Syntactic Tagging of Tamil Corpus

K.Rajan,
(Department of Computer Technology,
Muthiah Polytechnic College, Annamalainagar, Tamilnadu, India).
M.Ganesan,
(CAS in Linguistics, Annamalai University, Annamalainagar) and .
V.Ramalingam
(Department of Computer Science and Engineering,
Annamalai University, Annamalainagar, Tamilnadu, India)

Abstract

Language is one of the fundamental aspects of human behaviour and is crucial component of our lives. Computational models of language are useful for scientific purposes for exploring the nature of linguistics communications- and for practical purposes- for enabling effective human-machine communication [J.Allen.1995 P.1]. The goal of these computational models of language is to develop computer programs that could perform various tasks involving natural language. Natural language interfaces to computer allow complex systems to be accessible to everyone. The user is expected to interact in natural language. A natural language system must use considerable knowledge about the structure of the language, the possible semantics, the goals of the user, and the great deal of general world language. The syntactic structure deals with syntactic relations among the words in the sentence , where as the semantic structures deals with the meaning of the sentence.

Studies of language usage, focus on a particular linguistic structure, investigating the ways in which similar structures occur in different contexts and different functions. Computer corpora are bodies of natural language material which are stored in machine readable form. Corpus can be used to provide more useful information on sentences, words, morphemes, etc. It could be achieved by adding linguistic information to the text in the corpus. The practice of adding interpretative information to an existing corpus of written language is called annotation or **tagging**. The tagging can be done at different levels.

Hand coding of corpus is a laborious, error prone, and time-consuming task, and here are a number of advantages, especially regarding speed and consistency, with developing ways to perform the tagging, as far as possible, automatically.

While we are far from being able to do automatic tagging on the level of semantics or pragmatics. The current state of the art of automatic tagging allows part-of-speech tagging with some success. Consequently, this paper will concentrate on automatic tagging methods applied to phrases and clauses. We used, for our research, the corpus developed at CIIL, Mysore, India.

This paper , *"SYNTACTIC TAGGING OF TAMIL CORPUS"* , presents the methods to identify the syntactic structure of Tamil sentences and to develop appropriate algorithm to label the grammatical categories at phrase/clause and sentence level. Also discusses, how this structure is directly used in applications such as grammar checking in word–processing systems; automatic translation of natural language text; question answering systems; information retrieval systems; lexicographic applications for building dictionaries, thesaurus, etc.

Keywords: Corpus Analysis, Syntactic Tagging, Tamil text analysis, Automatic tagging.