# 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

February 20, 2012

---

## 1 Introduction

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.
WSD relies heavily on previously acquired knowledge, generally created in a time consuming and expensive process. Therefore, substantial amount of knowledge is currently available in English only, while for Hindi and other Indian languages, it is quite scarce.
Cross-Lingual methods have therefore been used for achieving WSD in languages other than English, where the vast amount of sense data available in English is used as a tool to make sense of texts in other languages for machines.

## 2 WordNet

WordNet[2] 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[4] 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[3] developed a method to disambiguate words using dictionary definitions (gloss). 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. But the availability of WordNet like packages for languages other than English is low, and even the packages available (including the Hindi Wordnet) are not very exhaustive. WSD for such languages, therefore, has been performed using a parallel corpus approach, instead of the knowledge based approach. This method requires the availability of large amounts of cross-lingual accurate sense-tagged texts, which again are not available for Hindi.

# 4 Proposed Approach

Keeping in mind the lack of reliable linguistic resouces for Hindi, we propose a method which is not limited to parallel corpora only. We plan to utilize Wikipedia articles to generate comparable corpora. This can generate a large amount of data, as the number of Wikipedia articles in Hindi is over 100,000. We plan to use a method similar to [5], using extended Lesk to get sense of ambiguous words from WordNet and transferring the sense to the Hindi text. To achieve the above goal, we will try to map the synset obtained from the (English)WordNet to a synset in Hindi Wordnet. As an ambiguous word would probably be used in the same sense as the context of the texts in both the languages is the same, we use the most frequently assigned sense of the word in the English text as a heuristic for determining the sense of the corresponding words in Hindi.
This can be used to annotate words with POS and sense information. This work may be used to build a sense-tagged corpus for Hindi which can be used in supervised approaches.

# References


[1] Satanjeev Banerjee and Ted Pedersen. Extended gloss overlaps as a measure of semantic relatedness. In *IJCAI'03*, pages 805–810, 2003.

[2] Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. Bradford Books, 1998.

[3] 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.

[4] 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.

[5] 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.