Development of Punjabi WordNet

Ashish Narang · R. K. Sharma · Parteek Kumar

Received: 28 June 2013 / Accepted: 21 November 2013 / Published online: 11 December 2013
© CSI Publications 2013

Abstract Natural language processing (NLP) tasks such as word sense disambiguation, machine translation (MT) and part-of-speech tagging etc. require large scale lexical resources. These lexical resources have already been developed for digitally advanced languages, such as English, but yet to be developed for widely spoken but digitally young languages, such as Punjabi. WordNet is one such lexical resource that can be used for variety of NLP tasks ranging from digital dictionary to automated MT. The basic building block of WordNet is synset, a word sense with which one or more synonymous words are associated. Each synset in WordNet is linked to other synsets using lexical and semantic relations. Lexical relations are between word forms, whereas semantic relations exist between two whole synsets. This paper presents lexical and semantic relations of Punjabi WordNet, including synonymy, hypernymy/hyponymy, meronymy/holonymy, entailment and troponymy etc. It also illustrates the process of creation of Punjabi synsets from Hindi synsets, design of synsets and semantic databases for implementation of semantic relations for Punjabi WordNet.

Keywords Punjabi WordNet · Semantic relations · Lexical relations

A. Narang
Thapar University, Patiala, India
e-mail: ashish.narang6789@gmail.com

R. K. Sharma
School of Mathematics and Computer Applications,
Thapar University, Patiala, India
e-mail: rksharma@thapar.edu

P. Kumar (✉)
Department of Computer Science and Engineering,
Thapar University, Patiala, India
e-mail: parteek.bhatia@thapar.edu

1 Introduction

WordNet can be defined as 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 can be used for concept. Besides synset data, a WordNet also maintains lexical and semantic relations. Lexical relations like antonymy and gradation are between the words in a language, whereas semantic relations like hypernymy, hyponymy, meronymy, holonymy, entailment, troponymy and causation are between concepts in a language. WordNet is considered to be a most important resource available to researchers in computational linguistics, text analysis, and many other related areas. Major applications of WordNet include word sense disambiguation (WSD) [1], text categorization [2] and generic text summarization etc.

This paper is divided into seven sections. Section 2 gives a brief account on the syntactical features of Punjabi language. Section 3 provides an overview of related work that has been done in this field. Section 4 presents semantic and lexical relations used in WordNet with respect to Punjabi Language. Section 5 of this paper covers the process of Punjabi synsets creation from Hindi synsets and its database design. Section 6 demonstrates semantic relations for Punjabi WordNet and Sect. 7 concludes the work presented in this paper.

2 Punjabi language

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 (., °) [3]. Punjabi is the language used by hundreds of millions of people in India, and is also the language used by Punjabis around the world. There are relatively less efforts in the field of computerization and development of lexical resources of this language. This is the motivation to develop Punjabi WordNet under Indradhanush Project sponsored by MoCIT, under the guidance of IIT, Bombay.

3 Related work

English WordNet is the first WordNet created in this field. Its development began in 1985 and is being maintained at the Cognitive Science Laboratory of Princeton University [4]. 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 deals with Dutch, Italian, Spanish, German, French, Czech, and Estonian languages [5]. In 2004, BalkaNet WordNet project has been initiated for development of WordNets for Bulgarian, Greek, Romanian, Serbian and Turkish languages [6].

In India, Hindi WordNet has been developed by IIT, Bombay. Later on, Hindi WordNet was extended to Marathi WordNet. Currently, IndoWordNet project, a linked structure of Indian languages is in progress in India [7]. Moreover, Indradhanush Project, aims at developing WordNets for seven major Indian languages; Bengali, Gujarati, Kashmiri, Konkani, Oriya, Punjabi and Urdu. All Indian language WordNets are being created using expansion approach from Hindi WordNet.

4 WordNet relations

WordNet contains standard information found in dictionaries and thesauri. An additional feature of WordNet is its information about the relationships between words and synsets. The synsets and words in WordNet are linked through two types of relations, i.e., semantic and lexical relations. Semantic relations are the relations between two whole synsets. For example, hypernymy/hyponymy 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 thirteen semantic relations, namely, hypernymy, hyponymy, meronymy, holonymy, entailment, causation, troponymy, ability link, capability link, functional link, attribute, modifies noun and modifies verb that have been considered for WordNet development.

Lexical relations are the relations between members of two different synsets. For example, consider two synsets given in (3) and (4).

{ਮੋਟਾ *mōṭā* 'fat', ਭਾਰੀ *bhārī* 'fat', ਸਥੂਲ *sathūl* 'fat', ਥੂਲ *thūl* 'fat', ਨਿੱਗਰ *niggar* 'fat', ਵਜਨੀ *vajnī* 'fat'}

{ਪਤਲਾ *patlā* 'thin', ਦੁਬਲਾ *dublā* 'thin', ਕਮਜ਼ੋਰ *kamzōr* 'thin', ਮਾੜਾ *māṛā* 'thin'}

Here, synsets in (3) and (4) are opposites, but these 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. But, the words ਮੋਟਾ *mōṭā* 'fat' and ਮਾੜਾ *māṛā* 'thin' does not share antonym relation. In next sub-sections, all WordNet relations have been explained from perspective of Punjabi language.

4.1 Synonymy

Synonymy means similarity of meaning. This relation is used to represent the words that have similar meanings. The relation is symmetric, i.e., if 'x' is similar to 'y', then 'y' is also similar to 'x' [8]. For example, for the word ਘਰ *ghar* 'house' some synonymous synsets are ਘਰ *ghar* 'house', ਮਕਾਨ *makān* 'house' and ਇਮਾਰਤ *imārat* 'house'.

4.2 Hypernymy/hyponymy

Hypernymy is the semantic relation between two synsets to capture super-set hood. Similarly, hyponymy is the semantic relation between two synsets to capture sub-set hood. The hyponymy relation is transitive and asymmetrical. Hypernymy is the reverse of hyponymy. For example, ਕਬੂਤਰ *kabūtar* 'pigeon' inherits the features from superset ਪੰਛੀ *pañchī* 'bird'. Here, ਕਬੂਤਰ *kabūtar* 'pigeon' is hyponym of ਪੰਛੀ *pañchī* 'bird' and ਪੰਛੀ *pañchī* 'bird' is hypernym of ਕਬੂਤਰ *kabūtar* 'pigeon' as shown in Fig. 1.

4.3 Meronymy/holonymy

It is also a semantic relation between two synsets. If concepts 'x' and 'y' are related in such a manner that 'x' is one of the constituent of 'y', then 'x' is the meronym of 'y' and 'y' is the

Fig. 1 Hypernymy/hyponymy relation

Fig. 2 Meronymy/holonymy relation

holonym of 'x'. The meronymy relation is transitive and asymmetrical. Holonymy is the reverse of meronymy. For example, ਸਿਰ *sir* 'head' is a part of ਸ਼ਰੀਰ *sharir* 'body'. Here, ਸਿਰ *sir* 'head' is meronym of ਸ਼ਰੀਰ *sharir* 'body' and ਸ਼ਰੀਰ *sharir* 'body' is holonym of ਸਿਰ *sir* 'head' as shown in Fig. 2.

4.4 Causative relation

The causative relation links the causative verbs and base verbs. It shows interdependency between both of these. For example, ਖਵਾਉਣਾ *khavāunā* 'to make someone to eat' is a causative verb of ਖਾਣਾ *khānā* 'eat'.

4.5 Entailment

It is a semantic relationship between two verbs. A verb 'x' entails a verb 'y', if the meaning of 'y' follows logically and is strictly included in the meaning 'x'. The relation of entailment is unilateral, i.e., it is one way relation. For example, ਘਰਾੜੇ ਮਾਰਨਾ *gharārē mārṇā* 'snoring' entails ਸੋਨਾ *sōnā* 'sleeping', but ਸੋਨਾ *sōnā* 'sleeping' does not entail ਘਰਾੜੇ ਮਾਰਨਾ *gharārē mārṇā* 'snoring'.

4.6 Troponymy

It is a semantic relation between two verbs when one is specific manner elaboration of another. It shows manner of an action, i.e., 'x' is a troponym of 'y', if to 'x' is to 'y' in some manner. For example, ਦਹਾੜਨਾ *dahārṇā* 'to roar' is troponym of ਬੋਲਨਾ *bōlnā* 'to speak'.

Fig. 3 Creation of Punjabi synsets from Hindi synsets

4.7 Ability link

The ability link specifies the inherited features of a nominal concept. For example, {ਤੈਰਨਾ *tairnā* 'swim'} is ability of {ਮਛਲੀ *machlī* 'fish', ਮੱਛੀ *macchī* 'fish'}.

4.8 Capability link

The capability link specifies the acquired features of a nominal concept. For example, ਤੈਰਨਾ *tairnā* 'swim' is capability of {ਆਦਮੀ *ādmī* 'person', ਬੰਦਾ *bandā* 'person'}.

4.9 Function link

The relation function link specifies the function of a nominal concept. For example, {ਪੜ੍ਹਾਉਣਾ *par hāunā* 'teach', ਗਿਆਨ ਦੇਣਾ *giān dēnā* 'teach'} is a function of {ਅਧਿਆਪਕ *adhiāpak* 'teacher', ਗੁਰੂ *gurū* 'teacher', ਮਾਸਟਰ *māstar* 'teacher'}.

4.10 Attribute

It shows linkage between noun and an adjective. For example, ਖੰਬਦਾਰ *khambhdār* 'having wings' is an attribute of {ਪੰਛੀ *pañchī* 'bird', ਪਰਿੰਦਾ *parindā* 'bird'}.

4.11 Modifies noun

Certain adjectives can modify certain nouns only. Such adjectives and nouns are linked in Punjabi WordNet by the relation modifies noun. For example, {ਸੁਪੁੱਤਰ *suputtar* 'son with good

character', ਚੰਗਾ ਪੁੱਤਰ *caṅgā puttār* 'son with good character'} can be used with {ਆਦਮੀ *ādmī* 'person', ਬੰਦਾ *bandā* 'person'}.

4.12 Modifies verb

Certain adverbs can be used only with specific verbs. Modifies verb is the relation that specifies connection between such words. For example, words ਕਦੇ *kadē* 'sometimes' and ਕਿਸੀ ਸਮੇਂ *kisī samēṃ* 'sometimes' can be used with ਕੰਮ ਕਰਨਾ *kamm karnā* 'to work'.

5 Implementation of Punjabi WordNet

For implementation of semantic relations, MySQL has been used as back-end and JSP has been used as front-end to build a web application. The process of creation of synsets has been discussed in next subsection.

5.1 Punjabi synsets creation process

Punjabi WordNet has been developed using expansion approach from Hindi WordNet. For creation of synsets from Hindi WordNet with expansion approach, a tool named as IL-MultiDic has been developed by IIT Bombay as shown in Fig. 3.

The synsets created by this tool are stored in a text file and later these synsets are ported to MySQL database. The designing of database has been discussed in next subsection.

Table 1 Synset database design

S. no.	Table name	Columns	Description of table
1.	wn_synset	synset_id concept_definition category_id	It maintains details of a synset (concept in a language)
2.	wn_word	word_id word	To store unique words of a language. It holds the vocabulary of a language
3.	wn_synset_words	synset_id word_id word_priority	It maintains synonymous words in a synset, which are used to describe a concept in a language
4.	wn_synset_example	synset_id example_content example_priority	It maintains example sentences for a concept/synset

Table 2 Semantic database design

S. no.	Table name	Columns	Description of table
1.	wn_rel_hyper_hypo	parent_synset_id child_synset_id	It maintains information of hypernymy and hyponymy relation
2.	wn_rel_mero_holo	whole_synset_id part_synset_id	It maintains information of meronymy and holonymy relation
3.	wn_rel_troponymy	synset_id troponym_synset_id	It maintains information of troponymy semantic relation
4.	wn_rel_entailment	synset_id entailed_synset_id	It maintains information of entailment semantic relation

Fig. 4 Web interface for Punjabi WordNet

Fig. 5 Search results for word ਘਰ *ghar* 'house'

5.2 Synset database design

In WordNet database, synset information has been stored in four database tables as shown in Table 1 [9].

5.3 Semantic database design

In IndoWordNet development process, synset_id has been fixed across all the languages, being considered. Thus, semantic relations become language independent in this process. The information about these relations is stored in semantic database consisting of four database tables as shown in Table 2 [9].

6 Web application for Punjabi WordNet

A web application has been developed using JSP as front-end and MySQL as back-end for implementation of semantic relations in Punjabi WordNet. As shown in Fig. 4, a user can enter the word that is to be searched in WordNet.

The system first finds *word_id* of given word from **wn_word** table then maps the *word_id*'s to the *synset_id*'s from the **wn_synset_word** table. It returns all the *synset_id*'s corresponding to that word. For each of the resulting *synset_id*'s, system finds all *word_id*'s from **wn_synset_word** table. To create a synonym set, system finds all words from **wn_word** table after matching the

The screenshot shows the Punjabi WordNet interface. At the top, there is a header with the logo and title 'PUNJABI WORDNET A Lexical Database for Punjabi'. Below the header, there is a search bar with the word 'ਘਰ' entered. The interface displays a list of synsets for the word 'ਘਰ'. The selected synset is 1901, which is a NOUN. The interface also displays synonyms, a concept, and examples in Punjabi.

Fig. 6 Hypernymy for word ਘਰ *ghar* 'house'

word_id's for a particular *synset_id*. In order to find description and example for each *synset_id*, system searches **wn_synset** and **wn_synset_example** tables, respectively. For example, search results for word ਘਰ *ghar* 'house' is shown in Fig. 5.

To view a hypernymy relation, system finds the *child_synset_ids* from the **wn_rel_hyper_hypo** table for specific *parent_synset_id*. For example, semantic relation hypernymy for word ਘਰ *ghar* 'house' is shown in Fig. 6.

Similarly, all other semantic relations have been implemented in Punjabi WordNet. Punjabi WordNet is available online for public feedback and use at <http://www.punjabiwordnet.in>, while IndowordNet, integrated WordNet for all major Indian languages is available at <http://www.cfilt.iitb.ac.in/indowordnet/> [8].

7 Conclusion

WordNet is a large lexical database of a language in which words are grouped together according to their similarity of meanings. Each concept in a language represents a synset. WordNet also maintains lexical and semantic relations. Lexical relations like antonymy and gradation are between the words in a language, whereas semantic relations like hypernymy, hyponymy, meronymy, holonymy, entailment, troponymy and causation are between concepts in a language. In Punjabi WordNet information about synsets has been stored in four database tables, i.e., **wn_synset**, **wn_word**, **wn_synset_words** and **wn_synset_example**.

Punjabi WordNet is available online for public feedback at <http://www.punjabiwordnet.in>.

References

- Chakrabarty A, Purkayastha B, Gavshinde L (2010) Knowledge-based contextual overlap keen ideas for word sense disambiguation using WordNet. In: Online proceedings of 3rd national workshop on IndoWordNet under the aegis of the 8th international conference on natural language processing (ICON 2010), Kharagpur, India
- Gabrilovich E, Markovitch S (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
- Kaur R, Sharma R, Preet S, Bhatia P (2010) Punjabi WordNet relations and categorization of synsets. In: 3rd national workshop on IndoWordNet under the aegis of the 8th international conference on natural language processing (ICON 2010), Kharagpur, India
- Miller G, Richard B, Christiane F, Derek G, Katherine J (1990) Introduction to WordNet: an on-line lexical database. *Int J Lexicogr* 3(4):235–244
- Vossen P (1998) EuroWordNet: a multilingual database with lexical semantic networks. *J Comput Linguist* 25(4):628–630
- Tufis D, Cristea D, Stamou S (2004) Balkanet: aims, methods, results and perspectives: a general overview. *Romanian J Sci Technol Inf* 7(1):9–43
- Bhattacharya P (2010) IndoWordNet. In: lexical resources engineering conference, Malta
- Hindi WordNet from CFILT, IIT Bombay, technical report (2007). <http://www.cfilt.iitb.ac.in/indowordnet/>
- (2011) Database design by Goa University, Goa, technical report