WORD SENSE DISAMBIGUATION IN THE HINDI LANGUAGE: NEURAL NETWORK APPROACH

Sailendra Kumar[1], Rakesh Kumar[2]
E-Mail Id: sailendra.patel.2014@gmail.com, rakesh_rbl@rediffmail.com
Department of Computer Science, Assam University Silchar, Assam, India

Abstract- Hindi is the national language of India. A massive number of peoples share, retrieve, and access documents in the Hindi language. Hindi Word Sense Disambiguation (HWSD) system used to extract ambiguity from the Hindi language. "Word Sense Disambiguation (WSD) eliminates ambiguity and you can easily understand the meaning of a specific ambiguous word used in sentence". It comes up as a field of research in computational linguistics and it helps in learning the real concept of the words appearing in a particular context. Humans can easily use the WSD technique to distinguish the different meanings and can speak a better language. However, computers may find it difficult to deal with the WSD technique. There are different approaches using which it has become easy to carry out the complete procedure. The four main approaches, which are commonly used, are knowledge-based, Supervised, Semi-Supervised, and Unsupervised. Hence, it improves the computer's performance and you can learn the true importance of search engine optimization. It also helps in collecting information and helps in dealing with different software's. If you are looking for a voice assistant this method works the best and you can explore the best form of machine learning. It comes up with an organized neural network and the algorithms help in detecting the differences easily and you would get accurate results. There is an inner layer of the network with nodes and you can recognize the binary values, which are set according to the frequency of the context words followed by the ambiguous words. On the other hand, there is an outer layer too consisting of the nodes, which has a similarity to the senses of the ambiguous words. "In this paper, we describe different approaches used in WSD, resources required for disambiguation tasks, and a review of previous works for the Hindi language".
Keywords: Word Sense Disambiguation, Neural Network, Machine Learning.

1. INTRODUCTION

The Hindi language is written and spoken by the majority of Indians. Hindi, is vague, which makes it difficult to communicate. "The use of the information technology, we needs system called WSD to eliminate ambiguity from a single word, or from all words, in order to use Hindi language easily and effectively on the web". The internet is used by all to share and find knowledge. However, the data is available in natural languages. Natural languages, as we all know, are vague. Ambiguity is defined as a concept that can be interpreted in two or more ways. So, in order to make the most of information technology, we need to eliminate ambiguity from sentences using a technique called Word Sense. The role of WSD is to determine the correct interpretation of a word in a given context. It is a key challenge in Natural Language Processing as well as an open research area in the field of NLP. Consider the following two Hindi sentences as an example. राहुल यात्रा में सोना पसंद करते हैं। and सोना बहुत महंगा धातु है। The ambiguous word here is सोना, it can interpreted as either "Sleep" or "Gold" depends on context. It seems that determining the precise meaning of a given word is a simple task. By using their common sense, humans can easily detect the correct interpretation of a word provided a context. However, it is a challenging task for computers because it necessitates the processing of a large volume of unstructured data found in natural languages and translating it into a data structure that must be analyzed in order to determine the correct interpretation. The aim of WSD is to compare a meaning repository, like Hindi WordNet, Machine readable dictionary, etc., with a context to get a sense of the ambiguous word, or all words. The conceptual model for WSD is shown in fig. 1.1.

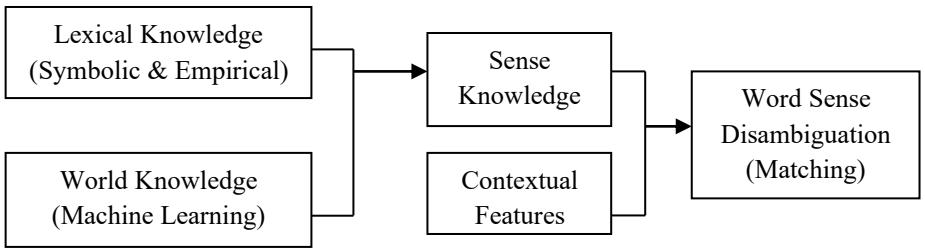

Fig. 1.1 The Conceptual Model for Word Sense Disambiguation

1.1 Lexical Knowledge

Information that can be conveyed in words is referred to as lexical knowledge. Despite how circular this might seem, we believe it offers a perfectly logical starting point, since it posits communicability as the most distinguishing feature of lexical understanding, following a long philosophical tradition.

1.2 World Knowledge

Knowing the meanings of words, understanding the relationships between words (word schema), and possessing linguistic knowledge about words are all examples of word knowledge. Understanding (background knowledge) of several different topics and disciplines (domains) and how they relate to one another constitutes world knowledge.

1.3 Sense Knowledge

The part of experience that can be directly linked to how the senses work. Sensing, as it is generally known, is the capacity to respond to such stimuli on an automatic or preconscious level; as such, it enables intelligence, despite the fact that it "knows" nothing in and of itself.

1.4 Contextual Features

POS tags, local collocations, bag-of-words, named entities, and a predicate-argument structure were all used to create the WSD scheme. The following features were extracted from the OntoNotes corpus.
Word knowledge can be obtained from manually sense-tagged examples and sense knowledge can be obtained from dictionaries that provide definitions and lexical knowledge for each sense. The context in which an ambiguous word appears may provide contextual information. As a result, disambiguation is accomplished by comparing a contextual function to sense information [2],[3],[4],[5].

2. APPLICATIONS OF HINDI WORD SENSE DISAMBIGUATION

2.1 Machine Translation

Machine translation is the most significant application of HWSD. For example, depends on context, the Hindi word 'सोना' may be translated into English as 'Gold' or 'Sleep,' resulting in an incorrect sentence in another language [6].

2.2 Information Retrieval

It's primarily used to find useful information on the internet. There are some unclear terms in the query used to retrieve documents. As a result, for proper question disambiguation, Words may aid in the retrieval of pertinent information [6].

2.3 Speech Synthesis

It is necessary for correct phonetization of ambiguous words in speech synthesis [6].

2.4 Text Processing

Text processing is the process of correcting a word's spelling. WSD can recognize word senses based on meaning and perform spelling correction [6].

2.5 Grammatical Analysis

The HWSD is needed for accurate part of speech tagging (POS), as words may have different POS depends on the background [6].

3. RESOURCES FOR HINDI WORD SENSE DISAMBIGUATION

WSD is built on the foundation of information. It takes knowledge to assign a meaning to an expression. We give detailed introduction of required resources for task of HWSD.

3.1 Structured Resources

The following are the structured resources available for HWSD.

3.1.1 Hindi WordNet

Since it contains rich semantic networks of concepts, Hindi wordnet is now the most common source of information for HWSD after machine readable dictionary [4].

3.1.2 Machine Readable dictionaries

It contains list of meanings, examples, and definitions. For example, in Hindi, khoj etc.

3.1.3 Thesauri

Thesauri involves a connection between words like Synonyms and antonyms, as well as a variety of other relationships.

3.2 Unstructured Resources

3.2.1 Corpora

Corpora are collections of texts that are required to learn language models. For both supervised and unsupervised approaches to WSD, corpora are needed. It can be raw corpora or meaning annotated corpora.

3.2.2 Raw Corpora

Since it lacks senses associated with the phrases, it is mostly used for unsupervised approaches. "The Central Institute of Indian Languages has developed the Hindi Corpus" [7].

3.2.3 Sense-Annotated Corpora

For the supervised approach, sense-annotated corpora are used because they have meaning associated with ambiguous words, and classification is based on these sense-annotated corpora. However, creating sense-tagged corpora for resource-scarce languages like Hindi necessitates a lot of manual effort.

4. APPROACHES AND METHODS TO HWSD

There are two ways to disambiguate words that are followed for Word Sense Disambiguation, machine learning methods and knowledge-based methods. In this system is trained to perform the task of defining the meaning of words. In Knowledge based approach, it needs external lexical resources like Word web, dictionary, synonym finder etc.

4.1 Machine Learning Methods

"Initial input is that the word to be disambiguated referred to as target word, and the text during which it's embedded, referred to as context". In this approach options are themselves served by the words. The value of feature is that the range of times the word occurs within the region closes the target word. The region is commonly a hard and fast window with target word as center. There are three kinds of techniques of machine learning based approaches are supervised techniques, unattended techniques, and semi-supervised techniques [1].

4.2 Supervised Approach

"It infers a classifier from manually sense-annotated data sets using machine-learning techniques". Typically, the classifier is focused on a single word and performs a classification task to allocate the specific meaning of each instance of that word. Usually, training set used for learn classifier is contains a set of examples in which a specific target word is tagged with sense from sense inventory of reference dictionary. "A representative set of classified instances taken from the same distribution as the test set is used to train the word sense disambiguation system" [2].

4.2.1 Neural Networks

Neural networks processes knowledge based on computational model of connectionist approach. Target output is included in input like input features. Based on the desired outputs, the dataset divided in non-overlapping sets. As the network detects new input pairs the weights are changed such that output's unit giving the target output has the larger activation [8].

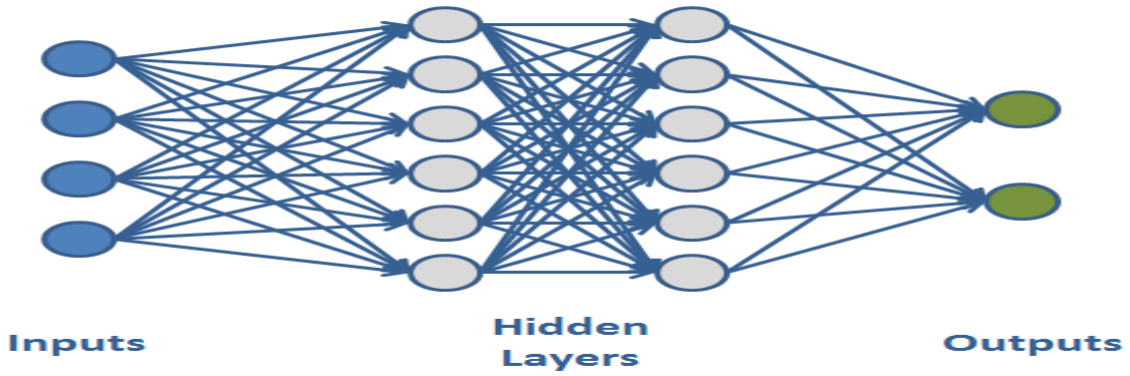

Fig. 4.1 Neural Network Conceptual Model

Normally, neural networks are read from left to right. In this first layer is where inputs are entered. There are one or more than one internal layers (referred to as hidden layers) that perform math and a final layer that comprises all possible outputs. The input is the data that is first fed into the neural network from the source. Its aim is to provide data to the network so that it can make a decision or make a prediction based on the information it receives. In most cases, the neural network model accepts real-valued inputs, which should be fed into a neuron in the input layer.

4.3 Un-supervised Approach

For HWSD, an un-supervised approach, unlike a supervised approach, doesn't require hand labeled awareness of meaning information in large scale resources. It's founded on fact that's words with similar meanings would be surrounded by words with similar meanings. "The task is to classify the new occurrence to the derived clusters, and word senses are derived by creating clusters of occurrences of words. Instead of assigning meaning labels, this method detects clusters".

4.3.1 Context Clustering

It's based on clustering technique, in which, "background vectors are generated first, and then clustered to determine the meaning of the word. It uses vector space as a word space, with only words as dimensions". "A word in a corpus will also be denoted as a vector in this system, and the number of times it occurs within its meaning will be counted". Co-occurrence matrix is established and used similarities measures. After that, every clustering strategy used to implement discrimination [9],[10].

4.3.2 Word Clustering

Words with identical definitions are assigned to same cluster with using these techniques. One of the methods listed in was to find a word sequence that similar with target words. Syntactical dependency determines how similar the words are. If W contains terms that are identical to w_m, a tree is created with only one node wm at

first, and a node w_i will have a child node wm when w_i is discovered to be the word with the similar meaning of w_m [11].

4.3.3 Co-occurrence Graphs

"This approach generates a co-occurrence graph with a vertex V and an edge E, where V indicates the words in the document and E is inserted if the words co-occur in the same paragraph or text according to syntax. For a given target word, the graph is first formed, followed by the graph's adjacency matrix. The Markov clustering method is then used to determine the meaning of the word". Each edge of graph assigned weight, it is co-occurring frequency of those words. The formula for calculating the weight of edge m,n is [1],[12]:

$$w_{mn} = 1 - \max\{P(w_m|\, w_n),\, P(w_n|\, w_m)\} \ldots$$

Where $freq_{mn}$ is the co-occurrence frequency of terms w_m and w_n, and $freq_n$ is the occurrence frequency of w_n, and $P(w_m|w_n)$ is the $freq_{mn}/freq_n$. Weight of 0 occurred words, while words that occur infrequently are given a weight of 1. The target word's hubs with zero weight are connected, minimum spanning tree used graph to build. Spanning tree is determines the target word's true meaning.

4.4 Semi-supervised Approach

Information is present in semi-supervised learning techniques, In the supervised techniques, although there might be less information provided. Only critic information, not exact information, is available here. The system can tell that only specific about of target performance is correct and so. Semi-supervised or minimally supervised methods are popular because their ability of work with very little annotated reference data while outperforming fully un-supervised methods on large datasets. Different methods and approaches for extracting important characteristics from auxiliary data and clustering or annotating data with the information gleaned.

4.4 Knowledge-based Approach

They can use grammar rules for HWSD in a this approach based on machine-readable dictionaries in the formation of corpus, Hindi WorldNet, and so on. Main aim of Knowledge-based approach (Dictionary-based approach) WSD is to use knowledge tools to infer word meanings in context. Dictionaries, thesauri, ontologies, collocations are knowledge resources. Although the above methods perform deficient than supervised counter parts, they have the advantage of a wider range.

4.4.1 Overlap Based Approaches

Machine readable dictionary is used in this approach necessitates. It entails determining the various characteristics of ambiguous word senses, along with characteristics, meaning of words.

4.4.2 Lesk's algorithm

"W is defined as a word that creates disambiguation, C defined as a set of terms in the meaning array in the surrounding, S is defined as a senses for W, and B is defined as bag of words derived from glosses, synonyms, hyponyms, glosses of hyponyms, example sentences, hypernyms, glosses of hypernyms, meronyms, example sentences, hypernyms, glosses of hypernyms, meronyms Then, using the interaction similarity law, calculate the overlap and produce the meaning that is most likely to have the greatest overlap".

4.4.3 Walker's approach

The algorithm described as follows: each word in the thesaurus is allocated more than one or one subject categories. In different senses of the word, different subjects are allocated.

4.5 Selection Preferences

In selection Preferences find information about the likely relationships between word types, and use knowledge sources to denote common sense. Words senses are omitted in this method and only certain senses are chosen that are in accordance with common sense laws. "The basic idea behind this method is to count how many times a given word pair with syntactic relation appears in the corpus". Word senses will be listed based on this count. Other methods, such as conditional probability used to find this type of relationship between words [13].

CONCLUSION

We define a WSD for the Hindi language, its uses, and the various methods used for the disambiguation. We presented a survey of the various approaches available in HWSD, with a particular emphasis on Machine Learning Methods and Dictionary-based (knowledge-based) approaches. We concluded, the supervised approach outperforms the unsupervised approach, but one of its drawbacks is requires a large corpora without training is impossible, it can be overcome in un-supervised approach since it doesn't depend on such a massive scale resources for disambiguation. The Dictionary-based approach, on other hands, relies on knowledge sources to determine meanings of words in given contexts, as long as a machine-readable knowledge base is available.

REFERENCES

[1] Dixit, Vimal, Kamlesh Dutta, and Pardeep Singh. "Word Sense Disambiguation and Its Approaches." CPUH-Research Journal 1, no. 2 (2015): 54-58.
[2] Navigli, Roberto. "Word sense disambiguation: A survey." ACM computing surveys (CSUR) 41, no. 2 (2009): 1-69.

[3] Sharma, Dilip Kumar. "A comparative analysis of Hindi word sense disambiguation and its approaches." In International Conference on Computing, Communication & Automation, pp. 314-321. IEEE, 2015.
[4] Hindi Word Net, http://www.cfilt.iitb.ac.in/wordnet/webhwn/wn.php.
[5] Zhou, Xiaohua, and Hyoil Han. "Survey of Word Sense Disambiguation Approaches." In FLAIRS conference, pp. 307-313. 2005.
[6] Agirre, Eneko, and Philip Edmonds, eds. Word sense disambiguation: Algorithms and applications. Vol. 33. Springer Science & Business Media, 2007.
[7] Hindi Corpus, http://www.cfilt.iitb.ac.in/Downloads.html.
[8] Màrquez, Lluís, Gerard Escudero, David Martínez, and German Rigau. "Supervised corpus-based methods for WSD." In Word sense disambiguation, pp. 167-216. Springer, Dordrecht, 2007.
[9] Schütze, Hinrich. "Automatic word sense discrimination." Computational linguistics 24, no. 1 (1998): 97-123
[10] Mooney, Raymond J. "Comparative experiments on disambiguating word senses: An illustration of the role of bias in machine learning." arXiv preprint cmp-lg/9612001 (1996).
[11] Véronis, Jean. "Hyperlex: lexical cartography for information retrieval." Computer Speech & Language 18, no. 3 (2004): 223-252.
[12] Lin, Dekang. "Automatic retrieval and clustering of similar words." In 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 2, pp. 768-774. 1998.
[13] Palta, E. Word Sense Disambiguation, M.Tech. dissertation, dept. CSE Indian IIT, Mumbai (2006).