ISSN: 2321-9653

INTERNATIONAL JOURNAL FOR RESEARCH IN APPLIED SCIENCE AND ENGINEERING TECHNOLOGY (IJRASET)

Study of Hindi Word Sense Disambiguation Based on Hindi WorldNet

Preeti Yadav[1], Mohd. Shahid Husain[2]

[1] *Department of Computer Science, M.J.P. Rohilkhand University, Bareilly,India*

[2]*Department of Computer Science, Integral University, Lucknow, India*

Abstract:. In computational linguistics, word sense disambiguation (WSD) is an open problem of natural language processing, which governs the process of identifying which sense of a word (i.e. meaning) is used in a sentence, when the word has multiple meanings. In this paper, We introduce the reader to the motivations for solving the ambiguity of words and provide a description of the task. We overview supervised, unsupervised, and knowledge-based approaches with help of Hindi Wordnet prepared by IIT Bombay.

Keywords: Word Sense Disambiguation, Ambiguity for Humans and Computers, Hindi WordNet, approaches to WSD, WSD for Hindi languages, WSD applications.

I. WORD SENSE DISAMBIGUATION

In computational linguistics, word sense disambiguation (WSD) is an open problem of natural language processing, which governs the process of identifying which sense of a word (i.e. meaning) is used in a sentence, when the word has multiple meanings. The solution to this problem impacts other computer-related writing, such as discourse, improving relevance of search engines, anaphora resolution, coherence, inference and others. Word Sense Disambiguation (WSD) is defined as the task of finding the correct sense of the word in a context.

Human language is ambiguous, so that many words can be interpreted in multiple ways depending on the context in which they occur. For instance, consider the following sentences:[1]

(a) I can hear bass sounds.

(b) They like grilled bass.

The occurrences of the word bass in the two sentences clearly denote different meanings: low-frequency tones and a type of fish, respectively. Unfortunately, the identification of the specific meaning that a word assumes in context is only apparently simple. While most of the time humans do not even think about the ambiguities of language, machines need to process unstructured textual information and transform them into data structures which must be analyzed in order to determine the underlying meaning. The computational identification of meaning for words in context is called word sense disambiguation (WSD) [2].The task needs large amounts of word and word knowledge Let us consider the word **स्वच्छ** in the following Hindi sentence.

आज हर व्यक्ति पर्यावरण की बात करता है, प्रदूषण से बचाव के उपाय सोचता है। व्यक्ति स्वच्छ और प्रदूषण-मुक्त पर्यावरण में रहने के अधिकारों के प्रति सजग होने लगा है और अपने दायित्वों को समझने लगा है। वर्तमान में विश्व ग्लोबल वार्मिंग के सवालों से जूझ रहा है।

Adjective (5)

1. स्वच्छ, मेघहीन, निरभ्र, अनभ्र, मेघरहित, अनाकाश, अपघन, अमेघ, खुला, साफ़ - मेघ से रहित "रात का समय था और स्वच्छ गगन में तारे स्पष्ट दिखाई दे रहे थे"

2. साफ़, स्वच्छ - (द्रव) जिसमें तलछट न हो "हमेशा साफ़ पानी पीना चाहिए"

3. साफ, साफ़, स्वच्छ, खुला - जो घटा, कोहरे आदि से आच्छादित न हो "सुबह की अपेक्षा दोपहर को मौसम साफ़ था"

4. निर्मल, विमल, पवित्र, शुद्ध, स्वच्छ, साफ़, साफ, चंगा, साफ सुथरा, साफ-सुथरा, ताजा, ताज़ा, अमल, प्रांजल, विशुद्ध, पाकीजा, पाक़ीज़ा, पावित, अनाविल, अपंकिल, नफीस, नफ़ीस, अमनिया, अमलिन, शुक्र, अम्लान, सित, साफ़-सुथरा, अवदात, इद्ध - जिसमें

ISSN: 2321-9653

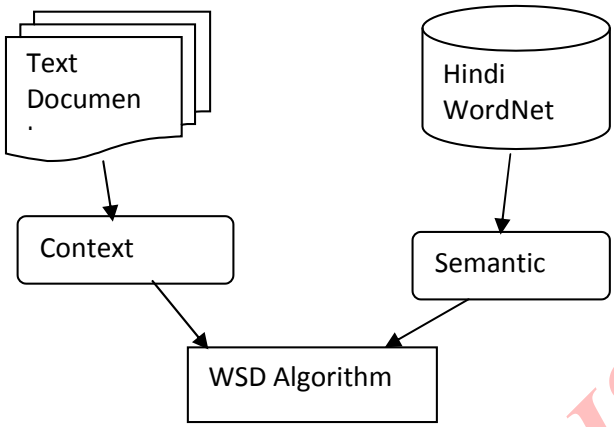
INTERNATIONAL JOURNAL FOR RESEARCH IN APPLIED SCIENCE AND ENGINEERING TECHNOLOGY (IJRASET)

किसी प्रकार का मल या दोष न हो "निर्मल मन से प्रभु को याद करो/ वातावरण शुद्ध होना चाहिए"

5. उज्वल, उज्ज्वल, उजला, साफ, स्वच्छ, साफ़, साधुजात, सित, अवदात, उजर, उजरा, उज्जर, उज्जल - जो प्रकाशमान हो "उसके कपड़े उज्ज्वल थे और वह किसी संभ्रांत घर का लग रहा था"

In this particular case, sense 1 is the most appropriate one, though sense 2, 3, 4and 5 too are relevant.

The conceptual model of WSD is:



II. ROLE OF WSD

Word sense disambiguation at ask of removing the ambiguity of word in context,

Is important for many NLP applications such as:

Machine Translation The term Machine Translation (MT) is the now traditional and standard name for computerised systems responsible for the production of translations from one natural language into another, with or without human assistance. Earlier names such as 'mechanical translation' and 'automatic translation' are now rarely used in English; but their equivalents in other languages are still common. [4]

For example-

सोना सोना चाहता है ।

It can be translated as-
Sona wants gold.
Or
Sona wants to sleep.
Or
Gold wants to sleep.
Or
Sleep wants gold.
Or
Gold wants Sona etc.

So in this way there is ambiguity for सोना because it is being interpreted of as gold means सोना or as sleep means 'नीद' or as Sona (the name) means 'सोना' .

Information Retrieval When searching for specific keywords, it is desirable to eliminate occurrences in documents where the word or words are used in an inappropriate sense; for example, when searching for sleep references, it is desirable to eliminate documents containing the word **नीद**' associated *Speech Processing* Speech recognition in computer domain involves various steps with issues attached with them. The steps required to make computers perform speech recognition are: Voice recording, word boundary detection, feature extraction, and recognition with the help of knowledge models.[8].

Example: "राजू ने रेल यात्रा के लिए एक महीने का पास निकाला" or "उसका कार्यालय पास ही है" or "मेरे पास एक गाय है"

Text Processing Text to Speech translation i.e, when words are pronounced in more than one way depending on their meaning. Example: मान can be disambiguated as weight of something or the honour.

III. APPROACHES OF WSD

As in all natural language processing, there are two main approaches to WSD – deep approaches and shallow approaches.

Deep Approaches: Deep approaches presume access to a comprehensive body of world knowledge. Knowledge, such as "दया एक सात्विक भावना है" or "दया भुवनेश्वर के पास से बहती है", here दया is Ambiguated by two meaning 'compassion' and 'name of river'. Then Deep approaches used to determine in which sense the word is used[3].

There are two types of Deep approach of Word Sense Disambiguation are:

Selectional restriction'- based approaches: They have frequently been cited as useful information for WSD. But it has been noted that their use is limited and that additional sources of knowledge are required for full and accurate WSD. Indeed, the exemplar for sense disambiguation in most computational settings is Katz and Fodor's use of Boolean

ISSN: 2321-9653

selection restrictions to constrain semantic interpretation. [13] For example खाना can be treated as food or to eat, only first sense is available in the context of "उसको आम खाना है"only second sense is applicable here as 'आम' species the selection restriction to eat in the context.

Approaches based on general reasoning with 'world knowledge': The Lesk algorithm may be identified as a starting point for resurgence of activity in this area that continues to this day. It selects a meaning for a particular target word by comparing the dictionary definitions of its possible senses with those of the other content words in the surrounding window of context. It is based on the intuition that word senses that are related to each other, are often defined in a dictionary using many of the same words. In particular, the Lesk's algorithm treats glosses as unordered bags of words, and simply counts the number of words that overlap between each sense of the target word and the senses of the other words in the sentence. This algorithm selects the sense of the target word that has the most overlaps with the senses of the surrounding words. [5]

Shallow Approaches: Shallow approaches don't try to understand the text. They just consider the surrounding words, using information such as: if दया has words भावना or दुख nearby , it probably in the sense of 'compassion'; if दया has worlds बहती or भुवनेश्वर nearby, it probably in the sense of 'river'

These rules can be automatically derived by the computer, using a training corpus of words tagged with their word senses. This approach, while theoretically not as powerful as deep approaches, gives superior results in practice, due to the computer's limited world knowledge.

Our paper is base on the Shallow approach methodology.

The different types of Shallow approaches of WSD are: Supervised methods, Semi-supervised, Unsupervised methods, Hybrid approach. *Supervised Techniques:* The learning here perform in supervision. Let us take the example of the learning process of a small child. The child doesn't know how to read/write. He/she is being taught by the parents at home and then by their teachers in school. The children are trained and modules to recognize the alphabets, numerals, etc. Their each and every action is supervised by the teacher. Actually, a child works on the basis of the output that he/she has to produce. Similarly, a word sense disambiguation system is learned from a representative set of labeled instances drawn from same distribution as test set to be used. In supervised learning, it is assumed that the correct (target) output values are known for each Input. So, actual output is compared with the target output, if there is a difference, an error signal should be generated by the system. This error signal helps the system to learn and reach to the desired or target output. *Unsupervised Technique:* In unsupervised learning technique, no supervision is provided. Let us consider an example of a tadpole. Learning is done by itself i.e. child fish learn to swim without any supervision. It is not taught by anyone. Thus its leaning process is independent and not supervised by a teacher. Unsupervised approaches to word sense disambiguation eschew the use of sense tagged data of any kind during the training. In this technique, feature vector representations of unlabeled instances are taken as input and are then grouped into clusters according to a similarity metric. These clusters are then labeled by hand with known word senses. Main disadvantage is that senses are not well defined. *Semi-Supervised Techniques:* In semi-supervised learning techniques, the information is present like in supervised but might be less information is given. Here only critic information is available, not the exact information. For example, the system may tell that only particular about of target output is correct and so. The semi-supervised or minimally supervised methods are gaining popularity because of their ability to get by with only a small amount of annotated reference data while often outperforming totally unsupervised methods on large data sets. There are a host of diverse methods and approaches, which learn important characteristics from auxiliary data and cluster or annotate data using the acquired information.

IV. HINDI WORDNET

Pushpak Bhattacharyya [3], The Hindi WordNet is a system for bringing together different lexical and semantic relations between the Hindi words. It organizes the lexical information in terms of word meanings and can be termed as a lexicon based on psycholinguistic principles. The design of the Hindi WordNet is inspired by the famous English WordNet.

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. This is done to remove ambiguity in cases where a single word has multiple meanings. Synsets are the basic building blocks of WordNet. The Hindi WordNet deals with the content words, or open class category of words. Thus, the Hindi WordNet contains the following category of words- Noun, Verb, Adjective and Adverb.

ISSN: 2321-9653

Each entry in the Hindi WordNet consists of following elements

1. Synset: It is a set of synonymous words. For example, "विद्यालय, पाठशाला, स्कूल" (vidyaalay, paathshaalaa, skuul) represents the concept of school as *an educational institution*. The words in the synset are arranged according to the frequency of usage.

2. Gloss: It describes the concept. It consists of two parts:

Text definition: It explains the concept denoted by the synset. For example, "वह स्थान जहाँ प्राथमिक या माध्यमिक स्तर की औपचारिक शिक्षा दी जाती है" (vah sthaan jahaa praathamik yaa maadhyamik star kii aupacaarik sikshaa dii jaatii hai) explains the concept of school as an educational institution.
Example sentence: It gives the usage of the words in the sentence. Generally, the words in a synset are replaceable in the sentence. For example,इस विद्यालय में पहली से पाँचवीं तक की शिक्षा दी जाती है" (is vidyaalay me pahalii se pancvii tak kii shikshaa dii jaatii hai) gives the usage for the words in the synset representing school as an educational institution. Each synset is mapped into some place in the ontology. A synset may have multiple parents. The ontology for the synset representing the concept school is shown in figure.

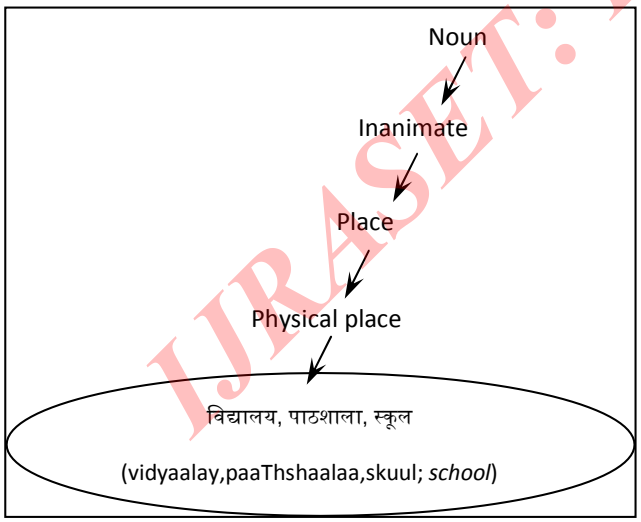


Figure IV.1. Ontology for the synset of school

Current Status of Hindi WordNet is still under construction. In the version 1.0 we have attempted to cover all the common concepts in Hindi. The present status is as follows:

Total unique words: 97441
Total Synsets: 38461
Linked Synsets: 25219
Last Updated: 31 Mar 2014

V. WSD ALGORITHMS

Manish Sinha, Mahesh Kumar Reddy .R, Pushpak Bhattacharyya , Prabhakar Pandey and Laxmi Kashyap[4], "Hindi Word Sense Disambiguation" that was the first attempt for an Indian language at automatic WSD. The use of the Wordnet for Hindi developed at IIT Bombay, which is a highly important lexical knowledge base for Hindi. The main idea is to compare the context of the word in a sentence with the contexts constructed from the Wordnet and chooses the winner. The output of the system is a particular synset number designating the sense of the word. The mentioned Wordnet contexts are built from the semantic relations and glosses, using the Application Programming Interface created around the lexical data. The evaluation has been done on the Hindi corpora provided by the Central Institute of Indian Languages and the results are encouraging. Currently the system disambiguates nouns. Work is on for other parts of speech too.
Neetu Mishra, Shashi Yadav and Tanveer J. Siddiqui[5], "An Unsupervised Approach to Hindi Word Sense Disambiguation" The algorithm learns a decision list using untagged instances. Some seed instances are provided manually. Stemming has been applied and stop words have been removed from the context. The list is then used for annotating an ambiguous word with its correct sense in a given context. The evaluation has been made on 20 ambiguous words with multiple senses as defined in Hindi WordNet.
Rohan[6], "Word Sense Disambiguation for Hindi Language" attempt to resolve the ambiguity by making the comparisons between the different senses of the word in the sentence with the words present in the synset form of the WordNet and the information related to these words in the form of parts-of-speech. This WordNet is considered to be the most important resource available to researchers in computational linguistics, text analysis and many related areas.
Avneet Kaur[7], "Development of an Approach for Disambiguating Ambiguous Hindi postposition" They have chosen to develop an efficient algorithm for disambiguating ambiguous postpositions present in the Hindi language. They are taking this problem with the case study of existing HindiPunjabi Machine Translation System. Thus the disambiguation will be done from the machine translation point of view. This is mainly used for removing the ambiguity from the corpus. N-gram algorithm is used for developing the

Hindi postpositions. N-gram algorithm is used for extracting the words from the corpus.

Ripul Gupta [8], "Speech Recognition for Hindi" Speech interface to computer is the next big step that computer science needs to take for general users. Speech recognition will play a important role in taking technology to them. The need is not only for speech interface, but speech interface in local languages. His goal is to create speech recognition software that can recognise Hindi words. That report takes a brief look at the basic building block of a speech recognition engine. That talks about implementation of different modules. Sound Recorder, Feature Extractor and HMM training and Recogniser modules have been described in details. The results of the experiments that were conducted are also provided. The report ends with a conclusion and Future plan.

Ravi Sinha and Rada Mihalcea[9] "Unsupervised Graph-based Word Sense Disambiguation Using Measures of Word Semantic Similarity" that paper describes an unsupervised graph-based method for word sense disambiguation, and presents comparative evaluations using several measures of word semantic similarity and several algorithms for graph centrality. The results indicate that the right combination of similarity metrics and graph centrality algorithms can lead to a performance competing with the state-of-the-art in unsupervised word sense disambiguation, as measured on standard data sets.

Siva Reddy, Abhilash Inumella, Rajeev Sangal, Soma Paul[10], "All Words Unsupervised Semantic Category Labeling for Hindi" they use the ontological categories defined in Hindi Wordnet as semantic category inventories. In this paper they present two unsupervised approaches namely Flat Semantic Category Labeler (FSCL) and Hierarchical Semantic Category Labeler (HSCL). The former method treats semantic categories as a flat list, whereas the latter one exploits the hierarchy among the semantic categories in a top down manner. Further their methods use simple probabilistic models, using which the category labelling becomes a simple table look up with little extra computation and thus opening the possibility of its use in real-time interactive systems.

R. Mahesh K. Sinha,[11] " Learning Disambiguation of Hindi Morpheme 'vaalaa' with a Sparse Corpus" The Hindi morpheme 'vaalaa' is very widely used as a suffix and also as a separate word. The common usage of this suffix is to denote an activity or profession of a person. This form of the usage has been borrowed in English with the spelling of 'wallah'. However, it has a large number of other interpretations depending upon the context in which it is used. That paper presents an account of different senses in which this morpheme is used in Hindi and presents a strategy for learning their disambiguation based on contextual features with sparse data using a semi-supervised method. They present a new technique of unifying learned instances using supervised training with limited data and computing matching index and bootstrapping the training set to deal with corpus sparseness. This study finds application in machine translation, information retrieval, text understanding and text summarization.

Parul Rastogi and Dr. S.K. Dwivedi[12], "Performance comparison of Word Sense Disambiguation (WSD) Algorithm on Hindi Language Supporting Search Engines", The major population of India use Hindi as a first language. The Hindi language web information retrieval is not in a satisfactory condition. Besides the other technical setbacks, the Hindi language search engines face the problem of sense ambiguity. Their WSD method is based on Highest Sense Count (HSC). That works well with Google. The objective of that paper is comparative analysis of the WSD algorithm results on the three Hindi language search engines- Google, Raftaar and Guruji. They have taken a test sample of 100 queries to check the performance level of the WSD algorithm on various search engines.

Mitesh M. Khapra, Pushpak Bhattacharyya, Shashank Chauhan and Soumya Nair,[13] "Domain Specific Iterative Word Sense Disambiguation in a Multilingual Setting", they work on Domain Specific Iterative Word Sense Disambiguation (WSD) for nouns, adjectives and adverbs in the trilingual setting of English, Hindi and Marathi The methodology proposed relies on dominant senses of words in specified domains. They combine corpus biases for senses along with information in wordnet graph structure to arrive at the sense decisions. To the best of our knowledge, that is the first attempt at a large scale multilingual WSD involving Indian languages and English.

VI. CONCLUSION

The words appeared in the polysemous context have different roles to determine the polysemous sense. We can extract the words in the polysemous context by a variety of ways and measure their degree of importance in determining the meaning of polysemy. In this process, we must not only consider the sentence collocation, but also should further consider the syntax and semantic to obtain more knowledge in line with human cognitive behavioral models. Co-occurrence words, collocation words and demonstratives have different degrees of constraint on determining the polysemy sense. Therefore they can be extracted from the corpus, dictionaries, and knowledge source to construct the unified disambiguation knowledge base, and then use them for disambiguation. But for the use of the knowledge base there are still some

shortcomings lead to disambiguation correct rate is not high. How to fully use the knowledge base to improve the disambiguation correct rate will be researched in our further study.

REFERENCES

1. Esha Palta, "Word Sense Disambiguation," M.Tech thesis, Indian Institute of Technology, Mumbai, CSE dept., India,2006.
2. "Word Sense Disambiguation",2009. http://en.wikipedia.org/wiki/Word-sense_disambiguation#Approaches_and_methods
3. Dr. Pushpak Bhattacharyya, "Hindi WordNet Data and Associated Software License Agreement", Indian Institute of Technology, Mumbai, CSE dept., Tchnical Report 2006.
4. Manish Sinha, Mahesh Kumar Reddy, R Pushpak Bhattacharyya, Prabhakar Pandey and Laxmi Kashyap, "Hindi Word Sense Disambiguation", *Indian Institute of Technology Bombay, Department of Computer Science and Engineering Mumbai*, 2008.
5. Neetu Mishra, Shashi Yadav and Tanveer J. Siddiqui, "An Unsupervised Approach to Hindi Word Sense Disambiguation," *Indian Institute of Information Technology, Allahabad. UP, India*, 2009.
6. Rohan, "Word Sense Disambiguation For Hindi language" *Thapar University Patiyala*, CSE Dept., India, 2007.
7. Avneet Kaur, "Development of an Approach for Disambiguating Ambiguous Hindi postposition," *International Journal of Computer Applications (0975 – 8887)*, vol.5, no.9, August 2010.
8. Ripul Gupta, "Speech Recognition for Hindi," M.Tech. thesis Indian Institute of Technology, Mumbai, CSE dept., India, 2007.
9. Ravi Sinha and Rada Mihalcea, "Unsupervised Graph-based Word Sense Disambiguation Using Measures of Word Semantic Similarity," *IEEE International Conference on Semantic Computing*, pp. 363 – 369, Sept. 2007.
10. Siva Reddy, Abhilash Inumella, Rajeev Sangal, Soma Paul, "All Words Unsupervised Semantic Category Labeling for Hindi" *Proceedings of the International Conference RANLP, Borovets, Bulgaria*, pages 365–369, September 2009.
11. R. Mahesh K. Sinha, "Learning Disambiguation of Hindi Morpheme 'vaalaa' with a Sparse Corpus," *International Conference on Machine Learning and Applications*, pp. 653 – 657, December 2009.
12. Parul Rastogi and Dr. S.K. Dwivedi, "Performance comparison of Word Sense Disambiguation (WSD) Algorithm on Hindi Language Supporting Search Engines", *International Journal of Computer Science Issues*, vol. 8, issue.2, March 2011.
13. Mitesh M. Khapra, Pushpak Bhattacharyya, Shashank Chauhan and Soumya Nair, "Domain Specific Iterative Word Sense Disambiguation in a Multilingual Setting ", *Proc. of ICON-2008: 6th International Conference on Natural Language Processing Macmillan Publishers, India*, 2008.