THESAURUS FOR INDIAN LANGUAGES AND CONVERSION RULES DURING DESIGN OF PUNJABI THESAURUS

Aarti Tayal¹, Dharma Veer Sharama²
¹²Department of Computer Science, Punjabi University, Patiala, INDIA
er.aarti.tayal@gmail.com¹
dveer@hotmail.com²

Abstract: This paper is an attempt to discuss thesaurus already available in Indian languages and conversion rules during design of Punjabi thesaurus. Basically this paper is divided into four sections. In first section try to give introduction about varieties of thesaurus and brief introduction to thesaurus, ontology and dictionary. In next section, provide information about already built thesaurus for Indian languages. After that, give details of conversion rules which is basic necessity to provide correct result when user work under different-different fonts.

Keywords: thesaurus, ontology, dictionary, Indian languages, conversion rules

INTRODUCTION

Thesaurus’ can mean a number of different language resources, useful for a range of different language engineering purposes. The term "thesaurus" has its etymological root in the Latin word thesaurus, which in turn comes from the Greek word thesaurós. In both cases, the meaning was treasure or repository of words.

Varieties of Thesaurus with Definitions

One definition of Thesaurus is “a book lists words in group of synonyms and related concept” [1]. The area of word “related concept” in definition is very vast. It may be antonyms, related term, near term, broader term etc. This definition generally follows when basic necessity is composing text in any language. Another definition is “A thesaurus can be define as a controlled vocabulary arranged in a known order in which equivalence, hierarchical, and associative relationships among terms are clearly displayed and identified by standardized relationship indicators for purpose of improved retrieval”. The need for vocabulary control arises from two basic features of natural language, namely [2]:

• Two or more words or terms can be used to represent a single concept

Example: देव / देवी

• Two or more words that have the same spelling can represent different concepts

Example: बीज (force) बीविस बीउ बीउ बीविस (power) टाइम बीविस बीउ बीउ बीउ
The varieties include at least the following [3]:

- Roget Roget, Macquarie and others, produced, as books, to help writers with word selection
- WordNet WordNet and EuroWordNet
- IR-manual Thesauruses produced manually for use in information retrieval systems
- Automatic ‘Automatic thesauruses’, produced by processing corpora, with similarity between words measured (directly or indirectly) by co-occurrence.

Brief Introduction to Dictionary, Thesaurus and Ontology

A dictionary is a book that contains all the possible acknowledged words in a language. It is used to give explanations of words, show spelling, give different parts of speech like nouns and verbs; it gives the origin of words and also antonyms or opposite gender of a word. It is a book listing and explaining the words of a language [4].

A thesaurus is also a type of dictionary which gives associated words like synonyms or antonyms. A thesaurus is a book of words all grouped together because of the similarities in their meaning and origin. A thesaurus is a documentary tool used in the field of information representation and retrieval that represents a field of specific knowledge through its conceptual structure. This conceptual structure provides a semantic organization by making explicit the conceptual relations and restricting the meaning of the terms that represent them. The field of knowledge is structured based on hierarchical, associative equivalence-based conceptual relations. A thesaurus is used by both professional computer users and end-users [4].

Ontology is a formal, explicit representation of the conceptual structure of a field of knowledge. Ontology is a semantic support for words that are described as linguistic objects in a lexical or terminological database. The conceptual relations represented in ontology are extremely varied and depend on the field of knowledge to be structured. An ontology is constructed with the aim of sharing and reusing stored information, which, having been formalized, can be interpreted by both persons and computer programmers.

Introduction to Thesaurus In Terms Of Ontology and Dictionary

A thesaurus could be organized in terms of ontology - a hierarchy of concepts, and the words are structured into groups that convey a specific meaning. The difference between a dictionary and a thesaurus, therefore, is more of structure and organization rather than that of content. Both the dictionary and the thesaurus contain words of a given language and their meanings [1].
THESAURUS FOR INDIAN LANGUAGES

Thesaurus is developed for many Indian languages by experts. A brief introduction to all those thesaurus is following but still many Indian language exist for which thesaurus not exist. Development of thesaurus for Indian language is tough because of two major reasons.

- The alphabet set for Indian Languages is very large when compared to the any English language.
- Indian Languages has a complex script for representation.

**Thesaurus for Telugu**

Telugu is the second most spoken language in India, one of the twenty-two official languages of the Republic of India and one of the official languages of the state of Andhra Pradesh. Telugu has a vast and rich literature dating back to many centuries. Yet there is no widely available electronic thesaurus till date. In this work, a thesaurus for Telugu was generated automatically starting from two English-Telugu dictionaries. One was of these was developed by C.P.Brown and the other was developed as a part of a machine aided translation project. These dictionaries give more or less substitutable equivalents rather than elaborate descriptions or precise definitions [1].

More than 30,000 root words were extracted from the above two bilingual dictionaries. For each word in Telugu, corresponding synonyms are listed based on their category and also its sense in English. Total number of Telugu words is 30361. Average number of synonyms per word is 1.39. This could be higher if the dictionaries gave more number of equivalents. Maximum number of synonyms for a word in the thesaurus is 28. Maximum categories for a word are 5. The synset with maximum number of synonyms for a word in particular category is 9. The total number of synsets found in the thesaurus is 27558.

**Thesaurus for Assamese**

In Gauhati University, Digital Assamese Thesaurus project going on. Project Sponsored by UGC (Total Project value: 9.26 lacs). This project will output a structured digital Assamese Thesaurus which will be integrated with an interface enabling cross lingual information retrieval more efficient and meaningful.

**Thesaurus for Oriya**

Utkal University, Bhubaneswar developed e-Dictionary system. The basic objective of this system is to provide an efficient user friendly and reliable tool for searching of words. The system is designed by using the object-oriented paradigm to increase its extensibility, robustness and reusability for dynamic use in different application. Initially, the system is successfully running over 27000 Oriya words and 20000 English words. Search Engine of each language has been designed to handle the misspelled words and gives some most accurate suggestive words. The system is developed in file management system through Java and Java Swing for both the Windows and Linux operating system. It also provides sufficient interface to use in other application like Spell Checker, Thesaurus, Grammar Checker and Machine Translation for Oriya language [5].

**Thesaurus for Kannada**

Kannada, a language spoken by more than 50 million people and with vast and rich literature dating back to many centuries. A thesaurus for Kannada was generated automatically from an English-Kannada dictionary. This dictionary was developed by the author for the purpose of machine translation from English to Kannada. As such, the dictionary gave more or less substitutable equivalents rather than elaborate descriptions or precise definitions. Further, the dictionary was designed to contain a large number of synonyms since the choice of translated words is best left to the human post editor looking at the output of the machine and he must be given adequate choices to select from. These form the ideal conditions for automatic construction of thesauri by using technique from dictionary.

**Thesaurus in Gujarati**

Experts of state Gujarat developed GujaratiLexicon Digital Dictionary which is a desktop application. It includes Gujarati-Gujarati Dictionary, Gujarati-English Dictionary, English-Gujarati Dictionary, Opposites, Thesaurus, Idioms, Proverbs and Phrases. GujaratiLexicon Language Resources has been enriched by Gujarati dictionaries like bhagwadgondal, Bruhad, Sarth, Narm etc and other important language resources. When user click on word in document, application will shows all result related to that word including Gujarati - Gujarati meaning, Gujarati- English, opposite, phrases, thesaurus, idioms etc [6].

**Thesaurus in Hindi**

The three-volume Penguin English–Hindi/Hindi–English Dictionary and Thesaurus is a landmark in bilingual lexicography. Today, just as more Hindi-speakers than ever before are eager to master English, a large number of Indians and non-Indians are learning Hindi through the medium of English. Global communication and educational systems today demand a bilingual dictionary and thesaurus that covers a wide spectrum of social and cultural terms, both Indian and non-Indian. The Penguin English–Hindi/Hindi–English Dictionary and Thesaurus is a resourceful cross-cultural handbook that provides numerous references to help understand and appreciate the sense behind a word or concept in either language. When required, it provides short indicative definitions, examples, samples and references to similar and opposite concepts to further help absorb the import behind a word [7].

**DEVELOPMENT OF PUNJABI THESAURUS**

We tried to develop Thesaurus for Punjabi language, mother tongue of Punjab. Punjabi is the language of the Punjab - the land of five rivers - of northern India and Pakistan. Primarily written in three distinct scripts, a unique feature of the language is that, along with Lahnda and the Western Pahari dialects, it is the only modern Indo-European language spoken in South-East Asia which is tonal in nature. It is recognized as one of the several national languages of India and Pakistan, and approximately forty-five million people speak Punjabi as either a first or second language [8].

**Difficulty in Development of Punjabi Thesaurus**

For Punjabi thesaurus, main thing is preparation of Punjabi words database along with their synonyms and antonyms. There are two ways either developer can add Punjabi thesaurus as option to already context menu or to menu bar. After
that major work remains only retrieval of synonyms/antonyms from database corresponding to word in current document through Punjabi thesaurus option. But it is not enough, to provide correct output there is need of conversion rules. Because there are large number of ASCII based fonts currently available for Punjabi language. The availability of too many fonts makes thesaurus difficult as each font makes use of different keyboard mapping. Development of a font independent thesaurus system is a big challenge. Every user when write document has its own font. To make use of font independent thesaurus system, requirement to convert the written word in document by user into standard font first and then get synonym list from database. To provide correct result to user, the reverse conversion from standard format to user document depended font is essential.

**Conversion Rules**

- Main problem in Punjabi is key mapping of words in Unicode and non Unicode font is different. For example if you write मॅलिगा in Unicode font then key combination is "मॅलिगा" and in most of other non Unicode font, key combination is "मॅलिगा". Now if user uses non Unicode font, write word मॅलिगा and wants synonyms corresponding to this word. Process undertaken is first covert current word in Unicode font, search for converted word into database and retrieve synonyms corresponding to word. To implement first step, requirement is conversion to be follow letter wise. Letter मॅ is easily convert to Unicode font letter मॅ but when thesaurus facility encounter next letter फॅ in non Unicode font then need of attention because in Unicode font letter फॅ first must be appear. So when फॅ comes, store into temporary array say A and picks next letter first which is फॅ and convert it into Unicode font, result of this conversion store into temporary array say B. Then thesaurus facility retrieves फॅ from temporary array A and converts into Unicode font. Final step is concatenates result of this conversion with result store in array B. now it becomes मॅ + फॅ instead of मॅ + फॅ. Continue process of conversion with other letter of remaining word also [9].

- Conversion rule related to half letter like if there is word दिग्वि in any Non Unicode font it will make दिग्वि + दिग्वि + दिग्वि + दिग्वि and in Unicode same word when write; key combination is दिग्वि + दिग्वि + दिग्वि + दिग्वि. In this case, problem is same discuss in above problem. To give list of synonyms and antonyms to user, thesaurus facility first converts दिग्वि in non Unicode font to Unicode font. For this postpones the conversion of first letter of word फॅ in non Unicode font and continue with other letter till . Then convert फॅ in Unicode font and concatenate result of this conversion with conversion of letters फॅ, फॅ, and फॅ. After that continue conversion process with remaining letters of word.

- Conversion rules related to vowel letter like if there is word आमा then key combination in Non- Unicode font is आमा + आमा + आमा and in Unicode font is आमा + आमा + आमा respectively. In Unicode font आमा can never write with combination of आमा. So problem again arises when user write this word using non- Unicode font and wants synonyms corresponding to this letter. Thesaurus facility unable to give result for this input word because database in Unicode font. To solve this problem, try to make function in which there will be condition set when thesaurus facility encounter letter आ and आ in any non-Unicode font then replace it with आमा.

- If there is word लाबी and need to get synonyms of this word. There are 5 synonyms of this word which are लाबी, लाबी, लाबी, लाबी and लाबी. If user click on word लाबी it display as लाबी which is wrong result. Database is in Unicode font. Word लाबी in Sukhmani font first convert to Unicode font and then perform search for this word in database. When there will exact match find, all synonyms display as result which are also in Unicode. When user click on word लाबी to replace with लाबी, conversion take place to convert लाबी in Unicode font to लाबी in Sukhmani font. Problem arises because of difference of key combination in Unicode and Non- Unicode font. To provide correct result there must follow conversion rule whenever encounter आ at current position and आ at next position, concatenate both and reduce length of word by 1.

- When there is word लाबी and from option of synonyms for this word, user click on word लाबी, it display as लाबी which is incorrect. Solution to this problem is first conversion take place for letter आ, आ and आ. Then insert आ at length-2 position because after conversion of letter आ, आ and आ word length become 2. Conversion then continues with rest of letters. One more thing here to note, same solution will be follow when there are आ and आ half letter present at foot of any other letter with आ at starting position.

- In the above case, लाबी was destination word but suppose it is source word for which user want synonyms then again there problem arises. There are many other words which have same structure like लाबी, लाबी, लाबी, लाबी etc. For all these words problem is presence आ and आ together with one letter at starting position of word. Solution to this problem is manually. For each word need to add code separately which is not best solution but no any other way to solve this problem is available. Suppose user wants synonyms of word लाबी, then solution is whenever compiler encounters letter आ at next position of आ at previous position आ and आ is at 1st position then put आ in temporary array say A and shift at 3rd position of word which is आ in this case. Continue conversion for examining letters of word which is आ, आ and आ. Then in last, concatenates these conversion result with
In simple language, it is not wrong to say that a word is in Asses font. So before search for result corresponding to such words includes sanyukat akhar, there are chances of errors like consider words which can be at any position in word. During my research in Computer Science area I was able to find many words in which sanyukat akhar at starting position and make two cases depend upon position which are only three Sanyukat akhar ([5], [7], [9]) at foot position of other consonants then the program unable to give synonyms because database store in Unicode font by using conversion rule that whenever deal with these type of words, handle them as separate case.

CONCLUSION

There are also unique Sanyukat akhar ([2], [3], [4]) in which sanyukat akhar at starting position and i can be at any position in word. During my research in Computer Science area I was able to find many words in which sanyukat akhar at starting position and make two cases depend upon position which are only three Sanyukat akhar ([5], [7], [9]) at foot position of other consonants then the program unable to give synonyms because database store in Unicode font by using conversion rule that whenever deal with these type of words, handle them as separate case.

REFERENCES:

er

eReferences