Synset Ranking of Hindi WordNet

Sudha Bhingardive, Rajita Shukla, Jaya Saraswati, Laxmi Kashyap, Dhirendra Singh and Pushpak Bhattacharyya

Department of Computer Science and Engineering,
Indian Institute of Technology Bombay, India.
{sudha,rajita,jayas,yupu,dhirendra,pb}@cse.iitb.ac.in

Abstract

Word Sense Disambiguation (WSD) is one of the open problems in the area of natural language processing. Various supervised, unsupervised and knowledge based approaches have been proposed for automatically determining the sense of a word in a particular context. It has been observed that such approaches often find it difficult to beat the WordNet First Sense (WFS) baseline which assigns the sense irrespective of context. In this paper, we present our work on creating the WFS baseline for Hindi language by manually ranking the synsets of Hindi WordNet. A ranking tool is developed where human experts can see the frequency of the word senses in the sense-tagged corpora and have been asked to rank the senses of a word by using this information and also his/her intuition. The accuracy of WFS baseline is tested on several standard datasets. F-score is found to be 60%, 65% and 55% on Health, Tourism and News datasets respectively. The created rankings can also be used in other NLP applications *viz.*, Machine Translation, Information Retrieval, Text Summarization, *etc.*

Keywords: Synset Ranking, WordNet, Hindi WordNet, Word Sense Disambiguation, WordNet First Sense, NLP

1. Introduction

Word Sense Disambiguation (WSD) is the ability to identify the meaning of words in context in a computational manner (Navigli, 2009). It is one of the toughest areas in natural language processing (NLP). Recently, a lot of research has been done for making powerful WSD systems with supervised, semi-supervised and unsupervised techniques. In WSD, the heuristics of choosing the most frequent sense is often found to be very hard for any WSD system. The WordNet First Sense (WFS) baseline is the most powerful baseline in WSD, even though it does not consider the context while assigning the senses. This baseline can be created by considering the sense-annotated statistics. For English, WFS baseline is created by using the frequencies of word senses from the sense-annotated SemCor corpus. Senses that have not occurred in SemCor are ordered arbitrarily. This WFS baseline is a very stong baseline in English WSD. Considering both precision and recall, only 5 of 26 systems in the Senseval-3 English all-words task were able to beat this baseline. Our goal is to create a WFS baseline for Indian language WordNets. We focus on Hindi language as the synsets of Hindi WordNet are not ranked according to the actual usage. This is because Hindi Wordnet was built using a dictionary where words were picked up according to the alphabetical order.

The rest of the paper is organized as follows. Section 2 gives a detailed description of Hindi WordNet. Hindi WordNet synset ranking methodology is explained in section 3. Section 4 gives the statistics of the ranked synsets. Section 5 highlights the performance of WFS baseline on various domains. Discussion is given in section 6, followed by the conclusion.

2. Hindi WordNet

Hindi WordNet[1] (HWN) is developed for capturing the fine grained senses of Hindi language. It consists of synsets and semantic relations. It is a part of IndoWordNet[2] (Bhattacharyya, 2010) which is the most useful multilingual lexical resource in Indian languages. HWN, inspired by English Wordnet, is created manually using lexical knowledge from various dictionaries. At first, the most common day-to-day words from a monolingual dictionary (*Bhargava Adarsh Hindi Shabdkosh*, ed. P. Ramchand) were incorporated. As soon as the last letter was reached, the whole process was repeated with the next set of common words. This was done till all the words in that dictionary were incorporated in Hindi WordNet. Then words from other dictionaries (*Samantar Kosh*, ed. Arvind Kumar, *Nalanda Vishal Shabd Sagar*, ed. Shri.Navalji and *Lokbharti Brihat Pramanik Hindi Kosh* by Acharya Ramchandra Verma) were picked up.

The current statistics of HWN is given in Table 1. HWN is used in various NLP applications like Word Sense Disambiguation (Khapra et al., 2010; Bhingardive et al., 2013; Bhingardive et al., 2015), Information Retrieval (Atreya et al., 2013), Sentiment Analysis (Joshi et al., 2010; Balamurali et al., 2012; Kashyap and Balamurali, 2013), Machine Translation (Ramanathan et al., 2008; Kunchukuttan et al., 2012), *etc.*

[1] http://www.cfilt.iitb.ac.in/wordnet/webhwn/index.php

[2] Wordnets for Indian languages have been developed under the IndoWordNet umbrella. Wordnets are available in the following Indian languages: Assamese, Bodo, Bengali, English, Gujarati, Hindi, Kashmiri, Konkani, Kannada, Malayalam, Manipuri, Marathi, Nepali, Punjabi, Sanskrit, Tamil, Telugu and Urdu. These languages cover 3 different language families, namely, Indo Aryan, Sino-Tibetian and Dravidian. http://www.cfilt.iitb.ac.in/indowordnet/

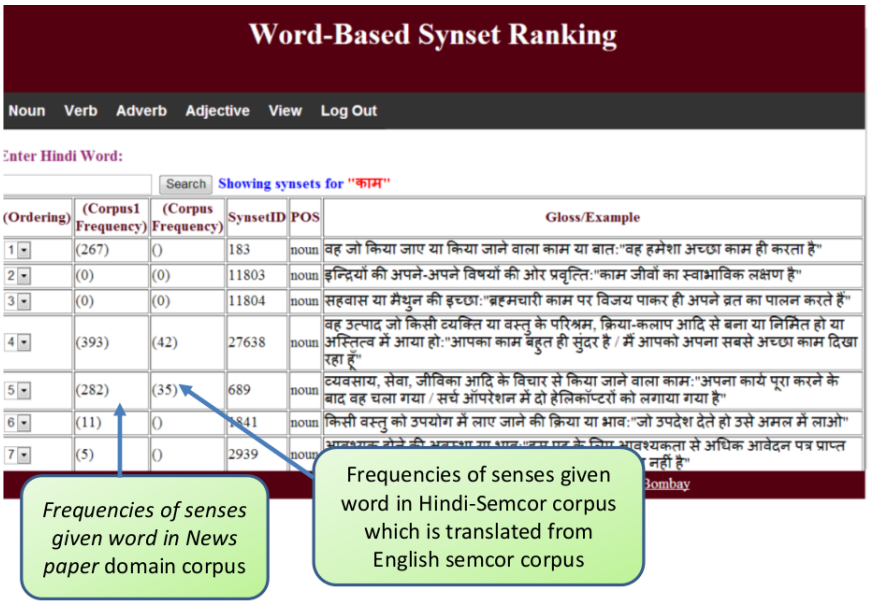

Figure 1: Synset Ranking Tool

POS	Synsets	Total Words	Polysemous Words
Noun	29104	78837	16516
Adjective	6178	18792	3575
Adverb	482	1936	218
Verb	6354	4816	1538
Total	39069	104381	21847

Table 1: Statistics of Hindi WordNet

3. Synset Ranking Methodology

For HWN synset ranking, we split the word-senses into three groups. Three human experts, who are native speakers of the language, were asked to rank the word-senses with the help of a synset ranking tool. This tool is developed for ranking the synsets of words of all POS categories. The screenshot of the tool is shown in figure 1. The tool provides the following functionalities to human experts.

- **Insert Ranking:** An input box is provided for the word and its POS. For a given input word and its POS, the tool displays all the synsets of that word extracted from Hindi WordNet with its default ranking. The tool also provides the frequencies of word-senses extracted from the sense-annotated corpus of various in-house datasets. Experts have been asked to rank the senses of a word based on this information and also his/her intuition. If the experts get confused or are unable to rank the synsets of a word, then he/she can skip the word from its ranking for the moment and move on to the next word.
- **Display Ranking:** An expert can see the already ranked synsets by providing a word and its POS.
- **Reset Ranking:** The experts have been given the facility of resetting the previous synset ranking of a word.
- **View Skipped Words:** All words which have been skipped by the experts are displayed for further discussion with other experts, leading to their ranking.
- **View Ranked Words:** The tool displays the words which are already ranked by the experts.
- **View Statistics:** The tool also provides statistics of the ranked synsets of words by all the experts who participated in the ranking process.

For some Hindi words, we can find different spelling variations. For example, the word ठंडा (*ThaMDaa*, cool) can be written as ठण्डा (*ThaNDaa*) or ठन्डा (*ThanDaa*). In such cases, the experts have been asked to rank the synsets of only one variation of such words. The same ranking will be given to the other variants of the word automatically.

4. Statistics of Synset Ranking

The statistics of synset ranking is shown in Table 2. As we can see in Table 2, we have ranked the synsets of 16516 nouns, 3575 adjecives, 208 adverbs and 1449 verbs till date. We are still in the process of ranking the newly-made synsets. These rankings have been made available on the website of CFILT[3].

[3]http://www.cfilt.iitb.ac.in/Downloads.html

POS	Words whose synsets are ranked
Noun	16516
Adjective	3575
Adverb	208
Verb	1449

Table 2: Statistics of ranked synsets

Dataset	Precision	Recall	F-score
Health	62.29	58.10	60.12
Tourism	67.81	64.07	65.88
News	58.32	52.53	55.28

Table 3: Performance of WFS baseline on WSD datasets

Algorithm	NOUN	ADV	ADJ	VERB	Overall
WFS (our)	58.69	**76.64**	**58.73**	**64.31**	**60.12**
EM-Context	59.82	67.80	56.66	60.38	59.63
EM	**60.68**	67.48	55.54	25.29	58.16
RB	35.52	45.08	35.42	17.93	33.13

Table 4: Performance of WSD algorithms on Health dataset

5. Performance on WSD task

In order to see how well the synsets are ranked, we check the performance of the WSD task. In this the first listed sense *i.e* WordNet First Sense (WFS) is given to all words irrespective of the context in which they appear in the corpus. We considered standard datasets[4] available freely for Hindi-Health, Hindi-Tourism, Hindi-News domains. The results are obtained in terms of precision, recall and f-score and are given in Table 3. F-score of WFS baseline on Health, Tourism and News domains was found to be 60%, 65% and 55% respectively.
We also compared this WFS baseline against some WSD algorithms as listed below:

- **EM-Context**: It is context-aware unsupervised WSD algorithm by Bhingardive et al., (2013) which uses Expectation Maximization (EM) algorithm for finding the sense distribution.
- **EM**: It is a basic EM based algorithm by Khapra et al., (2011) which does not consider context while finding the sense distribution.
- **RB**: It is the Random Baseline where senses are randomly assigned to words.

The results of these WSD algorithms are shown in Table 4 and Table 5. As we can see in the tables, WFS baseline beats all WSD algorithms even though it assigns senses irrespective of context. Hence, it is clear that HWN synset rankings given by human experts are of good quality and thus can be used in other NLP applications too.

6. Discussion

While ranking the HWN synsets, human experts faced some difficulties which are mentioned below. The solutions which were applied to such cases are also given.

- Some of the Hindi words (for instance नँधना - *nandhanaa*, to be harnessed) were very unfamiliar to the human experts and hence, made the ranking process difficult. In such cases, they took the help of dictionaries for ranking the synsets of such words.
- Synset ranking of highly polysemous words like निकलना (*nikalanaa*) (31 senses), निकालना (*nikaalanaa*) (31 senses), लगना (*laganaa*) (25 senses) , चढ़ना (*chadhana*) (21 senses), *etc* was found to be too tedious. For such words, the human experts were allowed to rank the top 10 most frequent senses of the words while rest of the senses were ranked according to the order given in the dictionary.

It was seen that most of the time the literal senses of a word were placed in ranks above the metaphorical or figurative uses. However, at times the ranking order did not adhere to the above mentioned criterion. In such cases the intuition of the expert and the usage of the word in common parlance took precedence. Here the rankings in dictionaries and Google search results were also ignored. For example, in the word अकड़ना (*akadana*) the figurative sense comes above the literal/physical sense of the word.

Hindi vocabulary has a number of foreign language words, mostly taken from English and these find a place in Hindi WordNet as well. The ranking of such words has been done based on the usage and it is observed that many times this may not necessarily match those found in English WordNet. An example of such a case is the word अकैडमी (*academy*), a word borrowed from English. The first sense assigned to this word in HWN (ID:10350) does not correspond to the first sense in EWN (ID:08296219). This is because the first sense found in English WordNet has negligible usage in India. The reason for this phenomenon may have historical roots.

During the synset ranking process, various HWN synsets have been validated. Some of the examples are listed below.

- **Insertion of synset members:** While ranking the synsets of words such as निकालना (*nikaalanaa*), the experts added बर्खास्त करना (*barkhaasta karanaa*) as a new synset member in the same synset (ID:11385).
- **Reordering of synset members:** In this, for example, the position of the word ठनना

[4]http://www.cfilt.iitb.ac.in/wsd/annotated_corpus

Algorithm	NOUN	ADV	ADJ	VERB	Overall
WFS (our)	**69.22**	**78.69**	53.85	**58.04**	**65.88**
EM-Context	62.90	62.54	53.63	52.49	59.77
EM	63.88	58.88	**55.71**	35.60	58.03
RB	33.83	38.76	37.68	18.49	32.45

Table 5: Performance of WSD algorithms on Tourism dataset

(*Thananaa*) in a synset (ID: 13494) has been changed from 2^{nd} position to 3^{rd} position in the synset. The updated order of the synset members is {अड़ना, उतारू होना, ठनना, अरना} (*aDanaa, utaaruu honaa, Thananaa, aranaa*).

- **Deletion of synset members:** Some synset members were deleted because they were found to be outliers due to the fine granularity of sense. For example, the word उड़ाना (*uDaanaa*) has been deleted from the synset members of synset (ID=11952).
- **Insertion of new synsets:** It was found that some frequently used senses were missing and thus were added. For example, for the word चढ़ाना (*chaDhaanaa*), a new synset has been added in the sense of कर्ज चढ़ाना (*karja chaDhaanaa*).
- **Deletion and merging of synsets:** During synset ranking, some synsets have been merged because of overlapping of senses. In this process some sysnets had to be deleted. For example, Synset (ID= 36173) of the word उड़ाना (*uDaanaa*) has been deleted as it was found to be same as of Synset (ID= 36981).
- **Correction of Hindi-English linkages:** Some Hindi-English linkages have been corrected during the ranking process. For example, the English linkage of the word निकालना (*nikaalanaa*) in the synset with ID: 11385 was found to be inaccurate. The correct English linkage is found to be *'depose, force_out force to leave (an office)'*.
- **Correction of semantic relations:** Semantic relations have also been corrected during the ranking process. For example, during synset ranking of the word काटना (*kaaTanaa*), experts came to know that synset (ID:7691) was wrongly linked to synset (ID: 245) via hypernymy relation. Such wrong relations have been corrected.

7. Conclusion

We presented our work on manually ranking the synsets of Hindi WordNet. Human experts ranked the synsets of a given word by using the synset ranking tool which is developed for the ranking purpose. The tool provides the information about words and their senses and also the frequencies of word-senses extracted from the sense-annotated corpus. The created rankings are evaluated on WSD task and it is observed that WSD, when assigning the first ranked sense i.e. WFS, can outperform the other WSD algorithms which have been proposed earlier. This process of ranking has led to the validation of HWN. The created rankings can also be used in other NLP applications *viz.,* Machine Translation, Information Retrieval, Sentiment Analysis, *etc.* The ranking tool created by us can be easily extended for ranking synsets of other Indian language wordnets.

8. Acknowledgments

We thank the CFILT members at IIT Bombay, TDIL and DeitY for their continued support.

9. Bibliographical References

Atreya, A., Kakde, Y., Bhattacharyya, P., and Ramakrishnan, G. (2013). Structure cognizant pseudo relevance feedback. In *the 6th International Joint Conference on Natural Language Processing (IJCNLP)*, Nagoya, Japan.

Balamurali, A. R., Joshi, A., and Bhattacharyya, P. (2012). Cross-lingual sentiment analysis for indian languages using linked wordnets. In *International Conference on Computational Linguistics (COLING)*, Mumbai, India. Citeseer.

Bhattacharyya, P. (2010). Indowordnet. In *Language Resources and Evaluation Conference (LREC)*, Malta.

Bhingardive, S., Shaikh, S., and Bhattacharyya, P. (2013). Neighbors help: Bilingual unsupervised wsd using context. In *Association for Computational Linguistics (ACL)*, Sofia, Bulgaria.

Bhingardive, S., Singh, D., V, R., Redkar, H. H., and Bhattacharyya, P. (2015). Unsupervised most frequent sense detection using word embeddings. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pages 1238–1243.

Joshi, A., Balamurali, A. R., and Bhattacharyya, P. (2010). A fall-back strategy for sentiment analysis in hindi: a case study. *Proceedings of the 8th International Conference on Natural Language Processing (ICON)*.

Kashyap, P. and Balamurali, A. R. (2013). The haves and the have-nots: Leveraging unlabelled corpora for sentiment analysis. In *Association for Computational Linguistics (ACL)*, Sofia, Bulgaria.

Khapra, M. M., Shah, S., Kedia, P., and Bhattacharyya, P. (2010). Domain-specific word sense

disambiguation combining corpus based and wordnet based parameters. In *5th International Conference on Global Wordnet (GWC)*, Mumbai, India.

Khapra, M. M., Joshi, S., Chatterjee, A., and Bhattacharyya, P. (2011). Together we can: Bilingual bootstrapping for wsd. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 561–569, Stroudsburg, PA, USA.

Kunchukuttan, A., Roy, S., Patel, P., Ladha, K., Gupta, S., Khapra, M. M., and Bhattacharyya, P. (2012). Experiences in resource generation for machine translation through crowdsourcing. In *Language Resources and Evaluation Conference (LREC)*, pages 384–391, Istanbul,Turkey.

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, 41(2):10.

Ramanathan, A., Hegde, J., Shah, R. M., Bhattacharyya, P., and Sasikumar, M. (2008). Simple syntactic and morphological processing can help english-hindi statistical machine translation. In *International Joint Conference on Natural Language Processing (IJCNLP)*, pages 513–520, Hyderabad, India.