Machine Translation Activities in India: A Survey

Akshar Bharati,
Rajeev Sangal, Dipti M Sharma, Amba P Kulkarni
Language Technologies Research Centre
International Institute of Information Technology, Hyderabad
{sangal,dipti,amba}@iiit.net

April 24, 2002

1 Introduction

This paper outlines the major MT related activities that are being carried out in India. The focus of activities is on building lexical resources for Indian languages and development of machine translation systems. A brief report of activities in speech processing systems, OCRs and knowledge extraction is also presented.

2 Lexical Resources

2.1 Corpora

A major effort was undertaken in mid 90s to develop balanced text corpora for 15 Indian languages, with funding from Ministry of Information Technology, Government of India. Under this project, a corpus of 3 million words for each of the 15 Indian Languages was created. [CIIL corpus]

Major newspapers in Indian languages have their publications on web, and form another source for the corpus. The news articles range from political news to sports, thereby covering a wide range of domains. Sizes of these corpora are also considerable. For example, webdunia has a corpus of tens of million words. [WEBDUNIA URL]

Sanskrit-Net project has been initiated to network all the sanskrit institutions and make available Sanskrit texts in e-form. [Sanskrit Net URL]

2.2 Dictionaries

In an effort to develop the language accessors among Indian languages, five pairs of bilingual dictionaries were developed by Akshar Bharati group and are available under GNU's General Public License (GPL) [Anusaaraka URL]. In addition two of the dictionaries Telugu-English and English-Telugu whose copyrights are over are also made available under GPL.

2.2.1 Collborative work model

English-Hindi dictionary called Shabdaanjali has been created through a collaborative effort involving many volunteers including school children, teachers, housewives, and retired persons. It has around 25000 entries.Each entry has a detailed division of senses with example sentences for each of the senses. The dictionary was co-ordinated and later edited by a small core team of two persons at International Institute of Information Technology, Hyderabad (IIITH). It is released freely under GPL[Dictionary URL].

The modern technology permits the incremental improvement and enhancement of the basic resources over a period of time. Therefore, though the dictionary has some lacunae and also is not of uniformly good quality, one can start using it. The feedback from different applications will help in improving it further.

2.2.2 Other ongoing activities

The successful completion of English-Hindi dictionary formed a base to undertake tasks of building few more specific lexical resources. Three major tasks were recommended at LRNLP 2001 viz.

- TransLexGram (Transfer Lexicon and Grammar)
- AnnCorra(Annotated corpus) treebank for each Indian language
- ShabdaSutra ('word thread' that connects various senses of a polysemous word giving rise to a 'formula' that faithfully and unambiguously represents the connection between these senses.)

These tasks are being carried out at several places covering a few of the Indian languages [Bharati, 2001].

2.3 Linguistic tools

2.3.1 Morphological Analysers

Morphological analysers for 6 Indian languages (Hindi, Telugu, Kannada, Marathi, Punjabi and Bangla) developed as part of Anusaaraka systems by Akshar Bharati group are available under GPL for free download and use [Anusaaraka URL]. Sanskrit morphological analysers have been developed with reasonable coverage (based on Paninian theory) by Indian Heritage Group at CDAC, and Academy of Sanskrit Research at Melkote.

2.3.2 Wordnet for Hindi

A wordnet for Hindi is being developed at Indian Institute of Technology, Bombay. The design of Hindi wordnet has been mainly inspired by the English Wordnet. However, it has some unique features such as graded antonymy and meronymy relationships. It also implements the Indian language specific phenomena like compound and conjunct verbs. Currently there are about 10,000 synsets. Recently work on Marathi wordnet has also started. The work of linking the Hindi Wordnet with the English wordnet and the Euro Wordnet is also going on.

Tamil thesaurus similar to wordnet has been developed at Tamil University, by S. Rajendran.

2.3.3 Other tools

A prototype of the parser based on Paninian theory for Indian languages was developed in early 90s [Bharati, 1995]. This parser is being enhanced at IIITH further for Hindi to widen the coverage.

Spell checkers are available for many Indian languages [TDIL URL].

3 Machine Translation and Language Accessors

A fully automatic high quality general purpose machine translation system is still a distant dream. Naturally one ends up with relaxing one or more of the constraints. For example a few groups in India have tried building MT systems for highly restricted domains. Another approach is to relax the 'fully automatic' constraint allowing for either pre-editing or post-editing by humans. Some of the groups in India are exploring this approach.

Both these approaches focus on translation. However the very purpose of any translation system is to have access to the information coded in other languages. Anusaaraka or the language accessor is the third approach that aims at access to the original text, thereby restricting the 'high quality' to the accuracy aspect only, thus sacrificing the ease of reading aspect.

In what follows we give a brief outline of the major efforts in building MT systems and language accessors.

3.1 Mantra System

The Mantra system translates appointment letters issued by government from English to Hindi. It is based on 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 official English documents. The system is being tested at different ministries [TDIL URL].

3.2 MaTra: English-Hindi Human aided MT

MaTra is a technology for helping human translators to translate from English to Indian languages (currently Hindi). The main focus in this project is on the innovative use of man-machine synergy to simplify a traditionally hard problem. One of the key features of MaTra is an intuitive structure-editor that the user can use to verify, correct and disambiguate the system's analysis of the source sentence, thus allowing a single correct translation to be produced. Currently, an advanced prototype that can handle simple sentences exists. Work is on to extend the range of sentences and to productionize the system [NCST URL].

3.3 Anglabharati Approach

A demo system for translation of public health campaign documents has been developed by Electronics Research and Development Center of India, Noida. The system uses the Anglabharati approach developed at IIT, Kanpur, which is based on pattern directed rule based system with context free grammar like structure for English. The system attempts

to integrate example-based approach with rule-based. At the end, human being post edits the output to correct the ill-formed sentences [TDIL URL].

3.4 UNL based MT

The Machine Translation effort at Indian Institute of Technology, Bombay is interlingua based, and uses Universal Networking Language (UNL) as an intermediate language. English/Hindi to UNL analysers and UNL to Hindi/Marathi generators are ready. The work on Marathi to UNL analyser has been started. All these are rule and lexicon driven. Currently each system has about 5000 rules covering wide ranging language phenomena. The English-Hindi MT system using UNL systems uses a dictionary of concepts for Hindi which has currently around 80,000 entries in it.

3.5 Anusaaraka or Language accessor

Anusaaraka systems or the language accessors are based on the principle of 'information preservation'. As a consequence, the anusaaraka output follows the grammar of source language. Hence before using anusaaraka systems to access the information, the reader has to undergo a short training to read and understand the output[Bharati, 2002]

Anusaaraka provides 'glosses' in target language for each meaningful lexical unit. There are cases where the meaning is too general or too specific. Such cases are handled by introducing some special notation to either narrow down or widen the meaning. An attempt is made to find the underlying thread (called 'shabda suutra' or 'word thread') that connects different senses of the polysemous word. A kind of formula ('suutra' also means a formula in Sanskrit) is then evolved that faithfully and unambiguously represents the connection between these different senses.

The current version of English-Hindi anusaaraka uses shabdaanjali, a 25000 English-Hindi dictionary, GCIDE (GNU's Collaborative International Dictionary of English) and the English Wordnet. POS taggers and WASP workbench for word sense disambiguation (WSD) are also used to get the most likely sense of a word. However, these modules being fragile, and since anusaaraka aims at information preservation, the output is presented in several layers. The topmost layer is the one in which the POS information and the WSD rules are used. In case of difficulty in understanding the text, reader can choose to go to the bottom layers that contain more faithful information.

Since the anusaaraka output follows the source language grammar it helps in identifying the areas that are important from MT point of view, thus acting as a stepping stone towards building a fully automatic MT system.

Beta versions of five anusaaraka systems (Telugu, Kannada, Marathi, Punjabi, Bangla into Hindi) were released by Akshar Bharati group in 1998 under GPL [Anusaaraka URL]. The current focus of the group is on building English-Hindi anusaaraka.

3.6 Example Based Machine Translation

Major work is going on at IIITH on example based machine translation and automatic acquisition of transfer grammars and transfer lexicon. Tools have been developed for cleaning and alignment of texts, sentences and chunks in parallel corpora.

4 Knowledge Extraction and Management

In the field of knowledge extraction and management the major thrust at IIITH is on developing technologies for text-mining. Development of customizable search engines, document summarization tools, text categorization tool and information retrieval tools are some of the current projects under progress. The tools make use of technologies available for English and for various Indian languages.

5 Speech Processing

Restricted domain Text to Speech(TTS) systems are now ready for a number of specialized domains for Hindi and Telugu at IIITH. Various projects with wide range of applications such as TTS for unrestricted domain for Telugu and Hindi, speaker verification system, etc. are under way.

A prototype for Text to Speech synthesis for Hindi is also developed by Central Electronics Engineering Research Institute, Delhi [TDIL URL].

6 OCRs for Indian Languages

Optical Character Recognition systems for Devanagari and Bangla have been developed at ISI Calcutta. The Devanagari OCR has an accuracy of around 98%. Bangla OCR has slightly lower performance. The Devanagari OCR is expected to be released commercially shortly.

Some work is in progress on Punjabi OCR at Thapar Institute of Engineering and Technology, Patiala.

7 Associations and Conferences

NLP Association of India (NLPAI) has been formed recently to organize the research community in NLP field for greater interaction and exchange of research ideas, sharing of language data and software, etc.[NLPAI addresses].

Many SIGs are also formed as part of NLPAI to plan activities in different sub-areas of NLP. A major conference on NLP called ICON-2002 is planned for December 2002 [ICON URL].

8 Conclusion

In conclusion, the NLP related activities in Indian languages are mainly centered around development of machine translation systems, language accessors and building of suitable lexical resources. There are also some encouraging efforts in speech processing, knowledge extraction, and OCRs.

References

1. Anusaaraka URL: http://www.iiit.net/ltrc/index.html

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

3. Bharati, Akshar, et.al, Anusaaraka: Overcoming the Language Barrier in India, To appear in "Anuvad", Sage Publishers, New Delhi,2002 (Available from Anusaaraka URL.)

4. Bharati, Akshar, Dipti M Sharma, Vineet Chaitanya, Amba P Kulkarni, Rajeev Sangal, Durgesh D Rao, "LERIL: Collaborative Effort for Creating Lexical Resources", In Proc. of Workshop on Language Resources in Asian Languages, together with 6th NLP Pacific Rim Symposium, Tokyo, 30 Nov 2001.

5. CIIL URL: http://www.ciil.org/

6. Dictionary URL: http://www.iiit.net/ltrc/Dictionaries/Dict_Frame.html

7. ICON URL: http://www.iiit.net/conferences/icon2002.html

8. LRNLP-2001: Recommendation of Workshop on Lexical Resources for Natural Language Processing for Indian Languages, Hyderabad, January 2001. (lr_egroup@iiit.net)

9. NLPAI Addresses: sangal@iiit.net, uday@ciil.stpmy.soft.net

10. NCST URL: http://www.ncst.ernet.in/ matra/

11. Sanskrit Net URL: http://www.sansknet.org/

12. TDIL URL: http://www.tdil.gov.in/

13. WEBDUNIA URL: http://www.webdunia.com/