Gujarati Word Sense Disambiguation using Genetic Algorithm

Zankhana B. Vaishnav

Natural language processing is related to human-computer interaction, where several challenges involve natural language understanding. Word sense disambiguation problem (WSD) consists in the computational assignment of a meaning to a word according to a particular context in which it occurs. A word can have number of senses, which is termed as ambiguity. This word sense disambiguation is an intermediate task, but rather is necessary at one level to accomplish most natural language processing tasks.

Genetic Algorithms (GAs) have widely been investigated to solve hard optimization problems, including the Word Sense Disambiguation (WSD). Several approaches have been investigated for WSD in English, French, German and some Indo-Aryan languages like Hindi, Marathi, Malayalam, etc. However, research on WSD in Guajarati Language is relatively limited. In this research, an approach for Guajarati WSD using Genetic algorithm has been proposed which uses Knowledge based approach where Indo-Aryan WordNet for Guajarati is used as lexical database for WSD.

The proposed approach GWSD using Genetic algorithm encompasses many phases. The first one is pre-processing, where a text T is transformed into a bag of words in a pre-processing stage including operations such as tokenization, stop word removal and stemming. In the second phase, the lexical database of Gujarati WordNet is considered. In the third phase, genetic algorithm is applied on the domain. The basic goal of the research is to design architecture of genetic algorithm based system for word sense disambiguation. Dedicated and application specific generic operators such as crossover, mutation and selection are also to be designed with customized fitness functions to evaluate populations. The architecture is then experimented using the database available at WordNet. The architecture can also be extended for Indo-Aryan Languages where the Indo-Aryan WordNet for different languages can be used to disambiguate the ambiguous word.