्ति" or "जन". It is difficult to decide that which one is to choose? Thus thesaurus 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.
- **Same word having different synonyms in more than one context.** 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. आम (as mango) and आम (as common) also. Another example कर as tax having synonyms कर, ब्याज, सूद, शुल्क, महसूल, टैक्स in one context and in another context कर as rays कर, किरण, रश्मि, मरीचि, अंशु, मयूख To eliminate ambiguities in case a term may carry multiple meanings/Thesaurus should provide a set of interface function for external access [4]. 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.
- **To understand the meaning of a term:** Synonyms and Antonyms provide better way of boosting your words power. It will certainly help in broadening the horizons of knowledge of the readers as well as writers also. 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.
- **Homonyms words**. There are different ways to write the same word. A word may be grammatically spelled in more than one way and all the forms may be acceptable [5]. For example: पंजाबी = पंजाबी, डाक्टर = डॉक्टर= डौक्टर It may be possible that the substituted word is not the ideal synonym. This is the case when the word is not the synonym one but it is a Homonyms.
- 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. During construction of thesaurus few problems occurred from the view point of thesaurus makers like during mismanagement within the principle of planning, organizing, innovation and administrative mismanagement between the team due to which gathering process of thesaurus database is lack behind. This should handle carefully.

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.

VI. Conclusion

This paper provides the knowledge about the Hindi scripts and their effects involving in the applications of thesaurus. This paper also presented the devanagari character set and some important

points which are required to understand before building thesaurus for document retrieval.

References

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

[2] "Languages and scripts of India". [Online] Available : http://www.cs.colostate.edu/~malaiya/scripts.html [Accessed on 20 May 2011].

[3] "Census of India- Data on Language", [Online] Available : http://censusindia.gov.in/Census_Data_2001/Census_Data_Online/Language/parta.htm [Accessed on 22 May, 2011]

[4] Lu, C.: Lee, K.H.: Chen, H.Y. "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.

[5] Kusum Kumar, Arvind Kumar, "Sahaj samantar kosh-hindi", [Online] Available : http://www.infibeam.com/Books/info/kusum-kumar-arvind-kumar/arvind-sahaj-samantar-kosh-hindi/9788126711031.html [Accessed on 7 June 2011].

[6] Hosseinizadeh, A. "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.

[7] Kilgarriff, A. "Thesauruses for Natural language processing", In proceedings of NLPKE, IEEE Beijing, 2003, pp. 513.

[8] Hindi Alphabets (2011), [Online] Available : http://www.hindidevanagari.com/transliteration/xnagari_scheme.html [Accessed on 20 June 2011]