

Machine Translation in India: A Brief Survey

Durgesh Rao <durgesh@ncst.ernet.in>

**National Centre for Software Technology
Gulmohar Road 9, Juhu, Mumbai 400049, INDIA.**

Disclaimer: *This survey is based on the information available to the author, and is believed to be accurate at the time of writing, to the best of his knowledge and belief. No legal claim is made to the accuracy of the information. For the latest information on these projects, the interested reader should write to the contact person mentioned for each project.*

1. Background

Machine Translation is an important technology for localization, and is particularly relevant in a linguistically diverse country like India. In this document, we provide a brief survey of Machine Translation in India.

Human translation in India is a rich and ancient tradition. Works of philosophy, arts, mythology, religion, science and folklore have been translated among the ancient and modern Indian languages. Numerous classic works of art, ancient, medieval and modern, have also been translated between European and Indian languages since the 18th century.

In the current era, human translation finds application mainly in the administration, media and education, and to a lesser extent, in business, arts and science and technology.

India has a linguistically rich area—it has 18 constitutional languages, which are written in 10 different scripts. Hindi is the official language of the Union. English is very widely used in the media, commerce, science and technology and education. Many of the states have their own regional language, which is either Hindi or one of the other constitutional languages. Only about 5% of the population speaks English.

In such a situation, there is a big market for translation between English and the various Indian languages. Currently, this translation is essentially manual. Use of automation is largely restricted to word processing. Two specific examples of high volume manual translation are—translation of news from English into local languages, translation of annual reports of government departments and public sector units among, English, Hindi and the local language.

As is clear from above, the market is largest for translation from English into Indian languages, primarily Hindi. Hence, it is no surprise that a majority of the Indian Machine Translation (MT) systems are for English-Hindi translation.

As is well known, natural language processing presents many challenges, of which the biggest is the inherent ambiguity of natural language. MT systems have to deal with ambiguity, and various other NL phenomena. In addition, the

linguistic diversity between the source and target language makes MT a bigger challenge. This is particularly true of widely divergent languages such as English and Indian languages. The major structural difference between English and Indian languages can be summarized as follows.

English is a highly positional language with rudimentary morphology, and default sentence structure as SVO. Indian languages are highly inflectional, with a rich morphology, relatively free word order, and default sentence structure as SOV. In addition, there are many stylistic differences. For example, it is common to see very long sentences in English, using abstract concepts as the subjects of sentences, and stringing several clauses together (as in this sentence!). Such constructions are not natural in Indian languages, and present major difficulties in producing good translations.

As is recognized the world over, with the current state of art in MT, it is not possible to have Fully Automatic, High Quality, and General-Purpose Machine Translation. Practical systems need to handle ambiguity and the other complexities of natural language processing, by relaxing one or more of the above dimensions.

Thus, we can have automatic high-quality 'sub-language' systems for specific domains, or automatic general-purpose systems giving rough translation, or interactive general-purpose systems with pre or post editing. Indian MT systems have also adopted one of these strategies, as we will see.

Machine Translation in India is relatively young. The earliest efforts date from the late 80s and early 90s. The prominent among these are the projects at IIT Kanpur, University of Hyderabad, NCST Mumbai and CDAC Pune. The Technology Development in Indian Languages (TDIL), an initiative of the Department of IT, Ministry of Communications and Information Technology, Government of India, has played an instrumental role by funding these projects.

Since the mid and late 90's, a few more projects have been initiated—at IIT Bombay, IIIT Hyderabad, AU-KBC Centre Chennai and Jadavpur University Kolkata.

There are also a couple of efforts from the private sector - from Super Infosoft Pvt Ltd, and more recently, the IBM India Research Lab.

2. Major MT Projects in India

We now look at some of the major Indian MT projects in more detail. The parameters we look at are: language pair(s), formalism, strategy for handling complexity/ambiguity, and application domain(s), wherever this information is available (see the [disclaimer](#) at the top of this document).

1. Anglabharat (and Anubharati):

Anglabharati deals with machine translation from English to Indian languages, primarily Hindi, using a rule-based transfer approach. The primary strategy for handling ambiguity/complexity is post-editing—in case of ambiguity, the system retains all possible ambiguous constructs, and the user has to select the correct choices using a post-editing window to get the correct translation.

The system's approach and lexicon is general-purpose, but has been applied mainly in the domain of public health. The project is primarily based at IIT-Kanpur, in collaboration with ER&DCI, Noida, and has been funded by TDIL.

Anubharati is a recent project at IIT Kanpur, dealing with template-based machine translation from Hindi to English, using a variation of example-based machine translation. An early prototype has been developed and is being extended.

The contact person is Prof. RMK Sinha <rmk@cse.iitk.ac.in>.

2. Anusaaraka:

The focus in Anusaaraka is not mainly on machine translation, but on Language Access between Indian languages. Using principles of Paninian Grammar (PG), and exploiting the close similarity of Indian languages, an Anusaaraka essentially maps local word groups between the source and target languages. Where there are differences between the languages, the system introduces extra notation to preserve the information of the source language. Thus, the user needs some training to understand the output of the system. The project has developed Language Accessors from Punjabi, Bengali, Telugu, Kannada and Marathi into Hindi. The approach and lexicon is general, but the system has mainly been applied for children's stories. The project originated at IIT Kanpur, and later shifted mainly to the Centre for Applied Linguistics and Translation Studies (CALTS), Department of Humanities and Social Sciences, University of Hyderabad. It was funded by TDIL.

Of late, the Language Technology Research Centre (LTRC) at IIIT Hyderabad is attempting an English-Hindi Anusaaraka/MT system.

The contact persons are Prof. Rajeev Sangal <sangal@iiit.net> and Prof. G U Rao <guraosh@uohyd.ernet.in>.

3. MaTra:

MaTra is a Human-Assisted translation project for English to Indian languages, currently Hindi, essentially based on a transfer approach using a frame-like structured representation. The focus is on the innovative use of man-machine synergy—the user can visually inspect the analysis of the system, and provide disambiguation information using an intuitive GUI, allowing the system to produce a single correct translation. The system uses rule-bases and heuristics to resolve ambiguities to the extent possible – for example, a rule-base is used to map English prepositions into Hindi postpositions. The system can work in a fully automatic mode and produce rough translations for end users, but is primarily meant for translators, editors and content providers. Currently, it works for simple sentences, and work is on to extend the coverage to complex sentences. The MaTra lexicon and approach is general-purpose, but the system has been applied mainly in the domains of news, annual reports and technical phrases, and has been funded by TDIL.

The contact person is Durgesh Rao <durgesh@ncst.ernet.in>, and the MaTra Team contact is <matra@ncst.ernet.in>.

4. Mantra:

The Mantra project is based on the TAG formalism from University of Pennsylvania. A sub-language English-Hindi MT system has been developed for the domain of gazette notifications pertaining to government appointments. In addition to translating the content, the system can also preserve the formatting of input Word documents across the translation. The Mantra approach is general, but the lexicon/grammar has been limited to the sub-language of the domain. Recently, work has been initiated on other language pairs such as Hindi-English and Hindi-Bengali, as well as on extending to the domain of parliament proceeding summaries. The project has been funded by TDIL, and later by the Department of Official Languages.

The contact person is Dr Hemant Darbari <darbari@cdac.ernet.in>.

5. UCSG-based English-Kannada MT:

The CS Department at the Univ of Hyderabad has worked on an English-Kannada MT system, using the Universal Clause Structure Grammar (UCSG) formalism, also invented there. This is essentially a transfer-based approach, and has been applied to the domain of government circulars, and funded by the Karnataka government.

The contact person is Prof. K Narayana Murthy <knmcs@uohyd.ernet.in>.

6. UNL-based MT between English, Hindi and Marathi:

The Universal Networking Language (UNL) is an international project of the United Nations University, with an aim to create an Interlingua for all major human languages. IIT Bombay is the Indian participant in UNL, and is working on MT systems between English, Hindi and Marathi using the UNL formalism. This essentially uses an interlingual approach—the source language is converted into UNL using an ‘enconverter’, and then converted into the target language using a ‘deconverter’.

The contact person is Prof. Pushpak Bhattacharya <pb@cse.iitb.ac.in>.

7. Tamil-Hindi Anusaaraka and English-Tamil MT:

The Anna University KB Chandrasekhar Research Centre at Chennai was established recently, and is active in the area of Tamil NLP. A Tamil-Hindi language accessor has been built using the Anusaaraka formalism described above. Recently, the group has begun work on an English-Tamil MT system.

The contact person is Prof. CN Krishnan <cnkrish@au-kbc.org>.

8. English-Hindi MAT for news sentences:

The Jadavpur University at Kolkata has recently worked on a rule-based English-Hindi MAT for news sentences using the transfer approach.

The contact person is Prof. Sivaji Bandyopadhyay <ilidju@cal2.vsnl.net.in>.

9. **Anuvadak English-Hindi software:**

Super Infosoft Pvt Ltd is one of the very few private sector efforts in MT in India. They have been working on a software called Anuvadak, which is a general-purpose English-Hindi translation tool that supports post-editing.

The contact person is Mrs. Anjali Rowchowdhury

<anjalir@del16.vsnl.net.in>

10. **English-Hindi Statistical MT**

The IBM India Research Lab at New Delhi has recently initiated work on statistical MT between English and Indian languages, building on IBM's existing work on statistical MT.

The url is: <http://www.research.ibm.com/irl/projects/translation.html>

At a glance: Summary of major MT projects in India

Project	Languages	Domain/ Main Application	Approach/Formalism	Strategy
Anglabharati (IIT-K, ER&DCI-N)	Eng-IL (Hindi)	General (Health)	Transfer/Rules	Post-edit
Anusaaraka (IIT-K, UoH)	IL-IL (5IL->Hindi)	General (Children)	LWG mapping/PG	Post-edit
MaTra (NCST)	Eng-IL (Hindi)	General (News)	Transfer/Frames	Pre-edit
Mantra (CDAC)	Eng-IL (Hindi)	Govt. notifications	Transfer/XTAG	Post-edit
UCSG MAT (UoH)	Eng-IL (Kannada)	Govt. circulars	Transfer/UCSG	Post-edit
UNL MT (IIT-B)	Eng/IL (Hindi, Marathi)	General	Interlingual/UNL	Post-edit
Tamil Anusaaraka (AU-KBC)	IL-IL (Tamil- Hindi)	General (Children)	LWG mapping/PG	Post-edit
MAT (JadavpurU)	Eng-IL (Hindi)	News Sentences	Transfer/Rules	Post-edit
Anuvadak (Super Infosoft)	Eng-IL (Hindi)	General	N/A	Post-edit
StatMT (IBM)	Eng-IL	General	Statistical	Post-edit

3. Conclusion

MT is relatively new in India – about a decade old. In comparison with MT efforts in Europe and Japan, which are at least 3 decades old, it would seem that Indian MT has a long way to go. However, this can also be an advantage, because Indian researchers can learn from the experience of their global counterparts. There are close to a dozen projects now, with about 6 of them being in advanced prototype or technology transfer stage, and the rest having been newly initiated.

The Indian NLP/MT scene so far has been characterized by an acute scarcity of basic lexical resources such as corpora, MRDs, lexicons, thesauri and terminology banks. Also, the various MT groups have used different formalisms best suited to their specific applications, and hence there has been little sharing of resources among them.

These issues are being addressed now. There are governmental as well as voluntary efforts under way to develop common lexical resources, and to create forums for consolidating and coordinating NLP and MT efforts.

It appears that the exploratory phase of Indian MT is over, and the consolidation phase is about to begin, with the focus moving from proof-of-concept prototypes to productionization, deployment, collaborative resource sharing and evaluation.