

English To Marathi Translation Of Assertive Sentences

Abhay Adapanawar¹, Anita Garje², Purnima Thakare³, Prajakta Gundawar⁴, Priyanka Kulkarni⁵

Department of Information Technology, STES's Sinhgad Academy of Engineering,
Pune-48, Maharashtra, India.

ABSTRACT

This paper presents the conversion for simple English Assertive sentences to Marathi sentences. This is basically a machine translation. In this proposed system we are going through various processes such as morphological analysis, part of speech, local word grouping, for converting the meaning of simple assertive English sentence into corresponding Marathi sentence.

Here English to Marathi bilingual dictionary has been formed for the purpose of language translation. And English sentence are translated into the Marathi sentences using the rules that are produced earlier in the language translation system.

Keywords- Machine Translation, Natural Language Processing, Artificial Intelligence

1. Introduction

Marathi is one of the richest languages among all the languages exist in the world and one of the largely spoken languages in the world. More than 72 million people speak in Marathi as their native language. It is ranked 19th based on the number of speakers. Marathi is the mother language of India and also a large number of people in southern area of India (Maharashtra) speak and write in Marathi.

Marathi is a member of the Indo-Aryan languages. It is derived from Sanskrit. It is written left-to-right, top-to-bottom of page (same as English). Its vocabulary is akin to Sanskrit. Though the vocabularies are quite difficult at first, but to some extent there are similarities with English as exemplified by the following words in Table 1.

TABLE 1: Comparison of the similarities between different languages Major Headings

Language	Word			
English	month	mother	new	night
Sanskrit	mās	matar	nava	nakt
Marathi	mahina	mata	navin	ratra

2. Need of translation

Natural Language Processing (NLP) is quite a difficult task. There were many researches in the field of language translation but there is no fully

successful language translation machine so far. Since it is a Human Language Technology (HLT), there are lots of varieties and lots of opportunities for research. It is not possible to work on the whole language translation process together. Therefore, it is segmented into many parts[1].

But, the fact is, most of them choose a part of the source language to be translated to the target language[2]. For example, in our paper the source language is English and the target language is Marathi. There are varieties types of sentences both in English and Marathi language. In this paper, we have taken assertive sentences. We have mainly focused on the process of language translation and the effectiveness of the language translation.

3. Motivation for the work:

The people from rural area are not able understand high level of English. And also it's not possible for them to carry dictionary for conversion everywhere every time. Also these people are hesitating while using internet. Hence they can't able to help themselves for using services like internet banking or other.

Therefore, the purpose behind this project is to help people who had done their primary education but not up to the level of understanding each and every meaning of English word.

4. Literature Survey

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.[3]

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

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

4.1. 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.

4.2. 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 Assessors 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.

4.3. 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.

4.4. 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.

4.5. 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.

4.6. 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'. [4]

4.7. 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.

4.8. 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.

4.9. 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 software called Anuvadak, which is a general-purpose English-Hindi translation tool that supports post-editing.

4.10. 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.

5. Flow of translation

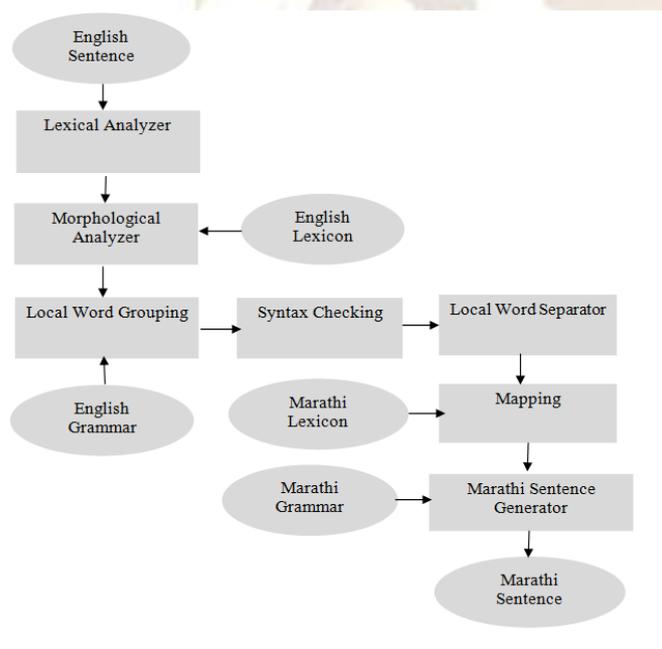


Fig 1 : flow of translation

5.1 Explanation of flow:

1. English sentence is taken as input
2. Lexical analyzer creates lexicons i.e it divides sentences into words.
3. Morphological analyzer is responsible for formation of morphemes and performs identification, analysis, description of structure of words.
4. Then with according to English grammar words are grouped.

5. The grouping of words forms the sentence and then Syntax checking of sentences according to language rules is done

6. Grouped words are separated by using local word separator for mapping from English to Marathi.

7. Separated words, corresponding Marathi match words are fetched from dictionary.

8. Finally, Marathi sentence is generated by applying Marathi grammar.

6. Features:

1. Can be used for localization purpose.
2. Can be useful for students in learning phase.
3. Useful for the people who have taken not understand English of higher level
4. No need to learn other languages.
5. Can be useful for document conversion.

7. Utilities:

1. Tokenization: User is able to see the separation of given sentences into words by lexical analyzer.
2. English tokenized words and their exact Marathi word are shown parallel to the user.
3. Parts of speech: User is able to see the parts of speech of each and every English words.

8. Conclusion

In this paper the system is proposed for user, where the user will get the English to Marathi translation of simple assertive sentences. The system also proposes that user will get the details of parts of speech of English words. The approach gives easy translation; hence it can be used for developing other language translators.

References:

- [1] Elaine Rich and Kevin Knight, Shivashankar Nair, *Artificial Intelligence* (Chapter Natural Language Processing), 3rd Edition, Tata McGraw-Hill, ISBN-10-0070087709, ISBN-13- 9780070087705
- [2] S. M. Chaware, Srikanta Rao, Domain Specific Information Retrieval in Multilingual Environment, *International Journal of Recent Trends in Engineering*, 2(4), 2009, 179-181
- [3] Bharati, Akshar, Vineet Chaitanya, Rajeev Sangal, *Natural Language Processing: A Paninian Perspective*, Prentice-Hall of India, 1995.
- [4] Sangal, Rajeev, Akshar Bharati, Dipti Misra Sharma, Lakshmi Bai, *Guidelines For POS And Chunk Annotation For Indian Languages*, December 2006.
- [5] Sangal, Rajeev, Dipti Misra Sharma, Lakshmi Bai, Karunesh Arora, *Developing Indian languages corpora: Standards and practice*, November 2006.