Malayalam Wordnet: A Relational Database Approach

Mujeeb Rehman O.
Department of Computer Science and Engineering,
Govt. Engg. College, Sreekrishnapuram,
Palakkad, India-678633

P. C. Reghu Raj
Department of Computer Science and Engineering,
Govt. Engg. College, Sreekrishnapuram,
Palakkad, India-678633

ABSTRACT: WordNet is a hierarchical information base in any language. A WordNet is implemented using indexed file system.Even though there are many languages in which we have good wordnets, Malayalam is not having an efficient wordnet. This paper suggests a method for implementing wordnet using database and shows how it is more efficient than the traditional system. We also discusses about the implementation details of a Malayalam wordnet using Mysql database.

Keywords – Malayalam WordNet, WordNet approaches, WordNet implementation

I. INTRODUCTION

In the area of Natural Language Processing, WordNet plays an important role. Wordnet is a semantic dictionary that was designed as a network following the idea that representing words and concepts as an interrelated system is consistent with evidence for the way speakers organize their own mental lexicons[1].
Nowadays , WordNets are available in many Languages. But in Malayalam there is not a good wordnet available yet. Frist, we look at some general properties of WordNet.

A. Properties of a wordnet-

A WordNet can provide the following information:

- Synonymy: This one is easy and links words that have similar meanings, e.g. happy and glad.
- Antonymy: The opposite of synonymy, e.g. happy and sad
- Hypernymy: Hypnernymy refers to a hierarchical relationship between words. For example, furniture is a hypernym of chair since every chair is a piece of furniture (but not vice-versa).
- Hyponymy: Hypnonymy is the opposite of hypernymy. Dog is a hyponym of canine since every dog is a canine.
- Meronymy: Meronymy refers to a part/whole relationship. For example, paper is a meronym of book, since paper is a part of a book.
- Troponymy: Troponymy is the semantic relationship of doing something in the manner of something else. For example, walk is a troponym of move and limp is a troponym of walk.
- Entailment: Entailment refers to the relationship between verbs where doing something requires doing something else. If you are snoring, you must be sleeping so sleeping is entailed by snoring[3]

B. Structure of WordNet

The wordnet is implemented using indexed file system. In a conventional wordnet have the files like adj.exc, data.adj, data.verb,index.noun etc.
For each syntactic category, two files are needed to represent the contents of the WordNet database - index. pos and data. pos , where pos is noun , verb , adj and adv . The other auxiliary files are used by the WordNet library's searching functions and are needed to run the various WordNet browsers.
Each index file is an alphabetized list of all the words found in WordNet in the corresponding part of speech. On each line, following the word, is a list of byte offsets (synset_offset s) in the corresponding data file, one for each synset containing the word. Words in the index file are in lower case only, regardless of how they were entered in

the lexicographer files. This folds various orthographic representations of the word into one line enabling database searches to be case insensitive.
A data file for a syntactic category contains information corresponding to the synsets that were specified in the lex icographer files with relational pointers resolved to synset_offset's. Each line corresponds to a synset. Pointers are followed and hierarchies traversed by moving from one synset to another via the synset_offset's.
The exception list files, pos .exc , are used to help the morphological processor find base forms from irregular inflections. The files sentidx.vrb and sents.vrb contain sentences illustrating the use of specific senses of some verbs. These files are used by the searching software in response to a request for verb sentence frames. Generic sentence frames are displayed when an illustrative sentence is not present.
The various database files are in ASCII formats that are easily read by both humans and machines. All fields, unless otherwise noted, are separated by one space character, and all lines are terminated by a newline character. Fields enclosed in italicized square brackets may not be present[3].

II. RELATED WORK

WordNet was created at the Cognitive Science Laboratory of Princeton University under the direction of psychology professor George A. Miller. Wordnet is still maintained by the Cognitive Science Laboratory[4]. Development began in 1985. Over the years, the project received funding from government agencies interested in machine translation. As of 2009, the WordNet team included the following members of the Cognitive Science Laboratory: George Armitage Miller, Christiane Fellbaum, Randee Tengi, Pamela Wakefield, Helen Langone and Benjamin R. Haskell. WordNet has been supported by grants from the National Science Foundation, DARPA, the Disruptive Technology Office (formerly the Advanced Research and Development Activity), and REFLEX. George Miller and Christiane Fellbaum were awarded the 2006 Antonio Zampolli Prize for their work with WordNet.
Today well established WordNets are available at many languages. All of them are implemented in indexed file system.
So this paper proposes an alternative method using Relational Database. Also implements a Malayalam WordNet.

III. APPLICATIONS OF WORDNET

WordNet has been used for a number of different purposes in information systems. Also wordnet can play an important role in most Natural Language processing(NLP) tasks in Natural Language Generation (NLG).

WordNet can play a crucial role in Word Sense Disambiguation(WSD). WSD is the process of determining the correct sense of an ambiguous word in a context. One of the important methods of WSD is dictionary based.In this we search a WordNet for all senses of an ambiguous word. Then only we can move forward.

Another important application of wordnet is in the scenario of semantic parsing. Where the properties like hypernymy, hyponymy relations are used for the better understanding of a word.

In the area of Information retrieval, wordnet can contribute with the synonymy property for improving the query. By using the synonyms of user query provided by the wordnet we can attain a better recall and precision of that information retrieval system.

IV. PROPOSED SYSTEM

This paper proposes an alternative method for storing the data in wordnet; relational database structure. In relational database the data are stored as attributes and relations.This proposed system is having unique tables for each property. The referring attribute is the word id. It is the id given to each sense in the word net. Some Advantages of using RDBMS over file System.

- Data can be easily accessed.
- Data can be shared.
- Data modeling can be flexible.
- Data storage and redundancy can be reduced.
- Data inconsistency can be avoided.
- Data Integrity can be maintained.
- Standards can be enforced.
- Security restrictions can be applied.

- Independence of physical storage and logical data design can be maintained.
- High-level data manipulation language (SQL) can be used to access and manipulate data.

C. *Database Design*

The back end is made using a relational database. Here the 'word_id' is given as primary key. System assigns unique 'word_id' for each definition. The database includes following tables:

- Sense Table
- Synonymy Table
- Tag Table
- Hypernymy Table
- Meronymy Table
- Antonymy Table
- Troponymy Table
- Entailment Table
- Example Table

The sense table is used to store the word and the sense of word, and this table assigns word_id for each sense. Synonymy Table is used for storing the word_id and Synonymy word id. So we can access the synonyms of a word

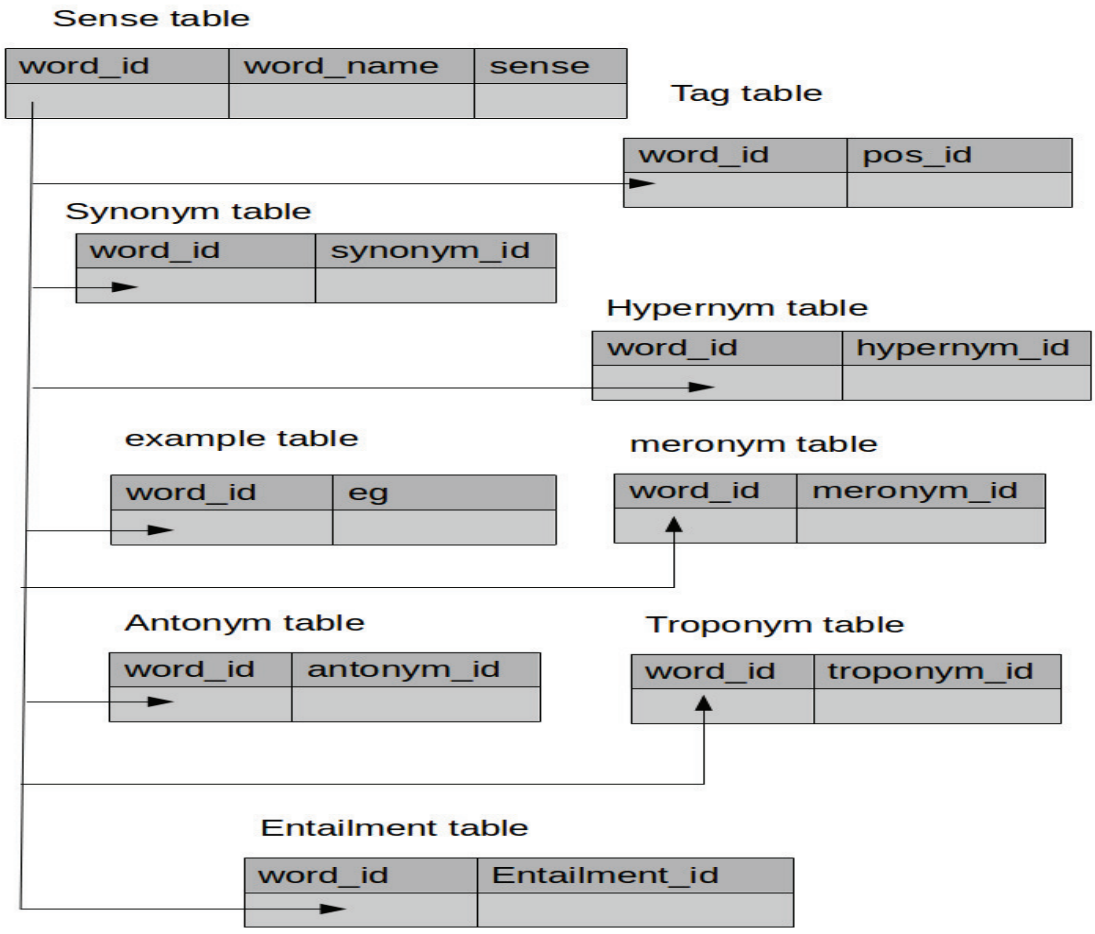

Fig 1. Database Design

easily. Similarly the Hypernymy, Meronymy, Antonymy,Troponymy and Entailment tables are used to store ids of Hypernymy, Meronymy, Antonymy, Troponymy and Entailment of a word. The tag table is used for POS tags and Example table gives the example context of a word.

D. Implementation and Result

Malayalam WordNet is implemented as a web based application. For implementation, Mysql database is used as the back-end, PHP as programming Language and Apache as web server.

The System can be divided into two modules: word insertion module and word search module. Using the word insertion module, any user can add a word and details like sense, synonymy etc. The block diagram is given in Fig 2.

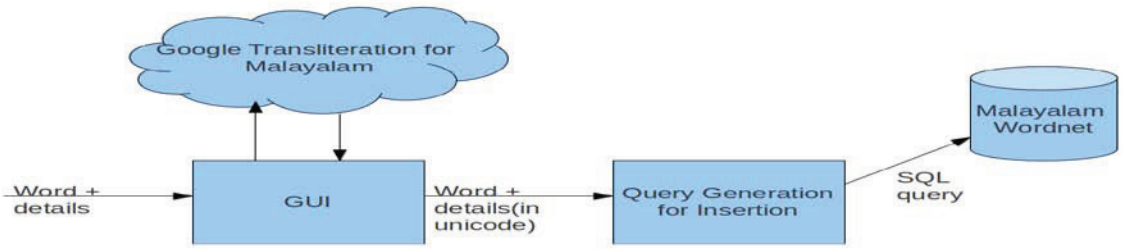

Fig 2. Word Insertion Module

As the figure(Fig 2) shows the word insertion module have several sub modules. The GUI module first receives the word and details as transliteration and then converts that word and details into Unicode format. This Unicode format is given to the query generation module. Query generation module prepares the SQL query fitted on Mysql-insert and given to Mysql database where we can store the word and details.

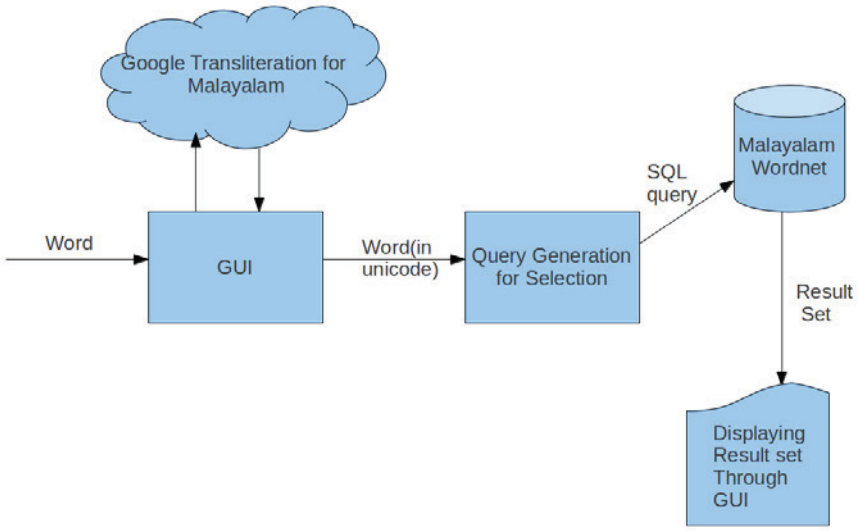

Fig 3.Word Search Module

In the Word search module(Fig 3), there are two submodules; GUI and query generation module. Here the user gives the word he wants to search. Then the GUI module interacts with Google transliteration cloud and convert the word into Unicode characters. The query generation module prepares a search query (using SELECT .... FROM ... WHERE) and gives to Mysql database. After executing the select query, retrieves the result set and displays it through the system.

After implementing the system using Mysql as back end and PHP as front end, added around 148342 words alond with their definitions. The system needed more word adding and voluntary work from those who is having deep knowledge in Malayalam vocabulary. The system can grow with the contributions from learned enthusiasts of Malayalam.

V. CONCLUSIONS & FUTURE SCOPE

This system implements an alternative technique for WordNet. It is a step towards an efficient WordNet for Malayalam language. This per demonstrate a working model for on-line Malayalam WordNet where even a naive user can add words, and is the first of its kind.

Even though this is an alternative method for traditional system, this system has limitation. The main one is the access of Mysql database from the front end programming language like Python, Java ,etc. A good interfacing library for Mysql WordNet is yet to be created.

REFERENCES

[1] Marin Dantchev, "WORDNET 2.1 Overview",EECS 595 / SI 661 & 761 / LING 541 Natural Language Processing Fall, 2006
[2] Satanjeev Banerjee, Ted Pedersen. ``Extended Gloss Overlaps as a Measure of Semantic Relatedness", Eight International Joint Conference on Artificial Intelligence, 2003
[3] http://www.shiffman.net/teaching/a2z/wordnet/ (Visited on :7July 2013)
[4] http://wordnet.princeton.edu/man/wndb.5WN.html(visited on :12July 2013)
[5] http://wordnet.princeton.edu/ (visited on :12 July 2013)
[6] http://olam.in/(visited on :12 July 2013)
[7] http://wiki.answers.com/Q/What_are_the_advantages_of_relational_database_over_f lat_file_database (visited on :25 July 2013)