COMPUTATIONAL LINGUISTICS IN INDIA: AN OVERVIEW

Akshar Bharati, Vineet Chaitanya, Rajeev Sangal
Language Technologies Research Centre
Indian Institute of Information Technology, Hyderabad
{sangal,vc}@iiit.net

1. Introduction

Computational linguistics activities in India are being carried out at many institutions. The activities are centred around development of machine translation systems and lexical resources.

2. Machine Translation

Four major efforts on machine translation in India are presented below. The first one is from one Indian language to another, the next three are from English to Hindi.

2.1. Anusaaraka Systems among Indian languages

In the anusaaraka systems, the load between the human reader and the machine is divided as follows: language-based analysis of the text is carried out by the machine, and knowledge-based analysis or interpretation is left to the reader. The machine uses a dictionary and grammar rules, to produce the output. Most importantly, it does not use world knowledge to interpret (or disambiguate), as it is an error prone task and involves guessing or inferring based on knowledge other than the text. Anusaaraka aims for perfect "information preservation". We relax the requirement that the output be grammatical. In fact, anusaaraka output follows the grammar of the source language (where the grammar rules differ, and cannot be applied with 100 percent confidence). This requires that the reader undergo a short training to read and understand the output.

Among Indian languages, which share vocabulary, grammar, pragmatics, etc. the task (and the training) is easier. For example, words in a language are ambiguous, but if the two languages are close, one is likely to find a one to one correspondence between words such that the meaning is carried across from the source language to target language. For example, for 80 percent of the Kannada words in the anusaaraka dictionary of 30,000 root words, there is a single equivalend Hindi word which covers the senses of the original Kannada word. Similarly, wherever the two languages differ in grammatical constructions, either an existing construction in the target language which expresses the same meaning is used, or a new construction is invented (or an old construction used with some special notation). For example, adjectival participial phrases in the south Indian languages are mapped to relative clauses in Hindi with the '*' notation (Bharati, 2000). Similarly, existing words in the target language may be given wider or narrower meaning (Narayana, 1994). Anusaarakas are available for use as email servers (anusaaraka, URL).

2.2. Mantra System

The Mantra system translates appointment letters in government from English to Hindi. It is based on synchronous Tree Adjoining Grammar and uses tree-transfer for translating from English to Hindi.

The system is tailored to deal with its narrow subject-domain. The grammar is specially designed to accept analyze and generate sentential constructions in "officialese". Similarly, the lexicon is suitably restricted to deal with meanings of English words as used in its subject-domain. The system is ready for use in its domain.

2.3. MaTra System

The Matra system is a tool for human aided machine translation from English to Hindi for news stories. It has a text categorisation component at the front, which determines the type of news story (political, terrorism, economic, etc.) before operating on the given story. Depending on the type of news, it uses an appropriate dictionary. For example, the word 'party' is usually a 'politicalentity' and not a 'social event', in political news.

The text categorisation component uses word-vectors and is easily trainable from pre-categorized news corpus. The parser tries to identify chunks (such as noun phrases, verb groups) but does not attempt to join them together. It requires considerable human assistance in analysing the input. Another novel component of the system is that given a complex English sentence, it breaks it up into simpler sentences, which are then analysed and used to generate Hindi. The system is under development and expected to be ready for use soon (Rao, 1998).

2.4. Anusaaraka System from English to Hindi

The English to Hindi anusaaraka system follows the basic principles of information preservation. It uses XTAG based super tagger and light dependency analyzer developed at University of Pennsylvania [Joshi, 94] for performing the analysis of the given English text. It distributes the load on man and machine in novel ways. The system produces several outputs corresponding to a given input. The simplest possible (and the most robust) output is based on the machine taking the load of lexicon, and leaving the load of syntax on man. Output based on the most detailed analysis of the English input text, uses a full parser and a bilingual dictionary. The parsing system is based on XTAG (consisting of super tagger and parser) wherein we have modified them for the task at hand. A user may read the output produced after the full analysis, but when he finds that the system has "obviously" gone wrong or failed to produce the output, he can always switch to a simpler output.

3. Corpora and Lexical Resources

3.1 Corpora for Indian Languages

Text Corpora for 12 Indian languages has been prepared with funding from Ministry of Information Technology, Govt. of India. Each corpus is of about 3-million words, consisting of randomly chosen text-pieces published from 1970 to 1980. The texts are categorized into: literature (novel, short story), science, social science, mass media etc. The corpus can be used remotely over the net or obtained on CDs (Corpora, URL).

3.2 Lexical Resources

A number of bilingual dictionaries among Indian languages have been developed for the purpose of machine translation, and are available "freely" under GPL. Collaborative creation of a very large English to Hindi lexical resource is underway. As a first step, dictionary with 25000 entries with example sentences illustrating each different sense of a word, has been released on the web (Dictionary, URL). Currently work is going on to refine it and to add contextual information for use in the anusaaraka system, by involving volunteers.

4. Linguistic Tools and Others

4.1. Morphological Analyzers

Morphological analyzers for 6 Indian languages developed as part of Anusaaraka systems are available for download and use (Anusaaraka,URL). Sanskrit morphological analyzers have been developed with reasonable coverage based on the Paninian theory by Ramanujan and Melkote.

4.2 Parsers

Besides the parsers mentioned above, a parsing formalism called UCSG identifies clause boundaries without using sub-categorization information.

4.3 others

Some work has also started on building search engines. However, missing are the terminological databases and thesauri. Spelling checkers are available for many languages. There is substantial work based on alternative theoretical models of language analysis. Most of this work is based on Paninian model (Bharati, 1995).

5. Conclusions

In conclusion, there is a large computational linguistic activity in Indian languages, mainly centred around machine translation and lexical resources. Most recently, a number of new projects have been started for Indian languages with Govt. funding, and are getting off the ground.

References:

Anusaaraka URL: http://www.iiit.net, http://www.tdil.gov.in

Bharati, Akshar, and Vineet Chaitanya and Rajeev Sangal, Natural Language Processing: A Paninian Perspective, Prentice-Hall of India, New Delhi, 1995,

Bharati, Akshar, et.al, Anusaaraka: Overcoming the Language Barrier in India, To appear in "Anuvad". (Available from anusaaraka URL.)

CDAC URL: http://www.cdac.org.in
Corpora URL: http://www.iiit.net
Dictionary URL: http://www.iiit.net

Narayana, V. N, Anusarak: A Device to Overcome the Language Barrier, PhD thesis, Dept. of CSE, IITKanpur, January 1994.

Rao, Durgesh, Pushpak Bhattacharya and Radhika Mamidi, "Natural Language Generation for English to Hindi Human-Aided Machine Translation", pp. 179-189, in KBCS-98, NCST, Mumbai.

Joshi, A.K. Tree Adjoining Grammar, In D. Dowty et.al. (eds.) Natural Language Parsing, Cambridge University Press, 1985.

Joshi, AK and Srinivas, B., Disambignation of Supertags: Almost Parsing, COLING, 1994.