Word Sense Disambiguation Using Selectional Restriction

Prity Bala

Apaji Institute, Banasthali Vidhyapith Newai,Rajesthan,India

***Abstract*-** Word sense disambiguation (WSD) is still an open research area in natural language processing and computational linguistics. It is from both theoretical and practical point of view. Here, the problem is to find the sense for word in given a context, It is a technique of natural language processing(NLP) ,which requires queries and documents in NLP or texts from Machine Translation (MT). MT is an automated system which involves Marathi, Urdu, Bengali, Punjabi, Hindi, and English etc. Most of the work has been completed in English, and now the focus has been shifted to other languages. The applications of WSD are disambiguation of content in information retrieval (IR), machine translation (MT), speech processing, lexicography, and text processing. In this paper, we have used knowledge based approach along with selectional restriction. It is used to block the formation of component word meanings representation that contains selectional restriction. We have developed a WSD tool using 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. In this paper we shall discuss the implementation of our tool and its evaluation.

***Index Terms*-** Word sense disambiguation, knowledge based approach using Selectional Restriction, Hindi wordnet, etc.

I. Introduction

A.Word Sense Disambiguation

Word sense disambiguation is the problem to find the sense of a word in natural language context, where the word have multiple 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 determined by other neighboring words. This is called as local context or sentential context. This task needs lot of words as well as world knowledge. A WSD is the process of identifying the sense of the word.

Example: गौतम बुद्ध ने गया में ज्ञान प्राप्त किया था

or राम स्कूल गया

Here, we can easily see that ,In first sentence the word' गया' refers to 'a place name' in the former sentences, whereas the word 'गया' refers to a 'verb'

Words do not have well-defined boundaries between their senses, and our task is to determine, which meaning of word is indented in a given text. This is one of the first problems that is encountered by any natural language processing system which is refered to as lexical ambiguity. WSD is a research area in NLP, which is very useful nowdays. It is the technique of natural language processing (NLP).It can be represented by tasks, performance, knowledge source, computational complexity, assumptions and application for WSD algorithms.WSD involve more word knowledge or common sense, which identifies Dictionary or thesauri. It is also helpful 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.

II. HISTORY

The problem of WSD was first formulated as a computational task during the early days of machine translation in the 1940s. In 1949 Warren Weaver, in his famous 1949 memorandum on machine translation, first introduced the problem in a computational context. Early researchers understood well the significance and difficulty of WSD in NLP. In 1960, Bar-Hillel, used the example to argue that WSD could not be solved by "electronic computer". In 1970s WSD was a subtask of semantic interpretation systems or model developed within the field of artificial intelligence, but since WSD systems were largely rule-based and hand-coded. By the 1980s large lexical resources, such as the Oxford Advanced Learner's Dictionary of Current English (OALD), became available: hand-coding was replaced with knowledge based approach automatically extracted from these resources, but word sense disambiguation remained still knowledge-based or dictionary-based. In the 1990s the statistical linguistics revolution swept through computational linguistics, and WSD became a paradigm problem on which to apply supervised machine learning techniques. The 2000s saw supervised techniques to reach a plateau in accuracy, and for this reason attention has been shifted to coarser-grained senses, domain adaptation, semi-supervised and unsupervised corpus-based systems, combinations of different methods or approachs, and the return of knowledge-based systems via graph-based methods. Still, supervised method continue to perform best hybrid Systems, minimizing or eliminating use of sense tagged text by taking the advantage of the Websystem.

III. WORDNET CONCEPT

A.Wordnet

Wordnet is a network of words linked by lexical and semantic relationship. Wordnets for Hindi and Marathi being built at IIT Bombay are amongst the first IL wordnets. Wordnet is an electronic large lexical database of English, and it is a combination of dictionary and thesaurus which is being 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 describes 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 antonymy) and the synsets (meaning to meaning or semantic relations e.g. hyponymy/hypernymy relation, meronymy relation). Wordnet has four types of part of speech (POS), such as noun, verb, adverb, adjective. POS tagging is the process of identifying lexical category of a word in a sentence on the basis of its context.

B. Synset

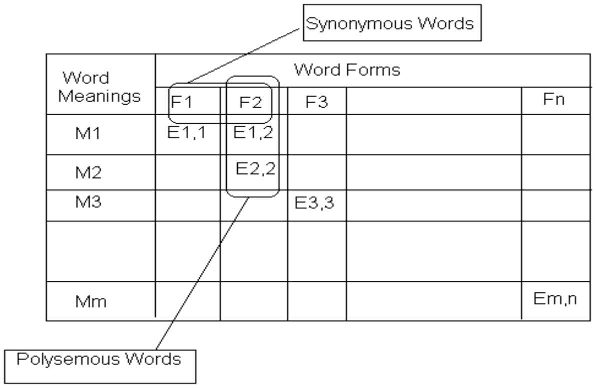
- The smallest unit in wordnet.
- A synonym set.
- Represent a specific meaning of a word.

Synsets are connected to one another through semantic and lexical relations. Each word meaning can be represented by a set of word-forms.It is called as synonym set or synset. Synsets are made by content words such as noun, verb, adjective, and adverb.

C. Lexical Matrix

The lexical matrix is a part of the language system.It refers the link between word form and the word 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 columns represent polysemy.

Table 1. Lexical Matrix



Word Meanings	Word Forms				
	F1	F2	F3		Fn
M1	E1,1	E1,2			
M2		E2,2			
M3			E3,3		
Mm					Em,n

For example the word 'कमल' of synset like {कमल,पंकज,सरोज,नीरज}gives the meanings'फूल' (पानी में होने वालेएक पौधे का फूल जो बहुत सुन्दर होता है) belongs to a synset, whose members from a row in the matrix, and the row numbers gives a ID to the synset.'कमल' has different meanings,(रक्त में पित्त वर्दंक के जमा हो जाने से उत्पन्न एक रोग जिससे शरीर व आँखे पीले पड़ जाते है) which comes in the column by the word.

IV. SEMANTIC RELATIONS IN WORDNET

The Hindi wordnet is inspired by the English wordnet, semantic relation use in structuring lexical data. They have been extensively used in wordnet and evaluated also, and they are mainly used to structure the lexicon such as, and the semantic relations is following below.

Types of semantic relations (It is based on POS):

- Hypernymy (kind-of):'घर ' is hypernym of 'गृह '
- Hyponymy (kind-of):'गृह ' is a hyponym of 'घर'
- Holonymy (part-of):'आवास' is a holonym of 'निवास'
- Meronymy (part-of):'बरामदा' is a meronym of 'आगंन'

For example, we have the synset {घर, गृह}. The hypernymy relation (Is-A) of it links to {आवास,निवास}. Its meronymy relation (Has-A) links to {आँगन, बरामदा} and hyponymy relation to {सराय} and {झोपड़ा}.

V. LITERATURE SURVEY

In the last 15 yaers, the NLP community has an increasing interest in machine learning based approaches for automated classification of word senses. This is evident from the number of supervised WSD approaches that have spawned. Today, the supervised approaches for WSD are the largest number of algorithms, used for disambiguation, Robert R.Korfhage [11]. Supervised WSD uses machine learning techniques on a sense data set to classify the senses of the words Fellbaum Christiane [3], there are a number of classifiers also called word senses that assign an appropriate sense to an instance of a single word. The supervised algorithms thus perform target-word WSD. Any algorithm uses certain features associated with a sense for training.Supervised algorithms trained a model based on the corpus provided to it. This corpus needs to be manually annotated, and the size of the corpus needs to be large enough in order to train a generalized model. Semi-supervised, also known as minimally supervised algorithms and it make some assumptions about the language and discourse in order to minimize these restrictions, Kieinberg.M.Jon [4]. The common threads of operation of these algorithms are these assumptions and the seeds used by them for word sense disambiguation purposes. However, these are fundamental overlap based algorithms which suffer from overlap sparsity, dictionary, thesauri, definitions being generally small in length, Yarowsky [8]. Supervised learning algorithms for WSD are mostly word specific classifiers.The requirement of a large training corpus renders these algorithms unsuitable for resource scarce languages, Fellbaum Christiane [3].

There are three approaches for WSD:

A.Supervised Approach

Supervised is based on a labeled training set and corpus, and it is a learning system, has a training set of featured-encoded inputs and their, appropriate sense label or category, because they can cope with the high dimensionally of the feature spaces, Mark Stevenson[7], basically, supervised approaches applied to the problem of WSD, and used to machine learning techniques for classifier from senses,Yarowsky[8].

B.Unsupervised Learning Approach

In approach, Stevenson, Mark Stevenson [7], unsupervised approach based on unlabeled corpora, and it is learning system, has a training set of feature encoded inputs but not their appropriate sense label or category. it only use the information available in raw text, do not use outside knowledge sources or manual annotations, unsupervised reduces WSD to the problem of finding the targeted words that occurs in the most similar contexts and placing them in the cluster, Agirre Eneko,and Rigau[14].

C.Knowledge Based Approach

This approach based on wordnet, and knowledge based approach to word sense disambiguation taken place, when experimental are conducted on extremely limited domains, Robert R.Korfhage[11], Here the knowledge resources are dictionary, thesauri, collocation, and ontology etc.In knowledge based approach to disambiguate ,we will determines to the target word, along with a context, Kleinberg.M.Jon[7].

VI. PROPOSED WORK

In our thesis, we are using knowledge based approach using selection restriction and developed a WSD tool with Hindi wordnet. Our system currently deals with POS tagger such as noun, verb, adjective, adverb.To given corpora to assign correct sense to the words.This is sense tagging, needs word sense disambiguation (WSD). It is highly important for Question Answering, Machine Translation, Text Mining tasks. Work is depend on to including words of other part of speech. We have taken the database of text files saved from Hindi Wordnet. It prepared by IIT, Bombay but in future, In directly, the database for Hindi language's WSD can use the database prepared for Hindi Wordnet.

Selection restriction s'Algorithim

Selectional preference used here is that proposed by Resnik, combining statistical linguistics and knowledge-based approaches. The basis of the approach is a probabilistic model capturing the co-occurrence and collocation behavior of predicates and conceptual classes in the taxonomy.

The prior distribution PrR(c) captures the probability of a word-class occurring as the argument in predicate-argument relation R, regardless of the identity of the predicate. For example, given the verb relationship, the prior probability for tends to be significantly higher than the prior probability for (insect). However, once the identity of the predicate is taken into account, the probabilities can change, if the verb is buzz, then the probability for (insect),it can be expected to be higher than its prior, and (person) will likely be lower. In probabilistic terms, it is the difference between this conditional or posterior distribution and the prior distribution that determines selectional restriction. Information theory provides an appropriate way to quantify the difference between the prior and posterior distributions, in the way of relative entropy. The model defines the selectional preference strength of a predicate as:

$$\mathrm{Sr}(p) = D(\mathrm{pr}(c \mid p) \| \mathrm{Pr}(c))$$
$$= \sum \mathrm{Pr}(c \mid p)\log \mathrm{Pr}(c \mid p) / \mathrm{Pr}(c \mid p) \quad (1)$$

Intuitively, Sr(p) measures the information, in bits, p predicate provides about the conceptual class of its argument. The better Pr(c) approximates Pr(c│p), the leas influence p is having on its argument, and therefore the less strong its selectional restriction or preference. Given this definition, a natural way to characterize the "semantic fit of a particular class as the argument to a predicate is by its relative contribution to the overall selectional restriction strength. In particular, word to classes that fit very well can be expected to have higher posterior probabilities, compared to their priors, as is the case for (insect) in. Formally, selection association is defined as given below:-

$$\mathrm{Ar}(p, c) = 1 / \mathrm{Sr}(p).\mathrm{Pr}(c \mid p)\log\mathrm{Pr}(c \mid p) / \mathrm{Pr}(c) \quad (1)$$

This model of selectional restriction or preferences has turned to make reasonable predictions about human judgments of argument plausibility obtained by psycholinguistic methods (Resnik, 1993a). The selectional association has also been used recently to explore apparent cases of syntactic or lexical system optionality (Paola Merlo, personal communication).

Selectional restriction depends on knowledge based algorithm. A knowledge based algorithm is one which depends upon the selectional restriction to restrict the number of meaning of a target word in given a context, and it is to determine word to word relation. The problem of lexical and syntactic ambiguity encountered in NLP. It is also called as selectional preferences. A selectional preferences or restriction are constraints on semantic type that a sense imposes on the words, which it combines usually through grammatical relation in sentences.

Now consider the following example:

Here, we have taken the word in Hindi as 'खाना', which consists of two senses, as the first 'खाना' and other 'भोजन'.In the given context, such as "मुझे आम खाना है".so, here the word 'खाना' is the same sense of the word 'खाना'. Thus, this technique shows any sense of the word depends on its context.What's that word makes sense. Namely the 'ordinary' word has been seen as a reference to the selection of what is the sense of the word.

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 are trained to perform the task of word sense disambiguation.

In this approach, what is learned to classifier that is used to assign examples to one of a fixed senses.WSD is refers to the knowledge resources to the senses of word in given a sentences. These are some knowledge resources such as dictionary, thesauri, ontology, glossaries, collocations, etc.

Knowledge based approach have a knowledge resources of machine readable dictionary(MRD), and selectional restriction(SR) in front of corpus, for example Wordnet. They may use grammar rules for word sense disambiguation. It is a fundamental component of WSD, and can be provide the data, in which associate sense with words.This is one of the most important knowledge based in natural language processing and has been used for syntactic lexical and word sense disambiguation, and the degree of the selectional restriction for a word combinations from a tagged corpus, based on the multiple regression model.

VII. RESULT

Synset format: The word 'प्रेम'

ID: 121(a unique number identifying a synset)
CATEGORY: NOUN (POS category of the words)
CONCEPT: अपने से छोटों के लिए प्रेम होना चाहिए (The part of the gloss that gives a brief summary of what the synset represents)
EXAMPLE: "चाचा नेहरू को बच्चों से बहुत ही स्नेह हुआ करता था" (one or more example of the word in the synset used in context)
SYNSET: स्नेह, नेह, लगाव, ममता (The set of synonymous word)

VIII. 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; firstly the corpus was tagged by part of speech (POS) tagger. A parsed, sense-tagged corpus determined by Hindi wordnet sense-tagged corpus.

The test for the verb-object relationship was constructed by a selectional restriction model on the corpus. The 100 verbs that select most strongly for their objects were identified, including verbs appearing only once in the corpus, test instances of the form (verb, object, correct sense),were the extracted from the merged test corpus, including all triples where verb was one of the 100 test verbs.

Here, we contained some words like {सोना, आम, गया, चैन, बाली}.These files contained 5463 words, out of which we could disambiguate 5236. The accuracy of our approach was 66.92%, which means that our system disambiguated correctly 3492 out of 5463 words.

IX. CONCLUSION

In this paper, we describe a knowledge based approach using selectional 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.
Here, 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.

REFERENCES

[1] Bhattacharya and Unny: Word sense disambiguation and text similarity measurement using wordnet. (2002) .

[2] Budanitsky and Hirst: Evaluating WorldNet based measures of lexical semantic relatedness.

[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] Dagan and Itai, Ido Dagan and Alon Itai.: Automatic acquisition of constraints for the resolution of anaphoric references and syntactic ambiguities. Proceedings of Coling-90, 3:162–167, (1990).

[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] Eric Brill: Unsupervised learning of disambiguation rules for part of speech tagging. In Natural Language Processing Using Very Large Corpora. Kluwer Academic Press (1997).

[10] Thomas M. Cover and Joy A. Thomas: Elements of Information Theory, Wiley Series in Telecommunications, Wiley, New York (1991).

[11] Robert R. Korfhage: Information Storage and Retrieval, Wiley, New York (1991).

[12] Shari Landes, C laudia Leacock, and Randee I. Tengi: Building semantic concordances, In WorldNet, an Electronic Lexical Database, pages 199–216, MIT Press, Cambridge MA (1998).

[13] Johnson M. Barnard K: Word sense disambiguation with pictures.(2005).

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

[15] A.Deepa Pushpak Bhattacharyya Ganesh Ramakrishnan, B. P. Prithviraj: Soft word sense disambiguation (1997).

[16]] P. Panda P. Bhattacharyya, S. Jha, and D. Narayan: A WorldNet for Hindi. (2001)

[17] Michael Susana: Word sense disambiguation for free-text indexing using a massive semantic network (1993)

AUTHORS

First Author – Prity Bala, Apaji Institute, Banasthali Vidhyapith Newai,Rajesthan,India, meethi.prity@gmail.com