Cross-Lingual Word Sense Disambiguation

Priyank Jaini
pjaini@iitk.ac.in
Department of Mathematics and Statistics
.

Ankit Agrawal
ankitag@iitk.ac.in
Department of Mathematics and Statistics
.

Mentor: Prof. Amitabha Mukerjee, Department of Computer Science and Engineering
Indian Institute of Technology, Kanpur

Final Project Report
15th April, 2013

Abstract

Word Sense Disambiguation using Cross-Lingual approach has been used successfully for languages like Farsi and Hindi. However, a comparable corpus in the form of Wikipedia articles available in English and Hindi has been used for such a task. This motivated us to further the approach and test the results when a parallel corpus is used. In this project, we specifically wanted to observe if the accuracy would include if we could get a sentence to sentence mapping across translations and then disambiguating the sense using cross lingual approach. The use of a parallel corpus helped us in getting a sentence aligned data across translations. The results show that when we used individual sentences only for disambiguating, the results increased as compared to those when whole texts were consider in a parallel corpus. The result between local disambiguation (when individual sentences were used) and global disambiguation (when the text was used as a whole) also show that local disambiguation scores more on accuracy as compared to global disambiguation.

1 Previous Work

In [1], they propose a cross lingual approach to tagging the word senses in Persian texts. The new approach makes use of English sense disambiguators, the Wikipedia articles in both English and Persian, and a newly developed lexical ontology, FarsNet. It overcomes the lack of knowledge resources and NLP tools for the Persian language [Yakovets et al 2011]. Our approach is also based primarily on [1]. We have used the idea they have suggested to achieve Cross-Lingual Disambiguation using Hindi-English corpus(both comparable and parallel).

2 Introduction

Word Sense Disambiguation (WSD) is defined as assigning the correct sense to a word according to its context. A word can have several meanings and the correct usage depends on the context eg[Wikipedia]:

1. I went fishing for some sea bass.

2. The bass line of the song is too weak.

In the first sentence bass means a type of fish whereas in the second sentence bass is used in the sense of sound. One of the major problems in WSD is knowledge acquisition bottleneck. WSD needs large amount of word and word knowledge. While this kind of knowledge is amply available in English through a lot of resources primarily WordNet[7], such resources for Hindi or other languages are not available. Therefore much of the research in WSD has been performed in English only. There are four main approaches for WSD namely

1. Dictionary and knowledge-based methods
2. Supervised methods
3. Semi-Supervised methods
4. Unsupervised methods

Of these approaches, Supervised learning has proven to be the best performer. However, there is a lack of such a resource in Hindi. This encouraged us to look into the possibility of using Cross-Lingual methods by exploiting the resources available in English to sense tag Hindi words. Our approach consisted of the following three steps for sense-tagging the data:

- English text Word Sense Disambiguation using WordNet
- Synset Mapping: Assigning the correct Hindi synset using the tagged English data
- Transfer of this synset to Hindi data to produce sense tagged Hindi data.

The project report is organised as follows: Section 2 discusses in brief some of the work which we studied and relates to the project, Section 3 outlines the major algorithms used in the project, evaluation method and the results have been presented in Section 4, Section 5 discusses some of our conclusions, the limitations and what future work can be done.

3 Related Work

A gloss is a textual definition of a synset with possibly some examples about its usage. In [4], a method was developed by Lesk to disambiguate words using dictionary definitions. They have proposed an algorithm which counts the number of words that are shared in the two definitions (glosses) and hence determines the relatedness. One of the major limitations of this is that definitions (dictionary) are generally brief and therefore might not always give accurate results. However, in [5] the algorithm proposed has been extended such that the gloss exploration technique include glosses of other concepts to which they are related according to a given context hierarchy. Hence, this method yields far more accurate results than that of [4]. Therefore, using WordNet and Extended Lesk Algorithm, mapping of words in English to their correct senses has been done with greater accuracy. However, packages like WordNet are not available for other languages. The packages that do exist such as the Indo-WordNet for Hindi are not exhaustive. In [3], Lefever et al have shown that Cross-Lingual approaches have produced more reliable results and offer distinct advantages for languages which lack large sense-annotated corpora. For such languages, target words can be disambiguated using translations in a language in which such resources are amply available.

4 Sense tagging using Cross Lingual Approach

This approach mainly consists of the following three steps:

1. English text Word Sense Disambiguation using WordNet
2. Synset Mapping: Assigning the correct Hindi synset using the tagged English data
3. Transfer of this synset to Hindi data to produce sense tagged Hindi data. It is to be noted here that the approach is only for nouns , therefore, we are able to tag senses of only nouns. Fig.1 illustrates the complete outline of the process.

4

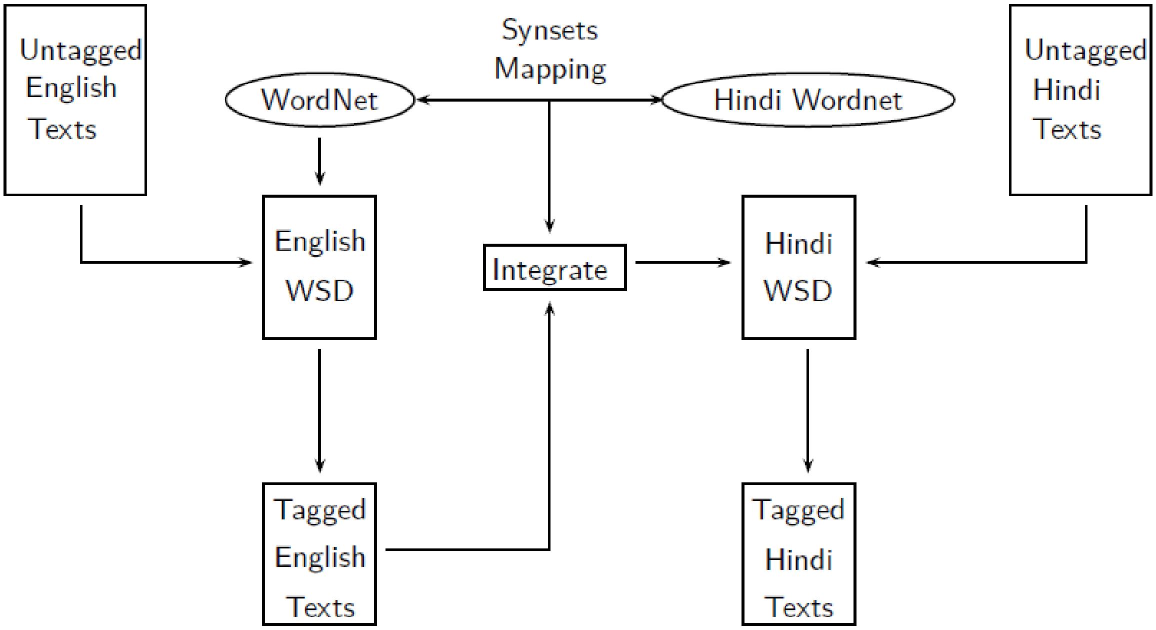


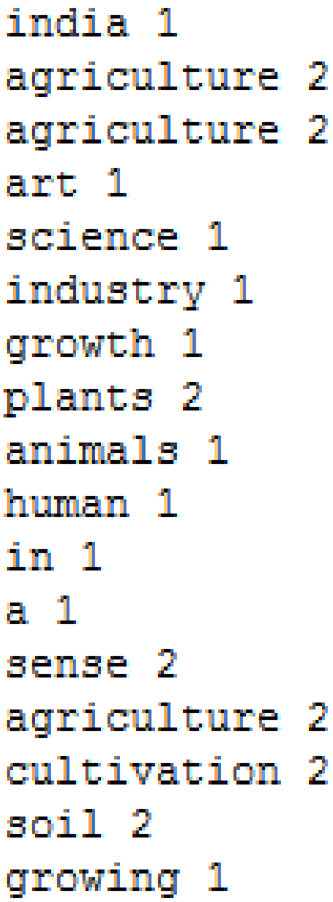
Figure 1: Complete outline taken from [1]

4.1 English text Word Sense Disambiguation using WordNet

In this step we generate sense tagged English text from the initial untagged text. As we had stated earlier, there are a lot of resources available in English for sense-tagging. In this project we have used the idea presented in [9], and used the Perl-based application Sense Relate[9] which makes use of the Lesk Algorithm for disambiguation. For this step we have used WordNet::SenseRelate::AllWords[9]. This program tags all the words irrespective of their part of speech. In the next step we discard all the non-nouns form since we have restricted ourselves to only nouns. The repeated nouns are then discarded and replaced with a single occurrence and the sense assigned to this single occurrence is that sense which appears more frequently. The output of this step is shown in Fig.2 and Fig.3. 4 4

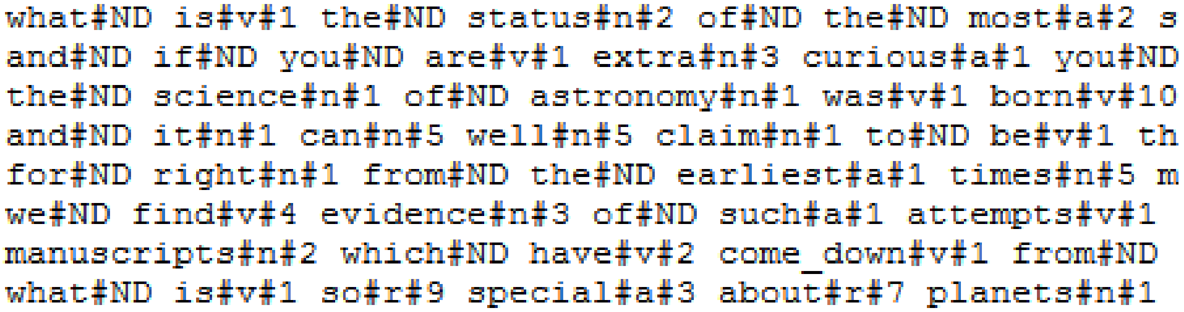
4.2 Synset Mapping: Assigning the correct Hindi synset using the tagged English data

After the first step, the English words tagged with the synsets in WordNet need to be mapped to the appropriate Hindi synset in the Hindi WordNet. This is achieved by making use of the algorithm proposed in [2]. The algorithm takes an English synset and maps it to the most appropriate Hindi synset. The algorithm works as follows: [11] states that the first word of the synset best describes a synset. Therefore, the first word is taken of the English synset and all its Hindi translations are found using a Hindi-English dictionary. The synsets in the Hindi WordNet consisting of all the above Hindi translations serve as candidate synsets. In the end the English synset would be mapped to one of these candidate synsets. The hypernymy hierarchies of each of the above candidate synsets are found and these are called candidate hypernyms. The hypernymy hierarchies of the original English word are also found. The Hindi translations of each of the above English hypernymy are obtained. These translations are then searched for in the candidate hierarchies. If a match is found, a relevant weight is assigned to the corresponding synset. The candidate synset is mapped to the English synset whose candidate hierarchy has the highest weight. The algorithm is illustrated in Fig:4. 4



```
india 1
agriculture 2
agriculture 2
art 1
science 1
industry 1
growth 1
plants 2
animals 1
human 1
in 1
a 1
sense 2
agriculture 2
cultivation 2
soil 2
growing 1
```

Figure 2: Only nouns tagged

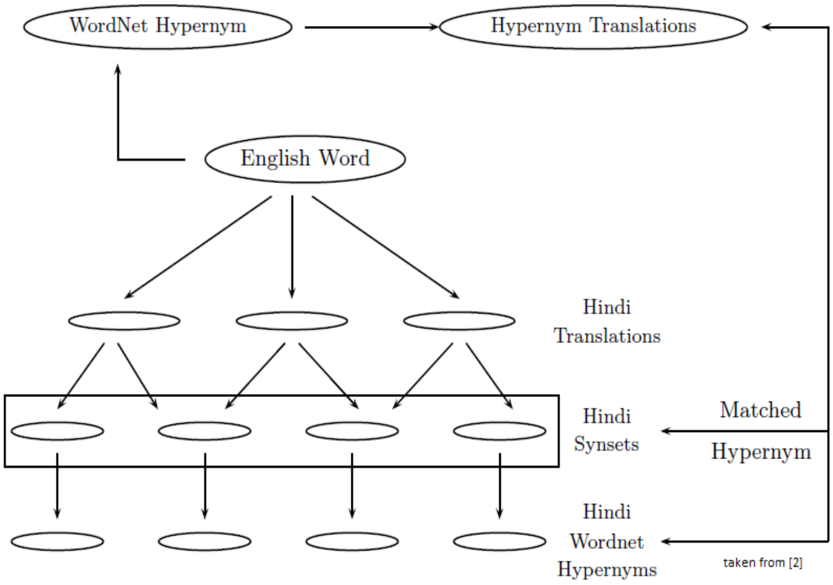


```
what#ND is#v#1 the#ND status#n#2 of#ND the#ND most#a#2 s
and#ND if#ND you#ND are#v#1 extra#n#3 curious#a#1 you#ND
the#ND science#n#1 of#ND astronomy#n#1 was#v#1 born#v#10
and#ND it#n#1 can#n#5 well#n#5 claim#n#1 to#ND be#v#1 th
for#ND right#n#1 from#ND the#ND earliest#a#1 times#n#5 m
we#ND find#v#4 evidence#n#3 of#ND such#a#1 attempts#v#1
manuscripts#n#2 which#ND have#v#2 come_down#v#1 from#ND
what#ND is#v#1 so#r#9 special#a#3 about#r#7 planets#n#1
```

Figure 3: English tagged data

4.3 Transfer of this synset to Hindi data to produce sense tagged Hindi data.

In the final step, the words in the Hindi text are to be tagged with the same sense as the corresponding English word. When the tagging is done, both the English sense and the Hindi Offset are labelled. The output is the shown in the following example: उपयोग had two senses and was correctly tagged in the following



taken from [2]

Figure 4: Synset Mapping

way: उपयोग [consumption#n#1]#2933 The synset for उपयोग in the Hindi WordNet was {खपत, इस्तेमाल, इस्तमाल, उपभोग, उपयोग, प्रयोग, ख़र्च, खर्चा, ख़रच, ख़र्चा, दोहन, उठान - काम में आने या लगने की क्रिया ”हमारे देष में चावल कि ख़पत ज्यादा होतीं है।” }

5 Results

The proposed approach was tested on a parallel corpus of 2000 lines(32,000 wrods) and on a comparable corpus of Hindi-English Wikipedia pages of a total of 6000 words. The results have been illustrated in Table.1. It is evident from the table that the accuracy was very high when the sentences were tagged individually.

Test Data	Correct Sense	Almost Accurate	Wrong Sense	Precision	Recall	F-Score
Parallel Corpus(Complete Text)	82.4%	5.2%	12.4%	0.0824	0.237	0.368
Parallel Corpus (Individual Lines)	91.2%	3.3%	5.5%	0.912	0.224	0.359
Comparable Corpus	73.3%	5.2%	20.5%	0.732	0.174	0.28

Table 1: Results

Also, the accuracy of the parallel corpus is higher than that of a comparable corpus. The system, however, suffered from a low recall in all the cases. This might be because of the relatively inefficient synset mapping method as the coverage of English words in the dictionary used is poor.

6 Conclusions and Future Work

The proposed approach clearly suggests that where a mapping between sentences across translations is available, the accuracy achieved is very high as compared to when a comparable corpus was used or the whole text data was used. However, the approach suffers from a poor recall. One of the reasons for this might be the limited number of synsets in the Hindi WordNet.

Some of the major limitations of this work is that this is restricted only for nouns. To consider other part-of speech, different relations amongst synsets in the WordNets would have to be used. Also, this approach is unable to handle morphology, which was also the reason for a low recall.

By increasing the recall, this work can be extended to create a sense-tagged data Hindi corpus. It would be interesting to compare results of this approach with those when a comparable corpus is used and the sentences in the data in the corpus are aligned first. There are a couple of approaches available for aligning sentences across translations. One of these is the Church and Gale Algorithm [12], which is language independent. A more recent work in [13], suggest a modification of the Church and Gale algorithm for aligning sentences. For a future work, such sentence alignment can be used on a comparable corpus followed by the same cross-lingual approach to get a sense-tagged data for a comparable corpus.

7 Acknowledgement

We would like to show our sincere gratitude towards Prof. Amitabha Mukerjee, Department of Computer Science and Engineering, IIT-Kanpur for his valuable guidance throughout the project. We are grateful to the Centre for Indian Language Technology (CFILT), IIT-Bombay for providing us the Hindi WordNet API, and the Universal Word-Hindi Dictionary. We are thankful to Prof. Pushpak Bhattacharyya, Department of Computer Science and Engineering, IIT-Bomay and his student Mr. Kashyap Popat for providing us with English -Hindi parallel corpus. We also thank Ondej Bojar , Pavel Strak , Daniel Zeman , Gaurav Jain, Om Prakesh Damani for providing us with a part of EMILLE corpus. We also extend our gratitude towards Prof. Ted Petersen, Department of Computer Science, University of Minnesota, for his Wordnet::Senserelate::Allwords package. We are also thankful to Mohit Sharma, Sophomore , Department of Computer Science and Engineering, IIT-Kanpur for showing interest in the project and discussing the details and possible modifications. We lastly thank all our friends for their support throughout the project.

8 References

[1]Bahareh Sarrafzadeh, Nikolay Yakovets, Nick Cercone, Aijun An. Cross Lingual Word Sense Disambiguation for Languages with Scarce Resources.[2011]

[2]J.Ramanand, Akshay Ukey, Brahm Kiran Singh and Pushpak Bhattacharyya. Mapping and structural analysis of multi-lingual wordnets. IEEE Data Engineering Bulletin,2007.

[3]Els Lefever and Veronique Hoste. SemEval-2010 Task 3:Cross-Lingual Word Sense Disambiguation.

[4] Michael Lesk.Automatic sense disambiguation using machine readable dictionaries:how to tell a pine cone from an ice cream cone. In SIGDOC86: Proceedings of the 5th annual international conference on Systems documentation,pages 24-26, New York, NY, USA, 1986.ACM

[5] Satanjeev Banerjee and Ted Pedersen. Extended gloss overlaps as a measure of semantic relatedness. In IJCAI03, pages 850-810,2003.

[6] Els Lefever and Veronique Hoste. Examining the validity of Cross-Lingual Word Sense Disambiguation[2011]

[7] http://wordnet.princeton.edu/

[8] Roberto Navigli. Word Sense Disambiguation-A Survey[2009]

[9] Ted Petersen and Varada Kolhatkar. WordNet::Senserelate::Allwords. A Broad coverage word sense tagger that maximises semantic relatedness. [2007]

[10] Debasri Chakrabarti,Dipak Kumar Narayan,Prabhakar Pandey,Pushpak Bhattacharyya.Experiences in building the Indo Word Net-A WordNet for Hindi.[2002]

[11] G.Miller, R.Beckwith,C.Fellbaum,D.Gross and K.Miller.Introduction to wordnet:An online lexical database.[1990]

[12] Gale, William A.; Church, Kenneth W. (1993), "A Program for Aligning Sentences in Bilingual Corpora".

[13] ”An Algorithm for Aligning Sentences in Bilingual Corpora Using Lexical Information” by Akshar Bharati, Sriram V Vamshi Krishna A, Rajeev Sangal, Sushma Bendre.[200