Chapter 1

Introduction

The thesis investigates the applications of machine learning techniques to Word Sense Disambiguation (WSD).

It is one of the most amazing and perplexing feats of human mind that we understand written and spoken communication in spite of enormous number of possibilities that exist because of multiple meanings of words that compose a sentence. Human beings can produce a correct sentence choosing words in their appropriate context with little effort. But the same problem becomes very hard and complex when it is sought to be automated. Therefore any system that proposes to implement Natural Language Processing (NLP) on a computer has to address very seriously the question of WSD. WSD can be said to be the Holy Grail of NLP in the sense that if it is solved other hard problems of NLP such as Machine Translation (MT), Information Retrieval (IR) , Question-Answering etc. can be successfully tackled to a considerable degree. Formally WSD is the task to examine word tokens in context and specify exactly which sense of each word is being used.

A WordNet[2][9] is used to obtain the senses of words for disambiguation. WordNet is a machine readable lexical database and it is a key resource that aids in several NLP tasks. The salient feature of WordNet is its emphasis on organizing lexical information in terms of word meanings/concepts in lieu of word forms. A key feature of WordNet is the concept of Synset where words are grouped according to similarity of their meanings. Synsets are linked to one another through well known lexical and semantic relations.

There are broadly three approaches to WSD. These are :- 1) Knowledge based approach, 2) Machine Learning based Approach and 3) Hybrid Approach. The

Knowledge based approaches acquire knowledge from manually created lexical resources like dictionary definitions[14]. Notable Knowledge based approaches are semantic relations[36],conceptual density methods[19], Walker algorithm[15] etc. Knowledge based approaches may use grammar rules as well as hand coded rules for disambiguation. In case of machine learning approach, systems are trained to perform the task of WSD. In these approaches, what is learned is a classifier which can be used to assign unseen examples to one of a fixed number of senses. The one thing that is common to machine learning approaches is emphasis on acquiring the knowledge needed for the task from the data itself rather than from human analysis. The machine learning approaches vary as to the nature of training material, how much material is needed , the degree of human intervention , the kind of linguistic knowledge used and the output produced. The Knowledge base for a hybrid approach is diverse in the sense that it obtains its knowledge from both the WordNet and the corpus.

In this thesis a hybrid WSD approach based algorithm has been proposed where the algorithm combinesCorpus parameters, WordNet parameters and a parameter obtained from Overlap based Approach to WSD.The experiments for this algorithm were done on two domains namely Tourism and Health . The languages for the texts in the corpora of these two domains were Bengali and Nepali.

Bengali or Bangla along with two other cognate languages, Assamese and Oriya, as well as Magadhi, Maithili and Bhojpuri in south-east zone forms a linguistic group. Their immediate source can be traced back to the Magadhi Prakrit or Eastern Prakrit which was brought to this area from Magadh (or Bihar) and the language of Gauda-Banga with other eastern languages developed from this through Magadh Apabhramsa[85]. Genetically Bengali is derived from Indo-Aryan (IA) or the Indic sub-branch of the Indo-Iranian branch of the Indo-European (IE) family of languages. It is a language native to Indian states of West Bengal, Tripura, southern part of the Indian state of Assam and Bangladesh. Bangla is the national language in Bangladesh and second most spoken language in India. With about 250 million native[83] and about 300 million total speakers worldwide, it is the seventh most spoken language in the world by total number of native speakers and the tenth most spoken language by total number of speakers[84].The Bengali script evolved from the Siddham, which belongs to the Brahmic scripts[85].

Nepali language is, by its characteristics, of Indo-Aryan origin and is spoken by approximately 45 million people in Nepal, where it is the official language and to a

large extent the medium of education. It is also spoken in parts of neighbouring countries (India, Bhutan and Myanmar). Nepali is written in the Devanagari alphabet. It is written phonetically, that is, the sounds correspond almost exactly to the written letters. Nepali has many loanwords from Arabic and Persian languages, as well as it has some Hindi and English borrowings[5].

1.1 Motivation

Most of the Indian languages are resource poor. NLP systems are few and far between in most of the Indian languages. With the successful completion of the Indo WordNet project, a basic infrastructure or a scaffolding has been provided for building advanced NLP systems like WSD system, Machine Translation(MT) system etc. Our primary motivation was to exploit the resources in IndoWordnet, primarily the Bengali WordNet and the Nepali WordNet, to build a WSD system which acquires the knowledge needed for disambiguation not only from the corpus but also from the WordNet. We had also noticed in our studies of WSD that Supervised WSD has a performance level which is higher than Unsupervised WSD. But Supervised WSD requires sense annotation during training which is a costly affair whereas Unsupervised WSD requires no sense annotation. Therefore another motivating factor was to design an algorithm whose knowledge base would be diverse (both from the corpora as well as WordNet) and which would also require less amount of sense annotation.

1.2 Objectives

The primary objectives of this thesis can be summarized as follows:

- To study and investigate important WSD techniques and algorithms.
- To study WordNet and different WordNet relations like Synonymy, Antonymy, Hypernymy-Hyponymy, Holonymy-Meronymy, Entailment, Troponymyetc
- To study the trie data structure and develop a lemmatizer for generating root words from morphed word forms in both Bengali and Nepali languages.

- To develop an efficient hybrid WSD algorithm whose knowledge base would be diverse (both from the corpora as well as WordNet) and which would also require less amount of sense annotation.

1.3 Methodology

- First extensive study was done on the various approaches to WSD and the corresponding standard algorithms.
- Pros and cons of various approaches and the pros and cons of various algorithms related to those approaches were also studied in depth
- Study of WordNet, IndoWordNet and different WordNet relations like Synonymy, Antonymy, Hypernymy-Hyponymy, Holonymy-Meronymy, Entailment, Troponymy were carried out.
- Datasets for two domains namely Tourism and Health in Bengali and Nepali languages were obtained from Technology Development for Indian Languages (TDIL), Ministry of Communications & Information Technology, Govt of India.
- The trie data structure was studied in depth and a lemmatizer for generating root words in Bengali and Nepali was built based on the trie data structure
- A new hybrid WSD technique was then proposed which derives its knowledge base from diverse sources such as the corpora as well as WordNet and which also requires less amount of sense annotation.

1.4 Contributions

- A systematic survey of WSD approaches and techniques, WordNet and its various relations, Indo WordNet and lemmatizers.
- A Hybrid algorithm for WSD experimented on 2 domains, namely, Tourism and Health.
- Encouraging result from experiments that if sense marked data from one Domain (for e.g. A) is available, then a WSD system can be built for another Domain(for e.g. B) by infusing some amount of data from Domain-B into Domain-A.

1.5 Thesis Outline

Chapter 2 makes a comprehensive survey of various WordNet approaches and methods. It also does a survey of various WordNet creation projects, lemmatizers and sense marked corpora.

Chapter 3 does a detailed study of WordNet, constituent elements of WordNet, semantical and lexical relations of WordNet, Then there is a discussion on Indo WordNet which is a linked lexical knowledge base of WordNet of eighteenof 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. The chapter concludes with a discussion on a Sense Marking tool for the languages constituting Indo WordNet and the issues that cropped up when using the Sense Marker tool.

Chapter 4 does a detailed study of various approaches to WSD and the methods and algorithms that embed these approaches. A comparative study of the algorithms related to various approaches to WSD was done vis-à-vis their efficiency, the dataset or the corpus on which the experiments were carried out and the reported baseline. The chapter ends with a discussion on the Performance metrics of a WSD system.

Chapter 5 makes a study of the various approaches to Lemmatization and introduces a lemmatizer based on trie data structure to generate the root words from morphed words in a corpus. The chapter ends with a brief discussion on Hopfield Network and energy function of a Hopfield Network.

Chapter 6 introduces the proposed hybrid algorithm Hy_WSDwhich combines information from multiple knowledge sources namely WordNet and Corpus. It uses a scoring function adapted from the energy function of a Hopfield Net which combines Corpus parameters, WordNet parameters and a parameter from Overlap based Approach to WSD to score the competing senses of a target polysemous word. The score with the highest value is the winner sense. The algorithm is experimented on three domains, namely, Tourism, Health and Mixed where Mixed domain consists of

texts belonging to a wide variety of subjects- short stories, abridged novels, newspaper articles, sports topics etc. The languages for the corpora were Bengali and Nepali.

Chapter 7 concludes the thesis. It summarises the findings of this thesis and points out contributions and future research directions.