International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue VI, June 2018- Available at www.ijraset.com

Interface Enabled Multilingual Information Retrieval using Structured Thesaurus Database

Dibyajyoti Sarmah[1] , Shikhar Kr. Sarma[2]
[1, 2]*Department of Information TechnologyInstitute of Science & Technology, Gauhati University,Guwahati – 14 Assam, India.*

***Abstract:** This paper presents a study about structure and development of digital thesaurus in Assamese Language. The focus of the paper is the various steps to developing a digital thesaurus which retrieves synonyms and other relative meaningful words from database. The primary objective of the work is to standardize the design and development of a structured and comprehensive digital thesaurus in Assamese language in Unicode format. Local language database and retrieval is essential for the development of communication skills in this global era. A digital thesaurus for Assamese languages with reference to three major context-language specific information retrieval, cross lingual (English-Assamese-Bodo) information retrieval using Wordnet and eventual usage for CLMT(cross lingual machine translation) is highlighted in the paper.*
***keywords:** Electronic Thesaurus,Cross-Lingual Information Retrieval,Wordnet,Multilingual Thesaurus Database.*

I. INTRODUCTION

In any natural language applications, word level meaning look-up plays an important role. Background dictionaries interfaced with the applications do the work as required for different technology enabled NLP activities. Thesaurus enhances efficiency and effectiveness of such applications by replacing the background word level dictionary with a comprehensive and structured thesaurus database. As the ICT applications are growing ,and with the emphasis from Government for using local language enabled information technology dissemination, development of language technology in local languages have started.

The computerized system helps in quick and better decision making which is in demand of the fast paced world with automaton of different process and system.

It gives two main advantages over human handled traditional paper based system. These are: classification systems into their user interfaces to provide support for query formulation ,collection browsing and other search tasks. Many interfaces have utilized graphical as well as two as two- or three-dimensional category hierarchies using the MsSH Thesaurus in English Language . TraversNet, MeSHBrowse,Cat-a-cone,Visual MeSH, and the integrated thesaurus-Results Browser are among the prototype thesaurus-enhanced interfaces. There are also some studies that have found that thesaurus-enhanced search interfaces can support users query formulation and expansion. It has been established globally that in order to facilitate cross-cultural communication in an increasingly global information society multilingual thesaurus can play a significant role. Search engine major like Google etc. have already integrated their interfaces with strong thesaurus in the respective language.

hesaurus has played an important role in modern information storage and retrieval systems. While initial proposals to utilize thesauri focused on their ability to ensure consistent analysis of documents during input to information retrieval systems, they have increasingly become vital as aids to effective retrieval. Milstead noted that in the near future, it appears likely that thesaurus will be used more during retrieval than at input. The move to increasing use of thesaurus as an aid to retrieval has expanded their functional span within information retrieval systems. As Aitchison et al. Have stated, the role of the thesaurus is changing, but it is likely to remain an important retrieval tool. This refocusing of the use of thesauri within information retrieval systems means that it is imperative that professionals take cognizance of the potential of thesauri as essential components of the largest information retrieval environment, namely the World Wide Wed

II. REVIEW OF RESEARCH AND DEVELOPMENT IN THE SUBJECT

Over the last decade a number of search engines,digital libraries and online initiatives have incorporated knowledge organization systems such as thesauri and

A. Objectives

The primary objective of the work is to standardize, design, and development of a structured and comprehensive digital thesaurus in Assamese language in Unicode format.

1) To configure the system architecture and the computational algorithms for the digital thesaurus in Assamese.

2) To develop algorithms which will serve for th
3) systematic and efficient mechanism for storage,
4) To develop a word level link list for Assamese-English in Unicode format.
5) Once the system architecture is finalized, to develop modules with proper flow of works integrated with the backend database, database interface, and user interface.
6) To integrate the three layers for the complete system development enabling cross lingual language specific information retrieval.

III. METHODOLOGY

A thesaurus has the following characteristics
If the same concept is expressed by two or more terms, one of these is selected as the preferred term. The relationship between preferred and non-preferred terms is an equivalence relationship.
A thesaurus is distinguished from an unstructured list of terms through use of hierarchical relationships based on degree or level of super-ordination and subordination of terms, where the superordinate preferred term represents a class or a whole, and the subordinate preferred terms refer to its members or parts.
Associative relationships are links between
preferred terms that are semantically or conceptually associated to such an extent that links between them should be made explicit.
Museum information systems contain many kinds of data that possess the above characteristics:
e.g. taxonomic names and associated classifications, classifications of cultural terminology such as the Art and Architecture Thesaurus, the Library of Congress Subject Headings, and gazetteers of political, social, and biological geography. This document examines alternative data structures for implementing these relationships in a SQL relational database management system, with particular attention paid to information retrieval flexibility and efficiency.

A. System Architecture

This overview diagram contains only a snippet of the whole Topic Map typology. To give a closer impression, figure1 shows all the topic types and association templates used to represent a classical thesaurus.
The design of the system will be based on the concept of rich prospect interfaces in which multiple representations can be used to allow access to a digital collection. In the proposed system, the aim will be to provide the user with the following spaces within the interface:

1) *Query space:* for formulating search statement
2) *Thesaurus space:* for browsing and navigating the thesaurus
3) *Document space:* for viewing representations.

The formulation and reformulation of queries is the primary goal of the interface proposed.

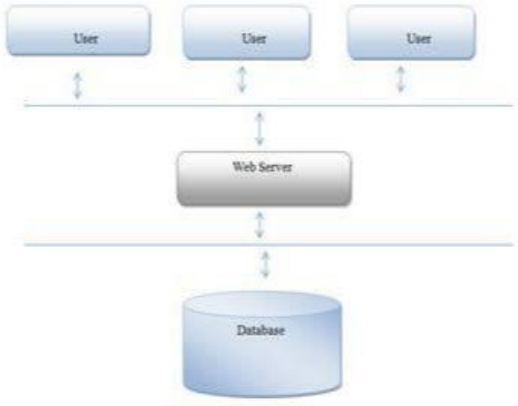


Figure 1. Architecture design

B. Basic Concepts

Before discussing data structures for implementing thesauri, properties of directed graphs and trees are introduced. These properties will be used to compare alternative data structures for implementing thesauri in a SQL relational database management system.

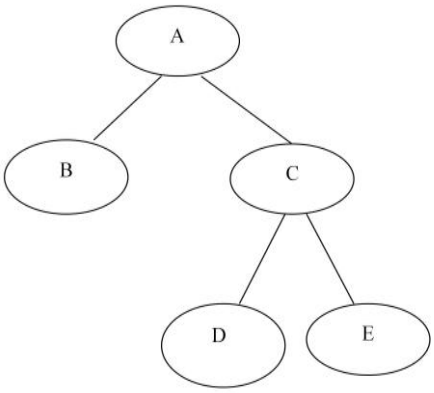
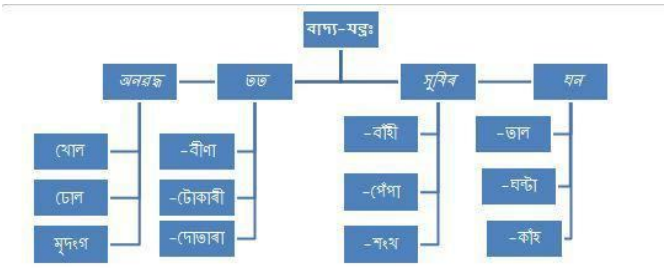


FIGURE2: Structure of Thesaurus

Figure 2. Tree with five nodes and four edges

The summary below presents definitions of the properties of trees needed for the evaluation of alternative data structures for storing and retrieving

hierarchical relationships of thesauri. A graph is a data structure that contains nodes connected by edges. In a directed graph, one node is designated the root of the graph and the graph is oriented so that the root lies at the top of the graph. Figure 1 is a graph with five nodes and four edges where the root is node A.



Assamese thesaurus example

Viewing a thesaurus as a directed graph, the terms in a thesaurus constitute the nodes and the super ordination and subordination relationships among terms constitute the edges. To describe super ordination and subordination relationships, it is necessary to specify the edge representation of the directed graph, i.e., pairs of adjacent nodes that are connected by an edge.

The node that lies above and is connected to another node in the directed graph is a parent node and the node that lies below the parent node is a child node. In Figure 1, node B is the parent of nodes D and E which are the child nodes of B. Parent / child node relationships are transitive. In Figure 2, because B is a child of A and C is a child of B then C is a child of A.

The nodes of the directed graph that have no nodes lying below them are designated leaf nodes. In Figure 1, nodes C, D, and E are leaf nodes. Sibling nodes are nodes that share the same parent. In Figure 1, nodes B and C are sibling nodes.

A cycle occurs in a directed graph if there exists more than one path between any pair of nodes. If there are no cycles in a directed graph, then there exists a unique path between any pair of nodes in that graph. A directed graph with no cycles is termed a tree. Hereafter, we discuss data structures for thesauri represented as trees.

The path between two nodes is the set of edges and nodes that connect these two nodes. The depth of a node is the number of edges separating that node from the root. The depth of node D in Figure 2 is 3.

A subgraph is a set that includes one parent node, all of its child nodes, and the edges connecting them. Nodes B, D, and E in Figure 1 are the members of

the subgraph for which node B is the parent node. The width of a subgraph is defined as the total number of child nodes descending from the parent node. For the subgraph in which the parent node is B, the width is 2 (Figure 2).

There are a number of ways to systematically order the nodes in a tree, each of which produces a method of traversal. One example, the preorder traversal, is achieved by enumerating the nodes in the order encountered by starting at the root node and then going to the next deeper, leftmost, unencountered node until a leaf node is reached (e.g., nodes A, B, D in Figure 1). After encountering a leaf node, return to its parent and continue the search as above (e.g., node E in Figure 2). Continue traversal by these rules until all nodes have been encountered. The full preorder traversal of the tree in Figure 2 is A, B, D, E and C.

C. Structure of Thesaurus Data

A thesaurus deals with terms and concepts and the relationships among and between these entities. A term is a linguistic entity, a character string with meaning in a given language. If the same character string has two meanings, we have two terms (homonyms); most thesauri use parenthetical qualifiers to make each string unique. The same character string occurring in two different languages represents two different terms, even if the meaning is the same.

A thesaurus captures a great many relationships among terms, between terms and concepts, and among concepts. Term-term relationships include A has morphological variant B (such as job and jobs), A has spelling variant B (such as labor and labour), and A has synonymous term (ST) B. More precisely, has morphological variant relates character strings that are derivative from the same stem. Has spelling variant relates stems and partitions the set of all stems into mutually exclusive groups.Each such group constitutes a normalized term; a preferred spelling variant can be selected to represent the term. (To keep matters simple, we mostly sidestep the spelling variant problem in this paper.) Has synonymous term relates normalized terms (preferred spelling variants) and partitions the set of normalized terms into mutually exclusive groups. A concept can be operationally defined as such a group of normalized terms. A preferred term can be selected from each group to uniquely designate the concept. All preferred terms may be used as descriptors, or descriptors may be further selected from the preferred terms. These considerations give rise to a status hierarchy among all terms (character strings).

The primary term-concept relationship is Term A designates Concept B; this relationship is implied by has synonymous term relationships, unless a thesaurus identifies concepts independently, for example through

class numbers or notations. Concept-concept relationships include A has broader term B, A has narrower term B, A has related term B. This simplified picture presents clear-cut distinctions, but reality is not that simple. Normalized terms often represent shades of meaning so that it is hard to tell whether two terms are synonyms or whether they represent closely related but different concepts. If the two concepts that are so closely related that to distinguish between them would not be useful for retrieval, some thesauri use the relationship equivalent term (ET) at least in their internal database (in the user version they may map ET to ST). The ET relation can be seen as partitioning the set of concepts into mutually exclusive groups. Each group corresponds to a newly formed ISAR (Information Storage And Retrieval) concept which is broader than any concept in the group. A preferred term can then be selected for each ISAR concept. The equivalent term relationship is at the borderline between term-term relationships and concept-concept relationships. For the term-based model (discussed below) this does not present a problem, but for the concept-based model one must decide whether to treat ET as a term-term relationship or a concept-concept relationship.

D. Data Structures For Thesaurus Databases

Some computer systems for thesaurus construction and maintenance use a record for every term with the information about the term, such as synonyms, broader, narrower, and related terms, stored in — usually repeating — data fields in the record. Information is stored in large packages, and to access or change any piece of information we must get into the appropriate package. Even for an individual thesaurus such a structure is inflexible. For an integrated thesaurus data base it is unwieldy. For example, comparing two records for the same term from two different thesauri requires cumbersome processing of the two records.

The relational approach to data base organization leads to a more elegant and efficient structure. Information is stored in individual pieces that can be arranged in different ways. For example, an employment RT labor relation is a piece of information that is stored by it. Combining two thesauri stored in this format can be accomplished simply by putting all the pieces of information into one data base and eliminating duplicates. This structure has an additional advantage: Relationship types are not defined as fields in a record (and thus fixed in the database structure), but they are simply data values in a relationship record; thus new relationship types can be introduced with ease.

E. Linguistic Input preparations and its difficulties

Our detailed discussions on lexicographic requirements of modern Assamese language boiled down to the following points:

After the rapid urbanization of the Assamese society from the sixties of the last century and gradual mechanization of agriculture, weaving and transport system, along with introduction of English medium schools even in remote areas of the state, there is a radical change in lexicographic requirements of the Assamese language.

agricultural training *BT* agricultural education *BT* vocational training *RT* agricultural extension employment *ST* jobs *RT* labor relations *RT* vocational training labor relations *ST* industrial relations *ST*, *BT*, and *RT* are field labels	T1 agricultural training T2 agricultural education T3 vocational training T4 agricultural extension T5 employment T6 jobs T7 labor relations T8 industrial relations Term file assigning term numbers (2-column table) Relational database structure	T1 BT T2 T1 BT T3 T1 RT T4 T5 ST T6 T5 RT T7 T5 RT T3 T7 ST T8 Relationship file using term nos. (3-column b. table)

Figure3. Data structure for a thesaurus

A large number of traditional rural words are enclosed in the Assamese dictionaries but are no more in active use, at the same time, a large number of Assamese words which have Sanskrit etymology or are newly coined and included in the scientific and administrative glossary are missing in these classical Assamese dictionaries.

While a large number of words found in the Assamese vaisnava scriptures have been collected, included and explained in the dictionaries, the compilers and editors have never totally neglected ancient words found only in Assamese historical literatures.
The worst sufferers at the present time are technical words and name of the implements used in rural daily life and cottage-industries.
The word (জুলুকি)juluki is translated as a "a kind of fishing basket" and a palo(পল') "is kind of a bamboo-basket for catching fish."A reader will naturally infer that palo is made of bamboo,but juluki is not .in reality a 'palo' has curved neck but juluki is weaven with a straight neck,both of them are made of bamboo and came and used for catching fishes in shallow waters.
Leaving aside Assamese words of class noun and proper noun,there are many implements in daily use in rural Assam which cannot be properly translated into english language. An ural(উৰল) can be easily translated

into English as a mortar,but what about dhenki(ঢ েঁিকী)?"A foot operated rice pounder" or "see-saw grinder" do not present a correct picture. A monolingual dictionary should deal with noun,adjective,verb etc,derived from the same word than the image it creates in the mind. Therefore the brain takes over and the appropriate grammatical forms automatically come to the mind. Therefore it was decided to economize space, and only one form of the word, be it noun, adjective, verb and rarely adverb would be included.
During our work in collection of words, another problem cropped up. Thousands of Assamese words were collected by us during compilation of our Anglo-Assamese dictionary which are mostly class noun or proper noun words, names of fishes, vegetables, implements, trees etc.

F. Queries

Implementation strategies for thesauri must facilitate data retrieval based on the properties of trees described above. The following query tyes should be accommodated:

1) find any node in the tree - this query meets the specification in for retrieving an individual term and displaying its relationships
2) find the path from the root to any node in the tree - this query provides the ability to display a thesaurus term in its hierarchical context.
3) find a subtree - this query meets the specification in for retrieval of terms at any level within a thesaurus .
4) find a set of nodes with an associated property - this query meets the specification in to retrieve set of terms in a thesaurus based on linked data .
5) find a set of data that have been organized using terms in a thesaurus - this query enables the retrieval of data sets that have been cataloged using a thesaurus .

These query types are not the only operations that may be performed on trees. A primary basis upon which to evaluate alternative data structures for implementing thesauri is the ability to perform these queries in SQL in a simple and flexible manner using little or no procedural code.

G. Relationship INDICATORS

To describe relationships among terms in a thesaurus, defines relationship indicators for hierarchical, equivalence, and associative relationships. These relationship indicators define a syntax for term relationships. To describe hierarchical relationships, the relationship indicators, "broader term" (BT) and "narrower term" (NT) are used. If, for two terms A and B, "A BT B "then "B NT A". The use of the BT - NT relationship indicators describe the edge representation for a thesaurus. Relationship indicators do not accommodate other properties of trees.
Equivalence relationship is described by the relationship indicators "use" (USE) and "use for" (UF). These indicators show the relationship between a preferred term and its synonyms. If, for two terms A and B, "A USE B", then "B UF A". Associative relationships are described by the relationship indicator "related term" (RT). Terms that are conceptually or semantically related are associated using RT in order to suggest alternative terms for indexing or for retrieval.
We have extended the set of relationship indicators used in to accommodate relationship types encountered in biology. Relationship indicators to accommodate the taxonomic names of hybrids, taxonomic names without authors, common names, and classificatory rank for the terms used in taxonomic names have been included (see Table 1). The relationship indicator, VT, is added in order to facilitate retrieval.

Relationship indicator abbreviation	Relationship indicator name
AVT	Alternate Valid Term
BT	Broader Term (Generic)
FHP	Female Hybrid Parent
HP	Hybrid Parent
MHP	Male Hybrid Parent
RT	Related Term
TCN	Taxonomic Common Name
TNNA	Taxonomic NameNo Author
TRE	Taxon Rank Epithet
TVN	Tree View Name
UFT	Used For Term
USE	Use
VT	Valid Term

Table 1. Relationship indicators

IV. CONCLUSION

In this paper, we have discussed the basic structure of Assamese digital thesaurus extracting information from database. Extracting information from database can be done for single words as well as a large list of words to create a database. The extracted information consists of the input word in Assamese language.

REFERENCES

[1] Sarma,Dr Shikhar Kr.,Moromi Gogoi,Rakesh Medhi and Utpal Saikia,Foundation and Structure of Developing an Assamese Wordnet**,** Global Wordnet Conference 2010, IIT Bombay
[2] , C. 1998. Wordnet: An Electronic Lexical Database. The MIT Press.
[3] Kamil, Bulke. 1997. An English-Hindi Dictionary (ed.). S. Chand & Co, New Delhi, India.
[4] Sarma, Shikhar Kr., Moromi Gogoi, Rakesh Medhi and Utpal Saikia, 2010. Foundation and Structure of Developing an Assamese Wordnet, Global Wordnet Conference, IIT Bombay.

[5] Sarma, Shikhar Kr.,Moromi Gogoi, Biswajit Brahma, Mane Bala Ramchiary,2010. A Wordnet for Bodo Language: Structure and Development, Global Wordnet Conference,IIT Bombay
[6] Deka Pranavjyoti, Jyoti Bilingual Thesaurus ASSAMESE & ENGLISH, An imprint of K B Publication, 200
[7] Deka Pranavjyoti, Anglo-Assamese Dictionary Jyoti Divashik Abhidhan, Assam Book Depot, 199
[8] Deka Pranavjyoti, Jyoti Anglo-Assamese bilingual dictionary cum Assamese thesaurus , Pranav Jyoti Deka, 199
[9] Sinha, Manish., Mahesh Reddy and Pushpak Bhattacharyya. 2006. An Approach towards Construction and Application of Multilingual Indo-WordNet.3rd Global Wordnet Conference (GWC 06), Jeju Island, Korea, January, 2006
[10] Kotoky Prafulla, Samarthasabdakosh, Jyoti Prakashan, 200
[11] Jalote Pankaj, An Integrated Approach to Software Engineering, Narosa Publishing House, 200
[12] William J. Black and Sabri El-Kateb,A Prototype English-Arabic Dictionary Based on WordNet
[13] Debasri Chakrabarti and Pushpak Bhattacharyya,Creation of English and Hindi Verb Hierarchiesand their Application to Hindi WordNet Building and English-Hindi M
[14] Nitin Verma and Pushpak Bhattacharyya, Automatic Lexicon Generation through WordNet
[15] Paolo Rosso1, Edgardo Ferretti2, Daniel Jiménez1, and Vicente Vidal1, Text Categorization and Information Retrieval Using WordNet Senses
[16] ha., S., D. Narayan, P. Pande, P. Bhattacharyya. 2001. A WordNet for Hindi.Workshop on LexicalResources in Natural Language Processing, Hyderabad,India, January, 200
[17] Murat Dhwaj Singh, Jayanta Chatterjee, and T.V. Prabhakar,Building Multilingual Agriculture Thesaurus in Indian Languages
[18] S.J. Darmoni, J. Grosjean, T. Merabti , N. Griffon,B. Dahamna, D. Dutoit,Combining WordNet and Crosslingual multi-terminology health portal to access health information
[19] Martin Doerr ,Semantic Problems of Thesaurus Mappin
[20] Amba, S., N. Narasimhamurthi, K.C. O'Kane and P.M. Turner (1996) "Automatic linking of thesauri". In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Konstanz:Hartung-Gorre, pp. 181-187
[21] Constantopoulos, P., M. Sintichakis (1997) "A Method for Monolingual Thesauri Merging". Proc. 20th International Conference on Research and Development in Information Retrieval, ACM SIGIR, July, Philadelphia, PA, USA
[22] Dachelet, R. (1997) "Multilingual querying and multilingual thesauri in Aquarelle", Technical Report, INRIA-Aquarelle, Mar
[23] Getty Information Institute (1996) Guidelines for Forming Language Equivalents: A Model Based on the Art&Architecture Thesaurus, International Terminology Working Group