The International Journal Of Engineering And Science (IJES)
||Volume||2 ||Issue|| 4 ||Pages|| 36-41||2013||
ISSN(e): 2319 – 1813 ISSN(p): 2319 – 1805

Knowledge Based Approach for Word Sense Disambiguation using Hindi Wordnet

Prity Bala
Department of Computer Science, Apaji Institute, Banasthali Vidyapith Newai, Rajesthan, India

Abstract

Word sense disambiguation (WSD) is an open research area in natural language processing and artificial intelligence. and it is also based on computational linguistics. It is of considerable theoretical and practical interest. WSD is to analyze word tokens in context and specify exactly,which sense of several word is being used.It can be declare as theortically motivated,one or more methods and techniques. WSD is a big problem of natural language processing (NLP) and artificial intelligence. But here, the problem is to search the sense for a word in given a context and lexicon relation. It is a technique of natural language processing in which requires queries and documents in NLP or texts from Machine Translation (MT). MT is an automatic translation. It is involving some languages such as Marathi, Urdu, Bengali, Punjabi, Hindi, and English etc. Most of the work has been completed in English, and now the point of convergence has shifted to others languages. The applications of WSD are disambiguation of content in information retrieval (IR), machine translation (MT), speech processing, lexicography, and text processing. In our paper, we are using knowledge based approach to WSD with Hindi language. A knowledge based approach use of external lexical resources suach as dictionary and thesauri . It involve incorporation of word knowledge to disambiguate words. We have developed a WSD tool using knowledge base d approach with Hindi wordnet. Wordnet is built from co-occurrence and collocation and it includes synset or synonyms which belong to either noun, verb, adjective, or adverb, these are called as part-of- speech (POS) tagger. In this paper we shall introduce the implementation of our tool and its evaluation.

Keywords-Word sense disambiguation, knowledge-based-approach, Hindi wordnet, etc.*

Date Of Submission: 20 April 2013 Date Of Publication: 5,May.2013

I. INTRODUCTION

Words are having different meanings, and it is based on context of the word usage in a sentences.Word sense disambiguation is the difficulty to find the sense of a word in given a natural language context, where the words have one or more meanings. The sense of a word in a text depends on the context in which it is used, the context of the ambiguous word is certify by the others neighboring words. This is called as local context or sentential context. This task needs a lot of words and word knowledge .A WSD is the process of detemine the sense of the word.
Example:

1.मैंने अपनी नीली कलम से पत्र लिखा है |

Here, the meaning of ‘कलम’is blue pen.

2.माली ने गुलाब के फूल की कलम को काटा |

Here, the meaning of ‘कलम’is graft cutting of rose.

3. नाई ने मेरी कलम को काट दिया |

Here, the meaning of ‘कलम’graft cutting of hair.

5. यह चित्र कलम द्वारा बनाया गया है |

Here, the meaning of ‘कलम’is brush.

WSD is determining an AI-complete problem, that is, a task whose solution is at least as hard as the most difficult problems in artificial intelligence. The problem of word sense disambiguation (WSD) has been discovered as AI-complete that are the problem which can be solved by resolving the entire difficult problem related with AI.

Words do not have well-defined boundaries between their word of senses, and our task is to determine which meaning of word is indented in a given context. This is the very first problem that is encountered by any natural language processing system which is referred to as lexical semantic ambiguity. WSD is a research area in NLP, which is very useful now days. It is the technique of natural language processing (NLP). It can be represented by task, performance, knowledge source; computational complexity, assumptions and application for WSD algorithms.WSD involve more words and word knowledge or common sense, which identifies Dictionary or Thesauri. It is also beneficial in many application such as information extraction (IE), information retrieval (IR), and speech recognition (SR). Word sense disambiguation is important for lexical knowledge and word knowledge. There are various approaches to WSD such as knowledge based approach, selection restriction based WSD, Machine learning approach to WSD, which include supervised approach, unsupervised approach, semi-supervised approach, and Hybrid approach. But in our paper we focus on knowledge based approach like selection restriction considerably with Hindi wordnet.

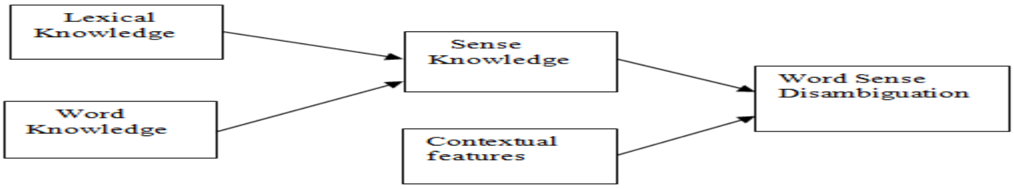

Fig 1: Conceptual model for WSD

II. LITERATURE SURVEY

2.1 Knowledge based approach or Dictionary based approach

Knowledge based approach is external lexical resources and it is a fundamental component of WSD. Knowledge sources provide a large amount of data which are necessary to associate senses with words Fellbaum Christiane [3]. They can vary from corpora of texts, either unlabeled or annotated with word senses, to machine-readable dictionaries, thesauri, collocations, Ontologies, etc. WSD is to knowledge resources to infer the senses of word in context. The knowledge resources consist like dictionary, Thesauri, ontology, and collocations etc**.** Knowledge based approach have a faith on knowledge resources of machine readable dictionaries (MRD) in form of corpus, WorldNet etc. they may use either grammatical information and rules for disambiguation.MRD like oxford English dictionary, Longman dictionary of ordinary contemporary English, Roget thesaurus and semantic networks which add lot a semantic relation like WorldNet Kieinberg.M.Jon [4] .The main knowledge based techniques namely the selection restriction, and structure approach. A review knowledge based approaches can be found also in manning and Mark Stevenson [7]. Knowledge sources used for WSD are either lexical knowledge or word knowledge, in which lexical knowledge released to the public, or world knowledge learned from a training corpus Mark Stevenson [7].

2.2 Supervised approach

Supervised approach based on a labeled training set, it is a learning system which has a training set of featured-encoded inputs and their appropriate sense label or category Fellbaum Christiane [3]. In the last 15 years, the NLP has an increased interest in machine learning approaches for automated classification of word senses Mark Stevenson [7]. Generally, supervised approaches to WSD have found better results than unsupervised methods, Kieinberg.M.Jon [4]. Supervised models fall into two classes, hidden models and explicit models based on whether or not the features are directly associated with the word sense in training set corpus, Yarowsky [8].A supervised approach only uses sense tagged corpora to train set the sense model, which makes it possible to link world knowledge to word sense Kieinberg.M.Jon [4].

2.3 Semi-Supervised or minimally supervised approach

These perform use of a secondary source of knowledge such as a small annotated corpus as seed data in a bootstrapping process, or a word-aligned corpus Agirre Eneko, and Rigau [5].Semi –supervised approach based on bootstrapping method. The bootstrapping approach starts from a small amount of data for each word, either manually-tagged training examples or a small number of surefire decision rules .The seeds data are used to train an initial classifier, using any supervised method Yarowsky [8].Other semi-supervised techniques or methods use a very large quantities of untagged corpora to provide co-occurrence information that supplements the tagged corpora. These techniques have the possible to help in the adaptation of supervised models to different domains Fellbaum Christiane [3].

2.4 Unsupervised approach

Unsupervised approach based on unlabeled corpora. It is learning system, which has a training set of feature encoded inputs but not their appropriate sense label or category. It is suited for online machine translation (MT) and information retrieval (IR). However, in theoretically, it has worse performance than the supervised approach because it relies on less knowledge. The types of lexical knowledge used consider sense frequency, sense glosses, concept trees Agiree and Rigau [5].Be Unsupervised methods have the possible to get over the knowledge acquisition bottleneck Fellbaum Christiane [3], that is, the lack of large-scale resources manually annotated with word senses. These approaches to WSD are based on the idea that the same sense of a word will have similar neighboring words. They are capable to induce word senses from input text by word occurrences, and then classifying new occurrences into the induced cluster.

III. WORDNET FUNDAMENTAL

3.1 Hindi WordNet

Wordnet is a network of words linked by lexical and semantic relation. Wordnet is a large lexical database of english.noun, verbs, adjective, and adverb. They are grouped into sets of synonym or synsets. Synsets are linked by means of conceptual-semantic and lexical relations. Wordnet have completed for Hindi and Marathi being built at IIT Bombay are amongst the first IL wordnets. Wordnet is also an electronic large lexical database of English, and it is a combination of dictionary and thesaurus in which created and maintained by cognitive science lab of Princeton university. In this way, the Hindi Wordnet is inspired by the English Wordnet. The Wordnet refers the lexical information in senses and set of words. In which is defining the meaning of the word in a specific text. Wordnet is the existence of various relations between the word forms (e.g. lexical relations, such as synonymy and antonyms) and the synonym or synsets (meaning to meaning or semantic relations e.g. hyponymy/hyponymy relation, metonymy relation). Wordnet has four types of part of speech (POS), such as noun, verb, and adverb, adjective. POS tagger is the process of identifying lexical category of a word in a sentence on the basis of its context.

3.2 Synset

synset is category of data elements that are examined semantically equivalent for the purposes of information retrieval. A collection of one or more words and phrases ("collocations") collectively referred to as "word forms" that can all divide the same meaning.

- The smallest unit in wordnet.
- A synonym set.
- Represent a specific meaning of a word.

Synsets are related to semantic and lexical relations**.** All word meaning can be represented by a set of words-forms. It is called as synonym sets or synsets. Synsets are built by contents words such as noun, verb, adjective, and adverb.

3.3 Lexical Matrix

The lexical matrix is a part of the language system .It refers the link between word form and meanings. The following table is representing the lexical matrix. It is called as lexical matrix. It shown of the lexical information by an organization. Word forms are imagined as headings for the columns and word meanings for the rows. Rows represent only synonymy while a column represents polysemy.

TABLE 1: THE CONCEPT OF LEXICAL MATRIX

Word-Meanings	***Words-Forms***				
	B1	B2	B3	……………………	Bn
A1	E1.1	E1.2			
A2		E2.2			
A3			E3.3		
..				……	
Am					Em.n

For example the word ‘खग’ of synset like **{आकाशचारी, सारंग}** gives the meanings ’सारंग’ (धातु का बना हुआ पतला हथियार जो धनुष से चलाया जाता है) belongs to a synset, whose members from a row in the matrix, and the row numbers gives a ID to the synset. 'खग' has different meanings, (पंख और चोचवाला द्रिपद जिसकी उत्पत्ति अंडे से होती है) which comes in the column by the word.

3.4 Semantic relations in Wordnet

The lexical matrix is based on an integral part of the human language system. It supports the link between word form and word meaning. The Hindi wordnet is inspired by the English wordnet, semantic relation use in structure lexical data. They have been extensively used in wordnet and estimate also, and they are mainly used to structure the lexicon such as, and the semantic relations are following below.
Types of semantic relations (It is based on POS):

- Hyponymy (kind-of): 'जल' is hyponym of 'पानी'
- Hyponymy (kind-of): 'पानी' is a hyponym of 'जल'
- Holonymy (part-of): 'अंबु' is a homonym of 'अम्बु'
- Meronymy (part-of): 'सलिल' is a meronym of 'उदक'

For example, we have the synset {जल, पानी}. The hyponymy relation (Is-A) of it links to {अंबु, अम्बु}. Its meronymy relation (Has-A) links to {सलिल, उदक} and hyponymy relation to {पय} and {वारि}.

IV. RELATED WORK

In our paper, we are using knowledge based system with Hindi language. And, I have developed a WSD tool using Hindi wordnet. But, currently our system deals with part-of-speech (POS) tagger which gives us tags such as noun, verb, adjective, and adverb. In this way t o a given corpus to assign correct sense to the word. This is called as sense tagging, so these needs word sense disambiguation (WSD). It is highly important for Question Answering, Machine Translation, Information retrieval, Text Mining tasks. Work is totally depending on to including words of other part of speech (POS). We have taken the database of text files saved from Hindi Wordnet. It prepared by IIT Bombay, but in future, the database for Hindi language's WSD can use the database prepared for Hindi Wordnet directly. But in future, the database for Hindi language's WSD can use the database prepared for Hindi Wordnet directly.

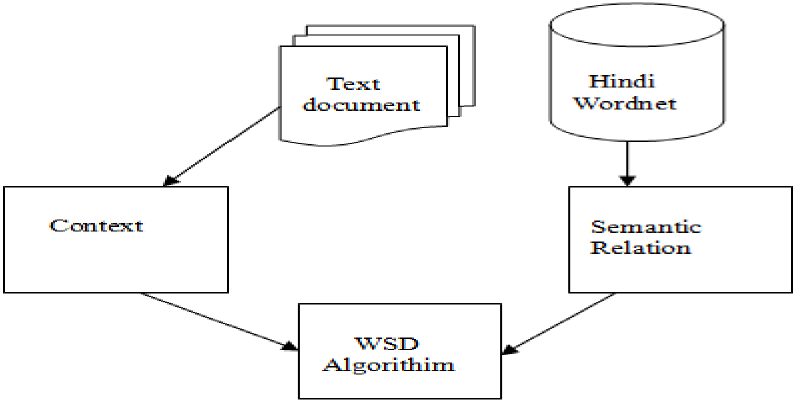

Fig 2: Word Sense Disambiguation document

4.1Methodology: our approach to WSD

Here, we define a statistical technique for assigning senses to words in Hindi. A word is committed a sense with the use of given below:
[1] The context in which it has been considered
[2] The information in the Hindi Wordnet, and
[3] The overlap between these two pieces of information. The sense with the maximum overlap is the winner sense.

4.2 WSD (Selection restriction Algorithm): Finding the word's Correct Sense-
A knowledge based algorithm is one which attempt of Selection preferences to restrict the number of meanings of a target word occurring in context. Selection restrictions are constraints on semantic type that a word sense imposes on the words with which it connect usually through grammatical relationships in sentences.
Now consider the following example:-

‘खाना’ can be treated as food or to eat, only first sense is available in the context of **”मुझे आम खाना है”** only second sense is available here as **‘आम’** specifies the selection restriction to eat in given a context.

An another example, In Hindi word **‘चैन’** is refers to as rest or as chain for neck, only first sense is available in the context of **“मै आज अपना काम चैन से कर रहा हूँ“,**only second sense is available here as **‘काम**’specifies the selection restriction to rest in a context. The determination of the semantic appropriateness of the association provided by a word to word and word to class relation is the way to learn selection restriction. The elementary measures of this kind are frequent count. P1and P2 is combining of words and syntactic relation R. the number of detail(R, p1, p2) in a corpus of parsed text. Count (R1, p1, p2).the other estimation of the semantic appropriates of a word to word relation is the depending on probability is the conditional probability of word p1 given to the word p2.

R: P (p1/p2, R) =count (p1, p2, R)/count (p2, R)

Considerable techniques has been devised for measure of selection association. The above approaches performance large corpora and model the selection restriction of predicates by combining checked, frequencies with knowledge about the semantic classes of their arguments. The disambiguation is performed with different means based on the strength of selection restriction towards a certain conceptual class. The selection restriction approach to disambiguation has many requirements to be very useful in large scale practical application using with wordnet and it have been developed part of speech (POS) tagger. These systems are designed to make minimal assumption about what information will be available from the processes. The knowledge based approach uses of external lexical resources like dictionary or thesauri. In knowledge based approach, system is trained to perform the task of word sense disambiguation.

V. RESULT

Synset format: The word ‘फूल’

ID: 124(a unique number identifying a synset).
CATEGORY: NOUN (POS category of the words).

CONCEPT: माली ने बगींचे में एक फूल का पेड़ लगाया (The part of the gloss that gives a brief summary of what the synset represents).

EXAMPLE: **“**यह पुष्प की माला भगवान के लिए बनाई गई है**”** (one or more example of the word in the synset used in context).

SYNSET: पुष्प, सुमन, प्रसून**,** कुसुम, (The set of synonymous word).

VI. EVALUATION

In our paper, the evaluation of WSD, We developed the WSD tool followed by Hindi wordnet. We used a small corpus with word occurrences and collocations, this approach used to evaluate word sense disambiguation are precision and recall. Precision is determine as the proportion of correctly classified instances of those classified or it is the percentage of words that are tagged correctly, out of the words addressed by our system, while recall is the proportion of correctly classified instances of total instances or it is the percentage of words that are tagged correctly, out of all words in the test set. Hence, the value of recall is forevermore less than that of precision unless all instances are sense tagged by our system. First of all, we have Consider a test set of 100 words and suppose that 85 words are experimental by the system and out of these 50 words are correctly disambiguated. Then the precision and recall can be calculated as:
Precision=50/85
Recall=50/100

VII. CONCLUSION

In this paper, we focused on Hindi language. The main goal of this paper is to give aim the people working in the field of NLP, who want to study about WSD. According to the area to which major some words in context are sense tagged, WSD tasks fall into two types:

(1)Firstly, tag all considerable words (nouns, verbs, adjectives and adverbs),
(2) Tag some considerable words (usually nouns or verbs)
In this paper, we describe a knowledge based approach using selection restriction for Hindi language. Our methods can be improved by parts of speech (POS) and Hindi wordnet .Manually; we have added these links in the Wordnet database available in MySQL format for some words. This method gives a multiple occurrences for the word in given a context, if a word occurs multiple times in different senses in the same text, it is high

likely that our methods would assign the same synset or synonyms to all its occurrences, for example the word ‘’occurs in the text with the meaning ‘धन’as well as ‘दौलत’ but the synonyms assigned to all occurrences of ‘धन’is{रुपया,पैसा,मूल्धन},since The wordnet relation system applied at training time and frequency data (no. of tagged senses)applied at runtime. The Hindi wordnet depend only the lexical data files distributed with wordnet not on any code. The accuracy of WSD highly depends on the part of speech (POS) tagger module. The efficiency of our work is limited due to the fact that, it can’t tag some words correctly with POS tagger.
For example consider the sentence:

Input: राम पढ़ रहा है (raam pdh rahaa haai)
raam read <verb string of continuity>
Output: राम_NNP पढ़ _VM रहा_VAUX हैVAUX

Here, we can easily notice that the word ‘पढ़’has been incorrectly tagged as VB (verb). Since the POS tagger plays an important role in the WSD. We need to improve the accuracy of the POS tagger in order to disambiguate a word correctly.

ACKNOWLEDGEMENTS

In our paper, firstly, we would like to thank: Mr. Nisheeth Joshi for sharing this paper. And Dr. Paul WO for sharing wall street journal corpus; Dr. Christopher Ting for installing and downloading wordnet and the undergraduate from Banasthali Vidyapith for preparing the sense-tagged corpus; and Prof. Mohanan for his fully support of the sense tagged.

REFERENCES

[1] Budanitsky and G. Hirst. 2006. Evaluating WordNet based Measures of Lexical Semantic Relatedness. Compututational Linguistics, 32:13–47

[2] Yarowsky. 1995. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In Proceedings
Of the 33rd annual meeting on Association for Computational Linguistics, ACL ’95, pages 189–196, Stroudsburg, PA, USA. Association for Computational Linguistics.

[3] Fellbaum Christiane, (ed.) WordNet: an Electronic Lexical Database, Cambridge, MIT press (1998).

[4] Kleinberg, M. Jon: Authoritative sources in a hyperlink environment. Proc. of ACM-SIAM Symposium on Discrete Algorithms (1998).

[5] Agirre Eneko and Rigau: Word sense disambiguation using conceptual density (1996).

[6] Stevenson and Wilks, Mark Stevenson, and Yogic, Wilks: The interaction of knowledge sources in word sense disambiguation. Computational Linguistics, 27(3):321–349, (2001).

[7] Stevenson, Mark Stevenson: Word Sense Disambiguation: The Case for Combining Knowledge Sources, CSLI Publications, Stanford, CA (2003).

[8] Yarowsky, David Yarowsky: Unsupervised word sense disambiguation rivaling supervised methods, Proceedings of the ACL. (1995).

[9] Navigli and M. Lapata. 2010. An Experimental Study of Graph Connectivity for Unsupervised Word Sense Disambiguation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 32:678–692.

[10] R. Navigli. 2007. Structural Semantic Interconnections: a Knowledge-Based WSD Algorithm, its Evaluation and Applications. Ph.D. thesis, University of Rome”La Sapienza.