Machine Translation as related to Tamil

B Kumara Shanmugam
AU-KBC Research Centre, M.I.T.,
Anna University, Chennai, India.

Abstract

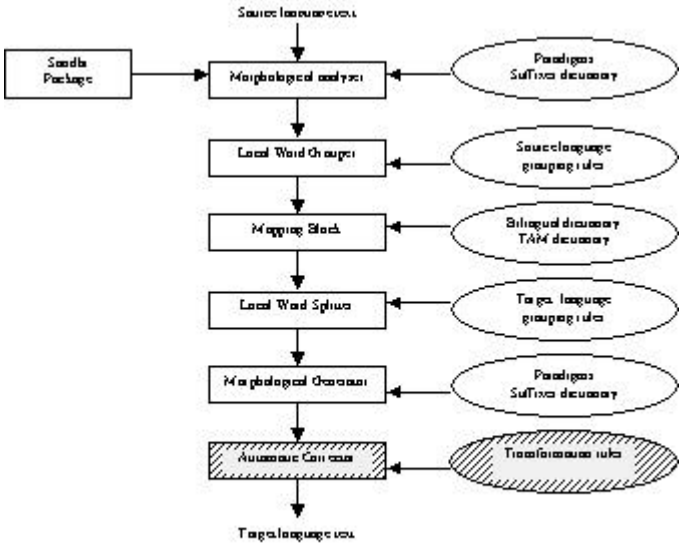
Machine Translation is a key technology in Language Processing that can play a lead role in information access and communication. As India is a country with eighteen official languages, machine translation applications have immense potential in the Indian market. Tamil is a widely used language both in India and outside India. Building translation system from and to Tamil helps the Tamil community all over the world in accessing the information in Tamil. In this presentation we will discuss the resources and tools that has to developed and the complexities involved in building translation systems involving Tamil. We will also discuss the applications of MT systems.

Machine Translation is a complex system that does different kinds of processing over the text at various levels. At the source language side we need tools like morphological analyser, POS tagger and Parser. The morphological analyser splits the words into meaningful units, the POS tagger assigns the right part-of-speech for the word and the Parser gives us the syntactic structure of the sentence. To develop these, we require dictionary, morphological rules and syntax rules. While rule based can be used reflecting the knowledge of the language community, it has been often noticed that such knowledge driven methods encounter problems with real text and that data driven approaches have usually outperformed. To enable the development of such data driven methods, large corpus both raw and annotated are required.

After the analysis of the source language input text, it must be transformed into a form appropriate for generation of text in the target language. This requires a transfer component as well as a target language sentence generation tool. The development of these tools requires aligned parallel corpus of the source and target languages. We also need bilingual dictionaries to do lexical transfer. While doing a lexical transfer for words with multiple senses, we might require a word sense disambiguation tool for disambiguating the sense. The discussion will focus on the special aspects of Machine Translation in Tamil.

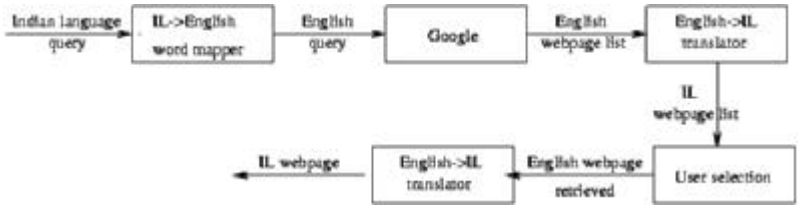
In contrast to English and other languages, Indian languages and Tamil more specifically, is rich in inflectional and derivational morphology that increases the complexity of morphological analysis. Developing a parser for Tamil is also non-trivial because of its free word order property. As in other languages Tamil also contains polysemous words that requires a collocation dictionary to disambiguate them.

After having discussed the various components of an MT system, and the resources that might be needed to build such tools, we will turn our focus on the discussion of an Tamil – Hindi Machine Aided Translation system. We will first discuss the methodology adopted in the system and then the current state of the system as well as the various components and the planned future enhancements.



Tamil – Hindi Machine Aided Translation

In conclusion, we would like to remark that the MT systems have tremendous potential for meaningful applications. Machine Translation systems would be very much useful to the rural masses in accessing the web content in their local language (This is the focus of "Web Bharati", a collaborative project we are planning to embark on with special emphasis on Tamil).



WEB BHARATI
An Application for Accessing Web Content in Indian Languages

In the communication domain it could be useful in translating Government documents, Business documents, Emails and so on. In the process of building the technology for Machine Translation we can get many spin-offs from that like Morph Analyser, POS tagger, parser, Word Sense Disambiguator etc. These tools could be useful in developing a variety of other NLP applications like Word Processor with spell checker and grammar checker, Information Retrieval, Text summarization & Categorization, Information Extraction and so on.