Abstract

It is one of the most amazing 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. It is equally amazing that we produce a correct sentence choosing words in their appropriate context. But when a computer has to produce the correct sense of a word in a sentence from among a number of competing senses , the problem does not remain as easy as it was in the case of human mind. To the contrary it becomes an absolutely critical problem. Formally, Word Sense Disambiguation (WSD) is a process to obtain the sense of target words or all words (All word WSD more difficult) against a sense repository like the WordNet or a Thesaurus using the context in which the word appears. Word Sense Disambiguation (WSD) is one of the cardinal and absolutely critical problems of Natural Language Processing (NLP). In fact WSD is often referred to as 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.

There are several ways to tackle WSD. Broadly they are referred to as the Knowledge based Approaches, Machine Learning based approaches and Hybrid Approaches. These approaches have their advantages and disadvantages which have been enumerated , at length , in this thesis.

The focus of this thesis is on a Hybrid mechanism to tackle the problem of WSD by combining, in parts, the three approaches namely Overlap, Supervised and Unsupervised approach. The hybrid mechanism combines information from multiple knowledge sources and uses those information to predict the correct sense from a number of competing senses of a target polysemous word by using a scoring function. While predicting the sense of a target word, the Hybrid Algorithm developed in this thesis takes into consideration the semantic relations of the WordNet and the information garnered from implementing a knowledge based algorithm on the

collocation vector of the target word. Our experiments were done on the Bengali and Nepali corpus of two specific domains, namely, Tourism and Health. The corpus was obtained from Technology Development of Indian Languages (TDIL), a project of Ministry of Communications and Information Technology, Govt of India. Our experiments have shown encouraging results compared to the other algorithms in this area.