Computer Assisted Translation System- An Indian Perspective

HEMANT DARBARI
Applied Artificial Intelligence Group
Center for Development of Advanced Computing
Pune University Campus, PUNE - 411 007 (INDIA)
darbari@cdac.ernet.in

Abstract

Work in the area of Machine Translation has been going on for several decades and it was only during the early 90s that a promising translation technology began to emerge with advanced researches in the field of Artificial Intelligence and Computational Linguistics. This held the promise of successfully developing usable Machine Translation Systems in certain well-defined domains. C-DAC took up this challenge, as we felt that India, being a multi-lingual and multi-cultural country with a population of approximately 950 million people and 18 constitutionally recognized languages, needs a translation system for instant transfer of information and knowledge.

The other groups who are working in this area of English to Hindi Translation are National Center for Software Technology (NCST), who are working on translation of News Stories and Electronics Research & Development Center of India (ER&DCI). who have developed the Machine Assisted Translation System for the Health Domain. A major project on Indian Languages to Indian Languages Translation (Anusaaraka) is also under development at University of Hyderabad.

1. Introduction

Language processing demands intelligence – artificial or otherwise. This may seem like a truism, but in fact it is difficult to convince oneself that something as easy as talking and listening demands very high computational sophistication. Just like seeing, using language seems effortless.

2. MANTRA-MAchiNe assisted TRAnslation tool

Center for Development of Advanced Computing (C-DAC) is a premier National Institute of the Department of Electronics (DOE), Government of India. C-DAC is committed to design, develop and deliver Advanced Computing Solutions for Human Advancement. The Applied Artificial Intelligence (AAI) Group at C-DAC is working on some of the fundamental applications in the field of Natural Language Processing, Machine Translation, Intelligent Language Teaching and Decision Support Systems. MANTRA is one of its major achievements. This system translates the text from English to Hindi in the domain of Personnel Administration.

The motivation for taking up this challenge was that in order to achieve national unity and integration in the face of the linguistic and cultural diversity, the founding fathers of our constitution had identified Hindi as the Official Language of the Indian Union. According to the Official Language Act, all Central Government communications have to be made simultaneously available both in Hindi and English, as English continues to be the associate official language. Accordingly the bulk of official business is initiated and conducted in English. Presently, the translation work is executed manually by a large network of translators positioned in all Government Departments and Public Sector Undertakings. However, the translators find it difficult to cope with the massive translation requirement leading to inordinate delays.

In order to overcome this problem, an early initiative was taken by the AAI group when it received funds from DOE and United Nations Development Program (UNDP) under the program ‘Knowledge Based Computer System’. We started exploring possibilities in Natural Language Processing and two parsers were developed using the Augmented Transition Network (ATN) and Tree Adjoining Grammar (TAG) formalisms. We compared their suitability for three areas namely Natural Language

Understanding, Natural Language User Interfaces and Machine Translation.

2.1 Approach

The strategy adopted in our translation system is NOT Word to Word NOR Rule to Rule BUT Lexical Tree to Lexical Tree.

Having built a TAG parser (VYAKARTA) that could handle English, Hindi, Gujarati, Sanskrit and German, we scouted for a relevant application. Translation in the Indian context was a more pressing concern. We, therefore, chose English-Hindi pair in the domain of Official Language, used in Central Government Departments, as the first real life application. Accordingly, a prototype translation system was decided upon, built and progressively refined, which was named MANTRA.

While initiating the MANTRA project we were aware that the English-Hindi language pair we had chosen for translation belonged to two different language families and, therefore, were dissimilar in structure and style which would pose altogether different kinds of problems and challenges. Hence we had to evolve some innovative computational and grammatical solutions.

The VYAKARTA an Integrated Parser for Natural Languages uses the Tree Adjoining Grammar (TAG) formalism. It simultaneously creates the syntactic parse and the functional description of the given sentences. Unification based feature checks are also done. The parser uses Early Style Bottom Up Parsing Technique, which can be viewed as processing elements in a parallel-processing paradigm. VYAKARTA uses the sub-language concept for its definition and can handle fairly complex sentence structures at a near real time on a IBM-Pentium machines.

The output of the parser is a derivation tree. The derivation tree is send to the Generator, which generates the Hindi output sentence with the help of English Lexicon. Hindi Lexicon & Transfer Lexicon.

For creating the Grammar an efficient tool is developed which is called KOSHAKAR in which a windowing system was developed which presents the relevant information in an easy to use and interactive way. The entry of the trees associated with each word into the lexicon was facilitated by a graphical tree acquisition module, which helps the user to visualize the tree in the pretty print form. This module also allows easy editing of the trees.

MANTRA was demonstrated to the Department of Official Language (DOL), Government of India and several other organizations and institutions. Consequently DOL sponsored a project entitled “Computer Assisted Translation System for Administrative Purposes” in 1996. The specific domain chosen for this purpose was the Gazette Notifications on appointments in the Government of India. The domain was significant because as all Government Orders and Notifications become the legal documents for compliance from the date of publication in the Gazette of India. This package is named MANTRA-Rajbhasha.

In this endeavor, all our efforts were directed towards two major goals: (a) accuracy of translation and (b) speed. Accuracy-wise, we had to create smart tools for handling transfer grammar and translation standards including equivalent words, expressions, phrases and styles in the target language. A lot of effort was put in to optimize the grammar with a view to obtaining a single correct parse and hence a single translated output. Speed-wise, we had to make innovative use of corpus analysis, alter the parsing algorithm, design efficient Data Structure and introduce run-time frequency-based rearrangement of the grammar, which substantially reduced the parsing and generation time.

Therefore the overall objectives of MANTRA-Rajbhasha, which we set before us, were:

- Instant dissemination of knowledge and information through on-line translation.
- Standardization and uniformity in the use of translation equivalents, expressions and styles.
- Increasing the efficiency of translation by providing maximum utilities and user friendly tools used in the translation like on-line Dictionary and Thesaurus and dynamic expansion of lexicon by the user.
- To help the Government bodies to execute and promote Official Language through the help of the modern IT
- To provide the translation facilities through all the three solutions: desktop, network and Web-based translation system to be installed in various ministries and departments.

The results of MANTRA-Rajbhasha have been extensively field tested and evaluated by experts and users. The accuracy of translation has been adjudged as over 95% within the specified domain.

While developing MANTRA we did not confine ourselves to the short-term objective of developing a working model but we had the vision of its enormous potentialities and its capability to expand and penetrate fully in the society supported by the state-of-the-art technological advancements.

Web-based Technology is fast becoming a very important part of our social communication network and an integral part of our lives. Dissemination of information through this channel can be a very effective tool for changing the socio-economic life of the common man in the society. With the state-of-art Internet Technology available today it is possible to reach the masses by providing them the required information on any topic of their interest and practical use in their own regional languages. It will enable the technology to reach their homes instead of their reaching the technology.

The new IT policy of the Government of India propose that the facility of Internet and web-site visit will be made available to the common man including the farmers all over the country. New advances in the various fields of agriculture are taking place very fast of which our farmers are unaware. In order to maintain the productivity level and quality of the food grains our farmers must remain abreast with the latest information on the new brands of seeds, new methods of cultivation and disease control, new products and other related areas. When implemented all the web-site information will be available to him in English, which obviously he would not be able to understand.

We are aiming at making this information available to him in his own language, in the present case Hindi, which will provide the Internet user the facility of obtaining an instant on-line translation, translation on mirror site or annotations in Hindi depending upon the nature and relevance of the information structure and content. This will not only create an interest in the farmer and rural community in the Internet but also make him computer literate. To begin with, rural community centers, Block Development Officers BDOs' or other specified centers can act as catalysts or platforms where semi-literate and as a source of information. The proper use of this technology can change the Agriculture face of the country. We can yield a better quality crop, which can stand in International market.

Information (documents) available on Internet is generally available in the form of HTML (Hyper Text Markup Language) pages written in English language. The idea here is to select a most frequently used Web-pages/HTML document from some web site, using some Internet browser (Internet Explorer, Netscape etc) save it in your computer and submit it to the Translation Server. The document will be translated into Hindi and sent back to the sender as an HTML page.

We are looking forward to expand MANTRA to various other domains including Banking sector Corporate Web Sites and to other Indian Languages.

3. Human-Aided Machine Translation from English to Hindi for News Stories

This project aimed at developing a practical system for human-aided machine translation of news summaries from English to Hindi. This application is both challenging and practically relevant. Most of the news generated in India is in English. Local language newspapers face the difficult task of translating this news manually within strict time constraints, and are therefore forced to use only a fraction of what they get. A translation system that can increase the productivity of news translators would be an important practical contribution. This project is being conducted at the National Center for Software Technology (NCST), and is funded by the Technology Development in India-Languages (TDIL) initiative of the Department of Electronics (DoE), Government of India.

Translating news from English to Hindi is challenging for the following reasons:

- News can be about any topic, so the domain is very vast
- News tends to involve long, compound-complex sentences

3.1 Approach

NCST is taking an pragmatic and engineering approach to a very hard problem. This involves using an intuitive user interface and taking advantage of man-machine synergy to tackle some of the hard problems of NLP. A number of technologies have been developed under the project to simplify the task. Some of these technologies and subsystems are briefly described below.

3.2 Technologies and subsystems developed

- **Automatic Categorization of News Items**

This is a subsystem that automatically classifies a news item into a pre-defined set of categories and sub-categories using a combination of statistical and knowledge-based methods.

Example:
knowing that a story is a political story helps us disambiguate the sense of the word ‘party’ as ‘political entity’ and not ‘social event’.

- **Semi-automatic Simplification of Complex Sentences**

This is a rule-based subsystem that simplifies some typical complex news sentences into simpler ones, to make the rest of the translation task easier. Example:
“Five people, including three women, were killed in an accident near Mumbai”
becomes
“Five people were killed in an accident near Mumbai. They included three women”, which is easier to translate.

- **Special Phrase Recognition**

This is a knowledge-based subsystem that recognizes time and place phrases. Phrases are recognized using a combination of grammar rules and knowledge-base lookup and then tagged as time, name or place phrases. Phrases that are recognized and tagged properly include phrases like “near Bangalore”, “the day before yesterday”, as well as complex phrases such as “the Kang Rian village under the Phillaur police station”.

- **Interactive Structure Editor**

This is a core subsystem that creates the internal representation of the English input, with the help of the user, and an intuitive GUI. The user does not need to know English grammar, but has to visually validate information about which part of the input sentence talks about which other part, represented in a hierarchical slot-value structure using simple slot names such as ‘who’, ‘what’ and ‘more-info’.

- **Hindi Generation**

This subsystem takes the structured internal representation created by the interactive structure editor, and performs structural, syntactic and morphological transfer of the input using knowledge-based rules, and the appropriate grammar and lexicon, to generate the Hindi output.

4. Machine Aided Translation System for the Public Health Campaign Material

ANGLABHARTI- A Machine Aided Translation system is a software for translation from English to Hindi for Public Health Campaign related materials. The software gives 85% parsing and about 60% correct translation output. The software also provides the editing facilities for correction of ill-formed sentences. The system comprises about 9500 root words dictionary. This work was a joint effort of Indian Institute of Technology, Kanpur and ER&DCI, Noida.

A prototype system for translation from English to Hindi was developed under the Anglabharti project started in 1991 and a functional prototype was made available. Now they have expanded the system to cover a larger domain in the Public Health domain.

4.1 Approach

This project heralds the major design consideration for providing a practical aid for translation which provides attributes for getting 90% of the task done by the machine and 10% is left for the Human post editing.

4.2 Major Modules of Anglabharti System

- **Morphological Analyser:**

The morphological analyser reads the source language sentence and analyses the word to give the root word and additional syntactic and category information, meanings in various languages, semantic tags and disambiguation rules of that word if it has multiple meanings.

- **Rule Base:**

This contains rules for mapping structures of sentences from English to Hindi. The pattern transformations is surface-tree to surface-tree, bypassing the task of getting the deep tree of the sentence to be translated.

- **Sense Disambiguator:**

This module is responsible for picking up the correct sense of each word in the source language. The approach used here is rule-to-rule semantic interpretation.

- **Generator:**

The generator takes the input from the output of the rule base i.e. the constituents of a sentence in the form of root words with their grammatical and semantic information for parser and produces sentence in the target language. The generator takes care of preposition disambiguation, ‘karaka’ disambiguation and decides the verbal agreements which are language dependent.

Currently as the software is Public Health Domain specific, the Software will be beneficial to the Central Health Bureau, Vigyan Prasar to name the few. Future technology tools that can be envisaged are Multilingual

Dictionary, Corporate Management, Translation to different domains and languages, Lexicon Management etc.

5. ANUSAARAKA System

Machine Translation Systems are extremely difficult to build. Translation is a creative process in which the translator has to interpret the text, something that is very hard for the machine to do. In spite of the difficulty of MT, the Anusaaraka can be used to overcome the language barrier in India today. This work started at Indian Institute of Technology, Kanpur and now further development is under progress at University of Hyderabad which was funded by TDIL, DOE and currently supported by Satyam Computers.

5.1 Approach:

Anusaaraka systems among Indian languages are designed by noting the following two key features:

In the Anusaaraka approach, the load between the reader and the machine is divided in such a way that the machine handles the aspects, which are difficult for the reader, and aspects which are easy for the reader are left to him. Specifically, reader would have difficulty learning the vocabulary of the language, while he would be good at using general background knowledge needed to interpret any text. On the other hand, the machine is good at "memorising" an entire dictionary, grammar rules, etc. but poor at using background knowledge. Thus, the work is divided, in which the language-based analysis of the text is carried out by the machine, and knowledge-based analysis or interpretation is left to the reader.

Among Indian languages, which share vocabulary, grammar, pragmatics, etc. the task is easier. For example, in general, the words in a language are ambiguous, but if the languages are close to each other, one is likely to find a one to one correspondence between words where the meaning is carried across from source language to target language. For example, for 80 percent of the Kannada words in the current Anusaaraka dictionary of 30000 root words, there is a single equivalent Hindi word which covers the senses of the original Kannada word.

In the Anusaaraka approach, the reader is given an image of the source text in the target language by faithfully representing whatever is actually contained in the source language text. So the task boils down to presenting the information to the user in an appropriate form. We relax the requirement that the output in the target language should be grammatical. The emphasis shifts to comprehensibility. The answer is to deviate from the target language in a systematic manner.

First, new notation is invented and incorporated. For example. Hindi has the post-position marker 'ko', which functions both as accusative marker as well as dative marker. We distinguish between them by putting a diacritic mark (backquote). Thus, existing words in the Target language may be given wider or narrower meaning.

Second, we may relax some of the conditions in the target language. For example, we give up agreement in our "dialect" of the target language. The principle behind the systematic deviations is simple: the output follows the grammar of the source language. In the case of agreement, to state it more precisely, the output follows the agreement rules of the source language, therefore, the output in the target language appears to be without agreement. Some of the constructions of the source language may also get introduced in the target language. (Actually, as the constructions are largely common across the two languages, a new construction is noticed only when the source language has a construction, which is somewhat different from the target language.)

Sometimes, language bridges might be built between constructions in the source language, which are not there in the target language. A different construction but which can express the same information in the target language is chosen, with some additional notation, if necessary. For example, adjectival participial phrases in the South Indian languages are mapped to relative clauses with the 'jo*' notation.

Because of the reasons mentioned above, some amount of training will be needed on the part of the reader to read and understand the output. This training will include teaching of notation, some salient features of the source language, and is likely to be about 10% of the time needed to learn a new language. For example, among Indian languages it could be of a few weeks duration, depending on the proficiency desired. It could also occur informally as the reader uses the system and reads its output, so the formal training could be small.

5.2 Applications

Anusaaraka can be used in a variety of situations. Here we give some examples:

a) A reader wants to read an e-mail message or a document quickly, to find out its gross contents.

The reader can run Anusaaraka on the source and read the output directly. He might not be proficient in the use of Anusaaraka, but since the reader motivation is high, he might be willing to put in the effort using the online help.

b) a publisher wants to translate a literary work and publish it.

The Anusaaraka output will have to be post-edited by a person, to make it grammatically correct, stylistically proper, etc. The post-edited output can be published.

c) A scholar wants to find out about what an original work or epic actually says where the original is in a language, which he does not know.

Translation is available, but he wants to see for himself as to what the epic says and what the translator has interpreted. He can read the epic directly through the Anusaaraka. As the machine does not interpret, and presents an image of the contents, he is able to see the original without the translator's interpretation.

6. Conclusion

With Internet becoming popular and reaching to the masses, there is an increasing interest in the area of Natural Language Processing and Computer Assisted Translation in particular, large scale projects are initiated in India also which had led to some practical usable systems. The Beta version of MANTRA-Rajbhasha, which does the translation of Indian Gazette Notification for Appointment/Transfer/Promotion etc., is now put to use in the various Indian Ministries/Government Departments to get the feedback. Anusaaraka is also put on the Internet, to which any user can send the text and can get the translated text back.

The Department of Electronics, Government of India has also initiated TDIL program (Technology Development of Indian Languages) in 1991. The long-term objectives of the program are:

i) to develop Information Technology (IT) tools to facilitate human machine interaction and information processing in Indian Languages and development of multilingual knowledge systems: and

ii) to promote use of Information Technology tools for language studies and research.

Research in these areas will continue to contribute to other facilities such as Automatic Speech Translation, Intelligent Information Retrieval and Indian Language Localization of Web pages.

7. References

Joshi, A. (1975) “Tree Adjoining Grammars”. In the Technical Report, University of Pennsylvania. Philadelphia, USA.

Schabes Y. (1994) “Left to right parsing of lexicalised Tree-Adjoining Grammars”. In the Computational Intelligence Journal.

XTAG Research Group of IRCS (1997) “A Lexicalized Tree Adjoining Grammar for English”. In the Technical Report, University of Pennsylvania, Philadelphia. USA.

Darbari H. (1990) “An efficient Knowledge Rep. and Acquisition Technique for Intelligent Systems”. In the Ph.D thesis, Dept.of CSE University of Roorkee, Roorkee India.

Bharati A., Chaitanya V., Sangal R. (1995), “Natural Language Processing: A Paninian Perspective”. Published in the Prentice-Hall of India.

V.N. Narayana (1994) , “Anusaaraka: A Device to Overcome the Language Barrier”. In the Ph.D. thesis, Dept. of CSE, I.I.T. Kanpur, India.

Deshpande W. R., Oberoi S. S. (1994) “Machine Translation: State of the Art”. In the Journal of Electronics Information & Planning, DOE, New Delhi, India