Use of Sense Marking for Improving WordNet Coverage

Neha R Prabhugaonkar
DCST, Goa University
Taleigao Plateau, Goa.
nehapgaonkar.1920@gmail.com

Jyoti D Pawar
DCST, Goa University
Taleigao Plateau, Goa.
jyotidpawar@gmail.com

Abstract

WordNet is a crucial resource that aids in several Natural Language Processing (NLP) tasks. The WordNet development activity for 18 Indian languages has been initiated in INDIA by the IndoWordNet[1] consortium using the expansion approach with the Hindi WordNet developed by IIT Bombay, as the source. After linking 20K synsets, it was decided that each of these languages should find the coverage of their respective language WordNets by using sense marker tool released by IIT Bombay.

The sense marking activity mainly helped in validation of WordNet and improving the WordNet coverage. In this paper, the various effects that sense marking activity had on the Konkani[2] language WordNet development are presented.

Keywords: sense marking, IndoWordNet, word sense disambiguation, annotation, coverage, challenges in sense marking.

1 Introduction

The IndoWordNet consortium in India is working towards the development of a multilingual WordNet which includes 18 Indian languages using the expansion approach with Hindi as source language. The IndoWordNet is a multilingual WordNet which links WordNets of different Indian languages on a common identication number called as synset Id given to each concept (Bhattacharyya, 2010).

[1] http://www.cfilt.iitb.ac.in/indowordnet/

[2] Konkani is an Indo-Aryan language and is spoken on the west coast of India. It is one of the 22 scheduled languages mentioned in 8th schedule of the Indian Constitution and the state language of the Indian state of Goa and minority language in Maharashtra, Karnataka, Kerala

Synset (Fellbaum, 1998) is composed of a gloss describing a 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. Currently, 11 language WordNets out of 18 of the IndoWordNet have created more than 20K concepts. As of now this covers around 40-50 percent of the day to day vocabulary of the respective languages. Currently, the Konkani WordNet contains 32063 concepts and more than 43200 unique words representing these concepts.

Sense marking is a task to tag each word of the corpus accurately with the WordNet sense or lexicon. In order to train machine understand the written language and thus to ensure speedy and high quality translation, a huge amount of data needs to be sense tagged precisely by humans using a standard lexicon. A word may have multiple senses and to identify which particular sense has been used in the given context, word sense disambiguation becomes a critical inevitability (Sarawati et al., 2010). In a given text, the occurrence of a particular word will correspond to only one sense and the nearby words provide strong and consistent evidence to the sense of a target word.

Language	No. of Files used	Total No. of words	Total No. of tagged words	Percentage
Bengali	11	163360	32952	20.17
Gujarati	101	337094	112884	33.49
Konkani	625	213415	103456	48.48
Kashmiri	350	98350	42290	43.00
Punjabi	45	138735	60182	43.38
Odiya	120	236125	100285	42.27
Urdu	10	100000	68689	68.69

Table 1: Sense marking status

One of the tasks in the first phase of WordNet

development was to sense mark a minimum 100K words. The source of the corpus used for sense tagging was local newspaper. The Sense Marker Tool developed by IIT Bombay was used for the sense marking activity. The table 1 shows the sense marking statistics.

The rest of the paper is organized as follows section 2 describes the Sense Marker Tool usage and the procedure used for sense-marking. The experiences of sense marking and the challenges faced are discussed in section 3. Section 4 gives the details about how the challenges were overcome and the results obtained. Section 5 gives the details about how sense marking activity helped in improving the quality of the WordNet, followed by the conclusion and future work.

2 Procedure Used for Sense Marking

The Sense Marker Tool developed by IIT Bombay was used in the sense marking task. It helps the lexicographer to efficiently tag the words. Since WordNet contains only open-class words, Sense Marker Tool is used to tag only nouns, verbs, adjectives, and adverbs; that is to say, only about 50 percent of the words in the corpus are semantically tagged. The following procedure was followed while sense marking the corpus -

- Examine each word of the text in its context of use and decide which WordNet sense was intended. In order to facilitate this task, the tool displays the word to be tagged in its context, along with the WordNet synsets for all of the senses of that word.

- Indicate the appropriate sense to the word by selecting the correct sense from the list of possible senses.

While sense marking there were situations when either the sense of the word was not found or the existing sense was not sufficient to provide the correct sense.

The main cases encountered by the lexicographers while sense marking, are listed below -

1. **Marking the word with exact sense:** The ideal situation is when the exact sense is available for the corpus word. Here, the lexicographer applying his/her language knowledge has to select the correct sense from the list of possible senses displayed by the tool.

2. **Marking the word using hypernymy:** When the exact sense is not found, the word can be tagged with its hypernymy depending on the context of the word.

3. **Marking the word with closest sense:** Sometimes the exact sense of a word is not present in the WordNet. If closest sense is available and if the lexicographer has knowledge about its existence, then he/she can assign the tag for the word with the closest sense.

4. **Creating a new sense for the word:** There are two situations when the lexicographer needs to create new sense for the word

   - If the sense of the word is not present in the WordNet. This is obvious in cases of language specific, culture specific words, species names or multiwords. Therefore it was decided that a new sense should be created for them.
   - If the sense of the word is not appropriate in the context.

5. **Marking the corpus word with the exact sense even if the sense/concept does not have the word in its synonyms set:** The word is tagged with the appropriate synset and later the word is added to the synset.

The coverage C of language vocabulary by the WordNet is measured by the following formula -

- ***Equation 1:*** *C = M*100/N, where M is the total number of words tagged and N is the total number of words in the corpus*

- ***Equation 2:*** *c = m * 100/n, where m is the total number of unique words tagged and n is the total No. of unique words in the corpus*

Equation 1 measures the coverage of more frequent words. If a frequently occurring word is covered in the WordNet then the count will increase. For Konkani language, this percentage was 48.48 percent.

Equation 2 measures the coverage of the vocabulary. If the number of words in the WordNet is high then the count will increase. For Konkani language, this percentage was 53.2 percent.

3 Challenges faced while sense marking

The main challenges faced were handling of compound words, multi-word expressions, language specific words, word with affixes, etc. They can be grouped under following heads -

3.1 Tool related challenges:

The challenges faced due to the limitations of the Sense Marker Tool are as follows:

1. There is no feature in the Sense Marker Tool to add a new synset directly to the synset file.

2. If two lexicographers are involved in the sense marking activity and both come across a same synset which is not found in the WordNet then both may end up creating a new sense. This may result in duplication of work.

3. Though the sense distinctions in the WordNet are quite fine-grained, there have been cases when the senses provided there have been inadequate or may contain some errors.

4. There is no feature in the tool to update the synset content in case of any issues like ambiguity, POS mismatch, false positive or false negative in the synonymous set, spelling mistakes, etc.

The only solution was to keep track of the information about the synsets to be created and words to be added to the existing synsets and then modify the WordNet accordingly at one place by the lexicographers. But this was a tedious and time consuming task.

3.2 Culture-Specific words

For sense marking we used corpus from the Konkani newspaper, Sunaparant. It is more likely that culture specific words occur more frequently in the corpus and these are not found in the WordNet. Examples of the frequently occurring concept specific words in Konkani newspaper corpus are:

- taraMgAM- noun, decorated pole with symbol of tutelary divinity on its top.

- huddameWI- noun, special kind of curry made with black grams and fenu-greek.

- Sigamo- noun, festival celebrated to welcome the spring which starts Holy festival.

Similarly, we have come across many such words belonging to domains such as cuisines, dance, festivals, culture and traditions, household items, etc. For the purpose of marking such words with a proper sense, it is of utmost important that the senses are to be created for them.

3.3 Named Entity Issue

It is more natural to come across many named entities such as places, companies, organizations, persons, locations, school names, personalities, etc. since the newspaper corpus was used for sense marking and news often contains such information which is not available in the WordNet.

3.4 Multi-words in the corpus

The newspaper corpus contains news on politics and critics, description on places, environment, health topics, and hence one can come across many multi-word expressions of the type compound verbs, compound nouns, idioms, echo-words, reduplication, etc. Currently, the WordNet does not store multi-word expressions. Creation of synsets for such words was also a challenging task for the lexicographers.

3.5 Words with affixes

In Konkani, one can come across a suffix like (kAr- suffix used for male), (kaAn suffix used for female) which gives different meaning to the words it is attached to. For example, (BAjI vegetable) when (kAr) is attached to it, it conveys the sense - the man selling vegetables. Similarly, when (kaAn) is attached to it, it conveys the sense the woman selling vegetables which results in the new word obtained from (BAjI). Such occurrences are quite huge in number in the corpus. However, these kinds of words are not found in the respective WordNets for the reason that all the words with the suffixes have not been incorporated.

3.6 Other challenges

Other situations where sense marking was difficult are listed below -

- The newspaper also contains many words belonging to Hindi and Marathi vocabulary. This is because Hindi and Marathi are sister languages of Konkani.

- Sometimes the newspaper articles describe information about a movie or a play, which

often use Hindi or Marathi terms. This may be because of the influence of these languages on the people. Tagging such words was also a challenge.

- Similarly we came across many foreign words in the corpus. Foreign words are those words written in a script other than our own script.

- Sense marking abbreviations and acronyms was also a difficult task as WordNet does not cover all the acronyms and abbreviations.

4 Methodologies used and Results Obtained

To overcome the challenges discussed above the following two methods were used

- **Method 1:** For each polysemous word, extract all sentences from the corpus in which that word occurs, categorize the instances and write definitions for each sense, and create a pointer between each instance of the word and its appropriate sense in the lexicon (Miller et. al, 1993). The advantage of this method was that concentrating on a single word should produce better definitions (Miller et al., 1993).

- **Method 2:** The alternative method is the sequential approach that starts with the corpus and proceeds through it word by word. This procedure has the advantage of immediately identifying deficiencies in the lexicon: not only missing words but also missing senses and inadequate senses, identifying the false positives and false negatives, etc.

The results obtained by using the combination of the above two approaches are given below -

1. Around 130 synsets were linked to Hindi WordNet and 86 new synsets having high frequency of occurrence in the corpus including concept/language specific synsets were created as a result an additional 1952 words were sense tagged.

2. Similarly, there were some synonyms which were found relevant to the context and were regarded as false negatives i.e. words which should have been present in the synset. Such words were added to the existing synsets. Additional 134 words were added which resulted in tagging of additional 380 words.

3. After analyzing the untagged words, we came across 11774 named entities in the corpus which were not available in the WordNet. It was decided that the proper noun part of the word would not be tagged, but the common noun part would be tagged. This decision helped in tagging additional 180 words.

The above methods helped in improving the WordNet coverage of Konkani language from 48.48 percent to 51.5 percent.

5 Role of Sense marking to improve WordNet Quality

The sense marking activity played a vital role in improving the quality of the WordNet in the following ways:

- Spelling errors, category mismatch were corrected and also the synsets with incomplete concept definition were improved.

- Words which had variations in spellings were added to the synsets.

- The synsets belonging to a language or language-specific synsets which covers a wide range of day-to-day language were added to the WordNet.

- Missed words (false negatives) which should have been present in the synset were added to the existing synsets.

- During sense-marking, false positives i.e. the words which were found to be irrelevant to the synsets were identified and deleted from the respective synsets.

6 Conclusion

In this paper we have discussed the importance of Sense marking activity in the WordNet development cycle. The various challenges faced, methods adopted and results obtained while sense marking have been presented. The sense marked data will act as a resource to aid in speedy and efficient machine translation, for developing and testing procedures for the automatic sense resolution in context. Our future work will be to sense mark domain specific data and to attempt to further improve the WordNet coverage and quality.

Acknowledgments

This work has been carried out as a part of the Indradhanush WordNet Project (11(13)/2010-HCC(TDIL), dated 3-8-2010) jointly carried out by nine institutions. We wish to express our gratitude to the funding agency DeitY, Govt. of India and also all the members of the Indradhanush Consortium.

References

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database. The MIT Press.*

George A. Miller, Claudia Leacock, Randee Tengi, Ross T. Bunker. 1993. *A Semantic Concordance, Pro-ceedings of the workshop on Human Language Technology, page 303–308. Stroudsburg, PA, USA, Association for Computational Linguistics.*

Jaya Sarawati, Rajita Shukla, Sonal Pathade, Tina Solanki, Pushpak Bhattacharyya. 2010. *Challenges in Multilingual Domain-Specific Sense-marking, Principles, Construction and Application of MultilingualWordNets, Proceedings of the 5th GlobalWordNet Conference,* Mumbai- India.

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