Word Sense Disambiguation Algorithms in Hindi

Drishti Wali Nirbhay Modhe

Department Of Computer Science and Engineering,
Indian Institute of Technology, Kanpur

भारतीय प्रौद्योगिकी संस्थान कान्पुर

Introduction

Word sense disambiguation (WSD) is an important and open problem in natural language processing. A substantial amount of research has been done in WSD in English, with the state of the art systems having an accuracy of up to 72.9% in Senseval-3 (2004) [1]. The first attempt at WSD in Hindi, performed solely on nouns, resulted in an accuracy ranging from 40% to 70%[2].

Traditional WSD Methods used in English

- **Lesk's algorithm** is a frequently used knowledge-based method of WSD which makes use of overlap of words occurring in context, with words present in each senses, along with other sources such as synonyms, hypernyms, homonyms, meronyms, example sentences, gloss of hypernyms and homonyms[3].
- In 1995, **Yarowsky** proposed an unsupervised method of WSD using decision lists. He introduced the famous "one sense per discourse" property which he and many others in future used for their disambiguation algorithms[4].

Resources and Corpus

- The **HindMonoCorp 0.5** is a Hindi-only segmented, tokenized corpus with morphological tags. It consists of 365 million lemmatized Hindi words[5].
- The **Hindi Wordnet** developed by IIT Bombay is a lexical resource which incorporates the different semantic relations between words in Hindi. These relations include synonyms, homonymy, hypernymy, meronymy[6].

Methodology

- **Word Vectorization** : Mikolov's Word2vec model was used, which implemented a skip-gram approach to train word vectors for predicting words given a context. The weights trained on the neural network represent the vectors of each word. The model was trained on the corpus to obtain 100 dimensional vectors for each of the 0.7 million words in the vocabulary.

- **Sense Extraction** : Given a word to be disambiguated, all of the senses of the word are obtained from the Hindi Wordnet (Java API). Each sense is represented as a collection of it's synsets, hypernyms, homonyms and gloss.

- **Sense Vectorization** : The vector which represents each sense is the average of all the word vectors in its collection which have a cosine similarity with the original word above a certain threshold[7].

- **Context Vectorization** : To disambiguate a sentence containing this word, its context words having cosine similarity with this target word above a threshold are used to create the context vector by averaging over the individual word vectors.

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

- **Sense Allocation** : The cosine similarity values are computed between the context vector and each sense vector. The sense which has the highest cosine similarity value is allocated sense for disambiguation.

- **Comparison Baseline :** The results were compared with the standard Lesk's algorithm using words from Hypernyms, Homonyms, Synsets, gloss and example sentences.

Results

Test Sentences	Sense Vector	Baseline
वस्तु की साफ **सफाई** जरूरी होती है ।	Correct	Correct
अदालत की **सफाई** जरूरी है ।	Correct	Incorrect
हमें **सफाई** देने की जरूरत नही है ।	Correct	Incorrect
शक होने पर उसे **सफाई** देनी पडी ।	Incorrect	Incorrect

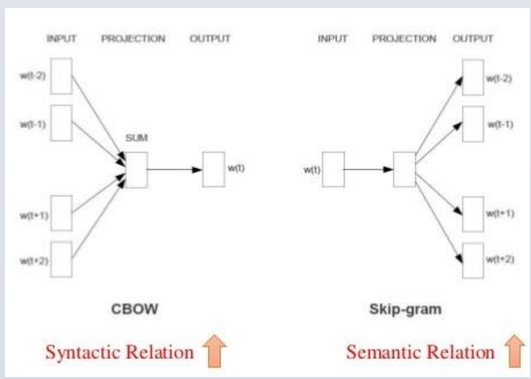
Given a sentence and a word to be disambiguated the algorithm provides a number corresponding to its allocated sense. The large corpus ensured coverage of almost all Hindi words in the vocabulary. The vectors provided an accurate representations of all words in the 100 dimensional space as observed by the correspondence of high cosine similarity with high similarity of words.

Conclusion

- The Sense Vector method outperforms Lesk's algorithm in almost all cases.
- Fine tuning of the threshold cosine similarity value can increase the accuracy of disambiguation. The threshold value of 0.3 was observed to be the best for our test cases.
- Averaging over vectors to obtain sense and context vectors does not seem like the most optimal representation of the senses.

Insights

- It was observed that the accuracy depends on the length of the sentence of the word to be disambiguated.
- The length of the gloss plays and important role, particularly in Lesk's algorithm, for disambiguation.
- There is a scarcity of large Hindi sense-tagged data, whose availability would enable us to use the state of the art supervised algorithms, which are currently implemented only in English.



Source: http://www.slideshare.net/ssuser9cc1bd/piji-li-dltm

Future Work

- A better representation of the sense and context vectors can be obtained by learning the contribution of different word vectors from the collection of that sense. This may be done using a neural network.
- Alternatively, metrics other that cosine distance can be used as a measure of similarity.
- Clustering is a viable unsupervised method for WSD in under resourced languages like Hindi. It can be used for comparison of performance with our method.
- Different adaptations of Lesk's algorithm can be used to improve upon the baseline performance.

References

[1] Mihalcea, R., Chklovsky, T., & Kilgarriff, A. (2004). ITRI-04-09 The Senseval-3 English lexical sample task. *Information Technology*, *25*, 28.

[2] Kashyap, Prabhakar Pandey Laxmi. "Hindi Word Sense Disambiguation."

[3] Lesk, M. (1986, June). 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* (pp. 24-26). ACM.

[4] Yarowsky, D. (1995, June). Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics* (pp. 189-196). Association for Computational Linguistics.

[5] Bojar, O., Diatka, V., Rychlý, P., Straňák, P., Suchomel, V., Tamchyna, A., & Zeman, D. (2014). HindEnCorp-Hindi-English and Hindi-only Corpus for Machine Translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*.

[6]Hindi Wordnet, IIT Bombay. http://www.cfilt.iitb.ac.in/wordnet/webhwn/index.php

[7] Chen, X., Liu, Z., & Sun, M. (2014). A unified model for word sense representation and disambiguation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1025-1035).