Cross-Lingual Word Sense Disambiguation using Wordnets and Context based Mapping

Prabhat Pandey, Rahul Arora
Advisor: Dr. Amitabha Mukerjee
{prabhatp,arorar,amit}@cse.iitk.ac.in
Department of Computer Science and Engineering,
IIT Kanpur, India

April 13, 2012

Abstract

Word Sense Disambiguation (referred to as WSD henceforth) is the task of finding the appropriate sense of a word used in a given sentence, when the word may have multiple senses. For example consider these two sentences -

Mary walked along the **bank** of the river.
HarborBank is the richest **bank** in the city.

It can be noticed that the word **bank** refers to 'river-side' in first sentence and 'financial institution' in the second sentence. Similarly in the following sentences -

रमेश को दिन में सोना पसन्द है
सोना एक कीमती पदार्थ है

The Hindi word सोना refers to 'sleep' in the first sentence while it points to 'gold' in the second sentence.

There are basically four conventional approaches to WSD - knowledge-based, supervised, semi-supervised and unsupervised. In the recent times, cross-lingual approaches have shown some good results for languages with scarce resources. In this paper, we propose a cross-lingual approach similar to [11] for Hindi language. This approach make use of Wikipedia articles which are present both in English as well as Hindi, WordNet[6] and Hindi Wordnet[7].

1 Introduction

Out of the four conventional approaches to WSD, supervised methods have been shown to be the leading performer. But languages like Hindi lack such accurate and large sense-tagged corpus required for supervised approaches. It encourage to investgate for cross-lingual approach which can make use of novel sense disambiguation systems for resources-rich language to disamiguate senses in resources-lacked languages via parallel texts.

[11] describes a novel cross-lingual approach using Wikipedia artciles as comparable corpus for Persian WSD. In this paper we have tried to work on similar guidelines for Hindi WSD. Though Wikipedia pages are not direct translations, they may be conidered as comparable corpus as the contexts are same in the two pages. This approach can be used to build sense-tagged corpus for Hindi which can be further used for supervised and semi-suprvised WSD systems.

We have followed a three step strategy - 1.Disambiguating words in English text, 2.Finding correct Hindi synset for the tagged English synset,and 3.Tagging the words in Hindi text with the mapped sense label. This approach attained a reasonably good accuracy and was able to dismbiguate almost half of the polysemous Hindi words, but it suffered a low recall when applied for noun words.

The roadmap is as follows: Section 2 describes the Wordnets Section 3 covers related work, Section 4 outlines the cross-lingual approach, evaluation and results are discussed in Section 5 and finally Section 6 points out the limitations and flaws and suggests direction for future works.

2 Wordnets

WordNet[6] is a lexical database for the English language. It groups English words into sets of synonyms called synsets, provides short, general definitions, and records the various semantic relations between these synonym sets. WordNet has been used in many knowledge based WSD applications over the years.
Hindi Wordnet[7] was developed on the lines of the (English)WordNet, providing semantic relations between Hindi words. In the Hindi Wordnet, the words are grouped together according to their similarity of meanings. Two words that can be interchanged in a context are synonymous in that context. For each word there is a synonym set, or synset, in the Hindi Wordnet, representing one lexical concept.

2.1 Relations in WordNet (Some examples)

- **Hyponymy and Hypernymy**: These relations express super-set hood and sub-set hood respectively. e.g. car and vehicle
- **Meronymy and Holonymy**: These relations express part-whole relation. e.g. root and tree
- **Entailment**: It is a relationship between verbs such that one's truth follows from the other. e.g. snore and sleep

Other relations like Antonymy, Gradation and Linkages are also defined in both the Hindi Wordnet as well as (English)WordNet.

3 Related Work

Lesk[5] developed a knowledge-based method to disambiguate words using dictionary definitions (gloss) in 1986. The algortithm counts the number of words that are shared between two glosses to determine the relateness of two senses. To disambiguate a polysemous word, the gloss of each word is compared to the glosses of the phrase in which it occurs. Its major limitation is that dictionary definitions are generally brief, and not enough to pinpoint the correct sense in which the word has been used in.

The Extended Lesk algorithm[1] extends the gloss exploration technique to include the glosses of other concepts to which they are related according to a given concept heirarchy. Thus it gives a better prediction of the sense that the word was used in in the reference text. With the use of WordNet and Extended Lesk algorithm, English words have been mapped to their correct senses with a high degree of accuracy.

Recent studies[3] have shown that the Cross-Lingual approaches have produced more relaible and finer sense distinctions and offers advantage to languages for which we do not have large sense-annotated corpora or sense inventories. In cross-lingual approaches, target words are disambiguated by labelling them with appropriate translation in other language. For example, [4] uses sentence aligned parallel corpus (Europarl[2]) to train WSD classifier and

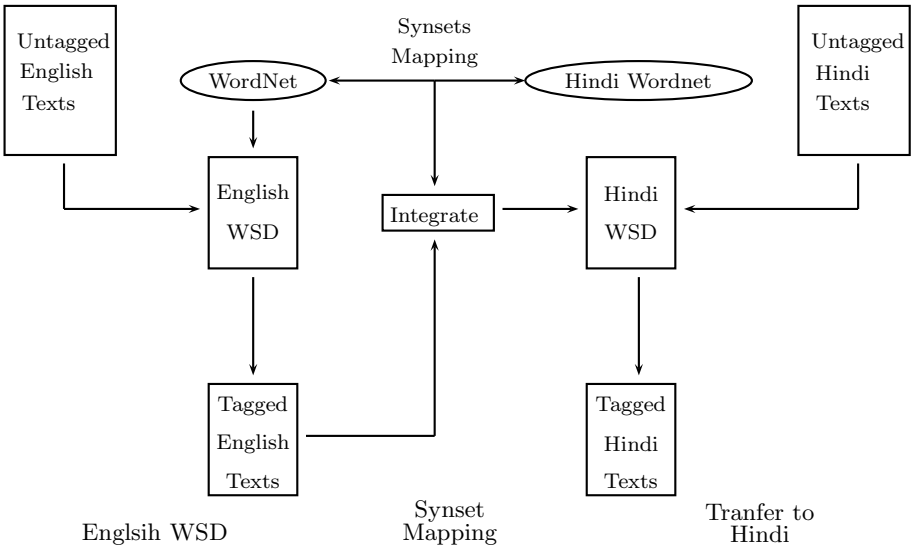

Fig.1. Outline of Cross-Lingual Approach[1]

then the sense inventory was constructed by using word alignment tools like GIZA++[8] and extracting out translations of target ambiguous words. However, there is a lack of such sentence aligned parallel corpus for languages like Hindi.

4 Cross-Lingual Approach

This approach consists of three parts - English Word Sense Disambiguation, Synset Mappings and English to Hindi Transfer. We created comparable corpus by utilizing Wikipedia pages which are available in both English and Hindi language. Figure 1 indicates the whole approach. Note that this approach focuses only nouns.

4.1 English Word Sense Disambiguation

The first step produces sense-tagged text from the raw untagged English text using English WSD tools. As mentioned earlier, we have various optimal systems for tagging English words. In our system, we used Perl-based application SenseRelate[9] which uses Lesk Algorithm for disambiguation. We selected WordNet::SenseRelate::AllWords[9] which tags all the words in the input text and it is not restricted to any specific target word. The non-noun forms were thrown out as the noun words are the focus of our approach. Among the repeated words, the most frequent sense for the word is chosen and the other senses were modified to the most frequent one.

4.2 Synset Mapping

After the first step, English words are tagged with the synsets in WordNet whcih needs to be mapped to the appropriate synset in the Hindi Wordnet. A novel algorithm has been proposed in [10] to match synsets of WordNet to synsets in Hindi Wordnet. The algorithm takes an English synset as input and produces the best possible Hindi synset. The alogrithm is as follows-

- As the first word in a synset best describes a synset[6], the first word is extracted out and all its traslations are found in English-Hindi dictionary.
- All the synsets in Hindi Wordnet which contains the above translations are searched which serves as *candidate synsets*.

[1]Figure taken from [11]

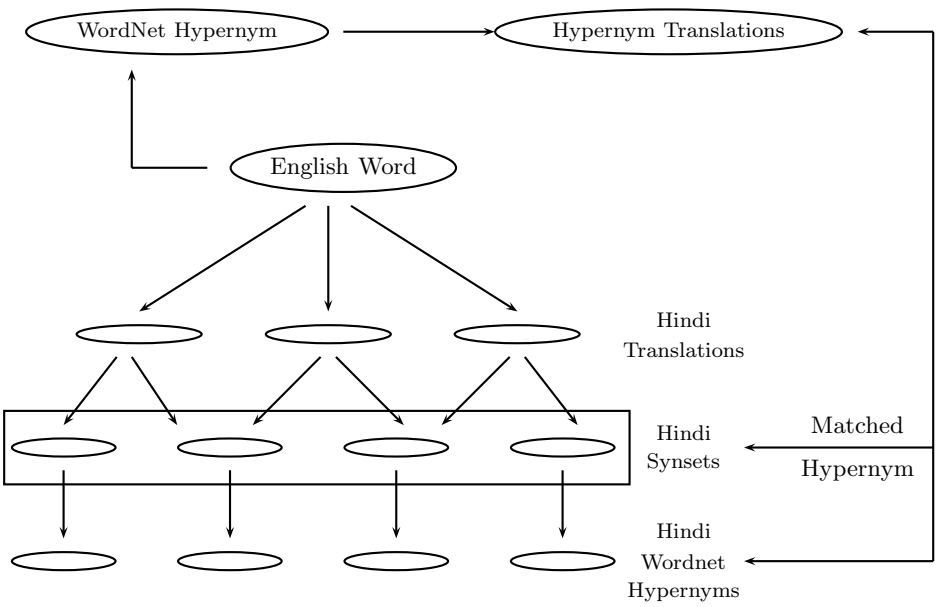

Fig.2. Outline of Synset Mapping Algorithm

- The hypernymy hierarchies of each *candidate synsets* are obtained which are the *candidate hypernyms*.
- The hypernymy hierarchy of the original English synset in WordNet is obtained.
- The Hindi translations of all the words in the above English hypernymy are obtained.
- Then the translations are searched for match in *candidate hypernymy* and the weights of *candidate hypernymy* are incremented whenever a match is found.
- Finally, that *candidate synset* is selected whose corresponding *candidate hypernymy* has the highest weight.

The above algorithm is indicated in Figure 2.

4.3 English to Hindi Transfer

In the final step, the words in the Hindi texts are to be labelled with the same sense as that of corresponding English Words. All the words in the Hindi synset obtained from the second step are searched in the corresponding Hindi Wikipedia article and it is assigned the same sense label as that of its English counterpart. Both the English sense tag and Hindi Wordnet *Offset* are labelled. For example, उत्पादन was tagged in the following way -

उत्पादन[output#n#1]#3790

It can be seen that three cases are posible here -

1. The English word is polysemous while the equivalent Hindi word is monosemous.
   For eg. - Guinea गिनी

   Here, 'guinea' has primarily three senses in WordNet- 'Republic of Guinea', 'coin' and 'foul'. But the equivalent Hindi word for the first sense, गिनी has only one sense. Following synset was obtained when Synset Mapping was applied for the first sense of 'guinea' -

   [गिनी , गिनी_गणराज्य , फ्रांसीसी_गिनी]

2. Both the English word and the Hindi Word are polysemous.
   For eg. - Party पार्टी

   The English word 'Party' has multiple senses such as 'polictical organisation' or 'function, a social occasion' and the corresponding Hindi word, पार्टी too has more than one senses. पार्टी may refer to any of the two senses of 'party'. Following synset was obtained when Synset Mapping was applied for the first sense of 'party' -

   [मंडली , टोली , संघ , पार्टी , मण्डली , संघात]

3. English Word is monosemous while equivalent Hindi word is polysemous.
   For eg. - Mango आम

   'Mango' has ony one sense as the 'fruit mango'. But the Hindi word for 'Mango' - आम has two senses - फ़ल ('fruit') and साधारण ('common'). Following synset was obtained when Synset Mapping was applied for 'mango' -

   [आम , रसाल , आम्र , अम्ब , अंब , प्रियाम्बु]

Most of the instances of the first case were correctly matched while the last two case resulted in the least accuracy. But in this particular instance of 'Mango' and आम, it matched to the correct sense may be because the other sense of आम, i.e., साधारण ('common') is adjective form and the first step of Synset Mapping algorithm discards all those translations which are in non-noun form.

5 Results

The proposed approach have been tested on five Wikipedia articels - 'Agriculture, Biodiversity, Conservation_biology, Ganges and India'. Each of these artciles contained atleast 1000 Hindi words. The results obtained are illustrated in Table 1. A total of 137, 97, 125, 99 and 113(in same order as in Table 1) distinct words were tagged in the five articles. The English synset tag and the Hindi synset tag sometimes differed in the senses and so, only the Hindi sense tag was considered for evaluation. As mentioned earlier, monosemous Hindi words were almost correctly tagged, we have analyzed the performance of polysemous Hindi Words seperately in Table 2.

Sl.No.	Wikipedia Artcle	Correct Sense	Almost Accurate	Wrong Sense	Precision[2]	Recall	F-Score[3]
1.	Agriculture	80.29%	5.11%	14.60%	0.80	0.17	0.28
2.	Biodiversity	83.83%	4.04%	12.12%	0.84	0.16	0.27
3.	Conservation_biology	74.40%	5.60%	20.00%	0.74	0.20	0.31
4.	Ganges	84.00%	4.00%	12.00%	0.84	0.15	0.25
5.	India	83.18%	3.54%	13.27%	0.83	0.20	0.32

Table 1. Results of all the tagged words

[2] $Precision = NumberofWordsTaggedCorrectly/TotalNumberofWordsTagged$
[3] $F - Score = 2.(Precision.Recall)/(Precision + Recall)$

Sl.No.	Wikipedia Article	Correct Sense	Almost Accurate	Wrong Sense
1.	Agriculture	38.89%	11.11%	50.00%
2.	Biodiversity	44.44%	14.81%	40.74%
3.	Conservation_biology	45.45%	9.10%	45.45%
4.	Ganges	41.66%	20.83%	37.50%
5.	India	57.69%	3.84%	38.46%

Table 2. Results of tagged polysemous words

It is evident from Table 2 that around half of the polysemous words are assigned almost correct senses. It failed to correctly assign the words whose senses are closely related. For example, जीवन was tagged as आयु(age) where it was meant to be ज़िन्दगी(life). These two senses are close enough and usually used in the same context. On few occasions, a word was incorrectly tagged to its noun sense where it had non-noun sense, like कर in the sense of करना(verb 'do'), was tagged as टैक्स(revenue). Relatively narrow coverage[4] of Hindi Wordnet, also caused incorrect tagging. For example, प्रकाश was tagged with *Offset* - 2037, which is the sense related to उजाला(light), but there is no sense defined for the other common usage of प्रकाश - ध्यान केंद्रित करना(to focus) in the Hindi wordnet.
The system suffered a low recall because of relatively inefficient synset mapping method as the coverage of English Words are poor in the dictionary used. A better synset mapping tool could have resulted into higher recall as most of the times we did not get a mapping from English synset to its relevant synset in Hindi Wordnet.

6 Conclusion and Future Work

The proposed approach assigns correct sense to words with a high degree of accuracy but offers a poor recall as such. This approach make use of semantic relations defined in Wordnets to map synsets, whereas [11] have used the inter-lingual relations connecting Persian synsets(from FarsNet[12]) to English synsets(in WordNet), as the synsets were developed in FarsNet on the lines of synsets defined in WordNet while the synsets in Hindi Wordnet were developed independently.

This approach does not care of morphology which too caused a low recall. For example, जानवरों, the pl. form of 'जानवर' could not be tagged by the system. Morphology handling can also help to stretch this approach to non-noun words, especially verbs. Moreover, the synset mapping uses hypernymy relation between synsets, which is possible for nouns only. For other parts of speech, other semantic relations must be taken into consideration. For example, for mapping between verb synsets, entailment and troponymy relations can prove to be useful and may be explored in the future. By impoving the recall, this work can be extended to create a sense-tagged Hindi corpus and a map between WordNet and Hindi Wordnet. There is also scope to improve the sense-tagging in the english texts, by applying algorithms other than the Lesk algorithm for the English WSD part.

Acknowledgements

We would like to show our sincere gratitude to our advisor, Professor Amitabha Mukerjee, for providing us the chance to work on this project, and for acting as a guiding light throughout the duration of the project.
Further, we are grateful to the Center for Indian Language Technology (CFILT), IIT-Bombay, for providing us the Hindi Wordnet API, as well as the Universal Word-Hindi Dictionary, both of which form an integral part of this work. We would also like to thank Professor Ted Petersen of the Department of Computer Science, University of Minnesota, for his Wordnet::Senserelate::Allwords package, without which we could not have achieved our objectives. We are also indebted to many of our colleagues for their valuable comments, and moral support in times of distress.

[4] 28,687 synsets are present in Hindi Wordnet as compared to 117,659 synsets in WordNet

References

[1] Satanjeev Banerjee and Ted Pedersen. Extended gloss overlaps as a measure of semantic relatedness. In *IJCAI'03: Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, pages 805–810, 2003.

[2] Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT-Summit*, Phuket, Thailand, 2005.

[3] Els Lefever and Veronique Hoste. Semeval-2010 task 3: Cross-lingual word sense disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010*, pages 15–20, Uppsala,Sweden, 2010.

[4] Els Lefever and Veronique Hoste. Examining the validity of cross-lingual word sense disambiguation. In *CICLing 2011: Proceedings of the Conference on Computational Linguistics and Intelligent Text Processing*, Tokyo,Japan, 2011.

[5] Michael Lesk. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *SIGDOC '86: Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26, New York, NY, USA, 1986. ACM.

[6] G. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. Introduction to wordnet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–244, 1990.

[7] Dipak Narayan, Debasri Chakrabarty, Prabhakar Pande, and Pushpak Bhattacharyya. An experience in building the indo-wordnet - a wordnet for hindi. In *GWC'02: Proceedings of the First International Conference on Global WordNet*, Mysore,India, 2002.

[8] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

[9] Ted Pederson and Varada Kolhatkar. Wordnet:: Senserelate:: Allwords: A broad coverage word sense tagger that maximizes semantic relatedness. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Demonstration Session*, pages 17–20, 2009.

[10] J. Ramanand, Akshay Ukey, Brahm Kiran Singh, and Pushpak Bhattacharyya. Mapping and structural analysis of multi-lingual wordnets. *IEEE Data Engineering Bulletin*, 30(1):30–44, 2007.

[11] Bahareh Sarrafzadeh, Nikolay Yakovets, Nick Cercone, and Aijun An. Cross-lingual word sense disambiguation for languages with scarce resources. In *Canadian Conference on AI'11*, pages 347–358, 2011.

[12] Mehrnoush Shamsfard, Akbar Hesabi, Nick Cercone, Hakimeh Fadaei, Niloofar Mansoory, Ali Famian, Somayeh Bagherbeigi, Elham Fekri, Maliheh Monshizadeh, and S. Mostafa Assi. Semi automatic development of farsnet: The persian wordnet. In *Proceedings of 5th Global WordNet Conference*, Mumbai,India, 2010.