Comparative Analysis of Word Sense Disambiguation Algorithms in Hindi

Drishti Wali (13266), Nirbhay Modhe (13444)

Abstract

Word sense disambiguation is the automatic identification of the correct meaning of a polysemous word in a given context. It continues to be an open problem in natural language processing. While high levels of accuracy have been achieved for English in WSD, however the lack of resources have stunted significant progress for the same in Hindi. In this project we aim to compare traditional methods of WSD with a sense vector method using the Hindi Wordnet developed by IIT Bombay. The traditional methods would included Yarowsky's semi-supervised approach, Lesk's knowledge-based approach and an unsupervised multi-prototype vector space model for clustering.

1 Introduction

1.1 Problem Definition

Given a Hindi sentence with polysemous words, we intend to automatically tag the target words with their correct sense according to the context in which they occur.

1.2 Motivation

Word sense disambiguation is essential for other natural language processing tasks like machine translation, multi-lingual search engines, semantics and discourse analysis. Little work has been done in this area for Hindi due to lack of lexical resources. We aim to extend the current well-known and other traditional methods for WSD to Hindi.

1.3 Previous Work

- WSD for nouns has been done using the IIT Bombay Wordnet based on an approach which maximizes the overlap of the different senses with the context of the word to be disambiguated. [7]
- A bilingual unsupervised approach of WSD has been used aimed at disambiguating verbs in Hindi. [3]

2 Approach

2.1 Sense vector method [4]

Starting with an unannotated corpus, the Skip-gram model is used to assign a vector to each word. Similarly, a vector is assigned to each sense of the word obtained from the Wordnet using their glosses. The sense with the maximum cosine similarity to the context vector is assigned as the sense of the current target word.

2.2 Lesk's Algorithm [2, 5]

For each word to be disambiguated, the maximum word overlap is found between the word context and the senses obtained from the Wordnet, and this sense is assigned to the word.

2.3 Multi prototype clustering [6]

Using either BOW or Skip-gram model, each word is vectorized. Subsequently, clustering is performed using K-means to obtain prototype vectors which is the centroid of each cluster. New words are disambiguated by maximizing their similarity with the prototype vectors. The clusters can be mapped to specific senses in the Wordnet to compare the results with the other methods.

2.4 Yarowsky's Algorithm [8]

This semi-supervised approach begins with assigning seed collocations which closely represent the concept for each sense of a word to be disambiguated. A decision list algorithm is used to expand the seed clusters reliably, until the the clusters stabilize. Hence, we obtain a classifier which can be used to disambiguate the target word.

3 Resources

1. Hindi Wordnet developed by IIT Bombay. [1]
2. Data Sets
    - Central Institute of Indian Languages
    - Center for Indian Language Technology
    - Wikipedia translated articles, text dump, embeddings.

References

[1] Hindi wordnet from center for indian language technology solutions, created by iit bombay, mumbai, india.

[2] Satanjeev Banerjee. *Adapting the Lesk algorithm for word sense disambiguation to WordNet.* PhD thesis, University of Minnesota Duluth, 2002.

[3] Sudha Bhingardive, Samiulla Shaikh, and Pushpak Bhattacharyya. Neighbors help: Bilingual unsupervised wsd using context. 2013.

[4] Xinxiong Chen, Zhiyuan Liu, and Maosong Sun. A unified model for word sense representation and disambiguation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1025–1035, 2014.

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

[6] Joseph Reisinger and Raymond J Mooney. Multi-prototype vector-space models of word meaning. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 109–117. Association for Computational Linguistics, 2010.

[7] Manish Sinha, Mahesh Kumar, Prabhakar Pande, Laxmi Kashyap, and Pushpak Bhattacharyya. Hindi word sense disambiguation. In *International Symposium on Machine Translation, Natural Language Processing and Translation Support Systems, Delhi, India*, 2004.

[8] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pages 189–196. Association for Computational Linguistics, 1995.