WordNet and its Applications

Project Report

Submitted in partial fulfillment of the requirements for the degree of

Bachelor of Technology

by

DIPTESH KANOJIA

Roll No: 0944910004

under the guidance of

Prof. (Dr.) Pushpak Bhattacharyya

Department of Computer Science and Engineering
Indian Institute of Technology, Bombay

August 2011

Acknowledgement

I would like to express my profound gratitude to Prof. (Dr.) Pushpak Bhattacharyya for his invaluable guidance and support during the course of the project. I am highly indebted to him for his kind help. I thank my senior research fellows, Salil Joshi and Arindam Chatterjee for the minutely detailed discussions we had on WordNet & Word Sense Disambiguation. I thank Mitesh Khapra and other fellow researchers for their support and suggestions. I would also like to thank Hindi WordNet member Mrs. Laxmi Kashyap for help with linguistics related problems. This report could not have been complete without all of them.

I would also like to thank Mr. Sameer Awasthi, (HOD), Computer Science and Engineering Department, K.S. Jain Institute of Engineering and Technology, for his constant moral support, help, and guidance.

Diptesh Kanojia
(0944910004)

Abstract

A WordNet is a rich repository of knowledge of words and plays a key role in many Information Retrieval and Language Processing applications. The English WordNet was the first of its kind and became central to most of the work in Natural Language Processing. A similar resource for Hindi is being developed at CFILT[1], IIT Bombay.

Word Sense Disambiguation [WSD] has become the heart of the Natural Language Processing problems. With increasing demand for cross lingual information retrieval and overall growth of machine learning applications, the task of identification of senses of words has become critical. Over past half century, many different approaches have been tried out throughout the world. Various structures have been introduced to assist these approaches, and the evaluation strategy is varied.

This report describes the study of WordNet and its applications and our work on Error Analysis of the IWSD algorithm. Our analysis is divided in various parts, *viz.* POS based, Ontology based, Ablation test on IWSD parameters, aim being the importance of contextual part of the scoring function of IWSD algorithm. A detailed error analysis on *context based* and *context agnostic* tagging has also been done.

[1] Resource Center for Indian Language Technology Solutions

Table of Contents

Chapter 1

Introduction

A WordNet is a large lexical resource for a particular language. It can be described as a computerized dictionary of synonyms, thesaurus, lexical database and taxonomy of concepts. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (*synsets*), each expressing a distinct concept (a detailed description regarding a synset is given in Section 2.2). Synsets are interlinked by means of conceptual semantic and lexical relations. The resulting network of meaningfully related words and concepts makes it a useful resource for computational linguistics and language processing. In short, *a WordNet is a machine readable lexical database organized by semantics*. It brings together various semantic and lexical relations between words. It organizes the lexical information in terms of word meanings and can be termed as a lexicon based on psycholinguistic principles [1].

1.1 English WordNet

The first WordNet that was developed was for English language. It was developed at the Princeton University under the direction of psychology professor George A. Miller. This WordNet distinguished between nouns, verbs, adjectives and adverbs, because they follow different grammatical rules. Every *synset* contains a group of synonymous words or collocations. Different senses of a word are in different synsets. The meaning of a synset is elaborated using short definition (followed by an example) called as *gloss*. Following are examples of English synsets, as represented in the Princeton WordNet:

ID: 779881 : {probe, examine}:question or examine thoroughly and closely

ID: 1281142 : {earth} :connect to the earth: ``earth the circuit''

ID: 2681617 : {weather, endure, brave, brave_out}:face or endure with courage: ``She braved the elements''

These synsets are further connected to other synsets and form a network through a number of semantic relations like Hypernymy, Hyponymy, Holonymy, Meronymy *etc.*, depending on the type of words in the synset under consideration [1]. Currently the English WordNet (version 3.0) has around 117,659 synsets and a total of 206,941 unique word-sense pairs[1].

1.2 Hindi WordNet

The Hindi WordNet is being developed at CFILT, IIT Bombay. The design of Hindi WordNet is inspired from the English WordNet. Here the synset structure is maintained similar to that of English synset but the relations between the synsets include Troponymy, Antonymy, Gradation and Causative apart from the Hypernymy, Hyponymy, Holonymy and Meronymy relations [3]. Currently the Hindi WordNet (version 1.2) has around 34,209 synsets and 83,310 unique Hindi words[2]. Following are examples of Hindi synsets, as represented in the Hindi WordNet:

ID 121: {**स्नेह**, **नेह**, **ममता**} : अपने से छोटों, हमजोलियों आदि के प्रति हृदय में उठने वाला प्रेम

ID 24 : {**शेरनी**, **मादा बाघ**, **मादा व्याघ्र**, **बाघिन**, **व्याघ्री**} : मादा शेर *"शेरनी शेर से अधिक खूँखार होती है /*

1.3 Word Sense Disambiguation (WSD)

Word Sense Disambiguation (WSD) is one of the toughest problems today, not only in Natural Language Processing (NLP) but also in Artificial Intelligence (AI). The problem of WSD dates back to 1950s. It basically refers to the automatic disambiguation of word senses and it has been an interest and concern since the earliest days of treating languages computationally. According to experts, sense disambiguation is an "intermediate task", which is just a stepping stone to most Natural Language Processing tasks. Before moving into further detail, let us find out what exactly WSD is, and the various ways of looking at the problem, along with the various forms of the problem and *why the problem is favorite among NLP researchers*.

1.3.1 WSD Basics

Initially it is important to know the basic definitions of the WSD problem and its different types. The following section deals with the evolution of WSD as a problem.

1.3.2 Definitions

In light of the motivation behind WSD, let us have a look at the colloquial and formal definitions of WSD.

[1]as of May, 2010, according to http://wordnet.princeton.edu/wordnet/
[2]as of May, 2010, according to http://www.cfilt.iitb.ac.in/wordnet/webhwn/wn.php

1.3.2.1 General definition

WSD is the ability to identify the meaning of words in context in a computational manner. Given a set of words (e.g., a sentence or a bag of words), a technique is applied which makes use of one or more sources of knowledge to associate the most appropriate senses with words in context.

1.3.2.2 Formal definition

WSD is the task of assigning the appropriate sense(s) to all or some of the words in T, that is, to identify a mapping A from words to senses, such that $A(i) \in Senses_D(w_i)$, where $Senses_D(w_i)$ is the set of senses encoded in a dictionary D for word w_i, and A(i) is that subset of the senses of w_i which are appropriate in the context T. The mapping A can assign more than one sense to each word $w_i \in T$,although typically only the most appropriate sense is selected, that is, $|A(i)| = 1$. Where, T is a sequence of words $(w_1, w2, \ldots, w_n)$.

1.4 Variants of WSD

The problem of Word Sense Disambiguation can be broadly classified into two categories.

1.4.1 Lexical Sample [Targeted WSD]

Here, the system is required to disambiguate a restricted set of target words usually occurring one per sentence. It employs supervised techniques using hand-labeled instances as training set and then an unlabeled test set.

1.4.2 All-words WSD

Here, the system is expected to disambiguate all open-class words in a text (i.e., nouns, verbs, adjectives, and adverbs) therefore, they are termed as wide coverage systems to disambiguate all open-class words. Typically, such a system suffers from Data sparseness problem, as large knowledge sources are not available.

1.5 WSD: Heart of the NLP Problems

As mentioned earlier, due its applicability and hardness, WSD has become an area of keen interest and challenge for NLP researchers. Word Sense Disambiguation is truly the heart of all NLP problems. This very fact is illustrated in the figure below:

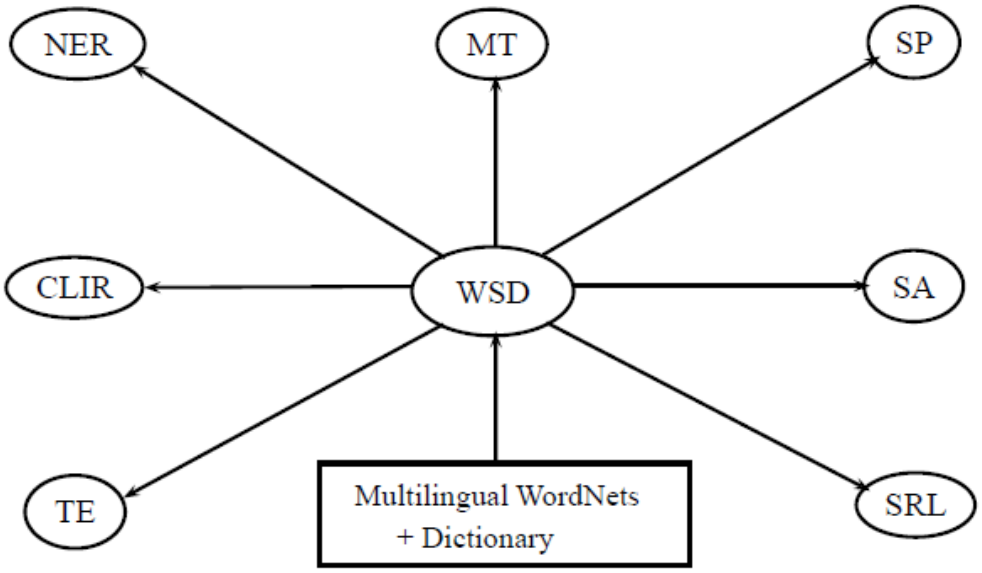

Figure 1.1: WSD - Heart of the NLP problems

- **SRL:** Semantic Role Labeling
- **TE:** Text Entailment
- **CLIR:** Cross Lingual Information Retrieval
- **NER:** Named Entity Recognition
- **MT:** Machine Translation
- **SP:** Shallow Parsing
- **SA:** Sentiment Analysis
- **WSD:** Word Sense Disambiguation

1.6 Report Outline

This report is organized as follows: Chapter 2 and Chapter 3 give a detailed overview of the English and Hindi WordNet respectively. A brief overview of the Multi Lingual Dictionary is given in Chapter 4, It also encompasses the study of a brief summary of Lesk's Algorithm and Extended Lesk's Algorithm. In the following chapters, examples are provided (whenever possible) in the context of Hindi and English WordNet. Chapter 5 contains a detailed explanation of the IWSD algorithm and its scoring function. The Error analysis performed on IWSD is described in detail in Chapter 6, followed by a detailed Discussion on the it in Chapter 7. The report ends with a chapter describing the conclusions and Future work.

Chapter 2

English WordNet

WordNet is a large lexical database of English, developed under the direction of George A. Miller. It is an on-line lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory. WordNet is a dictionary based on psycholinguistic principles. From a study conducted on a group of English speaking subjects as a part of psycholinguistic research, it was learnt that the syntactic categories are stored differently in the human brain. This formed the basis for the organization of WordNet into four different categories: *nouns*, *verbs*, *adjectives* and *adverbs*. All these categories are organized into synonym sets, each representing one underlying lexical concept [1]. In order to distinguish between a standard thesaurus from the WordNet, it is necessary to understand its design.

2.1 Lexical Matrix

In lexical semantics, a *word* is an association between a concept, which plays a semantic role and an utterance, which plays a syntactic role. This gives rise to many problems, one of which is the nature and organization of concepts that the words can express. The WordNet is a result of an attempt to answer this problem. In order to clear the ambiguity of what the word *word* is referring to - semantic role or syntactic role - two new terms were coined. The term *word form* is used to refer to the syntactic role and the term *word meaning* is used to refer to the semantic role. Now, the lexical semantics can be defined as a *mapping between a word form and a word meaning* [1]. This mapping is represented in the form of a matrix, called the Lexical Matrix, with the *word forms* along the columns and *word meanings* along the rows (refer Figure 2.1).

An entry in the Lexical Matrix implies that the *word form* in that column can be used to express the *word meaning* corresponding to that row. In Figure 2.1, the entry $E_{1,1}$ implies that the word form F_1 can be used to express the word meaning M_1. If there are two or more entries in a given column, then the word form corresponding to that column.

Word Meanings	Word Forms				
	F_1	F_2	F_3	...	F_n
M_1	$E_{1,1}$	$E_{1,2}$			
M_2		$E_{2,2}$			
M_3			$E_{3,3}$		
$\vdots$				$\ddots$	
M_m					$E_{m,n}$

Figure 2.1: Lexical Matrix [1]

is polysemous. If there are two or more entries in a given row, then the word forms are synonyms of each other [1].

2.2 Synset

As mentioned earlier, WordNet is a dictionary of word meanings/concepts. Hence there must be a standard representation of the concepts in order to simulate a lexical matrix on a machine. This representation is called a Synset [1]. It is assumed that, the person reading the synset already has the knowledge of the concept and the synset is merely a representation of the concept. For example, consider the word *chair*. This word could either mean *a seat for the person* or *a person who presides at a meeting*. To distinguish the two senses of the word *chair*, the word forms which also represent the same word meaning can be used along with the word *chair*. Hence the representation *{chair}* and *{chair, chairman, chairperson}* helps us to distinguish between the two different sense of *chair*. However, sometimes if the appropriate synonym is not available, the polysemy can be resolved with the help of a small *gloss* (a definition which explains the concept) and an example. This combination of the synonymous words and the gloss is called as a *Synset*. Often the word syntactic category or the *part-of-speech* information is also included in the synset. The following is an example of a synset from the Princeton English WordNet:

Synset ID : 10312324
Synonyms : president, chairman, chairwoman, chair, chairperson
Gloss : the officer who presides at the meetings of an organization
Example : "address your remarks to the chairperson"
POS : Noun

The core ideas behind the creation of the synset are: *minimality*, *coverage* and *replaceability*. Minimality means that the set of synonyms which uniquely identify the concept which the synset represents should be minimal. For example, excluding the word *bank_building* from the set *{bank, bank_building}* (bearing synset ID *2761218*) makes the word *bank* ambiguous. Coverage means that the synset should contain all the words that can represent the given concept. Replaceability means that the words in the synonymy set should be mutually replaceable in a given context.

2.3 Lexical & Semantic Relations

Another important aspect of the WordNet is that, it is organized by semantic relations. A semantic relation is a relation between word meanings. As meanings are represented through a synset, these relations are pointers between synsets. If two synsets S_1 and S_2 are related through a relation *R*, then the individual word forms belonging to these synsets are also related by the relation *R* [1]. There are various relations between the synsets in a WordNet. Some of the important relations are *Synonymy*, *Antonymy*, *Hyponymy/Hypernymy*, *Meronymy/Holonymy*.

2.3.1 Synonymy

Synonymy is the most important relation between two word forms. Two word forms are synonymous if the substitution of one for the other does not alter the meaning of a sentence in which the substitution is made [1]. This forms the basis for the creation of synsets in the WordNet. The replace ability of synonymous words also makes it necessary to partition the WordNet into various syntactic categories like *nouns*, *verbs*, *adjectives* and *adverbs*. This is because only words in different categories cannot be synonyms and hence are not interchangeable. Synonymy is a symmetric relation. If W_1 is a synonym of W_2 then W_2 is also a synonym of W_1.

2.3.2 Antonymy

Antonymy is an important and a difficult lexical relation. Antonym of a word *w* is sometimes *not-w*, but not always. For example, *rich* and *poor* are antonyms, but to say that someone is *not rich* does not necessarily imply that he is *poor* [1]. It is important to see that Antonymy is a lexical relation between word forms and not a semantic relation between word meanings. For example, *{rise, ascend}* and *{fall, descend}* may be conceptual opposites, but they are not antonyms. This is clear when one observes that though *rise-fall* and *ascend-descend* seem to be antonyms, *rise-descend* and *fall-ascend* are not antonyms.

2.3.3 Hyponymy/Hypernymy

Hyponymy and Hypernymy are two important semantic relations between word meanings in the WordNet. A synset S_1 is said to be a hyponym of another synset S_2 if the concept represented by S_1 *is a* kind of the concept represented by S_2 [1]. The opposite relation is called Hypernymy. Hence if S_1 is a hyponym of S_2 then S_2 is a hypernym of S_1. Both these relations are asymmetric and transitive in nature. For example, *oak* is a kind of *tree*. Hence the synset representing *oak* is a hyponym of synset representing *tree*.

2.3.4 Meronymy/Holonymy

Meronymy and Holonymy are also semantic relations. A synset S_1 is said to be a meronym of another synset S_2 if the concept represented by S_1 is a *part of* the concept represented by S_2 [1]. The opposite relation is called Holonymy. Even these relations are asymmetric and transitive in nature. For example, *cooker* is a part of *kitchen*. Hence synset representing *cooker* is a meronym of synset representing *kitchen*.

2.4 Graph Structure

Due to the transitive nature of the semantic relations Hyponymy, Hypernymy, Meronymy and Holonymy, huge graph like structure is formed, with nodes as the concepts/synsets. The nouns are stored in the WordNet in such a way that a lexical hierarchy is created using the semantic relations among various synsets [1]. Hence all hyponyms of a given synset inherit all its features. As we move down the noun hierarchy more and more features are added to the synsets. For example, consider the concept of a *{vehicle}* which has *{tyre, wheel}* as its meronymy and *{bicycle}* as its hyponym. From the WordNet hierarchy it automatically follows that *bicycle has tyre(s)/wheel(s)*. Similarly, *adjectives*, *verbs* and *adverbs* also form a graph like structure in the English WordNet. Figure 2.2 from [1] shows a network representation of three semantic relations among an illustrative variety of lexical concepts.

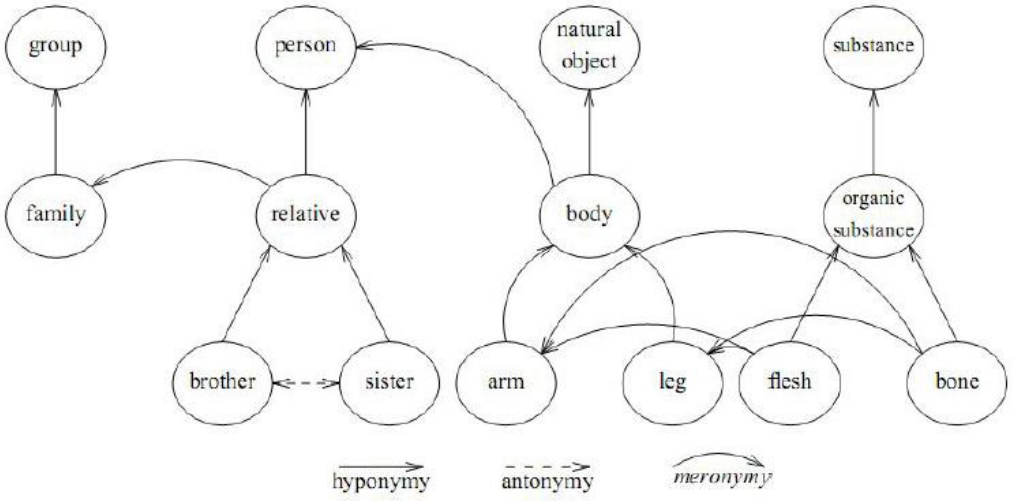

Figure 2.2: Network representation of semantic relations among lexical concepts [1]

Chapter 3

Hindi WordNet

The Hindi WordNet, inspired from the English WordNet, is being developed at CFILT, IIT Bombay. In the Hindi WordNet, the words are grouped together according to their similarity of meanings. For each word there is a synonym set in the Hindi WordNet representing one lexical concept. The Hindi WordNet deals with the content words or open class category of words. Thus, the Hindi WordNet also contains the four category of words - *nouns*, *verbs*, *adjectives* and *adverbs*.

3.1 Hindi Synset

The Hindi synset is also similar to English synset. Each entry in the Hindi WordNet consists of the following elements [3]:

1. **Synonyms**: It is a set of synonymous words representing the concept. The words in the synset are arranged according to the *frequency of usage*. For example, the Hindi words {विद्यालय, पाठशाला, स्कूल} represents the concept of *school* as an *educational institution*.

2. **Gloss**: It describes the concept through a formal definition, in Hindi. For example, 'वह स्थान जहाँ प्राथमिक, माध्यमिक या उच्च स्तर की औपचारिक शिक्षा दी जाती है', explains the concept of *school* as an *educational institution*.

3. **Example**: This gives the usage of the words in a sentence. Generally, the words in a synset are replaceable in a sentence. For example, 'इस विद्यालय में से एक पांचवी तक की शिक्षा दी जाती है', gives the usage for the words in the synset representing *school* as an *educational institution*.

4. **Ontology Position**: An ontology is a hierarchical organization of concepts, more specifically, a categorization of entities and actions. For each syntactic category, namely n*oun*, *verb*, *adjective* and *adverb*, a separate ontological hierarchy is present.

Each synset is mapped into some place in the ontology. A synset may have multiple parents. The ontology for the synset representing the concept of *school* is:
Noun → *Inanimate* → *Place* → *Physical Place* → {विद्यालय, पाठशाला, स्कूल}

3.2 Lexical and Semantic Relations

A WordNet is a word sense network. A word sense node in this network is a synset which is regarded as a basic object in the WordNet. Like English WordNet, Hindi WordNet also has synsets linked by the relations Hyponymy, Hypernymy, Meronymy, Holonymy and Antonymy. There are also some other relations not described in Section 2.3 of English WordNet, like *entailment* between verbs, *troponymy* (which also exist in English WordNet), *gradation* and *causative* relations. There are also some cross syntactic category or cross POS linkages like *ability*, *capability* and *function* links between nominal and verbal concepts, *attribute* and *modifies-noun* between nominal and adjectival concepts, *modifies-verb* and *derived-from* between verbal and adverbial concepts [3].

3.2.1 Entailment

Entailment refers to a relationship between two verbs. Any verb V_1 entails V_2 if the truth of V_2 follows logically from the truth of V_1. This relation is asymmetric. For example, खर्राटा लेना entails सोना [3].

3.2.2 Troponymy

Troponymy relation denotes a specific manner elaboration of another verb. It shows the manner of an action. *X* is a troponym of *Y* if to *X* is to *Y* in some manner. For example, {मुस्कुराना, मुस्कराना, मुस्काना} is a troponym of {हँसना} [3].

3.2.3 Gradation

Gradation is a lexical relation. It represents the intermediate concept between two opposite concepts. For example, सुबह and शाम are antonyms. But दोपहर is an intermediate state between them. So सुबह and दोपहर, दोपहर and शाम are related through Gradation relation [3].

3.2.4 Causative

In Hindi, there is a convention of forming causation by making morphological change in the base verb. The causative relation links the causative verbs and the base verbs and shows the interdependency between them. For example खिलाना is a causative verb of खाना [3].

3.2.5 Cross Part-of-Speech Linkage

There are some relations between synsets in Hindi WordNet which are across different POS category [3].

Linkages between Nominal and Verbal concepts

- **Ability Link**: This link specifies the inherited features of a nominal concept. This is a semantic relation. For example, {मछली, मच्छी, मत्स्य, मीन} and {तैरना, पैरना, पौड़ना, पौरना, हेलना} [3].
- **Capability Link**: This link specifies the acquired features of a nominal concept. This is a semantic relation. For example, {व्यक्ति, मानस, आदमी, शख्स, जन, बंदा, बन्दा} and {तैरना, पैरना, पौड़ना, पौरना, हेलना} [3].
- **Function Link**: This link specifies the function of a nominal concept. This is a semantic relation. For example, {अध्यापक, शिक्षक, आचार्य, गुरु, मास्टर} and {पढ़ाना} [3].

Linkage between Nominal and Adjectival concepts

- **Attribute**: This denotes the properties of a noun. It is a linkage between a noun and an adjective. This is a semantic relation. For example, {पक्षी, चिड़िया, पंछी, खग} and {पंखदार, पंखयुक्त, परदार} [3].
- **Modifies-Noun**: Certain adjectives can only modify certain nouns. Such adjectives and nouns are linked in the Hindi WordNet by this relation. For example, {सुपात्र, सत्पात्र, अच्छा पात्र} and {व्यक्ति, मानस, आदमी} [3].

Linkage between Adverbial and Verbal concepts

- **Modifies-Verb**: Certain adverbs can only go with certain verbs. This relation represents such a connection. For example, {कभी, किसी समय} and {काम करना} [3].
- **Derived Form**: This relation specifies the root form from which a particular word is derived. This relation can go from noun to adjective or vice versa, noun to verb and adjective to verb and aims to handle derivational morphology. This is a lexical relation. For example, {क्रमशः, कमानुसार, यथाक्रम} and {क्रम, सिलसिला, श्रंखला} [3].

Chapter 4

Multilingual dictionary framework and Knowledge based algorithms

Various research papers were read during the project for gaining knowledge related to wordnet which are summarized below in this particular chapter.

4.1 Multi Lingual Dictionary

The Multi Lingual Dictionary (*MultiDict*) is another project which is being lead by CFILT, IIT Bombay. The aim of this project is to overcome the multilingual resource scarcity barrier among Indian languages like Marathi, Konkani, Gujarati, Punjabi, Telugu, Tamil, Malayalam, Nepali, Boro and other Indian North-Eastern languages. In most of the Language Processing and Cross Lingual Information Retrieval tasks, bilingual dictionaries act as an important linguistic resource. But when working in a cross lingual domain, the number of bilingual dictionaries will be huge. For example, if the number of languages one is working is say 5, then the number of bilingual dictionaries required are 10. Creating so many dictionaries consumes a lot of linguistic effort, time, resources and involves implicit duplication of effort for each language. MultiDict tries to overcome this by adopting a framework which is based on the expansion approach of WordNet building [2].

4.1.1 MultiDict Framework

The MultiDict adopts the idea of expansion based approach of building a WordNet. Unlike a traditional bilingual dictionary, the MultiDict consists of synsets rather than words. The synsets corresponding a concept is created in a base language and a unique ID is assigned to them. Now corresponding to this synset, the same concept can be expressed in various other languages, all bearing the same synset ID [2]. By adopting such a design framework, the semantic features of each synset, the morpho-syntactic features, hyperrnymy, hyponymy kind of relations etc. are automatically applied across all languages.

4.2 Lesk's Algorithm

This method, also called gloss overlap, the simplest overlap based method, is named after its author [4]. The method uses a Machine Readable Dictionary, as the knowledge source. So the feature used here for the overlap purpose is the sense definition of words or the gloss. Given a pair of target words (w_1,w_2) to be disambiguated, the senses of the target words that have the

maximum overlap in their sense definitions are selected as the winner senses for w_1 and w_2. Putting it formally, given, senses $S_1 \in Senses(w_1)$ and $S_2 \in Senses(w_2)$, the score due to Lesk's algorithm can be calculated as follows:

$$score_{lesk}(S_1,S_2) = |gloss(S_1) \cap gloss(S_2))|$$

where gloss(S_i) is the bag of words in the textual definition of sense S_i of w_i. The senses which maximize the above value becomes the winner senses. However, this requires the calculation of $|Senses(w_1)| \cdot |Senses(w_2)|$ gloss overlaps.

More generically, we have to find $\prod^{n}_{i=1} |Senses(w_i)|$ overlaps. A different method that can be employed, is to find the overlap between the gloss of a target word and the glosses of the senses of the context words. This basically means to find the overlap between the sense bag and the context bag. The sense bag and the context bag can be defined as follows:

Sense Bag: gloss containing the words in the definition of a candidate sense of the target word.
Context Bag: gloss of each sense of each context word.

Formally, given a target word w, the following is computed for every sense in w:

$$score_{leskvar}(S) = |context(w) \cap gloss(S))|$$

Where, context(w) is the context bag and gloss(S) is the sense bag. The sense with the highest value becomes the winner sense.

The table below shows an example sentence and outcome of Lesk's algorithm.

4.3 Extended Lesk's Algorithm

Due to certain drawbacks of the Lesk's algorithm, a modification of the algorithm was introduced by Banerjee and Pedersen [5].

Sentence: On burning coal we get ash.		
	Ash	Coal
Sense 1	Trees of the olive family with pinnate leaves, thin furrowed bark and gray branches.	A piece of glowing carbon or **burnt** wood.
Sense 2	The **solid** residue left when **combustible** material is thoroughly **burned** or oxidized.	charcoal.
Sense 3	To convert into ash.	A black **solid combustible** substance formed by the partial decomposition of vegetable matter without free access to air and under the influence of moisture and often increased pressure and temperature that is widely used as a fuel for **burning**.
Sense 2 of ash would be the winner sense.		

Table 4.1: Table showing a run of Lesk's algorithm

Limitations of Lesk's Algorithm: Unfortunately enough, Lesk's approach is very sensitive to the exact words in the definition, so the absence of a certain word can change the results drastically. Furthermore, dictionary definitions are short and precise; hence they do not capture minute differences between senses. This is a serious limitation, as the Lesk's algorithm uses only dictionary gloss as features.

Modification introduced: Off late, Banerjee and Pedersen, introduced a measure based on the concept of extended gloss overlap, which expands the glosses of the words being compared by including glosses of words that are semantically to words in concern (e.g. hypernymy, meronymy, pertainymy, etc.). The relationships chosen, can be chosen from any combination of resources in theWordNet, which is another powerful knowledge resource introduced in section 2. For each sense S of a target word w we estimate its score as:

$$score_{ExtLesk}(S) = \sum s' : s \rightarrow s' \text{ or } s \equiv s' |context(w) \cap gloss(S))|$$

Where, context(w) is the context bag and gloss(S') is the sense bag of S or of any word related to S by a relation rel as shown in the formula. The scoring technique is parameterized and can take into account the gloss length (normalization) and can also include function words if necessary. This method thus uses a Machine Readable Dictionary and the WordNet as knowledge sources.

The method also scales an accuracy of 31.7%, which is high above the 23% accuracy of the basic Lesk's algorithm.

Chapter 5

IWSD – Iterative Word Sense Disambiguation

The algorithm, against which we propose our claim, is a supervised WSD algorithm, developed at IIT Bombay, called *Iterative WSD (IWSD)* [6]. The algorithm is greedy and uses a scoring function to disambiguate senses. The scoring function, the parameters based on which it has been designed and the basic algorithm are described in the following subsections.

5.1 Parameters for IWSD

Khapra et al. (2010) proposed a supervised algorithm for domain-specific WSD and showed that it beats the most frequent corpus sense and performs on par with other state of the art algorithms like PageRank. The various parameters used by Iterative WSD can be classified as:

Wordnet-dependent parameters

- *belongingness-to-dominant-concept*
- *conceptual-distance*
- *semantic-distance*

Corpus-dependent parameters

- *sense distributions*
- *corpus co-occurrences*.

5.2 Scoring function for IWSD

The scoring function of the IWSD algorithm integrates the WordNet-dependant parameters and the corpus-based parameters to rank the candidate senses of the target word. The scoring function is illustrated below:

$$S^* = \arg\max_i \theta_i V_i + \sum_{j \in J} W_{ij} V_i V_j \qquad (1)$$

Where,
J = *Set of disambiguated words*
θ_i = *BelongingnessToDominantConcept*(S_i)
Vi = $P(S_i|word)$
W_{ij} = *CorpusCooccurrence*(S_i,S_j) * 1 / *WNConceptualDistance*(S_i,S_j) * 1 / *WNSemanticGraphDistance*(S_i,S_j)

5.3 Algorithm

As stated earlier, IWSD is a greedy algorithm. The greedy nature of the algorithm can be explained through the steps followed by the algorithm.

Algorithm 1: performIterativeWSD (sentence)

1. Tag all monosemous words in the sentence.
2. Iteratively disambiguate the remaining words in the sentence in increasing order of their degree of polysemy.
3. At each stage select that sense for a word which maximizes the score given by Equation (1).

Monosemous words are used as the seed input for the algorithm but are not considered while calculating the precision and recall values. It is quite possible that a sentence may not contain any monosemous words in which case the algorithm will first disambiguate the least polysemous word in the sentence. In this case, the disambiguation will be performed only using the first term in the formula which represents the corpus bias (the second term will not be active as there are no previously disambiguated words).

The least polysemous word thus disambiguated will then act as the seed input to the algorithm. IWSD is clearly greedy. It bases its decisions on already disambiguated words, and ignores completely words with higher degree of polysemy. For example, while disambiguating bisemous words, the algorithm uses only the monosemous words and ignores completely the trisemous words and higher order polysemous words appearing in the context. This is illustrated in Figure 5.1.

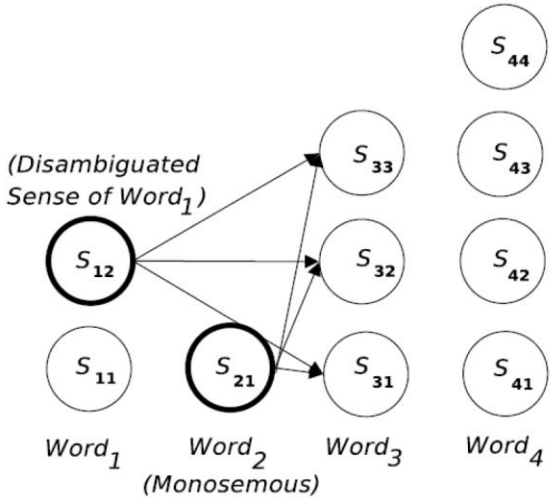

Figure 5.1 *IWSD Operation: Only previously disambiguated words and monosemous words are used while disambiguating* $Word_3$

5.4 Critique of IWSD

The accuracy of the IWSD algorithm is comparable to other state-of-the-art supervised algorithms. But, as shown in Table 1 below, the accuracy of IWSD algorithm is marginally better than MFS which tags the words with $P(S_i|word)$. This strongly indicates that all the parameters encompassed by the scoring function of the algorithm do not contribute sufficiently enough. Hence it became mandatory to get a more in-depth insight into the working of the algorithm, to check if the WordNet-dependant parameters actually have a role in the disambiguation process of IWSD.

	Precision	**Recall**	**F-Score**
MFS	79.57	78.52	79.04
IWSD	79.45	78.98	79.21

Table 1: MFS v/s IWSD

The study and analysis of the extent and manner of use of the WordNet-based parameters is important here, as these parameters account for the context-based information in the IWSD scoring function. From the acumen of skilled lexicographers, we learnt the kind of cognition techniques used in annotating words. The interaction with the lexicographers reveals that context-based knowledge is of paramount significance in the human annotation process.

Thus the degree of usage of the contextual evidence, which is induced in the IWSD scoring function through the WordNet-dependant parameters for sense disambiguation, would draw a comparison with human annotation techniques for sense marking.

Chapter 6

Error Analysis

We probed the IWSD algorithm for errors and analyzed them based on different categories; the experiment and its results are described in this chapter.

6.1 Experimental Setup

We carried our experiments on NEWS domain corpus collected from BBC news website[1]. The text was annotated manually by expert lexicographers. The various statistics pertaining to the total number of words, number of words per POS category and average degree of polysemy are described in Table 2.

We did a 4-fold cross validation of IWSD algorithm using this corpus.[2]

	Polysemous words	Monosemous words	Wordnet Polysemy	Corpus Polysemy
Noun	72225	61682	3.03	1.82
Verb	26436	4372	4.47	3.00
Adjective	15462	30122	2.68	2.03
Adverb	12907	10658	2.52	2.11
Overall	127030	106834	3.13	2.02

Table 2: Statistics of the News corpora

To investigate the shortcomings of IWSD and accordingly to find its resemblance with human annotation techniques we conducted the following experiments.

First, in order to obtain an idea of which POS categories are subject to errors, we performed a Part-of-Speech (POS) based analysis on error-prone words.

Second, we performed an ontology-based analysis on the nouns which are sense-marked incorrectly by IWSD, as nouns occupy a major part of the corpora and correspondingly account for maximum errors.

Third, we tweaked the IWSD scoring formula into a linear combination of the statistical parameters and contextual parameters as explained in Equation 2, and tested for varying values of α.

[2] www.bbc.co.in/hin

$$S^* = \alpha \arg\max_i \theta_i V_i + (1-\alpha)\sum_{j \in J} W_{ij} V_i V_j \qquad (2)$$

Fourth, to get further insight on the IWSD scoring function, we performed an ablation test on the parameters of the scoring function.

Fifth, we selected error-prone words which had *P(Si|word) > 0.6* and analyzed the root cause of such errors.

Finally, we conducted an experiment which we call *context-agnostic tagging*, where a skilled lexicographer was assigned the task of tagging around 7500 words without looking at the context.

6.2 Results

6.2.1 Part-of-Speech (POS)-based analysis

The accuracy of the IWSD algorithm on the News corpora in the average case is 79%. In order to analyze the type of errors being made by the algorithm, we first made an analysis based on POS-tag, *i.e.*, we analyzed the percentage (%) of occurrences of each POS in the corpus and consequently the percentage (%) of errors in each POS category.

Table1 and Chart1 highlight the occurrence of errors in each POS category.

TOTAL ERRORS	**25581**
ADVERBS	1971
VERBS	7789
NOUNS	12289
ADJECTIVES	3532

***Table1**: Distribution of errors across POSs*

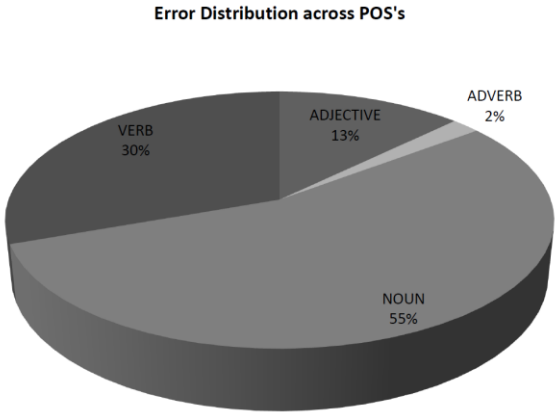

***Chart1**: Words under Generic Ontology*

6.2.2 Ontology-based analysis

The above analysis revealed that the errors occur mostly in case of nouns. To get a closer view of the scenario, we dived deeper and analyzed the nouns based on their ontology. The primary purpose of this experiment was to detect if errors pertaining to the News corpora occur in some particular i.e., which ontological category the error-prone nouns fall into. We obtained two cases in this experiment:

Case 1: *Words under Generic Ontological category* – Synsets that are higher in the WordNet hierarchy, or rather closer to the root are generic concepts. The distribution of words under generic synsets is exhibited in Chart 3.

कार्य (Action)	67
भौतिक स्थान (Physical Place)	62
शारीरिक कार्य (Physical)	47
व्यक्ति (Person)	32
समूह (Group)	31
अमूर्त (Abstract)	25
वस्तु (Object)	25
अवधि (Period)	17
प्राकृतिक वस्तु (Natural Object)	13
घटना (Event)	6
व्यक्तिवाचक संज्ञा (Proper Noun)	6
शारीरिक अवस्था (Physiological State)	5
स्थान (Place)	5
जानकारी (information)	4
संज्ञा (Noun)	3
जातिवाचक संज्ञा (Common Noun)	2
समय (Time)	2

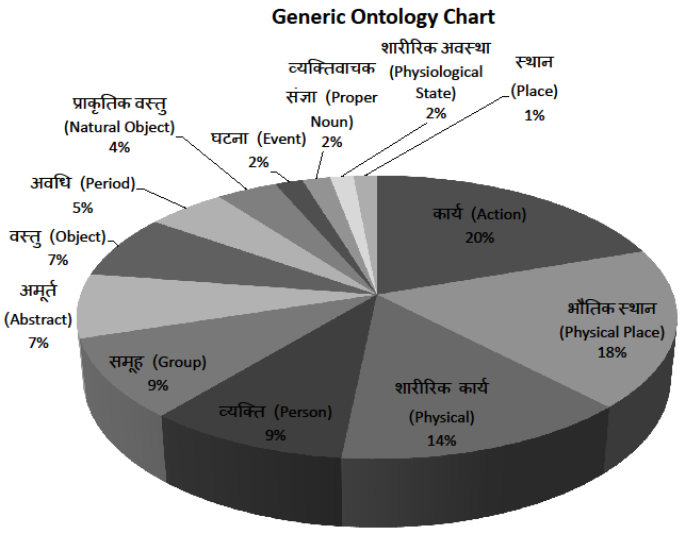

Chart2*: Words under Generic Ontology*

Case 2: *Words under Specific Ontological category* – Synsets that are deeper in the WordNet tree or rather closer to the root are generic. The distribution of words under generic synsets is exhibited in Chart 3.

मानवकृति (Artifact)	62
अवस्था (State)	46
भाग (Part of)	40
मनोवैज्ञानिक लक्षण (Psychological Feature)	31
ज्ञान (Cognition)	29
गुण (Quality)	21
गुणधर्म (property)	18
संज्ञापन (Communication)	18
संप्रेषण (Communication)	18
बोध (Perception)	15
मानसिक अवस्था (Mental State)	14
शारीरिक वस्तु (Anatomical)	14
असामाजिक कार्य (Anti-social)	6
भाषा (Language)	5
खाद्य (Edible)	4
भौतिक अवस्था (physical State)	4
सामाजिक कार्य (Social)	4
संकल्पना (concept)	4
कला (Art)	3
सामाजिक अवस्था (Social State)	3

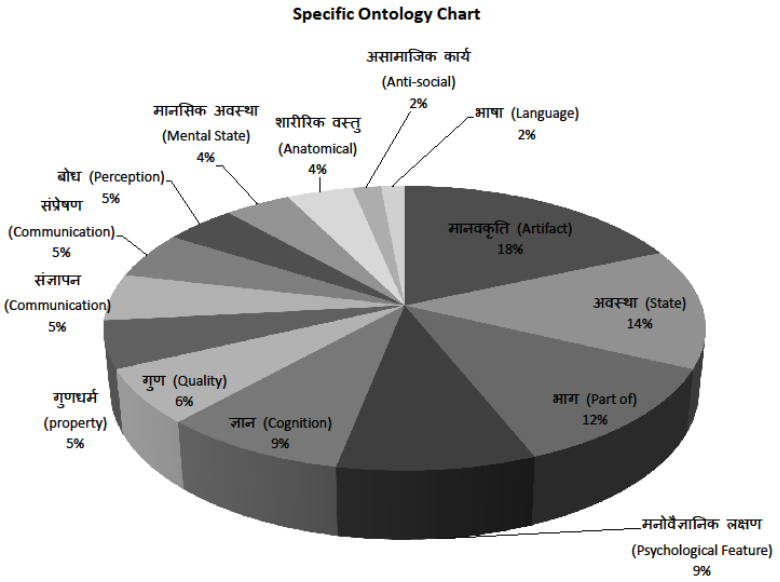

Chart3*: Words under Specific Ontology*

6.2.3 Alpha test results

Alpha	Precision	Recall	F-score
0	59.59%	58.84%	59.21
0.00001	79.48%	78.49%	78.98
0.0001	79.50%	78.51%	79
0.001	79.50%	78.51%	79.01
0.01	79.61%	78.62%	79.01
0.1	79.61%	78.62%	79.11
0.2	79.61%	78.62%	79.11
0.25	79.61%	78.62%	79.11
0.5	79.61%	78.62%	79.11
0.75	79.61%	78.62%	79.11
1	79.59%	78.60%	79.1

6.2.4 Ablation test on the IWSD parameters

An analysis into the parameters of the IWSD scoring function, could further give an insight into the working of the IWSD algorithm and the contribution of each of its parameters. We thus conducted an *ablation test* on the parameters of the IWSD scoring function.

Ablation Parameter	Precision	Recall	F-Score
θ	79.61%	78.62%	79.11%
$P(S\|W)$	59.59%	58.84%	59.21%
Corpus-Cooccurence	79.57%	78.58%	79.07%
$ConceptualDistance(S_i,S_j)$	79.50%	78.51%	79.01%
$SemanticSimilarity(S_i,S_j)$	79.61%	78.62%	79.11%

6.2.5 Different Scenarios in Sense Marking

Case 1: *Context Sensitive Human Tagging Vs. Context agnostic Human Tagging*

POS	No. of Errors	Token Count	Error Contribution (%)	Error Fraction (%)
Adjective	3532	15462	13.81	22.84
Adverb	1971	12907	7.7	15.27
Noun	12289	26436	48.04	46.49
Verb	7789	72225	30.45	10.78
Total	25581	127030	-	20.14

Table: Diptesh(CA) Vs. Akhlesh(CS)

Case 2: *Context Agnostic Human Tagging vs. Context agnostic Machine Tagging*

POS	# of Words	Precision	Recall	F-Score
NOUN	6175	84.29%	73.21%	78.36%
ADVERB	941	85.82%	83.88%	84.84%
ADJECTIVE	1323	71.27%	59.37%	64.78%
VERB	1983	67.78%	63.96%	65.82%
Overall	10422	79.59%	70.71%	**74.89%**

Table: Diptesh Vs. IWSD

POS	# of Words	Precision	Recall	F-Score
NOUN	6175	52.03%	53.41%	52.71%
ADVERB	941	61.48%	73.11%	66.80%
ADJECTIVE	1323	44.49%	22.90%	30.24%
VERB	1983	32.01%	21.58%	25.78%
Overall	10422	49.78%	45.26%	**47.41%**

Table: Diptesh Vs. WFS

POS	# of Words	Precision	Recall	F-Score
NOUN	6215	60.50%	62.09%	61.29%
ADVERB	944	73.89%	87.82%	80.25%
ADJECTIVE	1334	57.25%	29.61%	39.03%
VERB	1984	33.96%	22.88%	27.34%
Overall	10477	58.12%	52.85%	**55.36%**

Table: Akhlesh vs Diptesh

Chapter 7

Discussion

In the previous section, from the ablation test results, we have seen that the P(S|W) parameter is the essence of the IWSD scoring function. During the testing phase, this parameter is of paramount importance. Furthermore, this parameter is based on corporal evidence as a whole, rather than specific, case-wise conceptual evidence as is the case with the human annotation process. The IWSD algorithm or the machine thus follows an annotation methodology which differs from the human annotation method.

In this section we exhibit how the setting of the *P(S|W)* parameter after the training process becomes detrimental to obtaining the right sense, especially in cases where the target word is highly polysemous, does not have senses which are dominant and the domain is generic. This happens, as the *P(S|W)* parameter is context agnostic and in cases of highly polysemous words the senses are distributed across the corpora. Therefore, inclination towards one of these senses, without considering the context leads to errors in the sense disambiguation process.

We would also like to state that in cases of Domain-Specific WSD, bias towards a particular sense may be an intelligent method for disambiguation, but it is intuitive enough, that when WSD is performed on generic domains, where senses are not bent towards the domain, this ploy will become futile.

The following case studies, described below establish this fact:

Case Study 1:

The word "कहा" is not a dominant concept in the News corpus and has "22" senses in the Hindi WordNet. The value of the P(S|W) parameter for one of these "22" senses is set to "0.8157579303" after the training process, from corporal evidence and without considering the context. Hence, the IWSD algorithm fails to tag "कहा" with the right sense "596" times, as IWSD is biased towards one sense out of the "22".

For *e.g.*, one context of the word "कहा" contains words like "बीजेपी नेता ने कहा" which is a clear contextual evidence that the correct sense of "कहा" should be "बोलना" rather than "सूचना देना" as tagged by the IWSD algorithm.

Case Study 2:

The word "करने" is not a dominant concept in the News corpus and has "36" senses in the Hindi WordNet. The value of the P(S|W) parameter for one of these "36" senses is set to "0.9452961683" after the training process, from corporal evidence and without considering the context. Hence, the IWSD algorithm fails to tag "करने" with the right sense "85" times, as IWSD is biased towards one sense out of the "36".

For *e.g.*, one context of the word "करने"(to do) contains words like "शलाका सम्मान को स्वीकार करने से" which is a clear contextual evidence that the correct sense of "पहले" should be "स्वीकार करना" rather than "कार्य करना " as tagged by the IWSD algorithm.

Case Study 3:

The word "समय" is not a dominant concept in the News corpus and has "12" senses in the Hindi WordNet. The value of the *P(S|W)* parameter for one of these "12" senses is set to "0.6614987254" after the training process, from corporal evidence and without considering the context. Hence, the IWSD algorithm fails to tag "समय" with the right sense "190" times, as IWSD is biased towards one sense out of the "12".

For *e.g.*, one context of the word "समय" contains words like "जैसा कि समय बीत जाने" which is a clear contextual evidence that the correct sense of "समय" should be "काल, वक्त" rather than "वह समय जिसके बीच कोई विशेष बात हो" as tagged by the IWSD algorithm.

There are a number of such cases, where IWSD provides erroneous senses for words with high degree of polysemy. Some other cases have been listed in Table3.

Word	No. of Senses	No. of Occurences	POS	P(S|W)
करने	36	85	VERB	0.945296168
कर	36	78	VERB	0.945296168
कहा	22	596	VERB	0.81575793
कहना	22	85	VERB	0.81575793
दिनों	18	85	NOUN	0.644135177
लगा	16	79	ADJECTIVE	0.800000012
काम	14	243	NOUN	0.528150141
बच्चों	14	131	NOUN	0.515358388
बड़ी	14	122	ADJECTIVE	0.636882126
बनाए	13	142	VERB	0.635258377
बना	13	92	VERB	0.635258377
समय	12	190	NOUN	0.652068138
दूसरे	12	206	ADJECTIVE	0.541125536
बात	9	286	NOUN	0.509846807
दिन	9	163	NOUN	0.644135177

Chapter 8

Conclusion and Future Work

The work on WSD has been as old as work on language processing. The first steps in WSD dates back to 1950, roughly. In spite of several hurdles being overcome, WSD still remains a daunting task. This is because it deals with the identification of a semantic structure from unstructured textual sources, keeping in mind that all the complexities of the language are involved in it. Another major aspect for the hardness of WSD is the granularity of the sense distinctions being considered. *We agree that there are works being cited with accuracy above 90%. However, the story behind is that they are applied on very few words*, mostly nouns and have large sense distinctions.

Knowledge based methods have the advantage of having more knowledge injected into the system, which is highly desirable for disambiguation purposes. Moreover, they ensure that the domain ontologies and semantic interoperability are exploited. There are enriched knowledge sources being developed everyday, which adds to the enhancement of knowledge based approaches.

The discussion in the previous section implies clearly that even the words with a high value of P(S|W), have high error counts and IWSD has tagged them wrong when contextual part if removed. Though, P(S|W) plays a major role in disambiguating the sense of the word, but the contextual parts importance is proved by the experimental human context agnostic tagging where accuracy falls down low, compared to context sensitive tagging.

The question it raises*: If Humans can't tag without the context, how can machine?*

Machine is ultimately trained on the context sensitive annotated data already tagged by a human, in cases where it has a high accuracy.

A lot of work on WSD has been done in the last 50 years especially after the 90s. But it is apparently evident, that a dead end has been reached by researchers. Sometimes in the context of the next directions of research, the path traversed till now needs to be referred. We have made an honest attempt to provide that path for researchers, at least in part bringing unsupervised and knowledge based approaches for WSD into the perspective of last 50 years of work on the topic. We are aware that much could be added to what we have presented, we have tried to lay down the major work and a broad outline of in the field.

Future Work

There is no disagreement that WSD needs to show its relevance in vivo, i.e. in applications such as Information Retrieval and Machine Translation and in vitro, i.e. WSD as a standalone application. There are certain aspects to WSD that need to be looked at in the near future. First, the representation of word senses needs to be looked at and better resources that convey a better understanding of word senses need to be developed. Second, in spite of development of several machine learning approaches to WSD, the Knowledge acquisition bottleneck still remains a hurdle to cross.

Finally, great results have recently been obtained from domain-specific WSD, which would require greater attention in the years to come.

References

[1] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller, ``Five papers on WordNet'' in WordNet: An Electronic Lexical Database, Ed. MIT Press, May 1998.

[2] Rajat Mohanty, Pushpak Bhattacharyya, Prabhakar Pande, Shraddha Kalele, Mitesh Khapra and Aditya Sharma, Synset Based Multilingual Dictionary: Insights, Applications and Challenges, Global WordNet Conference (GWC08), Szeged, Hungary, January 22-25, 2008.

[3] Prabhakar Pandey, Laxmi Kashyap, Pushpak Bhattacharyya, Hindi WordNet Documentation (at http://www.cfilt.iitb.ac.in), CFILT, IIT Bombay, October, 2009.

[4] LESK, M. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In SIGDOC '86: Proceedings of the 5th annual international conference on Systems documentation (New York, NY, USA, 1986), ACM, pp. 24–26.

[5] BANERJEE, S., AND PEDERSEN, T. Extended gloss overlaps as ameasure of semantic relatedness. In Proc. of the 18th Int'l. Joint Conf. on Artificial Intelligence (2003), pp. 805–810.

[6] Mitesh Khapra, Pushpak Bhattacharyya, Shashank Chauhan, Soumya Nair and Aditya Sharma, *Domain Specific Iterative Word Sense Disambiguation in a Multilingual Setting*, International Conference on NLP (ICON08), Pune, India, December, 2008.