Gujarati WordNet – A Lexical Database

Sagar S Panchal
Dept. of Computer Engineering, KJSCE
Vidyanagar, VidyaVihar East
Mumbai, Maharashtra 400077, India

Parth P Shukla
Dept. of Computer Engineering, KJSCE
Vidyanagar, VidyaVihar East
Mumbai, Maharashtra 400077, India

Piyush R Panchal
Dept. of Computer Engineering, KJSCE
Vidyanagar, VidyaVihar East
Mumbai, Maharashtra 400077, India

Jayesh S Kolte
Dept. of Computer Engineering, KJSCE
Vidyanagar, VidyaVihar East
Mumbai, Maharashtra 400077, India

Bharathi H N
Dept. of Computer Engineering, KJSCE
Vidyanagar, VidyaVihar East
Mumbai, Maharashtra 400077, India

ABSTRACT

The English WordNet and the Hindi WordNet have inspired the Gujarati WordNet; this paper specifies the overview and the basic methods for developing a Gujarati WordNet.

General Terms

WordNet lies under the category of Natural Language Processing the result shows all the possible meanings, it depends on how humans think and how the language is exactly used.

Keywords

Development of the Gujarati WordNet and the basic knowledge about how it works and applications where Gujarati WordNet can be used.

1. INTRODUCTION

Gujarati WordNet is an online Lexical Database for Gujarati language. The result can be categorized into POS (Part Of Speech categories) like Noun, Verb, Adverb and Adjective, which are organized in a group of synonymy called as synsets [5][2]. WordNet shows all possible meanings of the searched word. The further result derives the relational category of the entry such as Hypernymy, Hyponymy, Antonymy, Holonymy, Meronymy, and Troponymy [1][2].

Data in the WordNet database is manually entered by a lexicographer [1]. The Lexicographer needs to produce all the correct and accurate relations for each word forms and word meanings in the WordNet database.

2. SEMANTIC RELATIONS

2.1 Synonymy

Synonymy can be two or more different words, which mean the same, or they have the same explanation in a particular language. Synonymy in WordNet uses the semantic similarity for the words [4]. Synonymy can be of any word category, whether it is Noun, Verb, Adverb or Adjective. The example for synonymy set in Gujarati can be {ભવન, ઇમારત, ઘર} Home.

Synonymy set are exactly similar in their meanings so that they can be easily replaced with each other in their usage, but if the word are not been replaced or interchanged then they cannot form a synonymy set.

2.2 Antonymy

Antonymy can be referred as the simple and easy relation in the WordNet but it is not so it has various difficulties. For instance if a person says, “I am not hungry” does not mean that he is “full” and he just ate something, although full and hungry are antonymy for each other [4]. The Antonymy also depends on the time and its occurrence. In Gujarati {તે “અમીર” નથી} they aren’t rich doesn’t mean they are very poor {બહુ “ગરીબ”} antonymy in a sentence has to be used with care.

2.3 Hypernymy and Hyponymy

Hypernymy and Hyponymy relations are attached to one another; one follows another in many kinds. The reversal of Hypernymy relation can be Hyponymy relation of other word. For instance {ઘર} ‘home’ is Hypernymy for ‘ghar’ and {મંદિર} Temple is Hyponymy for {ઘર} that is {ઘર} is a place where a person lives and {મંદિર} is a place where god lives. They are like set and superset in which set inherit all the features of the superset. The set can also have its own unique feature, which may not be present in the superset.

2.4 Meronymy and Holonymy

The Meronymy relation is transitive and symmetric while the Holonymy relation is the reversal of Meronymy [1]. The Meronymy relation of a word is Holonymy if and only if it is a part of Meronymy. For instance {લાલ} Red is a color is a Meronymy relation while {રંગ} Color is its Holonymy relation i.e. {લાલ} Red can only be {રંગ} Color. They are constructed as per component-object, stuff-object, member-collection, place-area, feature-activity, and phase-state.

2.5 Troponymy

Troponymy relation can be only found between two verbs. Like a verb {નસકોરા} Snoring can be only expressed if and only if a person is {સૂવું} Sleeping. So {સૂવું } Sleeping and {નસકોરા} Snoring are Troponymy relation of each other.

3. SYSTEM DESIGN

The Gujarati WordNet system consists of two parts - One is a place where the Lexicographer can enter the word into the database with their each and every detail and in their exact relational form [3]. And other is the place where the user can use the system to get the details about the words which they like to and also get their semantic relational details.

3.1 Database

The WordNet database contains various tables in which the word form and their relations are been stored. As it is an online lexical database system data is stored on XAMPP server with MySQL and the data is stored in UTF-8 (Universal Character Set **T**ransformation **F**ormat—**8**-bit) format so it can store and display word in Gujarati language. Each word entry in the database is stored with all the synonymy given with the same id or synset id in one table called tbl_all_words_gj. They are mapped to another table called as tbl_all_gujarati_synset_data where each details of the word is being stored such as synset, gloss, POS category [3].

Table: Sample records for Gujarati WordNet

synset_id	synset	gloss	category
1	મંદિર, દેવળ	એ પવિત્ર ગૃહ	NOUN
2	મંદિર, દેવાલય	કોઇ વિશેષ કે મહાન ઉદ્દેશને માટે સમર્પિત હોય	NOUN

A clear hierarchical structure of the relation is shown as per user demands all the data continuously flows to the interface as per user clicks. Then this word from the table is connected to the other relational tables in the database. Like for finding different

- For Hypernymy it will jump to tbl_hypernymy table search for the query and display the accurate result.
- For Hyponymy it will jump to tbl_hyponymy table search for the query and display the accurate result.
- For Meronymy it will jump to tbl_meronymy table search for the query and display the accurate result.
- For Holonymy it will jump to tbl_holonymy table search for the query and display the accurate result
- For Troponymy it will jump to tbl_troponymy table search for the query and display the accurate result.

4. USE OF GUJARATI WORDNET DATABASE

4.1 Stemmer

Stemmer is an application of the WordNet where the user is required to enter an inflected word and the system will run and eliminate the inflection and produce the result, which will show the root word [1]. For instance {છોકરાઓ} chokrao is the word with the inflection added so it will eliminate {ઓ} ‘o’ and the output will show only chokra {છોકરા}. Stemming is done with the help of a hierarchical tree like structure which will check for all the possibilities of the inflection that can be added to a word whether it is a prefix a suffix or a circumfix. The hierarchical tree structure helps us backtrack, the inflection until accurate results are not found [4]. Hierarchical approach can be followed to find the accurate results of the stemmer it goes as:

1. Check for word in database if found it is a root word
2. Eliminate {અ આ ઇ ઈ ઉ ઊ એ ઐ ઓ ઔ અં અઃ} if found
3. Eliminate {એ ઓ}
4. Replace {એ} with {આ}
5. Eliminate {ે ો}

4.2 Morph Analyzer

Morph Analyzer is another application of the WordNet. The user is required to enter an inflected word and the system will run and produce the result which will show the root word as well as the inflection added to the word and will show that whether it is a suffix a prefix or a circumfix [1]. For instance {છોકરાઓ} chokrao is the word with the inflection added so it will show {ઓ} ‘o’ is the inflection carried with the root word and the root word chokra {છોકરા}.

5. CONCLUSION

Gujarati WordNet can be served as online lexical system. It can be used as a very useful system for retrieval of information, can be used in Machine Translation and for Word Sense Disambiguation [3]. It can be a very useful tool for the person who has less knowledge to the language and wants to learn more about the Gujarati language. It can be also useful for a person who is building an application on Gujarati language to seek knowledge and develop a proper application or deeply study the language.

6. ACKNOWLEDGMENTS

Our thanks to the project Mentor Prof. Bharathi H Narayan and Dr. Sopan G Kolte to help us in the project and provide us with all the details and resource for the project and Mr. Diptesh Kanojia for helping in the entire process of development.

7. REFERENCES

[1] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller, ―Introduction to WordNet: An On-line Lexical Database‖, CSL Report 43, Princeton University Cognitive Science Laboratory,1990 (Revised August, 1993).

[2] D. Narayan, D. Chakrabarty, P. Pande, P. Bhattacharyya, ―An Experience in Building the Indo-WordNet – A WordNet for Hindi‖, In Proceedings of the First International Conference on Global WordNet (GWC 02), Mysore, India, 2002.

[3] Kalyanamalini Sahoo & V Eshwarchandra Vidyasagar, ―Kannada WordNet – A Lexical Database, Resource Center for Indian Language Technology Solutions, Department of Management Studies, Indian Institute of Science Bangalore – 560012, India.

[4] Arindam Chatterjee, Salil Rajeev Joshi, Mitesh M. Khapra, Pushpak Bhattacharyya, ―Introduction to Tools for IndoWordNet and Word Sense Disambiguation**",** Department of Computer Science and Engineering, Indian Institute of Technology Bombay Powai, Mumbai-400076 Maharashtra, India.

[5] http://www.cse.iitb.ac.in/~pb/papers/gwc12-gujarati-wn.pdf ‖Introduction to Gujarati wordnet (GCW12) IIT Bombay, Powai, Mumbai-400076 Maharashtra, India.