Chapter 2

Literature Review

The number of methods for WSD is quite large and varied. This chapter surveys previous work on WSD techniques and strategies. The sense repository for Supervised WSD is the WordNet. Hence a survey of works on building WordNet is presented in this chapter. The words in WordNet are in root forms. But the words in corpora are in root as well as inflected forms. So the words need to be stripped of their inflexions by a lemmatizer. Some amount of sense annotation is also required for Supervised WSD. Therefore a survey of Lemmatizers and sense marked corpus have also been done.

2.1 Introduction

Building of efficient WSD systems is a critical problem area of NLP. A systematic review of various approaches to WSD and corresponding algorithms associated with those approaches is presented in this thesis. Section 2.2 presents a review of building of WordNet, Section 2.3 presents a review of approaches to WSD and related algorithms, Section 2.3 presents a review of Lemmatizers and sense marked corpus and Section 2.4 concludes the chapter with chapter summary.

2.2 Multi Lingual Dictionary and WordNet

A model for developing a single dictionary for n languages, in which there are linked concepts expressed as synsets and not as words was proposed by R.Mohanty et al.[1]. For each concept, semantic features- which are universal- are used only once. As for morph-syntactic features, their incorporation would demand much less effort, if languages are grouped according to their families; in other words the model can take

advantage of the fact that close kinship languages share morpho-syntactic properties. The advantage of the proposed model is economy of labour and storage. Semantic features like [±Animate, ±Human, ±Masculine, etc.], are assigned to a nominal concept and not to any individual lexical item of any language. Similarly, the semantic features, such as [+Stative (e.g., know), +Activity (e.g., stroll), +Accomplishment (e.g., say), +Semelfactive (e.g., knock), +Achievement (e.g., win)] are assigned to a verbal concept. These semantic features are stored only once for each row and become applicable independent of any language. Consequently, lexical entries with highly enriched semantic features can be added to a dictionary for as many languages as required within a short span of time. English WordNet[2] is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. English WordNet was created in the Cognitive Science Laboratory of Princeton University under the direction of psychology professor George Armitage Miller starting in 1985 and has been directed in recent years by Christiane Fellbaum. EuroWordNet[4][11] was a 3-year project that developed a multilingual database with linked WordNets for 8 European languages: English, Dutch, Italian, Spanish, French, German, Czech and Estonian. Each WordNet is structured along the same lines as the Princeton WordNet[2]. English WordNet contains information about nouns, verbs, adjectives and adverbs in English and is organized around the notion of a synset. The Hindi WordNet[3] is the first of its kind in India. It 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. IndoWordNet[10] is a linked lexical knowledge base of WordNets of 18 of the scheduled languages of India, viz., Assamese, Bangla, Bodo, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Malayalam, Manipuri, Marathi, Nepali, Oriya, Punjabi, Sanskrit, Tamil, Telugu and Urdu. Indian languages form a very significant component of the languages landscape of the world. There are 4 streams of language typology operative in the Indian subcontinent- Indo European, Dravidian, TibetoBurman and Austro Asiatic. Many languages rank within top 10 in the world in terms of the population speaking them, e.g., Hindi-Urdu 5th, Bangla 7th, Marathi 12th and so on as per the List of languages by number of native speakers [83][84]. Creating

WordNets of Indian languages is therefore a highly important techno-scientific and linguistic project. Such project indeed took off in 2000 with Hindi WordNet being created by the Natural Language Processing group at the Center for Indian Language Technology (CFILT) in the Computer Science and Engineering Department at IIT Bombay.[3] It was made publicly available in 2006 under GNU license. The Hindi WordNet was created with support from the TDIL[6] project of Ministry of Communication and Information Technology, India and also partially from Ministry of Human Resources Development, India.

Wordnets of other languages of India then followed suit. The large nationwide project of building Indian language WordNets was called the IndoWordNet project. IndoWordNet is highly similar to EuroWordNet. However, the pivot language is Hindi which, of course, is linked to the English WordNet. Also typical Indian language phenomena like complex predicates and causative verbs are captured in IndoWordNet. IndoWordNet is publicly browsable. The Indian language wordnet building efforts forming the subcomponents of IndoWordNet project are: North East WordNet project, Dravidian WordNet Project and Indradhanush project all of which are funded by the TDIL project.

2.3 Approaches to WSD and related algorithms

WSD approaches proposed till date can be broadly classified as Knowledge Based Approaches, Machine Learning Based Approaches and Hybrid Approaches.

The Knowledge based approaches do not require any tagged or untagged corpus as no training is involved. But the performance of these algorithms are comparatively low because of their complete dependence on dictionary defined senses or WordNet defined senses where the number of words to describe or explicate any particular sense is, by nature, limited. However one redeeming feature of the above algorithms is that as far as computational complexity is concerned they are comparatively more efficient as they require only a machine readable dictionary or a Wordnet to be implemented. Selectional Preferences by P.Resnik [24]which falls under Knowledge based approach is concerned with the use of selectional constraints for automatic sense disambiguation in broad-coverage settings. The approach combines statistical and knowledge-based methods, but unlike manycorpus-based approaches to sense disambiguation it takes as its starting point the assumption that sense annotated

training text is not available. The Lesk algorithm [14] belonging to the class of Knowledge based approaches is based on the assumption that words in a given “neighbourhood" (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 neighbourhood. Another knowledge based approach proposed by AgirreEneko& GermanRigau [19] use the conceptual distance between the senses of the context words and the sense of the target word as a measure for disambiguation. They proposed a formula for conceptual distance which is directly proportional to the length of the path between two synsets in the WordNet graph and inversely proportional to the depth of their common ancestor in the WordNet hierarchy. Walker Algorithm [15] is a Thesaurus Based approach falling under Knowledge based approaches to WSD. It concerns with finding each sense of the target word from the thesaurus category to which that sense belongs and then calculating the score for each sense by using the context words. A context word will add one to the score of the sense if the thesaurus category of the word matches that of the sense. S.Banerjee[47] introduced an adaptation of Lesk algorithm where WordNet and not standard dictionary was considered as the knowledge base for the glosses. The algorithm compares the glosses between each pair of words in a window of context. An overlap is the longest sequence of one or more consecutive words that occurs in both the glosses. Each overlap found between two glosses contribute a score equal to the square of the number of words in the overlap and the candidate combination with the highest score is the winner.

The study of machine learning based algorithms namely supervised , semi-supervised as well as unsupervised approaches suggested that extracting “sense definitions” or “usage patterns” from corpora helps in improving the performance of WSD. However, most supervised algorithms which perform very well are not general purpose WSD systems, but word specific classifiers. Decision list algorithm proposed by Yarowsky[17] uses the “one sense per collocation" property of human languages for word sense disambiguation. The algorithm starts with a large, untagged corpus, in which it identifies examples of the given polysemous word, and stores all the relevant sentences as lines. For instance, Yarowsky uses the word "plant" in his 1995 paper[18] to demonstrate the algorithm. If it is assumed that there are two possible senses of the word, the next step is to identify a small number of seed collocations representative of each sense, give each sense a label (i.e. sense A and B), then assign

the appropriate label to all training examples containing the seed collocations. In this case, the words "life" and "manufacturing" are chosen as initial seed collocations for senses A and B respectively. The residual examples (85%–98% according to Yarowsky) remain untagged. A decision list algorithm is then used to identify other reliable collocations and the decision list is ranked by the log-likelihood ratio. The drawback of Decision list algorithm is that it is a word-specific classifier. A separate classifier needs to be trained for each word. Hwee.T.Ng and Hian.B.Lee [20] proposed an algorithm for word sense disambiguation using an exemplar-based learning algorithm. This approach integrates a diverse set of knowledge sources to disambiguate word sense, including part of speech of neighbouring words, morphological form, the unordered set of surrounding words, local collocations, and verb-object syntactic relation. Given a test sentence containing the ambiguous word, a test example is similarly constructed. The test example is then compared to all training examples and the k-closest training examples are selected. The sense which is most prevalent amongst these "k" examples is then selected as the correct sense. It is also a word-specific classifier and it will not work for unknown words which do not appear in the corpus.Y.K.Leeet.al[32] proposed a supervised algorithm where Support Vector Machine (SVM), a binary classifier, finds a hyperplane with the largest margin that separates training examples into two classes.

As SVMs are binary classifiers, a separate classifier is built for each sense of the word. The knowledge sources used included part-of-speech (POS) of neighbouring words, single words in the surrounding context, local collocations, and syntactic relations. Here, given a test sentence, a test example is constructed using the above features and fed as input to each binary classifier. The correct sense is selected based on the label returned by each classifier. Ciaramita et.al [48] proposed a supervised algorithm for WSD where perceptron based Hidden Markov Model(HMM) was used. A discriminative HMM was trained using the following features: POS of the word as well as POS of neighbouring words, Local collocations and Shape of the word and neighbouring words. The class space is reduced by using WordNet's super senses instead of actual senses. It performs well with Named Entity Recognition(NER) .A broad coverage classifier as the same knowledge sources can be used for all words belonging to super sense with this approach.

The study of semi-supervised approaches to WSD suggests that they work at par with their supervised counterparts although they need significantly less amount of tagged

data. Semi-Supervised decision list algorithm proposed by Yarowsky [18] trained the Decision List algorithm[17] using a small amount of seed data and then classified the entire sample set using the trained classifier. A new seed data was created by adding those members which were tagged as Sense-A or Sense-B with high probability. The classifier was retrained using the increased seed data.

Jean Veronis [29] proposed an unsupervised WSD algorithm Hyperlex that is capable of automatically determining word uses in a text base without recourse to a dictionary. The algorithm makes use ofthe specific properties of word co-occurrence graphs, which have "small world" properties. Unlike earlier dictionary-free methods based on word vectors, it can isolate highly infrequent uses (as rare as 1% of all occurrences) by detecting "hubs" and high-density components in the co-occurrence graphs. The basic assumption underlying the method proposed in [29] is that the different uses of a target word form highly interconnected "bundles" in a small world of co-occurrences, or in terms of graph theory, high density components. Use of small world properties was a first of its kind approach for automatically extracting corpus evidence. Hyperlex is a word-specific classifier and the algorithm would fail to distinguish between finer senses of a word (e.g. the medicinal and narcotic senses of “drug”).

D.Yarowsky [16] proposed an unsupervised algorithm called the WSD using ROGET’S THESAURUS CATEGORIES. The algorithm is underpinned by the following three observations:-1) Different conceptual classes of words, such as Animals or Machines tend to appear in recognizably different contexts. 2) Different word senses tend to belong to different conceptual classes (crane can be an ANIMAL or a MACHINE). 3) A context based discriminator for the conceptual classes can serve as a context based discriminator for the members of those classes. Furthermore, the context indicators for a Roget category (e.g. gear, piston and engine for the category TOOLS/MACHINERY)will also tend to be context indicators for the members of that category (such as the machinery sense of crane). The algorithm identifies “salient” words in the collective context of the thesaurus category and weigh those appropriately. The algorithm then predicts the appropriate category for an ambiguous word using the weights of words in its context.WSD using Roget’s Thesaurus categories. The algorithm was tested on a set of 12highly polysemous English words and reported 92% recall. But it was not tested on an “all word corpus”.

The method proposed by Dekang Lin [22] presented an algorithm that used the same knowledge sources to disambiguate different words. It is different from some corpus-

based algorithms which disambiguate a word with a classifier trained from previous usages of the same word. The algorithm does not require a sense-tagged corpus and exploits the fact that two different words are likely to have similar meanings if they occur in identical local contexts. Here similarity is directly proportional to the probability that the two words have the same super class (Hypernym) . This approach is a general purpose broad coverage approach and can even work for words which do not appear in the corpus. The method presented by P.Resnik et al.[28] is an unsupervised method for word sense disambiguation that exploits translation correspondences in parallel corpora. The technique takes advantage of the fact that cross-language lexicalizations of the same concept tend to be consistent, preserving some core element of its semantics, and yet also variable, reflecting differing translator preferences and the influence of context. Parallel corpora introduced an extra complication for evaluation, since it is difficult to find a corpus that is both sense tagged and parallel with another language; therefore the authors in [28] used pseudo-translations, created by machine translation systems, in order to make possible the evaluation of the approach against a standard test set. The key point here is translations can be considered as contextual indicators of the sense of the word. The method can distinguish even between finer senses of a word because even finer senses of a word get translated as distinct words. But the method[28] needs a word aligned parallel corpora which is difficult to get and it also requires an exceptionally large number of parameters need to be trained. Hybrid approaches like WSD using Structural Semantic Interconnections [36] use combinations of more than one knowledge sources (Wordnet as well as a small amount of tagged corpora). The algorithm proposed in [36] generates a labelled graph where the nodes are the synsets and the edges are the semantic relations present in Wordnet (for e.g. Hypernym-Hyponym relation, meronymy, modifies_nounetc) . A very high performing sense tagging rules in the above algorithm is the direct hypernymy path. This rule reads as follows: “if the word wj appears in the gloss of a synset Si, and if one of the synsets of wj, Sj, is the direct hypernym of Si, then, select Sj as the correct sense for wj. The method thus could capture important information encoded in wordnet as well as draw syntactic generalizations from minimally tagged corpora. R. Mihalcea et al [30] introduced a hybrid WSD method which uses a small data set for training purposes, and generalizes the concepts learned from the training data to disambiguate the words in the test data set. As a result, the algorithm does not need a separate classifier for

each word to be disambiguated. The hybrid approaches combine information obtained from multiple knowledge sources and they also use a very small amount of tagged data.

2.4 Lemmatizers and Sense tagged documents

The first automated stemmer which was developed specifically for IR/NLP applications was introduced by Lovins [49]. His approach consisted of the use of a manually developed list of 294 suffixes, each linked to 29 conditions, plus 35 transformation rules. For an input word, the suffix with an appropriate condition is checked and removed. Porter [50] developed the Porter stemming algorithm which became the most widely used stemming algorithm for English language. Later, he developed stemmers that covered Romance (French, Italian, Portuguese and Spanish), Germanic (Dutch and German) and Scandinavian languages (Danish, Norwegian and Swedish), as well as Finnish and Russian [51]. These stemmers were described in a very high level language known as Snowball. A number of statistical approaches have been developed for stemming. Significant works include: Goldsmith's unsupervised algorithm for learning morphology of a language based on the Minimum Description Length (MDL) framework [52]. Creutz uses probabilistic maximum a posteriori (MAP) formulation for unsupervised morpheme segmentation [54]. A few approaches are based on the application of Hidden Markov models [64] . In this technique, each word is considered to be composed of two parts "prefix" and "suffix". Here, HMM states are divided into two disjoint sets: Prefix state which generates the first part of the word and Suffix state which generates the last part of the word, if the word has a suffix. After a complete and trained HMM is available for a language, stemming can be performed directly. Plisson proposed the most widely accepted rule based approach for lemmatization [59]. It is based on the word endings, where suffixes are removed or added to get the normalized word form. Bart Jongejan [69]presented a method to automatically develop lemmatization rules to generate the lemma from the full form of a word and which can handle morphological changes in pre-, in- and suffixes alike. The lemmatizer was trained on Danish, Dutch, English, German, Greek, Icelandic, Norwegian, Polish, Slovene and Swedish full form-lemma pairs respectively. Lauri Karttunen [65]used a two level morphological analyser containing a large set of

morphophonemic rules. The work started in 1980 and the first implementation n LIST was available 3 years later. Tarek El-Shishtawy [62] proposed the first non statistical Arabic lemmatize algorithm . He makes use of different Arabic language knowledge resources to generate accurate lemma form and its relevant features that support IR purposes and a maximum accuracy of 94.8% is reported. Grzegorz Chrupala [56] presented a simple data-driven context-sensitive approach to lemmatizating word forms. Shortest Edit Script (SES) between reversed input and output strings is computed to achieve this task. An SES describes the transformations that have to be applied to the input string (word form) in order to convert it to the output string (lemma). As for lemmatizers for Indian languages, the earliest work by Ramanathan and Rao [57] used manually sorted suffix list and performed longest match stripping for building a Hindi stemmer. Majumdar et.al developed YASS: Yet Another Suffix Stripper[66]. Here conflation was viewed as a clustering problem with apriori unknown number of clusters. They suggested several distance measures rewarding long matching prefixes and penalizing early mismatches. In a work related to Affix Stacking languages like Marathi, [63] Finite State Machine (FSM) is used to develop a Marathi morphological Analyzer. In another approach, a Hindi Lemmatizer is proposed, where suffixes are stripped according to various rules and necessary addition of character(s) is/are done to get a proper root form [60]. P. Bhattacharya et al. [61] presented a human mediated lemmatizer based on the properties of a "trie" data structure which allows retrieving possible lemma of a given inflected word, with human help at critical steps.

Sense tagging is the task of tagging each word in the sentence with the correct sense of the word [70]. The availability of very large sense tagged corpora is a major resource for many natural language processing tasks. Yet, as of today, only few sense tagged corpora are publicly available. Petrolito and Bond [71] surveyed WordNet tagged corpora in terms of their accessibility and usefulness. Table 2.1 summarizes the corpora tagged with senses.

Among these sense tagged corpora listed in Table 2.1 some are explained below in further details. The English SemCor corpus is the first English corpus which is annotated with senses developed at Princeton University [72].The corpus consists of a subset of the Brown Corpus. A total of 700,000 words are available in the corpus and

more than 200,000 words have been tagged with senses. It is distributed Under the license of Princeton Wordnet

Table 2. 1: Corpora Tagged with Senses

Name of sense tagged corpus	Source Corpus	Total words	Total no.of words sense tagged	Language	Wordnet used as sense inventory	License	POS of words tagged with senses
English SemCor	Brown Corpus	700 k	200k all	English	WordNet 3.0	Wordnet	All
Jsemcor	English SemCor	380 k	58k	Japanese	Japanese WordNet	Wordnet	All
Multi SemCor	English SemCor	269 k	93k	Italian	MultiWordNet	CC BY 3.0	All
SemCor EnRo	English SemCor	176 k	48k	Romanian	BalkaNet	MSC . . .	All
BulSem Cor	English SemCor	101 k	99k	Bulgarian	BulNet	web only	All+
Eusem cor	English SemCor English SemCor	300 k	n/a	Basque	Basque WordNet	Web only	All
Spsem cor	English SemCor	850 k	23k	Spanish	Spanish WordNet 1.6	Web only	Noun, Verb
Dutch		500, 000	283k	Dutch	Cornetto	N/A	All

Semcor		k					
TüBa-D/Z Treebank	TüBa-D/Z Treebank8	1,365k	18k	Germany	GermaNet	None	Some, Noun, Verb
AQMAR Arabic SST	ArabicWikipedia articles	65k	32k	Arabic	WordNet	CC BY-SA 3.0	Noun, Verb
Hungarian WSD corpus	Hungarian National Corpus and Heti Világgazdaság (HVG) subcorpus	16k	5k	Hungarian	HuWN	None	Noun, Verb, Adjective
KPWr	Polish Corpus of Wroclaw University	438k	9k	Pol	Plwordnet	CC BY 3.0	Some
Gloss Corpus	Princeton WordNet Gloss Corpus	1621k	449k	English	WN 3.0	WordNet	Some
Groningen Meaning Bank	Groningen Meaning Bank (GMB) corpus	1020k	n/a	English	WN	None	All
DSO Corpus	Brown corpus and the Wall Street	n/a	193k	English	WN1.5	LDC	Noun, Verb

	Journal corpus						
Senseval 3		5k	2k	English	WN 1.7.1	None	All
Urdusense tagged corpus	CLE urdu digest corpus	100 k	17k	Urdu	Urdu WordNet	None	All

In the Princeton WordNet Gloss Corpus [71], the definitions (or glosses) of WordNet's synsets are manually linked to the context-appropriate sense in WordNet. Out of 1,621,12921 tokens available in the corpus 449,355 are sense tagged (330,499 manually + 118,856 automatically) on 656,066 taggable words and glosses (the tagged ones + 206,711 untagged). DSO corpus [73] consists of sentences from the Wall Street Journal corpus and the Brown corpus. This corpus includes about 192,800 instances of frequently used nouns (121) and verbs (70) of English. These occurrences have been manually tagged with WordNet 1.5 senses by 12 undergraduates from the Linguistics Program of the National University of Singapore. It is distributed on the Linguistic Data Consortium Catalogue23 (LDC) under different licences for LDC members and non-members.

Sense tagged corpora for languages other than English have also been developed. Japanese SemCor (JSEMCOR) [74] has been constructed by the method of annotation transfer. In this method, sense annotated corpus of source language is translated into the target language and sense annotations are also projected to the target language. Projection of sense is done using a WordNet in the target language. Target language WordNet is aligned with the source language WordNet which is used to sense tag the source language corpus. For the development of Japanese sense tagged corpus, English SemCor was used as the source corpus. Princeton (1.6) WordNet of English was the source WordNet and Japanese WordNet was the target WordNet. The final corpus contains 14,169 sentences with 150,555 content words of which 58,265 are tagged with senses. Japanese SemCor is also distributed under the License of Princeton WordNet.

DutchSemCor corpus [75] consists of superset of SoNaR, CGN, and the manually-selected web-snippets corpus. The source corpus were tokenised, part-of-speech tagged and lemmatised. Cornetto lexical database [76] was used as sense inventory for assigning senses to the instances in this corpus with a semi-automatic approach. In Dutch Semcor about 274,344 tokens for 2,874 lemmas manually tagged by two annotators with an inter annotator agreement (IAA) of 90%. About 132,666 tokens for 1,133 lemmas, manually annotated by a single annotator but agreeing with the WSD-system for IAA 44, 47,797,684 tokens have been tagged automatically by 3 WSD systems.

Similarly, Bulgarian Brown corpus used for the development of Bulgarian sense tagged corpus [77]. The corpus contains 500 excerpts each containing 100+words: total 63 440 words are available in the source corpus. The words were lemmatised and tagged with part of speech. Bulgarian WordNet [78] was used to assign the senses to the words in BulSemCor. The Bulgarian sense-annotated corpus contains total 45 562 semantically annotated words. Out of which 40,255 are single words and 2,177 are multi-word expressions. For the development of Urdu sense tagged corpus [79], CLE Urdu digest corpus has been used. The corpus consists of 100 k words annotated with part of speech tag. Other resources used to develop Urdu sense tagged corpus were Urdu wordlist, Urdu morphological analyser and Urdu WordNet. Urdu wordlist consists of 5000 high frequency content words. Version 1.0 of Urdu sense tagged corpus included 17006 sense tagged words with 2285 unique senses. An attempt for developing Chinese sense tagged corpus has been made [80]. Three components used for developing sense tagged corpus for Chinese language were a corpus, a lexicon and the linking between the lexicon and the Chinese Dictionary. The lexicon contains the description of 813 nouns and 132 verbs and 60,895 word instances have been tagged. This corpus consists of texts from People's Daily Newspaper. It is an official daily newspaper of Government of China.