Word Sense Disambiguation using Hindi WordNet

Rashish Tandon (Y6377)
Advisor: Dr. Amitabha Mukherjee
Dept. Of Computer Science and Engineering
IIT Kanpur
{rashish , amit} @iitk.ac.in

February 18, 2009

1 Introduction

Word Sense Disambiguation (WSD) is the task of finding the appropriate sense of a word used in a given sentence, when the word may have more than one sense. For example consider the sentences :

यह तो आम बात है । and यह आम मीठा है ।

It can be easily seen that the word आम refers to 'general' or 'usual' in the former sentence and to a 'mango' in the latter. Some words may not be as easy to disambiguate as others as they may have multiple senses that are close to each other. In some cases disambiguation may be impossible altogether using only the given sentence. For example, consider the sentence :

यह तो आम है ।

Given only the above sentence, one may translate it as "This is the norm" or "This is a mango",both having valid senses of the word आम. Thus, these scenarios require a look at the context of the discourse to disambiguate between the possible senses.

2 Hindi WordNet [4]

Certain WSD techniques involve the use of World Knowledge to perform the task of disambiguation. These may exploit Machine Readable Dictionaries or Thesauri. WordNet is one such framework that embeds World Knowledge and may be used for disambiguation.
Hindi WordNet defines a number of semantic relations for a hindi word based on each of its possible senses. It involves the representation of each Word Meaning as a set of word-forms known as synonym sets or synsets.These synsets are created for content words like Nouns, Verbs, Adverbs and Adjectives. It also defines a the following semantic relations to connect synsets.

- **Hypernym** - Y is a hypernym of X if every X IS-A (KIND-OF) Y.
- **Hyponym** - Y is a hyponym of X if every Y IS-A (KIND-OF) X.
- **Entailment** - the verb Y is entailed by X if by doing X you must be doing Y.
- **Troponym** - the verb Y is a troponym of the verb X if the activity Y is doing X in some manner.
- **Meronym** - Y is a meronym of X if Y is a part of X.
- **Holonym** - Y is a holonym of X if X is a part of Y.

As an example, the synset {घर ,गृह} has the hypernym relation to {आवास ,निवास} , a meronym relation to {बरामदा} and a hyponym relation to {झोपड़ी}.

3 Related Work

There are three primary approaches to WSD.

- Knowledge Based - Involve incorporation of world knowledge to diambiguate words.
- Supervised - Involve learning based on a sense tagged corpora by hand, and then using the rules learnt on untagged textual data.
- Unsupervised - Involve Clustering the words in the untagged data into different senses. It may not be possible here to tag each of these senses.

Lesk (in 1986) proposed an algorithm for WSD by counting the number of words shared between the sense definition and the context. The sense definition was obtained from a dictionary, so it only consisted of synonyms. The context was obtained by considering the words in the same sentence as the word to be disambiguated.

Most approaches today that use a Knowledge Base have as root, the Lesk's Algorithm, with changes made to the semantic relations used to obtain the sense definitions and the use of efficient metrics to compare the sense definitions and the contexts.

Yarowsky[6] proposes a solution to the problem of WSD using a thesaurus in a supervised learning setting. A collection of words is made from the sense tagged data for each category of the thesaurus and based on their occurrences, these words are appropriately weighted. Each word may be considered characteristic of the particular category it is in. The resulting weights are then used to predict the appropriate category for a word in the novel text.

Since, sense information is already present in WordNet,the problem of disambiguation is reduced to one of selection of an appropriate sense from a given list.

The WordNet can be represented as a graph, with Nodes representing synsets and the semantic relations representing links between these nodes. Sussna[5] exploits this by assigning weights based on the relation type (synonymy, hyperonymy, etc.) to WordNet links, and defines a metric which takes account of the number of arcs of the same type leaving a node and the depth of a given edge in the overall tree, to determine the distance between nodes. The sense that is chosen (node) is the one that minimizes the distance to all other nodes in its context.

Agirre and Rigau[2] use a measure based on the proximity of the text words in WordNet (conceptual density) to disambiguate the words. Conceptual distances between nodes are computed based on this measure and a similar approach as that of Sussna is utilized.

Ramakrishnan[3] proposes a soft WSD wherein a word, instead of being in only one sense, may be assigned to more than one sense with a measure for the degree of membership in each of the assigned senses. The senses of a word are determined using WordNet and their relevance is probabilistically determined through a Bayesian Belief Network.

Many modern approaches to WSD are hybrid systems i.e. they use both statistical methods on a given textual data (which may be sense tagged) and a Knowledge Base.

An extremely novel approach by Barnard and Johnson[1] attempts to disambiguate words using a Knowledge Base like WordNet and then finding the appropriate sense based on environment information obtained using Vision.

4 Proposed Approach

Most WSD algorithms have only been tested on English. I would initially like to implement these algorithms and evaluate their performance when applied to Hindi datasets.

I then hope to propose my own algorithm that exploits the structure of Hindi WordNet efficiently, perhaps uses statistical based learning, and avoids pitfalls of the implemented algorithms to achieve Word Sense Disambiguation with greater accuracy.

References

[1] Johnson M. Barnard K. Word sense disambiguation with pictures. 2005.

[2] Agirre Eneko and Rigau. Word sense disambiguation using conceptual density. 1996.

[3] A. Deepa Pushpak Bhattacharyya Ganesh Ramakrishnan, B. P. Prithviraj. Soft word sense disambiguation. 1997.

[**4**] P. Pande P. Bhattacharyya S. Jha, D. Narayan. A wordnet for hindi. 2001.

[**5**] Michael Sussna. Word sense disambiguation for free-text indexing using a massive semantic network. 1993.

[**6**] D. Yarowsky. Word sense disambiguation using statistical model of rogets categories trained on large corpora. 1992.