WORD SENSE DISAMBIGUATION (WSD)

Jagdeep Kaur
M.Tech Student
Department of Computer Engineering
Punjabi University
Patiala, India.

Abstract: ***Word Sense Disambiguation (WSD) is a natural classification problem: Given a word and its possible senses, as defined by a dictionary, classify an occurrence of the word in context into one or more of its sense classes. The features of the context (such as neighboring words) provide the evidence for classification. A rich variety of techniques have been researched, from dictionary-based methods that use the knowledge encoded in lexical resources, to supervised machine learning methods in which a classifier is trained for each distinct word on a corpus of manually sense-annotated examples, to completely unsupervised methods that cluster occurrences of words, thereby inducing word senses. Among these, supervised learning approaches have been the most successful algorithms to date. It is motivated by its use in many crucial applications such as Information retrieval, Information extraction, Machine Translation, Part of- Speech tagging, etc. This paper presents knowledge based methods for word sense disambiguation. We describe Lesk algorithm, which uses lexical database WordNet as knowledge base and Walker's algorithm.***

Keywords: ***Word Sense Disambiguation, Lesk algorithm, Walker algorithm, and WordNet.***

I. INTRODUCTION

Word sense disambiguation (WSD) is an open problem of natural language processing[1], which governs the process of analyzing meaning of particular word in given context using corpora, training data set, or lexical databases available and selecting right meaning i.e. right sense and assigning it to the word [2]. In 1940's WSD was developed as discrete field in computational linguistic due to fast research in of machine translation. In 1950's Weaver acknowledged that context is crucial and recognized the basic statistical character of the problem in proposing that statistical semantic studies should be undertaken as a necessary primary step. The automatic disambiguation of word senses has been an interest and concerned since the earliest days of computer treatment of languages in the 1950's. Then identifying work in estimating the degree of ambiguity in texts and bilingual dictionaries and applying simple statistical models. Sense disambiguation is an intermediate task which is not an end in itself, but rather is necessary at one level or another to accomplish most NLP task [8].

a) **Conceptual Model For WSD**

The reasons that WSD is difficult lie in two aspects. First, dictionary-based word sense definitions are ambiguous. Even if trained linguists manually tag the word sense, the inter-agreement is not as high as would be expected. That is, different annotators may assign different senses to the same instance. Second, WSD involves much world knowledge or common sense, which is difficult to verbalize in dictionaries. The conceptual model for WSD is shown in figure 1.

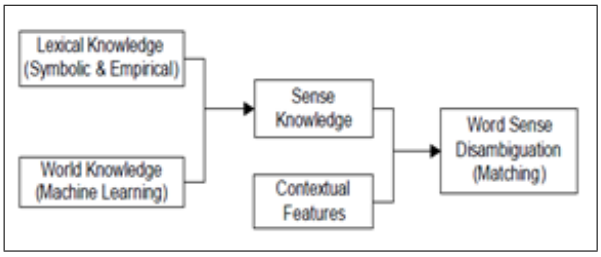

Figure 1: Conceptual model for WSD

Sense knowledge can be represented by a vector, called a sense knowledge vector (sense ID, features), where features can be either symbolic or empirical. Dictionaries provide the definition and partial lexical knowledge for each sense. However, dictionaries include little well-defined world knowledge (or common sense). An alternative is for a program to automatically learn world knowledge from manually sense-tagged examples, called a training corpus. In some cases of learning, contextual features are not directly associated with sense. Thus, we need some intermediate constructs to direct sense ID to features. Such knowledge is defined as "hidden knowledge"[4].

The word to be sense tagged always appears in a context. Context can be represented by a vector, called a context vector (word, features). Thus, we can disambiguate word sense by matching a sense knowledge vector and a context vector [4].

II. APPLICATIONS

Word sense disambiguation a task of removing the ambiguity of word in context, is important for many NLP applications such as:

1. **Information Retrieval**

WSD helps in improving term indexing in information

retrieval word senses improve retrieval performance if the senses are included as index terms. Thus, documents should not be ranked based on words alone, the documents should be ranked based on word senses, or based on a combination of word senses and words. For example: Using different indexes for keyword "Java" as "programming language", as "type of coffee", and as "location" will improve accuracy of an IR system..

2. **Machine Translation**
   WSD is important for Machine translations. It helps in better understanding of source language and generation of sentences in target language. It also affects lexical choice depending upon the usage context.
3. **Speech Processing and Part of Speech tagging**
   Speech recognition i.e., when processing homophones words which are spelled differently but pronounced the same way. For example: "base" and "bass" or "sealing" and "ceiling".
4. **Text Processing**
   Text to Speech translation i.e., when words are pronounced in more than one way depending on their meaning. For example: "lead" can be "in front of" or "type of metal".

III. METHODS

There are three basic approaches for WSD methods [2]:

1. **Supervised disambiguation**
   In supervised disambiguation method, the system is trained with manually created examples of correctly disambiguated words in context.
2. **The dictionary based or knowledge-based**
   Methods treat a dictionary as both the source of the sense inventory as well as a repository of information about words that can be exploited to distinguish their meanings in text. This paper presents WordNet as lexical database or sense inventory to access meanings and other information related to words [4].
3. **Unsupervised disambiguation**
   The unsupervised corpus-based methods of WSD are knowledge-lean, and do not rely on external knowledge sources such as machine readable dictionaries, concept hierarchies, or sense-tagged text [3].

IV. KNOWLEDGE BASED ALGORITHMS

The objective of knowledge-based or dictionary-based WSD is to exploit knowledge resources (such as dictionaries, thesauri, ontologies, collocations, etc) to infer the senses of words in context [6]. Knowledge based approach have a faith on knowledge resources of machine readable dictionaries in form of corpus, WordNet etc. they may use either grammar rules for disambiguation. A huge prominence of computer the large scale dictionaries are made available in form of MRD (machine readable dictionaries) like oxford English dictionary, Longman dictionary of ordinary contemporary English, Roget thesaurus and semantic networks which add more semantic relation like WorldNet, euro WordNet. These are all for English [8]. The main algorithms are described below.

1. **Lesk's Algorithm**
   This method is suggested by the scientist M.Lesk. According to him, a word is disambiguated by comparing the gloss of each of its senses to the glosses of every other word in the phrase. The sense whose gloss shares the largest number of words in common with the glosses of other words is selected as the correct sense. Given a two word context (w1, w2), the senses of the target words whose definitions have the highest overlap (i.e., words in common) are assumed to be the correct ones [11].

   The Lesk algorithm is based on the assumption that words in a given "neighborhood" (section of text) will tend to share a common topic. A simplified version of the Lesk algorithm is to compare the dictionary definition of an ambiguous word with the terms contained in its neighborhood. Versions have been adapted to use WordNet. An implementation might look like this [5]:

   - For every sense of the word being disambiguated one should count the amount of words that are in both neighborhood of that word and in the definition of each sense in a dictionary.
   - The sense that is to be chosen is the sense which has the biggest number of this count [11].

   Formally, given two words w1 and w2, the following score is computed for each pair of word senses S1 ϵ Senses(w1) and S2 ϵ Senses(w2) :

$$ScoreLesk(S1, S2) = |gloss(S1) \cap gloss(S2)| \qquad (1)$$

   Where gloss(Si) is the bag of words in the textual definition of sense Si of wi. The senses which maximize the above formula are assigned to the respective words. However, this requires the calculation of | Senses(w1) | · | Senses(w2) | gloss overlaps. If we extend the algorithm to a context of n words, we need to determine $\prod^n i = 1$ | Senses(wi) | overlaps. Given the exponential number of steps required, a variant of the Lesk algorithm is currently employed which identifies the sense of a word w whose textual definition has the highest overlap with the words in the context of w. Formally, given a target word w, the following score is computed for each sense S of w :

$$scoreLeskVar(S) = |context(w) \cap gloss(S)| \qquad (2)$$

   Where context(w) is the bag of all content words in a context window around the target word w [8].

   Recently, Banerjee and Pedersen [2003] introduced a measure of extended gloss overlap, which expands the glosses of the words being compared to include glosses of concepts that are known to be related through explicit relations in the dictionary (e.g.,hypernymy). The range of relationships used to extend the glosses is a parameter, and can be chosen from any combination of WordNet

Example: Two senses of ash :

Sense	Definition
s1: tree	D1: a tree of the olive family
s2:burned stuff	D2: the solid residue left when combustible material is burned

Table 1: Senses of ash

Disambiguation of ash using Lesk's algorithm :

Scores		Context
s1	s2	
0	1	This cigar burns slowly and creates a stiff ash.
1	0	The ash is one of the last trees to come into leaf.

Table 2: Disambiguation of ash using Lesk's algorithm

relations. For each sense S of a target word w we estimate its score as:

$$score_{ExtLesk}(S) = \sum_{S':S \xrightarrow{rel} S' \text{ or } S \equiv S'} | context(w) \cap gloss(S') |$$

Where context (w) is, as above, the bag of all content words in a context window around the target word w and gloss(S') is the bag of words in the textual definition of a sense S' which is either S itself or related to S through a relation rel. The overlap scoring mechanism is also parameterized and can be adjusted to take into account gloss length (i.e. normalization) or to include function words. The biggest drawback of this algorithm is that, dictionary definitions are often very short and do not have enough words for this algorithm to work well.

2. Walker's Algorithm It is a thesaurus based approach. In the year 1987, walker proposed an algorithm as follows. Considering a thesaurus each word is assigned to one or more subject categories in the thesaurus. There are several subjects are assigned with a word then it is assumed that they correspond to different senses of the word. Black applied walker's approach to five different words and achieved accuracies of 50Here first finds the thesaurus category to which that sense belongs. Then calculate the score for each sense by using the context words. A context will add 1 to the score of the sense if the thesaurus category of the word matches that of the sense [7].

Word	Sense	Thesaurus category
Bass	musical senses	Music
	Fish	animal, insect
Star	space object	Universe
	Celebrity	Entertainer
	Star shaped object	Insignia
Interest	Curiosity	Reasoning
	Advantage	Injustice
	Financial	Debt
	Share	Property

Table 3: Thesaurus Category of word senses

V. LEXICAL DATABASES FOR VARIOUS LANGUAGES

1. **WordNet**
WordNet is a large lexical database of English. WordNet is a machine-readable dictionary developed by George Miller and his colleagues at the Cognitive Science Laboratory at Princeton University. It is an online lexical database designed for use under program control; it provides a more effective combination of traditional lexicographic information and modern computing. WordNet are arranged semantically instead of alphabetically. Synonymous words are grouped together into synonym sets, called synsets.Each such synset represents a single distinct sense or concept. For example, in Wordnet, synset car, auto, automobile, machine, motorcar represents the concept of "4-wheeled motor vehicle; usually propelled by an internal combustion engine" [9]. WordNet is also freely and publicly available for download. WordNet's structure makes it a useful tool for computational linguistics and natural language processing [9].

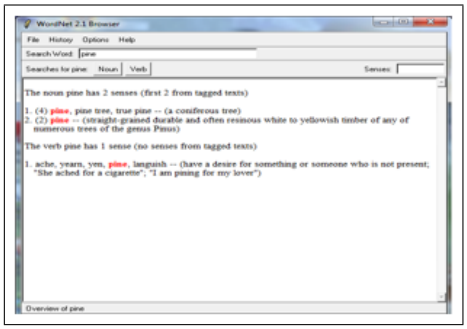

Figure 2: Snapshot of web interface for WordNet 2.1

2. **Hindi WordNet**
A lexical database for Hindi language: 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 [10].

Figure 3: Snapshot of web interface for Hindi WordNet (online available)

3. **Indo WordNet**
   A WordNet of Indian language: Seeing the enormous potential of wordnet, 16 out of 22 official languages of India, have started making their wordnets under the leadership of IIT Bombay. These languages are: Hindi, Marathi, Konkani, Sanskrit, Nepali, Kashimiri, Assamese, Tamil, Malyalam, Telugu, Kannad, Manipuri, Bodo, Bangla, Punjabi and Gujarati. These languages cover the length and breadth of India and are used by about 900 million people. IndoWordnet is a linked structure of wordnets of major Indian languages [10].

Figure 4: Snapshot of web interface for Indo WordNet

VI. CONCLUSION

Word sense disambiguation is a key problem to address in many applications in the areas of Natural Language Processing, Information Retrieval and others. WSD is typically configured as an intermediate task, either as a stand-alone module or properly integrated into an application (thus performing disambiguation implicitly). However, the success of WSD in real-world applications is still to be shown. Application oriented evaluation of WSD is an open research area, although different works and proposals have been published on the topic. We described WordNet that is also freely and publicly available for download. WordNet's structure makes it a useful tool for computational linguistics and natural language processing. In this paper we studied knowledge based algorithms. Knowledge based systems suffer from poor accuracies because of their complete dependence on dictionary defined senses, which don't provide enough clues. But the requirement of large corpus often renders learning algorithms unsuitable for resource poor languages, like Indian languages.

References

[1] Word Sense Disambiguation.

[2] Eneko Agirre Philip Edmonds. Introduction to Word Sense Disambiguation Algorithms and Applications. *Springer e-ISBN 978-1-4020-4809-2*, 2007.

[3] ACM Navigli. Word sense disambiguation: A survey. *Comput. Surv*, 41(2), 2009.

[4] Xiaohua Zhou, and Hyoil Han. Survey of Word Sense Disambiguation Approaches.

[5] Michael Lesk. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. *SIGDOC 86: Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26, 1986.

[6] Samit Kumar. Word Sense Disambiguation Using Association Rules: A Survey. *International Journal of Computer Technology and Electronics Engineering*, 2(2).

[7] Walker D. and Amsler R. . The Use of Machine Readable Dictionaries in Sublanguage Analysis in Analyzing Language in Restricted Domains, Grishman and Kittredge (eds). *LEA Press*, pages 69–83, 1986.

[8] J. Sreedhar. Word Sense Disambiguation: An Empirical Survey. *International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307*, 2(2), 2012.

[9] George A. Miller. WordNet: An online lexical database. *Comm. ACM*, 38(11):39–41, 1993.

[10] Pushpak Bhattacharyya. IndoWordnet. *Department of Computer Science and Engineering Indian Institute of Technology Bombay*.

[11] Satanjeev Banerjee and Ted Pedersen. An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet. *Lecture Notes In Computer Science ISBN 3-540-43219-1*, 2276:136–145, 2002.

