IT.ENG.3

SETRANS: A MACHINE TRANSLATOR FROM ENGLISH TO SINHALA

H. M. P. U. Herath, M. Z. Junaideen, D. Elkaduwe

Department of Computer Engineering,
Faculty of Engineering, University of Peradeniya

We present a syntax-based language model for natural language processing needs for Sinhala language. We specifically present an English to Sinhala translation scheme which makes use of the Stanford parser which is a free and open source software to parse the English sentence. The Sinhala language is not still properly analyzed and neither is data-gathered for a statistical approach whereas data on English is extensively studied and databases on part-of-speech tagged information are readily available. Therefore, we chose to use the existing software to parse English sentences and use the rule-based approach to generate the Sinhala translation of an English sentence.

The first phase of the translation scheme is the parsing of the English sentence in the English parser. We used the Stanford parser for this step. Stanford parser is a software that makes use of statistical methods to tag parts of a speech and to generate the most probable parsed tree.

The generated tree is then traversed level by level, and at each level, the part-of-speech sequence is matched against a database to come up with a possible reordering for the given sequence.

This reordered tree is then traversed to replace the English words with the most probable Sinhala translation. Either each word is translated individually or a phrase can be translated directly into a Sinhala phrase or a word. Unlike statistical translation methodology, these phrases are limited to the syntactic phrases that are usually recognized in usual natural language grammar entities.

The bilingual dictionary used to translate consists of a set of words, their part-of-speech tags and their Sinhala translated word or a phrase of words that gives the translation. It might also contain a set of English phrases, along with Sinhala translated word or a phrase of words that gives the correct translation.

Ambiguity problem is one problem yet to be resolved in this research. In natural languages, the same word is used with different meanings. This is called the ambiguity of the language and is problematic when it comes to machine translation.