An Efficient Database Design for IndoWordNet Development Using Hybrid Approach

Venkatesh Prabhu[2] *Shilpa Desai*[1] *Hanumant Redkar*[1]
Neha Prabhugaonkar[1] *Apurva Nagvenkar*[1] *Ramdas Karmali*[1]
(1) GOA UNIVERSITY, Taleigao - Goa
(2) THYWAY CREATIONS, Mapusa - Goa
venkateshprabhu@thywayindia.com, sndesai@gmail.com, hanumantredkar@gmail.com,
nehapgaonkar.1920@gmail.com, apurv.nagvenkar@gmail.com, rnk@unigoa.ac.in

ABSTRACT
WordNet is a crucial resource that aids in Natural Language Processing (NLP) tasks such as Machine Translation, Information Retrieval, Word Sense Disambiguation, Multi-lingual Dictionary creation, etc. The IndoWordNet is a multilingual WordNet which links WordNets of different Indian languages on a common identification number given to each concept. WordNet is designed to capture the vocabulary of a language and can be considered as a dictionary cum thesaurus and much more. WordNets for some Indian Languages are being developed using expansion approach.

In this paper we have discussed the details and our experiences during the evolution of this database design while working on the Indradhanush WordNet Project. The Indradhanush WordNet Project is working on the development of WordNets for seven Indian languages. Our database design gives an efficient plan for storage of WordNet data for all languages. In addition it extends the design to hold specific concepts for a language.

KEYWORDS: WordNet, IndoWordNet, synset, database design, expansion approach, semantic relation, lexical relation.

Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing (SANLP), pages 229–236,
COLING 2012, Mumbai, December 2012.

1 Introduction

1.1 WordNet and its storage methods

WordNet (Miller, 1993) maintains the concepts in a language, relations between concepts and their ontological details. The concept in a language is captured as a Synonym set called synset. The IndoWordNet is a multilingual WordNet which links WordNets of different Indian languages on a common identification number called as synset Id. A synset represents a unique concept in a language. Synset is composed of a Gloss describing the concept, example sentences and a set of synonym words that are used for the concept. Besides synset data, WordNet maintains many lexical and semantic relations. Lexical relations (Fellbaum, 1998) like antonymy, gradation are between words in a language whereas semantic relations like hypernymy, hyponymy are between concepts in a language. Ontology details for a synset are also maintained in a WordNet.

WordNet contains information about nouns, verbs, adjectives and adverbs and is organized around the notion of a synset. Earlier the WordNet data was stored in flat text files. This storage method was found insufficient for developing multi-lingual WordNet applications requiring random access to synsets or its constituents. This was the motivation for storage of data in a relational database. The design of the database was for a single language WordNet. The approach used to build the WordNet had an impact on how the data should be ideally stored. The various WordNet development approaches are discussed next.

1.2 WordNet development Approaches and its impact on the storage structure

Various approaches are followed in the construction of WordNets across the languages of the world. WordNets are constructed by following either the merge approach or the expansion approach (Vossen, 1998). The merge approach is also referred to as WordNet construction from *first principles* (Bhattacharyya, 2010). Here exhaustive sense repository of each word is first recorded. Then the lexicographers constructs a synset for each sense, obeying the three principles namely principle of minimality, coverage and replacebility (Bhattacharyya, 2010).

For many Indian languages, WordNets are constructed using the expansion model where Hindi WordNet synsets are taken as a source. The concepts provided along with the Hindi synsets are first conceived and appropriate concepts in target language are manually provided by the language experts. The target language synsets are then built based on the concepts created keeping in view the three principles mentioned above. The expansion approach gives rise to a multi-WordNet where a concept is given an id called synset-id and is present in all target language WordNets. One of our contributions is the storage structure of such a multi-WordNet. We maintain the data common to all languages in a central shared database for the multi-WordNet and the data which changes as per the language in separate databases for each language. The issues regarding the approach which affect the storage structure are discussed next.

1.3 Advantages and disadvantages of each approach

Both the merge and expansion approaches have their advantages and disadvantages. The advantages of using merge approach are: There is no distracting influence of another language, which happens when the lexicographer encounters culture and region specific concepts of the source language. The quality of the WordNet is good but the process is typically slow. The advantages of using expansion approach are: Using this approach, instead of creating the synset from the scratch, synsets are created by referring to existing WordNet of the related language. WordNet development process becomes faster as the gloss and synset of the source language is already available as reference. Also it has the advantage of being able to borrow the semantic relations of the given WordNet. That leads in saving an enormous amount of time.

The WordNets developed using expansion approach is very much influenced by the source language and may not reflect the richness of the target language. Therefore there was a need for Hybrid Approach to achieve perfection with respect to WordNet development time and quality. In order to overcome this disadvantage the IndoWordNet Community classified the synsets in the source language as Universal, Pan Indian, In family, language-specific, etc. Each language initially developed the synsets for Universal and Pan Indian synsets and then proceeded to develop synsets which were specific to their respective language viz. language-specific synsets.

Another contribution is that our storage structure also provides the flexibility to include concepts specific to a language or group of language. We also provide for a mechanism in our storage structure whereby the concept can be accepted by all languages or a group of languages.

The rest of the paper is organized as follows – section 2 introduces the IndoWordNet, section 3 presents tools for IndoWordNet Development and limitations of the existing tool. The details of IndoWordNet Database design and its strengths are presented in section 4. Section 5 presents the future perspectives and conclusions.

2 IndoWordNet

The IndoWordNet is a linked structure of WordNets of major Indian languages from Indo-Aryan, Dravidian and Sino-Tibetan families. These WordNets have been created by following the expansion approach from Hindi WordNet which was made available free for research in 2006. Since then a number of Indian languages have been creating their language WordNets (Pushpak Bhattacharyya, 2010). The issues faced while using the expansion approach which affect the storage structure are discussed next.

2.1 Challenges faced while using Expansion Approach

The challenges faced while using the Expansion Approach are:

1. *Linking of contextual words:* Using the expansion approach, certain synsets may totally get omitted because of the variety of shades of meaning of different words.
2. *Linking a concept which is not present in the source language.*
3. *Coining of words:* Another issue that remains to be resolved is how far the lexicographer can be given the liberty to coin new words.
4. *Coverage of synsets:* Though the meaning of many words are known to the people and are not found in common literature, but one may find some of the words possibly in poetry. Whether we have to cover them is also a question (Walawalikar et al., 2010).

Some of the challenges were addressed in the following ways:

1. The first and second challenge was addressed by creating language specific synsets and validating with experts before finalization.
2. The third challenge was addressed by following two approaches: first was by adapting the source language word and second approach was coining of new words.

3. The fourth challenge was addressed by adding such words towards the end of the synonymous set.

3 Tools for IndoWordNet Development

During the development of IndoWordNet we felt the need of various tools developed which drove us to reconsider the storage strategy used to store the IndoWordNet.

3.1 Existing tools

1. For the synset creation, we use the offline tool, Synset Creation Tool provided by IIT Bombay along with Hindi WordNet sysnets. This standalone interface allows users to view the Hindi synsets, concepts, example sentences on one side and simultaneously keying the target language synsets, concepts and example sentence. The tool also has the Princeton WordNet English synsets interlinked. The generated target language synset files which are stored in a flat text files have extension as .syns.
2. English-Hindi Linkage Tool is a heuristic based tool to link Hindi-English synsets.
3. Synset Categorization Tool is used to chose common linkable synsets across all languages by classifying them as Universal, Pan Indian, In family, language-specific, Rare, Synthesized and Core.
4. The Sense Marking Tool is used to find the synset coverage of a WordNet.

3.2 Limitations of existing tools with respect to the storage structure

The tools used for the development of WordNets using the Expansion approach were mostly based on flat files. Flat files have their own advantages but there are several disadvantages too. Some of the problems faced while working on the above tools were as follows:

1. Synset counting with respect to different criteria such as getting the synset count belonging to a specific grammatical category or range.
2. Merging synset files, finding common set of synsets
3. Security and Data integrity
4. Status of synsets – to know whether the synset is validated or not.
5. Additional data about synsets (meta data – source, useful links, video, audio, domain, images which gives additional information about the synset, etc).
6. Also it is difficult to store lexical and semantic relations between words and synsets in a flat file.

The major contribution presented in this paper is the storage design we implemented for IndoWordNet in form of solutions to the above problems. This design related solutions to the above problems are:

1. The data is stored in a systematic and classified manner.
2. Meta data feature - to store additional information about the synset.
3. Language-dependent information such as synset entries in the target language, lexical relations, etc. is maintained in the individual WordNets.
4. Language-independent information such as semantic relation, ontological details of a concept, etc. is stored only once and is made available to all the language-specific modules via the inter-lingual relations.
5. Centrally controlled system for issuing Ids in case of language specific synsets.

In order to provide support for a systematic and rapid expansion approach for development of good quality WordNets for languages, it was felt essential to have the data stored in databases in a systematic manner using normalised database design.

4 Database Design Details

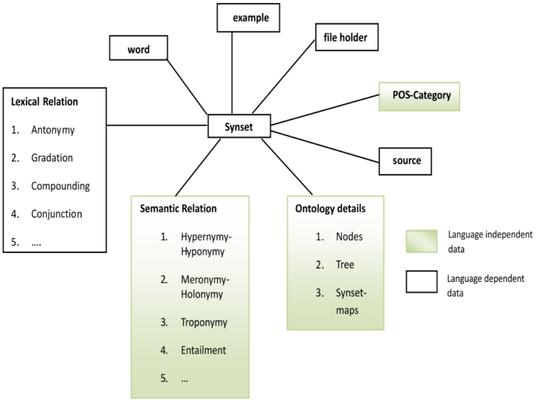

Figure 1: Database design depicting language dependent data and language independent data using colour coding.

The IndoWordNet database is designed to maintain the data for a WordNet in two databases. The common data such as semantic relations and ontology details for all languages are maintained in a common database called wordnet-master. The wordnet master maintains the data shared by all the languages. This database will keep tables which borrow the relations from the source WordNet (Hindi WordNet). The wordnet-master includes tables for semantic relations. It will include all ontology related tables in English.

The synset data for a language is maintained in a separate database for each language called wordnet-<respective-language>. Here respective-language is to be replaced by the actual language name for e.g. wordnet-konkani, wordnet-hindi, wordnet-marathi for Konkani, Hindi and Marathi languages respectively. Figure 1 shows the language dependent and language independent data represented in colour coding.

WordNet classifies word meanings into four basic lexical categories: nouns, verbs, adjectives and adverbs. Each synset in the WordNet is linked with other synsets through the well-known lexical and semantic relations. Semantic relations are between synsets and lexical relations are between words. These relations serve to organize the lexical knowledge base.

4.1 Salient features of using IndoWordNet Database design

Some of the salient features of using IndoWordNet Database design for storing synset data are as follows:

- Centralized database for common concepts and language databases for language concepts.

- Incorporates both merge-approach and expansion approach data
- Design provides for developing a centrally controlled system for issuing Ids in case of concepts specific to a language or language group.
- Open, scalable, normalized and modular design.
- Maintains language-specific relations in the WordNets.
- Achieves maximal compatibility across the different resources.
- Can be used for building the WordNets relatively independently (re)-using existing resources.
- Supports development of online and offline applications such as dictionary, synset creation tool, multilingual information retrieval, etc. Speeds up retrieval and processing time;
- Stores semantic, lexical relations,ontological details and Meta-data;
  All the above features make it a flexible design. The Entity-Relationship diagram for IndoWordNet database design is shown in Figure 2.

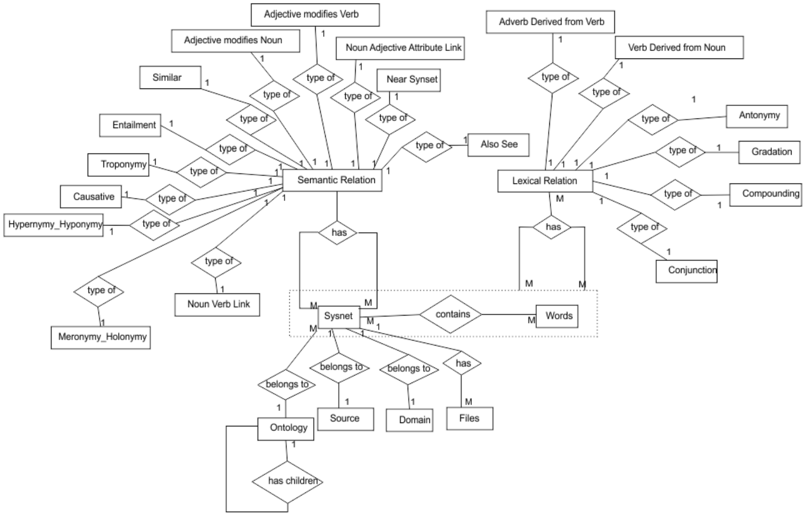

Figure 2: Entity Relationship diagram for IndoWordNet Database Design.

Conclusion and perspectives

The advantages of the IndoWordNet Database design are: Different WordNets can be compared and checked cross-linguistically which will make them more compatible. It will be possible to use the database for multilingual information retrieval, by expanding words in one language to related words in another language. The IndoWordNet Database design can be used for development of online and offline applications such as Multi-lingual Dictionary, WordNet for public use, etc. The IndoWordNet Application Programming Interfaces developed by Goa University which helps the developer to access and modify the IndoWordNet database is developed using IndoWordNet database schema.

References

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller (Revised August 1993). *Introduction to WordNet: An On-line Lexical Database.*

George A. Miller 1995. *WordNet: A Lexical Database for English.*

Fellbaum, C. (ed.). 1998. *WordNet: An Electronic Lexical Database, MIT Press.*

Pushpak Bhattacharyya, Christiane Fellbaum, Piek Vossen 2010. *Principles, Construction and Application of Multilingual WordNets, Proceedings of the 5th Global Word Net Conference (Mumbai-India), 2010.*

Pushpak Bhattacharyya, *IndoWordNet, Lexical Resources Engineering Conference 2010 (LREC2010), Malta, May, 2010*.

Shantaram Walawalikar, Shilpa Desai, Ramdas Karmali, Sushant Naik, Damodar Ghanekar, Chandralekha D'souza and Jyoti Pawar. *Experiences in Building the Konkani Word Net using the expansion Approach*. In Proceedings of the 5th Global WordNet Conference on Principles, Construction and Application of Multilingual WordNets (Mumbai-India), 2010.

Vossen P. (ed.). 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Dordrecht.