Cross-Lingual Word Sense Disambiguation

Ankit Agrawal,Priyank Jaini (Department of Mathematics and Statistics)
Advisor: Amitabha Mukerjee (Department of Computer Science and Engineering)
{ankitag,pjaini}@iitk.ac.in,amit@cse.iitk.ac.in
Indian Institute of Technology,Kanpur
Kanpur, India

March 10, 2013

1 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. To address WSD in multi-lingual scenario, cross-lingual approaches to WSD have been used wherein the vast amount of data available in English has been used as a tool to make sense of texts in other languages.

2 WordNet and Indo-worNet

WordNet[7] is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. A synset also contains definition of the words along with examples illustrating their usage.
Hindi WordNet[1] (Indo-WordNet) has been developed on the lines of WordNet(English). *It acts as a system which brings together different lexical*

relationships between the Hindi words. The Indo-WordNet is designed using the fact that basic lexical links exist between synonyms. The synsets in Indo-Wordnet and the semantic links between them are grouped according to different categories like nouns, adjectives,verbs, adverbs etc. Relations like Hyponymy, Hypernymy, Meronymy, Holonymy, Entailment ,Antonymy, Gradation and Linkages which are defined in WordNet are also defined in the Indo-WordNet.

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 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.However, 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 in [4] 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 great 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. Therefore, for addressing the problem of WSD for such languages, parallel corpus based approach is a better option. But, this method also requires a large amount of cross-lingual accurate sense-tagged texts, which are not available in Hindi.

4 Proposed Approach

We propose tagging of Hindi words and our approach would be limited only to nouns. For Hindi WSD the best approach is Cross-Lingual Approach since Hindi lacks a proper sentenced aligned parallel corpus. We will use Wikipedia articles available both in Hindi and English to create a corpora. This will help us generate a large amount of data since more than 100,000 such articles are available. We will essentially follow the approach used in[2] wherein we will get the sense of ambiguous word from WordNet and transfer it to Hindi text. Our approach mainly consists of three broad tasks:

- English Word Sense Disambiguation
- Synset Mapping between English and Hindi

- English to Hindi transfer

5 References

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

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

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

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

8. Roberto Navigli. Word Sense Disambiguation-A Survey.

9. Roberto Navigli and Simone Paolo Ponzetto. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network.