WORD SENSE DISAMBIGUATION ALGORITHMS IN HINDI

Drishti Wali (13266)
Nirbhay Modhe (13444)

Word Sense Disambiguation

The task of automatically assigning a sense to an ambiguous word according to the context in which it is present.

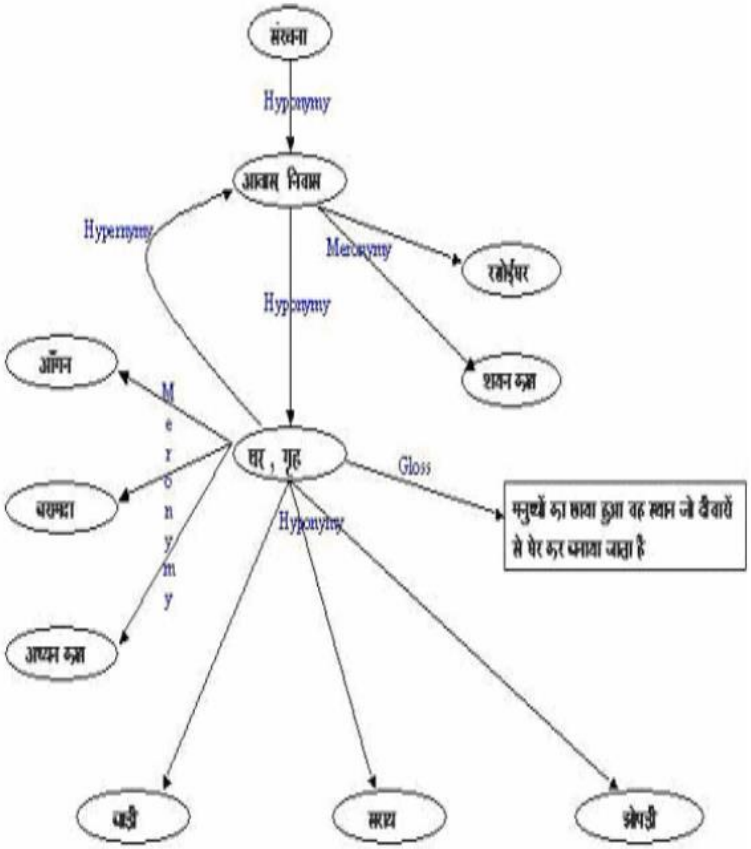
“सफाई”

Sense 1 - *"हर वस्तु की सफ़ाई जरूरी है"*

Sense 2 - *"हमें सफ़ाई देने की जरुरत नहीं है"*

Hindi Wordnet

- A lexical database which has semantic relations between words
- For each word, its different senses are provided
- Each sense has a
  - Synset
  - Gloss
  - Semantic Relations (Homonymy, Hypernymy)



Hindi Wordnet

Noun(3)

1. **सफ़ाई**, **सफाई**, **मार्जन**, **अवदान**, **अवधावन**, **उज्वलन**, **उज्ज्वलन** - साफ करने की क्रिया *"हर वस्तु की सफ़ाई जरूरी है ।"*

(R)(E)(A)(Be)(Bo)(G)(K)(Ka)(Ko)(M)(Ma)(Ml)(N)(O)(P)(S)(T)(Te)(U) (Close)

- Ontology Nodes (Close)
- Hyponymy (... is a kind of)
- Hypernymy (is a kind of ...)

2. **स्वच्छता**, **सफ़ाई**, **सफाई**, **शुद्धता**, **शुद्धि**, **निर्मलता**, **सुथरापन**, **साफ़-सफ़ाई**, **साफ-सफाई**, **उज्वलता**, **उज्ज्वलता**, **उजलापन**, **उज्वला**, **उज्ज्वला**, **उजलाई**, **उजराई**, **अमलता**, **पूति**, **धवलिमा** - स्वच्छ होने की अवस्था या भाव *"स्वच्छता बरतने से बीमारियाँ नहीं फैलतीं । / रासायनिक प्रक्रिया द्वारा जल की स्वच्छता बनाई रखी जा सकती है ।"*

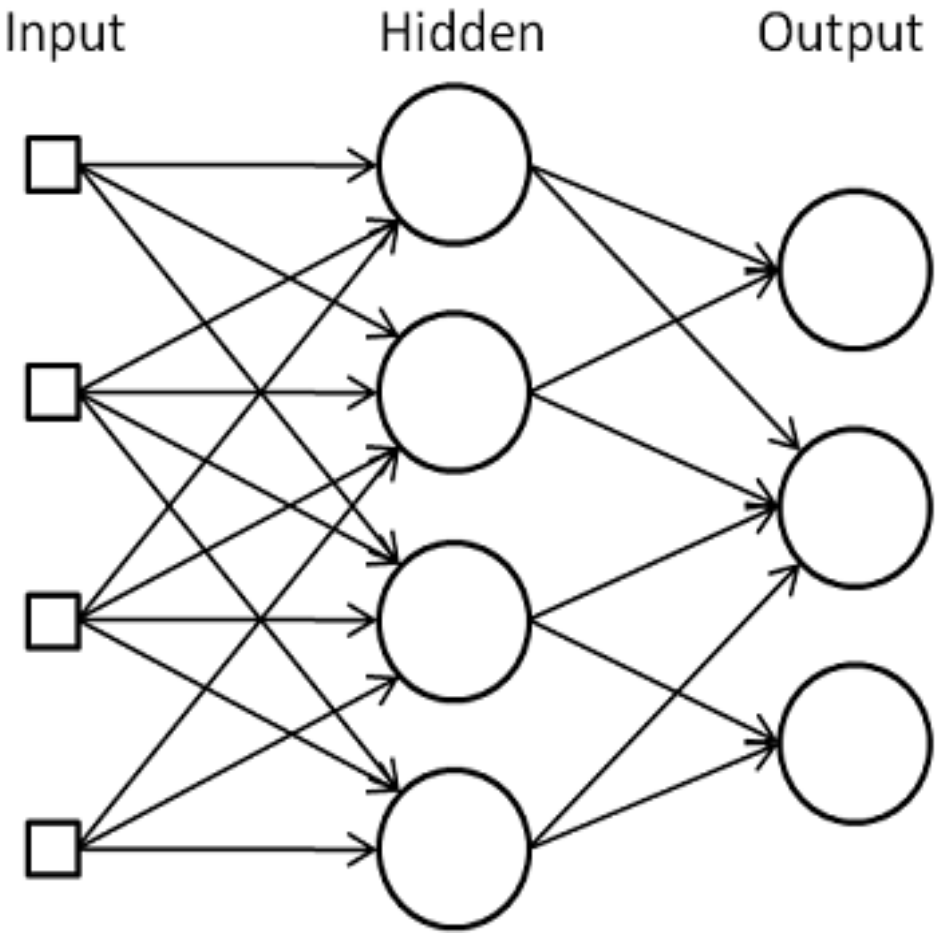
Relations and Languages

3. **सफ़ाई**, **सफाई**, **अभिवचन** - अभियुक्त आदि का अपनी निर्दोषिता प्रमाणित करने के लिए कुछ कहने की क्रिया *"उन्हें सफ़ाई देने का मौका ही नहीं मिला ।"*

Relations and Languages

Mikolov's Word2Vec Model

- Vectorizes a word using a skip-gram model
- The weights learnt form the "features" of the word according to its surroundings



Sense Vector Method

- Word2Vec (skip-gram) model is trained on a POS-tagged Hindi corpus
- For a word to be disambiguated, we create vector representations for each sense by averaging over the vectors of the relevant words in the gloss
- A vector representation for the word to be disambiguated is created by averaging the vectors of the relevant words in its neighbourhood
- We assign it the sense which has maximum cosine similarity with its vector

Older WSD Approaches

- Lesk's Algorithm : We assign that sense to the word which has maximum word-overlap

- Unsupervised clustering : Forming clusters using BOW (Bag Of Words) model for each sense of the word

THANK YOU

References

1. 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.
2. 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
3. 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.