Creation of Lexical Relations for IndoWordNet

Ashish Narang
CSED, Thapar University Patiala India.
ashish.narang6789@gmail.com

Rajendra Kumar Sharma SMCA, Thapar University Patiala India.
rkshar-ma@thapar.edu

Parteek Kumar
CSED, Thapar University Patiala India.
parteek.bhatia@thapar.edu

Abstract

WordNet is an electronic lexical database available on-line as a powerful resource to the researchers in the area of computational linguistics, text processing and other related areas. WordNet for Hindi language has already been developed by IIT, Bombay. The Indian languages WordNets are being created using expansion approach from Hindi WordNet under IndoWordNet project. In expansion approach, semantic relations are borrowed from the reference language, while the lexical relations need to be created for each language, as these relations are language dependent. This paper describes the process of creation of lexical relations like antonym, compounding, conjunction and gradation for IndoWordNet. A lexical creation tool has been presented in this paper with provision to create lexical relations in target language on the basis of relations created in Hindi WordNet and with another provision to create lexical relations in target language without referring to Hindi WordNet. It has been observed that lexical relations for target language can be created easily on the basis of relations created in Hindi WordNet for Hindi in-family languages, while for the languages that do not fall in the same family provision of creation of lexical relation without referring to Hindi WordNet can be used.

1 Introduction

WordNet is a large lexical database of a language. In WordNet, words are grouped together according to their similarity of meanings. WordNet maintains concepts in a language, relations between concepts and their ontological details. Each concept in a language represents a synset. Synsets are basic building blocks of WordNet. Synset is composed of gloss, example sentences and set of synonym words that are used for the concept. Besides synset data, a WordNet maintains lexical and semantic relations. Lexical relations like antonomy and gradation are between the words in a language whereas semantic relations like hypernymy, hyponymy, meronymy, holonymy, entailment, troponymy and casuation are between concepts in a language. WordNet structure makes it a useful tool for computational linguistics and natural language processing. The major applications of WordNet are text categorization (Gabrilovich and Markovitch, 2004), text summarization *(*Bellare *et al.,* 2004), word sense disambiguation (Banerjee and Pedersen, 2002) and machine translation *etc.*

Recognizing the immense importance of lexical resources arises the necessity for creation of IndoWordNet project. IndoWordNet is a linked structure of WordNets of major Indian languages from Indo-Aryan, Dravidian and Sino-Tibetan families. The WordNets for these languages are being created using the expansion approach from the Hindi WordNet which was made available free for research in 2006. Using expansion approach, there is advantage of being able to borrow the semantic relations of a given source WordNet (Bhattacharyya, 2010). Lexical relations cannot be borrowed from source WordNet using expansion approach as they are language dependent. In order to create lexical relations for IndoWordNet languages, a lexical creation tool has been proposed in this paper with a provision to create lexical relations from source WordNet and as well as to create lexical relations for those words which are not covered in source WordNet.

2 Related Work

English WordNet is the first WordNet created in this field. The development of English WordNet started in 1985 (Miller, 1985) at the Cognitive Science Laboratory of Princeton University. The success of English WordNet has inspired several projects that aim at constructing the WordNet for other languages or to develop multilingual WordNet. EuroWordNet is a system of semantic

network for European languages. The EuroWordNet project dealt with Dutch, Italian, Spanish, German, French, Czech, and Estonian languages (Vossen, 1998). BalkaNet WordNet project has developed WordNets for Bulgarian, Greek, Romanian, Serbian and Turkish languages (Tufis *et al.*, 2004).

In India, Hindi WordNet was developed in 2006 by IIT, Bombay. Later on Hindi WordNet was extended to Marathi WordNet. Currently IndoWordNet project, a linked structure of major Indian languages is in progress in India. Moreover, Indradhunsh Project a part of IndoWordNet project, aim at developing WordNets for seven major Indian languages, Bengali, Gujarati, Kashmiri, Konkani, Oriya, Punjabi and Urdu has been initiated in 2010. These Indian languages WordNets are being created using expansion approach from Hindi WordNet under the guidance of IIT, Bombay.

3 WordNet relations

WordNet contains the standard information found in dictionaries and thesauri. An additional feature of WordNet is its information about the relationships between words and synsets. The words and synsets in the WordNet are linked through two types of relations, *i.e.*, lexical and semantic relations. Lexical relation exists between the word forms while semantic relation exists between the concepts.

3.1 Semantic relations

Semantic relation is a relation between meanings, and since meanings can be represented by synsets, semantic relations can be considered as pointers between synsets (Tufis *et al.*, 2004). For example, hypernym/hyponym is a semantic relation. Consider two synsets given in (1) and (2).

{ਪੌਦਾ *paudā 'plant'*, ਬੂਟਾ *būṭā 'plant'*} ...(1)

{ਚਾਹ *cāh 'tea'*} ...(2)

Here, {ਪੌਦਾ *paudā 'plant'*, ਬੂਟਾ *būṭā 'plant'*} is hypernym of {ਚਾਹ *cāh 'tea'*} and {ਚਾਹ *cāh 'tea'*} is hyponym of {ਪੌਦਾ *paudā 'plant'*, ਬੂਟਾ *būṭā 'plant'*}. There are total thirteen semantic relations, namely, hypernymy, hyponymy, meronymy, holonymy, entailment, causation, troponymy, ability link, capability link, functional link, attributes, modifies noun and modifies verb exist in a WordNet.

Using expansion approach there is advantage of being able to borrow the semantic relations of a given WordNet. For example, consider two synsets in the source WordNet given in (3) and (4).

{चाय *chaie 'tea'*} ...(3)

{पौधा *paudha 'plant'*, पौदा *pauda 'plant'*} ...(4)

In Hindi WordNet (source), synset given by (4) is hypernymy of synset given by (3) and synset given by (3) is hyponym of synset given by (4). These two synsets also share hyperonymy/hyponymy relation in Punjabi (target) language. Since, the synset-id are kept same for all the languages, therefore, semantic relations from the source WordNet (Hindi) can be extended to all target languages with expansion approach.

3.2 Lexical Relations

Lexical relations are the relations between members of two different synsets. For example, consider two synsets given in (5) and (6).

{ਮੋਟਾ *mōṭā* '*fat'*, ਭਾਰੀ *bhārī 'fat'*, ਸਥੂਲ *sathūl 'fat'*, ਥੂਲ *thūl 'fat'*,ਵਜਨੀ *vajnī* '*fat'*} ...(5)

{ਪਤਲਾ *patlā 'thin'*, ਦੁਬਲਾ *dublā 'thin'*, ਕਮਜ਼ੋਰ *kamzōr 'thin'*, ਮਾੜਾ *māṛā 'thin'*} …(6)

Here, synsets (5) and (6) are opposites but they do not share antonym relation. Antonym relation exists between two words not between two synsets. Here, words ਮੋਟਾ *mōṭā 'fat'* and ਪਤਲਾ *patlā 'thin'* are in antonym relation.

4 Lexical creation tool

In order to create the lexical relations for all the participating languages of IndoWordNet project, a lexical creation tool has been designed with provision to create lexical relations in target language on the basis of relations created in Hindi WordNet and with another provision to create lexical relations in target language without referring to Hindi WordNet. Lexical creation tool can create the following lexical relations.

- Antonym
- Compounding
- Conjunction
- Gradation

In the subsequent subsection the lexical creation tool has been presented by considering Punjabi as target language. However, the system is able to handle all languages participating in IndoWordNet project.

4.1 Antonymy creation tool With reference to Hindi WordNet

Antonymy is a lexical relation that exists between a pair of words that represent opposite meaning. The antonyms for Hindi WordNet have already been created. Antonyms for the Punjabi WordNet can be created from the antonyms of Hindi WordNet, but database design of Punjabi WordNet is different from Hindi WordNet. There is a need to design an interface which can bridge the gap between two different database designs and create the antonyms for the Punjabi WordNet from Hindi WordNet. Algortihm 4.1 has been used for creation of antonyms from Hindi WordNet. The algorithm is developed using IndoWordNet database design (IndoWordNet Database design, 2011) and Hindi WordNet database design followed by IIT, Bombay.

Algorithm 4.1: Creation of Antonyms with reference to Hindi WordNet

1. Extract ***synset_id*** of source Hindi ***synset_word*** from ***HWN_DB*** table.
2. Extract ***word_ids*** from ***wn_synset_word*** table, for the ***synset_id*** found in step 1.
3. For each ***word_id*** found in step 2, extract the corresponding words in target language from ***wn_word*** table.
4. Extract ***synset_id*** of antonym Hindi ***synset_word*** from ***tbl_noun_anto_direction*** table.
5. Extract ***word_ids*** from ***wn_synset_word*** table, for the ***synset_id*** found in step 4.
6. For each ***word_id*** found in step 5, extract corresponding words in target language from ***wn_word*** table.

Description of Algorithm 4.1

For example, for the word पूर्व *purav 'east'* in Hindi, system searches for source word in ***tbl_noun_anto_direction*** table and extract corresponding ***synset_id***, *i.e.*, 6898 as shown in Figure 1. For the given ***synset_id*** 6898, system refers to ***wn_synset_word*** table to extract the ***word_ids*** as shown in step 1 of Figure 1. For each of the ***word_id*** found, system retrieves the corresponding words in target language, *i.e.*, Punjabi from ***wn_word*** table as shown step 2 of Figure 1. The similar approach has been followed to find the antonym words for antonym ***synset_id.*** A user interface has been designed in Java to provide the relevant information to the end user as shown in Figure 2.

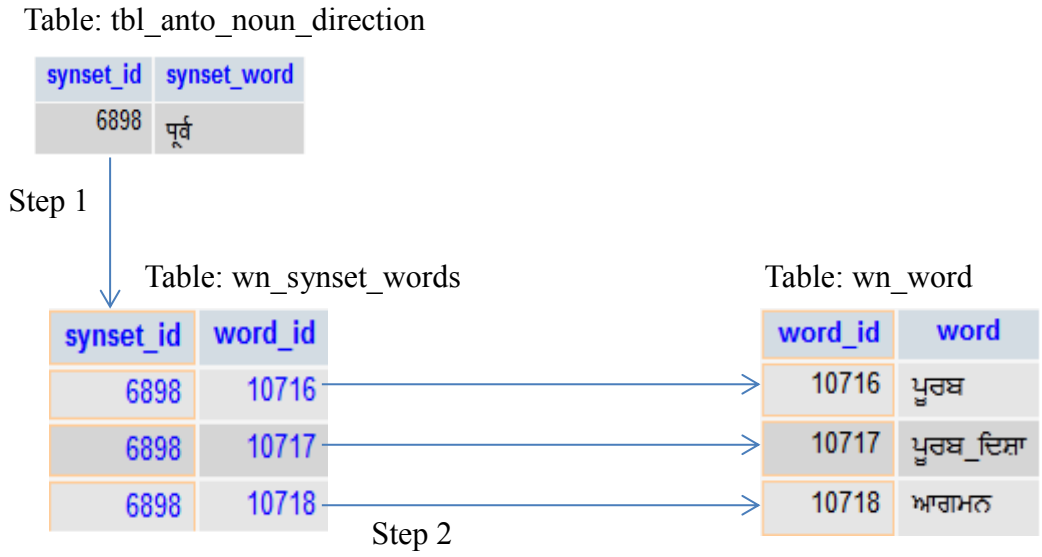

Figure 1: Extracting words corresponding to synset_id 6898

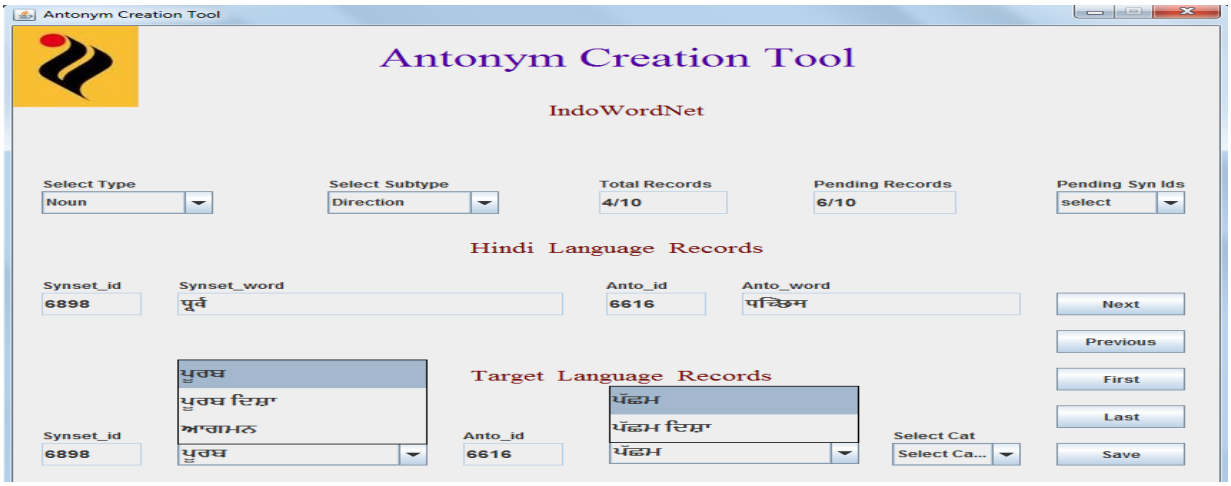

Figure 2: Interface for antonym creation tool with reference to Hindi WordNet

4.2 Antonymy creation tool without reference to Hindi WordNet

The antonym relation may also exist between the words which are not covered in Hindi WordNet, but may exist in the target language. This is a very common case for those Indian languages that do not belong to same language family as Hindi. There is need to design a tool which can create the antonyms for such words. The algorithm 4.2 has been designed for the creation of antonym for these words. The algorithm is developed using IndoWordNet database design (IndoWordNet Database design, 2011).

Algorithm 4.2: Creation of Antonyms without reference to Hindi WordNet

1. Extract ***word_id*** of the input word in target language from ***wn_word*** table.
2. Extract ***synset_ids*** from ***wn_synset_word*** table, for ***word_id*** found in step 1.
3. For each ***synset_id*** found in step 2, extract the corresponding concepts from ***wn_synset*** table.
4. Extract ***word_id*** of the input antonym word in target language from the ***wn_word*** table.
5. Extract ***synset_ids*** from ***wn_synset_word*** table, for ***word_id*** found in step 4.
6. For each ***synset_id*** found in step 5, extract the corresponding concepts from ***wn_synset*** table.

Description of algorithm 4.2

Let us consider an example for creation of antonym for input Punjabi word, ਚੰਗਾ *caṅgā 'good character'*, the system refers to ***wn_word*** table to extracts corresponding ***word_id*** as shown in Step 1 of Figure 3.

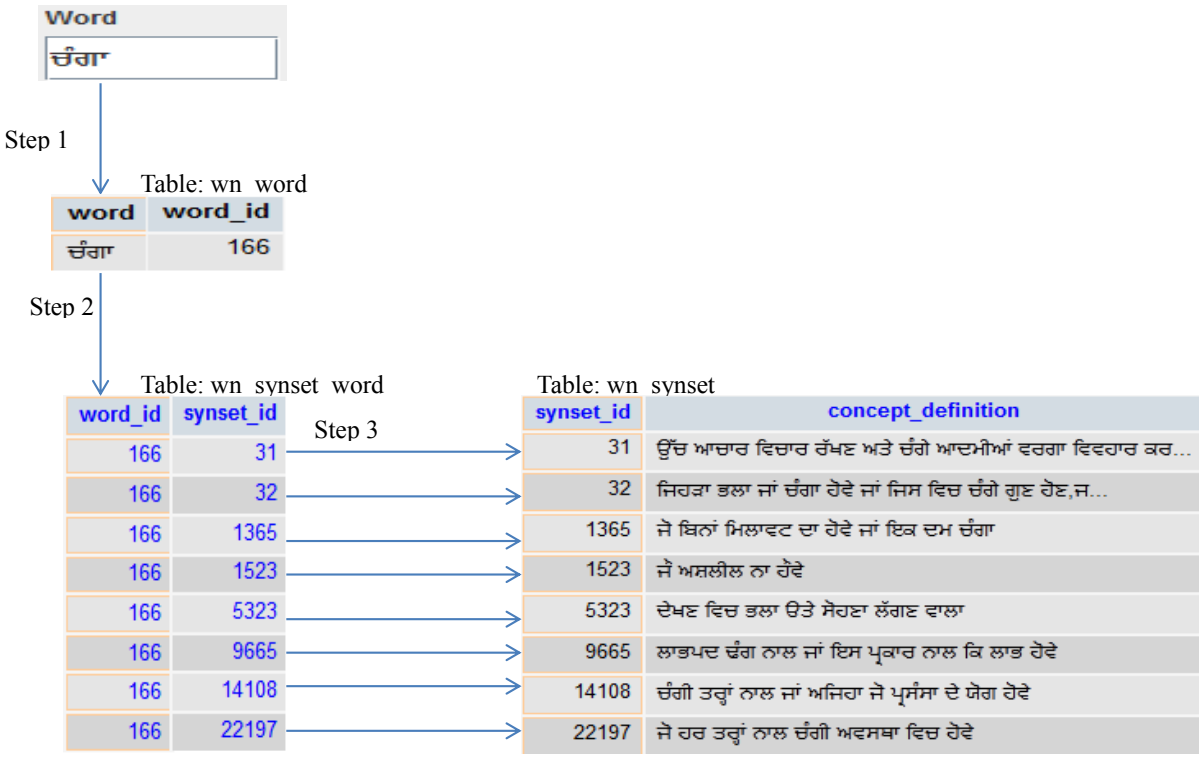

Figure 3: Extraction of concepts for the word ਚੰਗਾ *caṅgā 'good character'*

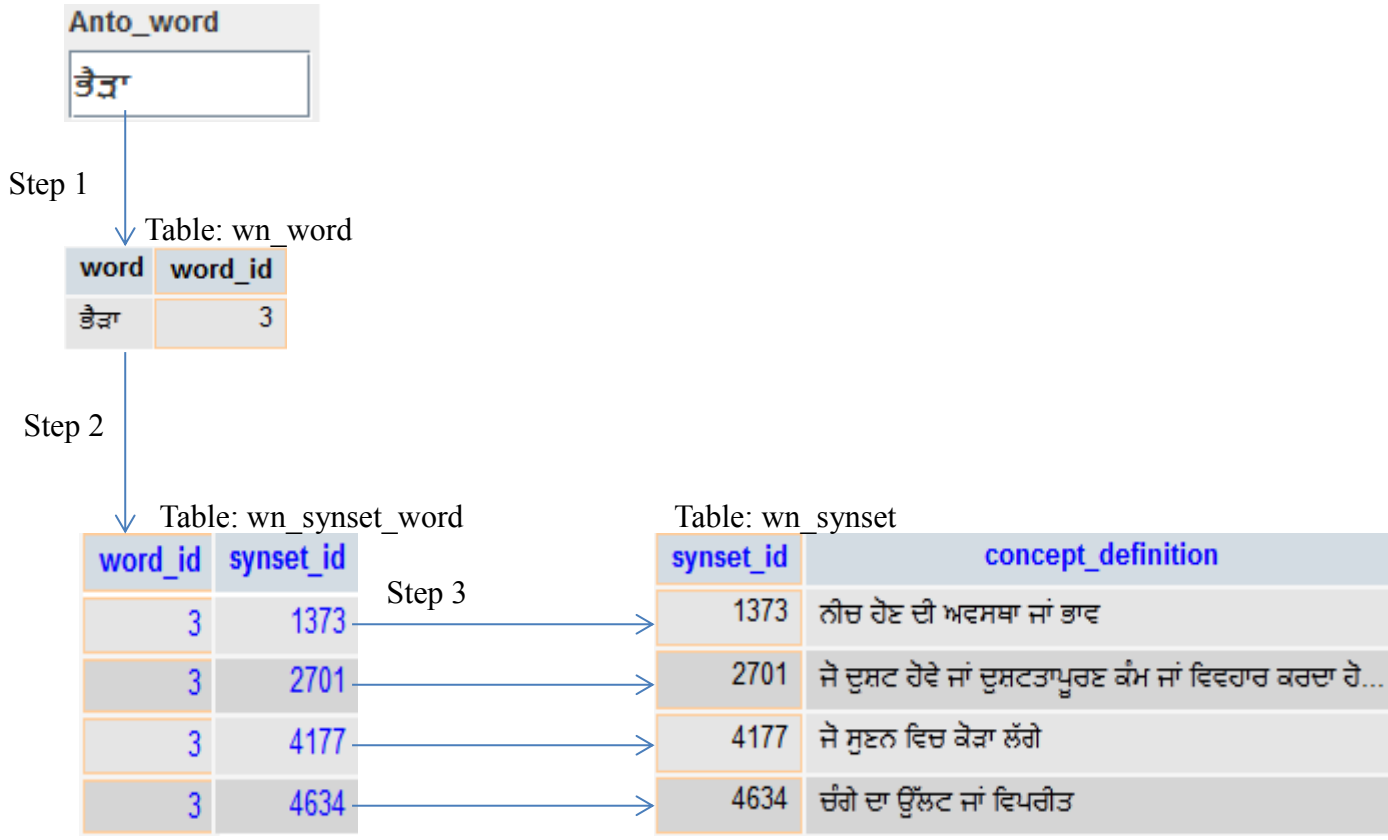

Figure 4: Extraction of concepts for the word ਭੈੜਾ *bhaiṛā 'charactterless'*

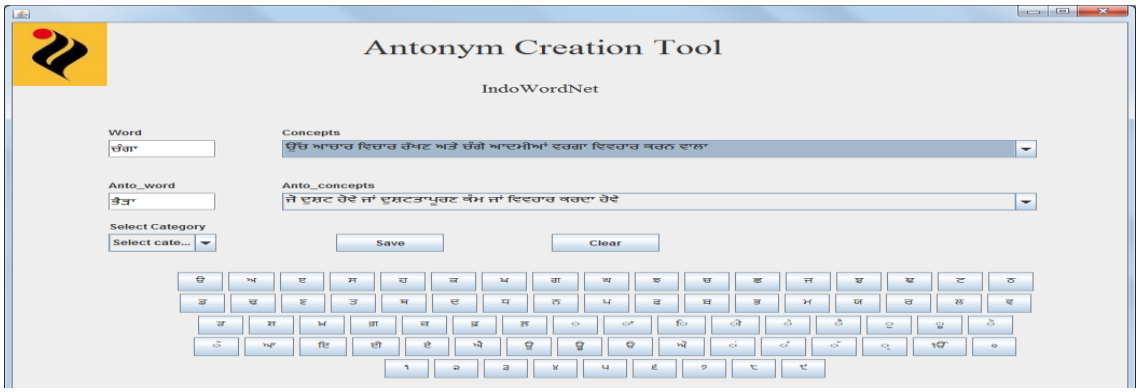

Figure 5: Interface for creation of antonyms without Hindi WordNet

For a given ***word_id***, system extracts ***synset_ids*** from ***wn_synset_word*** table. Concepts for extracted ***synset_id*** have been retrieved from ***wn_sysnset*** table as shown in step 3 given in Figure 3. The similar approach has been followed for corresponding input antonym word. The process of extraction of antonym word information is depicted in Figure 4.

A user interface has been designed in Java to provide relevant information to end user as shown in Figure 5.

4.3 Compounding creation tool

Compounding relation relates a compound word with its part word. A compound word is formed when two words are joined to form a new word**.** An interface has been designed to create such relations from compounding relations that already exist in Hindi WordNet. The tool reduces manual typing effort for the creation of compounding relation.

The snapshot of Compounding creation tool taking Hindi WordNet as basis is given in Figure 6.

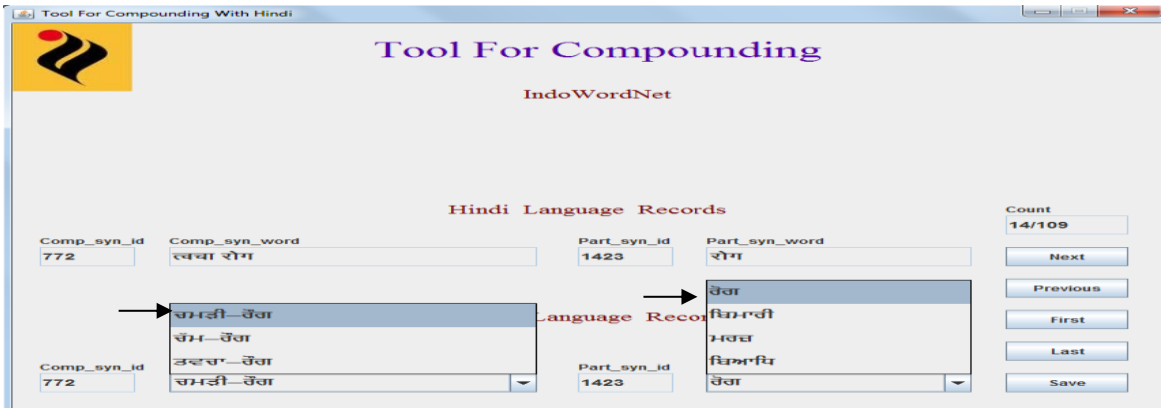

Figure 6: Compounding creation tool taking Hindi WordNet as basis

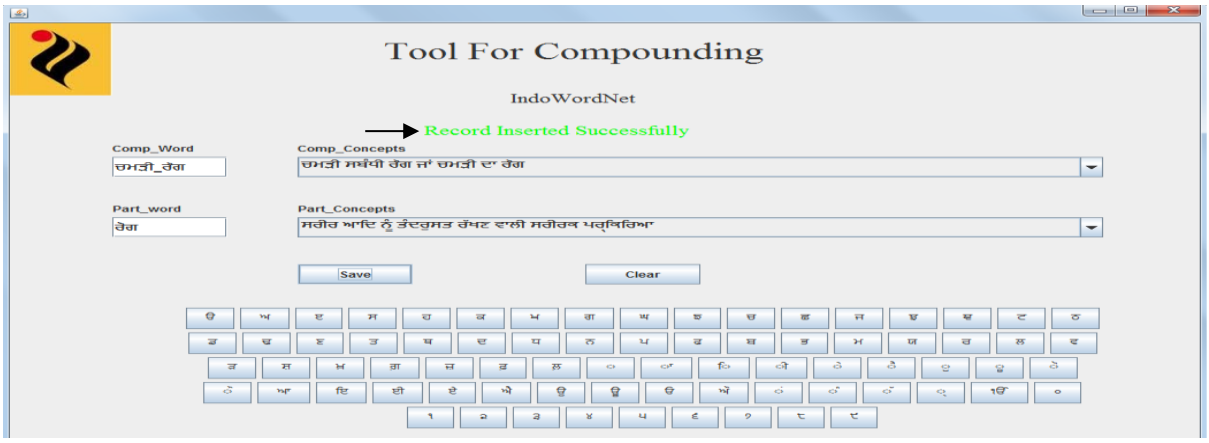

Figure 7: Compounding creation tool without taking Hindi WordNet as basis

A compounding relation may exist in target language between those words that are not covered by Hindi WordNet. For this a tool has been developed. The snapshot of compounding creation tool, without taking Hindi WordNet as basis is given in Figure 7.

4.4 Conjunction creation tool

Conjunction relation relates a conjunction word with its part word. The snapshot of conjunction creation tool taking Hindi WordNet as basis is given in Figure 8.

The snapshot of conjunction creation tool without taking Hindi WordNet as basis is given in Figure 9.

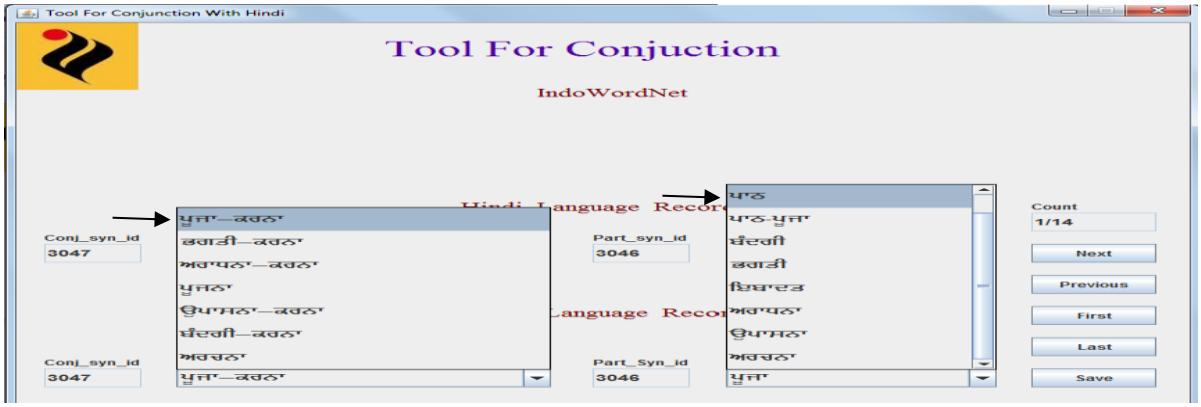

Figure 8: Conjunction creation tool taking Hindi WordNet as basis

Figure 9: Conjunction creation tool without taking Hindi WordNet as basis

4.5 Gradation creation tool

Gradation is a lexical relation that exists between three word forms. It represents the intermediate concept between two opposite concepts. The snapshot of gradation creation tool taking Hindi WordNet as basis is given in Figure 10.

The snapshot of conjunction creation tool with out taking Hindi WordNet as basis is given in Figure 11.

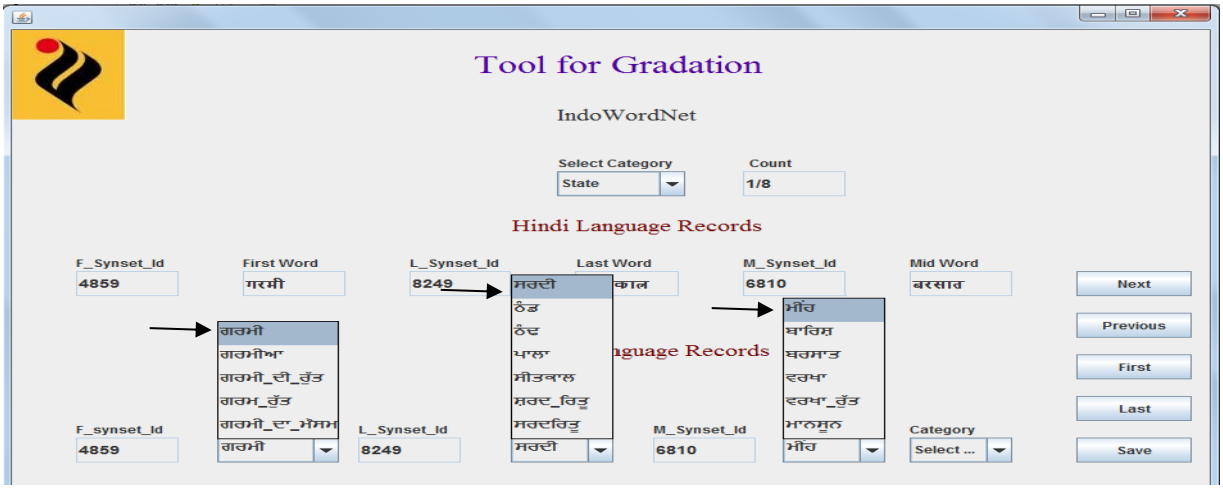

Figure 10: Gradation creation tool taking Hindi Wordnet as basis

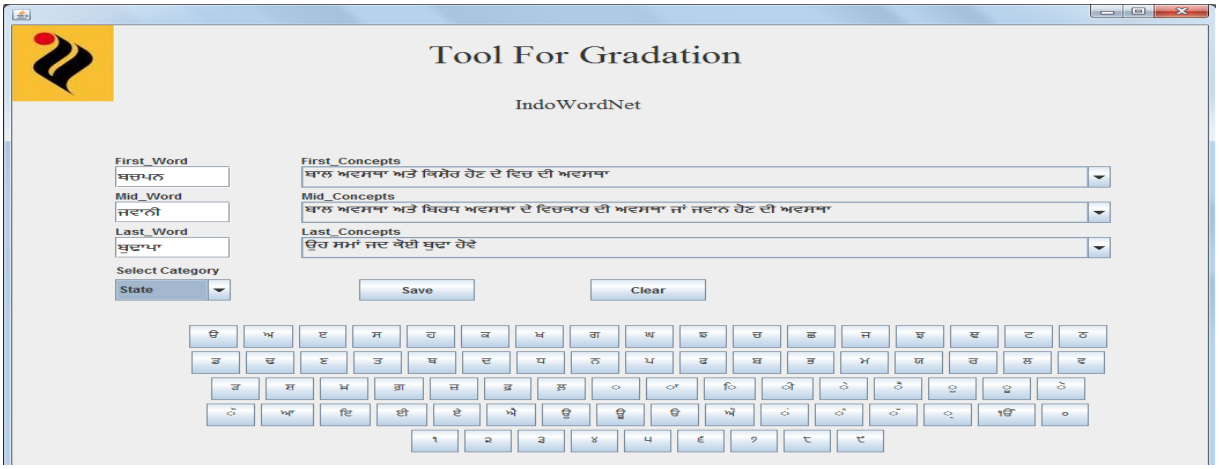

Figure 11: Gradation creation tool without taking Hindi WordNet as basis

5. Conclusion

Using expansion approach semantic relations are borrowed from the source language as they are same for all the languages. Lexical relations are language specific, so they cannot be borrowed from the source language. It has been observed that manual typing work can be reduced for Hindi in-family languages to a larger extent by creating lexical relations for target language on the basis of relations created in Hindi WordNet, while for languages that do not fall in the same family provision of creation of lexical relation without referring to Hindi WordNet will be helpful extensively.

Acknowledgements

This work has been carried out under research project titled “Development of Indradhanush: An Integrated WordNet for Bengali, Guajarati, Kashmiri, Konkani, Oriya, Punjabi and Urdu” under the leadership of IIT Bombay and Goa University. This project is sponsored by MoCIT, Govt. of India. We also acknowledge the contribution of Punjabi University, Patiala team for the development of Punjabi WordNet.

References

Dan Tufis, Dan Cristea and Sofia Stamou. 2004. Balkanet: Aims, methods, results and perspectives. A

general overview. *Romanian J. Sci. Tech. Inform.* vol.7 (1-2), pp: 9-43.

Evgeniy Gabrilovich and Shaul Markovitch. 2004. Text Categorization with Many Redundant Features: Using Aggressive Feature Selection to Make SVMs Competitive with C4.5. In *21st International Conference on Machine Learning*, Canada, pp: 321-328.

George A. Miller. 1985. WordNet: A Dictionary Browser. In *First International Conference on Information in Data*, University of Waterloo, Canada.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross and Katherine Miller. 1990. Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography*, pp: 235-244.

IndoWordNet Database Design. 2011. Tech. Rep. by Goa University, Goa.

Kedar Bellare, Anish D. Sarma, Atish D. Sarma, Navneet Loiwal, Vaibhav Mehta, Ganesh Ramakrishnan and Pushpak Bhattacharyya. 2004. Generic Text Summarization Using WordNet. In *Language Resources Engineering Conference*, Barcelona.

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

Pushpak Bhattacharyya. 2010. IndoWordNet. In *Lexical Resources Engineering Conference Malta.*

Rupinderdeep Kaur, Rajendra K. Sharma, Suman Preet, and Parteek Bhatia. 2010. Punjabi WordNet Relations and Categorization of Synsets, In *3rd IndoWordNet Workshop, IIT Kharagpur*.

Satanjeev Banerjee and Ted Pedersen. 2002. An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet. In *Third International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico City, pp: 1-10.