Punjabi WordNet Relations and Categorization of Synsets

Rupinderdeep Kaur
Computer Science Engineering Department, Thapar University, Patiala.
rupinderdeep@thapar.edu

R.K. Sharma
School of Mathematics and Computer Applications, Thapar University, Patiala.
rksharma@thapar.edu

Suman Preet

Department of Linguistics and Punjabi Lexicography, Punjabi University, Patiala.
virksumanpreet@yahoo.co.in

Parteek Bhatia
Computer Science Engineering Department, Thapar University, Patiala
parteek.bhatia@thapar.edu

Abstract

This paper describes an attempt to develop Punjabi WordNet by using expansion approach from Hindi WordNet under Indradhanush WordNet Project. The origin, symbols, morphological and syntactic characteristics of Punjabi Language are presented in this paper. The lexical semantic relations used in Punjabi WordNet are elaborated. The need for synset categorization and the results of this categorization for Punjabi Language is also presented in this paper.

Keywords: WSD, Punjabi WordNet, Indo WordNet, Universal Synsets.

1. Introduction

WordNet is a semantic lexicon for a language. It groups the words into sets of synonyms called synsets, provides short, general definitions, and records various semantic relations between these synonym sets. The purpose is twofold: to produce a combination of dictionary and thesaurus that is more intuitively usable, and to support automatic text analysis and artificial intelligence applications. WordNet is considered to be the most important resource available to researchers in computational linguistics, text analysis, and many related areas (Bhattacharyya *et al.,* 2010).

Punjabi is the language used by hundreds of millions of people in India, and is also the language used by Punjabis around the world. Surprisingly, little has been done in the field of computerization and lexical resources of this language. It is therefore motivating to develop a Punjabi WordNet under Indradhanush Project sponsored by MIT, India as an important lexical resource that discovers the richness of Punjabi language.

This paper is divided into 7 sections. Section 2 gives a brief account on the morphological and syntactical features of Punjabi language. Section 3 provides role of WordNet in Natural Language Processing (NLP). Section 4 presents an overview of relations used in WordNet with respect to Punjabi Language. Section 5 of this paper discusses the need of synset categorization and provides the results for this categorization for Punjabi Language. Section 6 concludes the work presented in this paper.

2. The Punjabi language

2.1 Origin and symbols

Punjabi language is world's 12th most widely spoken language. Punjabi

Language is used in both parts of Punjab, in India and also in Pakistan. Punjabi is syllabic in nature. It consists of 41 consonants called *vianjans*, 9 vowel symbols called *laga* or *matras* and 2 symbols for nasal sounds (. , °) (Meenu, 2007; Rupinderdeep, 2010).

2.2 Morphological characteristics

There are two genders in Punjabi Language: Masculine and Feminine. Every noun in Punjabi is assigned one of these genders. Both cardinal and ordinal numerals are found in Punjabi Language. Punjabi language has two types of affixes: Prefix and Suffix. Prefixes are less in number in comparison with suffixes. But both affixes are used in literature. There are two types of adjectives in Punjabi: inflected and uninflected. There are six types of Cases in Punjabi language, Nominative, Accusative, Instrumental, Dative, Ablative, and Locative.

2.3 Syntactic Characteristics

General syntactic structure of Punjabi language is Subject, Object and Verb (SOV). Punjabi sentences are mainly simple in structure but complex and compound sentences are also found in literature. Punjabi sentence structure is flexible. Depending on the context or mood of the speaker, it might vary. Punjabi sentences are mostly analytic in structure but the feature of synthesis is still found at dialectal level.

3. Role of WordNet in Natural Language Processing

WordNet is considered to be the most important resource available to researchers in computational linguistics, text analysis, and many related areas. Natural language processing is essential for dealing efficiently with the large quantities of text now available online. This will be specially useful for fact extraction and summarization, automated indexing and text categorization, and machine translation.

Assessment of semantic similarity has proved to be essential for a variety of Natural Language Processing (NLP) tasks, including syntactic disambiguation (either structural or functional), word sense disambiguation, selection of appropriate translation equivalent, assessment of lexical cohesion in texts for automatic summarization, query expansion and document indexing in Information Retrieval.

3.1 Word Sense Disambiguation

Word Sense Disambiguation (WSD) is regarded as one of the most interesting and longest-standing problems in natural language processing. It is the process of determining which sense of a word is the intended sense in a particular context.

A single word can be used in a language in various contexts and with different meaning in each context. For example, the word ਜੱਗ (jug) can correspond to various different meanings depending upon its usage in the sentence as given below.

ਜੱਗ -ਸਿਆਣਿਆਂ ਨੇ ਕਿਹਾ ਹੈ ਕੇ ਇਹ ਜੱਗ ਮਿਠਾ ਅੱਗਲਾ ਕਿੰਨ ਡਿੱਠਾ।

ਪਿੰਡਾਂ ਵਿੱਚ ਕਈ ਵਾਰ ਜੱਗ ਕੀਤਾ ਜਾਂਦਾ ਹੈ ਜਿੱਥੇ ਸਾਰੇ ਰੱਕ ਕੇ ਖਾਂਦੇ ਹਨ।

ਪਾਣੀ ਦਾ ਪੂਰਾ ਜੱਗ ਪੀ ਲਓ।

There are many usages of Word sense disambiguation. The most obvious application of Word sense disambiguation is Machine Translation. The machine translation process requires at least two stages, namely, understanding of the source language and generation of equivalent target language sentences. Word sense disambiguation is required in both stages since a word in the source language may have more than one possible translations in the target language.

4. Relations Used in Wordnet

WordNet groups sets of synonymous word senses into synonym sets or synsets. A word sense is a particular meaning of a word. A synset contains one or more synonymous word senses. WordNet is organized by semantic relations. Semantic relations can be represented as pointers between synsets. The central object in WordNet is a synset which is a set of synonyms. Each synset has a gloss (definition) associated with it.

4.1 Lexical and Semantic Relations

Lexical relations are the relations between members of two different synsets. For example: Antonymy is a lexical relation, {rise, ascend} and {fall, descend} are opposites but not antonyms. {rise} and {descend} are not antonyms.
Semantic relations are the relations between two whole synsets. For example: Hypernym/Hyponym relation. {organism, being} is hypernym of {plant, flora} and {plant, flora} synset is hyponym of {organism, being} (Shilpa and Parteek, 2007).
.

4.2 Synonymy

Synonymy means similarity of meaning. This relation is used to represent the words that have similar meanings. The relation is symmetric: if *x* is similar to *y*, then *y* is equally similar to *x*. Following words represent the synonymy relation between the words (Shilpa, 2007).

. For example the word ਆਜ਼ਾਦੀ(freedom) has synset ਸੁਤੰਤਰਤਾ, ਖਲਾਸੀ, ਖੁਲ, ਨਿਜਾਤ .

Similarly the following words have the synsets as follows:

ਅਦਬ – { ਅਭਿਨੰਦਨ, ਆਦਰ, ਆਦਰ ਭਾਓ, ਇੱਜਤ, ਸਤਿਕਾਰ, ਸ਼ਾਨ, ਕਦਰ, ਖ਼ਾਤਰ, ਪੂਜਾ, ਮਾਣ, ਮਾਣ ਤਾਣ, ਰਿਆਇਤ }

ਸਾਵਧਾਨ – { ਸੁਚੇਤ, ਹੁਸ਼ਿਆਰ, ਚੇਤਨਤਾ, ਚੌਕਸ, ਚੌਕੰਨਾ, ਜਾਗਦਾ, ਟਿਚਨ, ਫੁਰਤੀਲਾ }

4.3 Antonymy

Antonymy represents opposition of meanings. The words are antonyms if they are opposites in their meanings. Antonymy is a lexical relation between word forms, not a semantic relation between word meanings. For example, the word ਨੇੜੇ (near) has the antonym as ਦੂਰ (far).

4.4 Hypernymy/Hyponymy

This relation is called hyponymy/ hypernymy (variously called subordination/superordination, subset/ superset, or the ISA relation). *An x is a (kind of) y*. The relation can be represented by including in the synset a pointer to its superordinate, and including in other synset pointers to its hyponyms. Hyponymy is transitive and asymmetrical, and, it generates a hierarchical semantic structure, in which a hyponym is said to be below its superordinate. Such hierarchical representations are widely used in the construction of information retrieval systems.

For example, ਕਬੂਤਰ (Pigeon) inherits the features from superordinate ਪੰਛੀ (bird), but is distinguished from other ਪੰਛੀ (birds) by color, size and living conditions as shown in Fig. 1 (Shilpa and Parteek, 2007).

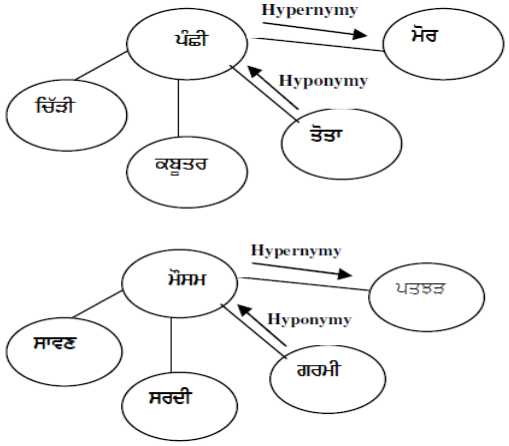

Figure 1: Hypernymy/Hyponymy Relations

4.5 Meronymy

Meronymy is a semantic relation—is the part-whole (or HAS A) relation, known as meronymy/holonymy. It is represented as *"y has an x (as a part)"* or "*An x is a part of y"*. The meronymic relation is transitive (with qualifications) and asymmetrical, and can be used to construct a part. These relations represent associations that form a complex network; knowing where a word is situated in that network is an important part of knowing the word's meaning.

For example, ਅੱਖਾਂ (eyes), ਬਾਂਹ (arm) and ਸਿਰ (head) are all parts of ਸ਼ਰੀਰ (body). This represents the meronymy/holonymy relation. ਸ਼ਰੀਰ (body) has a ਸਿਰ (head). ਸ਼ਰੀਰ (body) –meronym and ਸਿਰ (head) holonym as shown in Fig. 2 (Shilpa and Parteek, 2007).

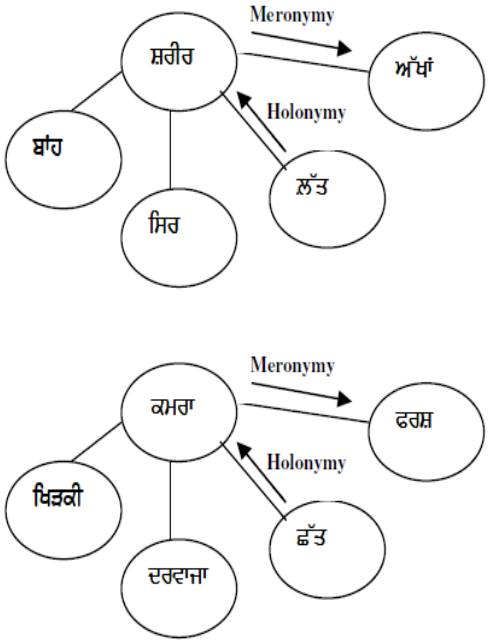

Figure 2: Meronymy/Holonymy Relations

4.6 Demonstration of Relations in WordNet

Fig. 3 shows all the relations like synonymy, hypernymy, hyponymy, meronymy etc for synset {ਘਰ(home) ,ਗ੍ਰਿਹ}. The hypernymy relation (Is-A) of it links to {ਨਿਵਾਸ, ਟਿਕਾਣਾ, ਰਹਾਇਸ਼, ਵਸੋਂ}. Its meronymy relation (Has-A) links to {ਚਬੂਤਰਾ}, {ਵਿਹੜਾ} and {ਕਮਰਾ} and hyponymy relation to {ਸਰਾਂ, ਧਰਮਸ਼ਾਲਾ}, {ਕੁਟਿਆ, ਕੁਲੀ} and {ਝੁੱਗੀ, ਝੋਪੜੀ} (Shilpa, 2007; Sinha *et al.*, 2004).

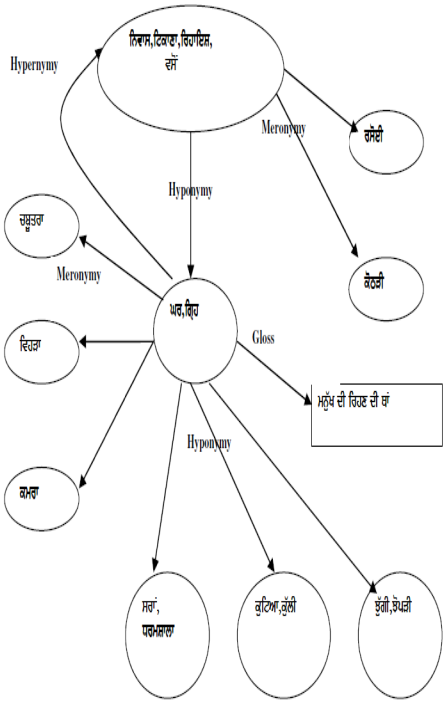

Figure 3: Relations for Synset of ਘਰ (Home)

5. Classification of Hindi WordNet Synsets

Hindi WordNet has approximately 35,000 synsets. In order to identify approximate 20,000 synsets whose concepts are available for most of Indian Languages or for a group of languages, the task of synset categorization had been started in Shillong workshop of IndoWordNet group. It had been decided that the synsets of the Hindi WordNet can be divided into different categories. These categories are as follows:

1. Universal: The synsets that are linkable across all languages of the world with natural and preferably indegenous lexemes/words are classified as Universal Synsets. For example: The synset of "Sun".
2. Pan-Indian: The synsets that are linkable across all languages of India with natural and preferably indegenous lexemes/words are classified as Pan-Indian. For example: "papad"; a kind of crispy food.
3. In-Family: The synsets that are linkable across all Indian languages belonging to a family (Indo-Aryan, Dravidian and Sino-Tibetan) with natural and preferably indegenous lexemes/words are consider as In-Family. For example: "bhatijaa"; brother's son; a naturally occurring expression in Indo Aryan family of languages.
4. Language Specific: There is no issue of linkage here. Only synsets expressing concepts that are common in a specific language have to be carefully included in the synset repository. Other languages can link to them only by constructing artificial phrases or through transliteration. For example: "bihu", the most important festival of Assam.
5. Rare: These are synsets expressing rare concepts. For example: Technical terms. These will necessarily be transliterated and a range bearing very large id numbers will be allocated for these.

It has been found from the experience of different WordNet teams that it is difficult to classify the Synsets by an individual team. Because every team has knowledge of their own language and they cannot comment on the existence of that concept in other languages or other in family languages. Thus, it is decided in the Ahemdabad Indradhanush WordNet workshop, that every team will mark all the Hindi synsets as Yes or No depending upon the existence of that concept in their own language. After marking all the synsets as Yes or No by all the WordNet teams, those IDs which are marked as Yes by all the teams will be considered as universal. Those synset IDs that are marked by group of languages as Yes will be considered as In-family and those which are marked by a particular language group as Yes will be considered as Language Specific.

Punjabi WordNet group has completed this task and it has been found that almost all the synsets are marked as 'yes' because the concepts that are in Hindi language exists in Punjabi language too as both the languages are from the same family. We have listed all the identified Hindi synsets having no equivalent concept in Punjabi language in table 1.

Table 1: Hindi Synsets having no equivalence concept in Punjabi

ID	CONCEPT	SYNSET	REASON
562	कनिष्ठा और मध्यमा के बीच की उँगली	अनामिका,उपकनिष्ठिका,अनामा	In Punjabi language, except from the thumb (ਅੰਗੂਠਾ) and the last finger (ਚੀਚੀ), we have same name (ਉਂਗਲ) for rest of the fingers.
952	अच्छी संतान प्राप्त करने की कामना से किया जाने वाला संस्कार जो गर्भाधान के तीसरे महीने में किया जाता है	पुंसवन संस्कार,पुंसवन	No such customs exists in our culture
3147	कर्मकांड में अनामिका में पहनने का कुश का छल्ला	पवित्री,पैंती,कुशमुद्रिका	No such concept exists
7650	पैर के अँगूठे में पहना जानेवाला छल्ला	अनवट	In Punjab, there exists no jewellery that is made to be worn in thumb of feet.
8091	पीठी में कुम्हड़े के टुकड़े मिलाकर बनाई हुई बरी	कुम्हड़ौरी	In Punjabi there is no concept of कुम्हड़े के टुकड़े
33079	अभ्युदय संबंधी	आभ्युदयिक	No such concept exists

6. Conclusion

In this paper, we have discussed the origin and characteristics of Punjabi language and role of WordNet in the NLP. Various relations used in wordnet like antonomy,

hypernymy *etc.* are discussed with respect to Punjabi. In order to identify the universal synsets from approximately 35,000 Hindi Synsets, they are classified into different categories. It has been found that almost all the synsets of Hindi WordNet has an existence of the concept in Punjabi language because both the languages are from the same family.

References

Bhattacharyya P., Fellbaum C. and Vossen P. 2010, *Principles, Construction and Application of Multlingual Wordnets,* Proceedings of the 5th Global Wordnet Conference*,* Mumbai, Narosa Publishing House, India.

Meenu Bhagat, 2007, *Spelling Error Pattern Analysis of Punjabi Typed Text*, M.E Thesis, Thapar University, Patiala.

Rupinderdeep Kaur, 2010, *Spell Checker for Gurmukhi Script*, M.E Thesis, Thapar University, Patiala.

Shilpa Rana, Parteek Bhatia, 2007 , *Punjabi WordNet-A Tool for Natural Language Processing*, in Second National Conference on Recent Advances and Future Trends in IT, RAFIT-2007.

Shilpa Rana, 2007, *Punjabi WordNet A tool for Natural Language Processing*, M.E Thesis, Thapar University, Patiala.

Sinha Manish, Kumar Mahesh, Pande Prabhakar, Kashyap Lakshmi and Bhattacharyya Pushpak, 2004, *Hindi word sense disambiguation*, Proceedings of International Symposium on Machine Translation, Natural Language Processing and Translation Support Systems, Delhi, India.