Indian Machine Translation Systems and Available Tools

Vikas Pandey[1*] Dr. M.V Padmavati[2] Dr. Ramesh Kumar[3]

[1] Dept. of Information Technology
Bhilai Institute of Technology
Durg , India

[2 , 3]Dept. of Computer Science and Engg.
Bhilai Institute of Technology
Durg , India

[1]vikas.pandey@bitdurg.ac.in, [2]vmetta@gmail.com, [3]rk_bitd@rediffmail.com

ABSTRACT

Language is the important means of communication for human race. India which is a morphologically rich and multi linguistic country due to which communication among people belonging to different states is major problem. Since India is moving towards Digital India where complete digitization and automation of every system is needed. Machine translation (MT) is a sub branch of Natural Language Processing(NLP).It is an automated system in which source language is inputted and the output will be a target language .In this paper an attempt has been made to survey various Indian machine translation systems and their approaches as well as to analyze various machine translation tools that can be helpful in implementation machine translation system.

Keywords: *Machine translation, Natural Language Processing, Digital India*

1. Introduction

India is having 30 recognized language and more than 2000 local dialects. There are 22 languages that comes under article 8 of our constitution. These are the official state languages through which various administrative work can be done. These languages are also becoming mode of communication between state and central government . Various national level exams are conducted through theses languages. There some languages that comes under article 8 like Marathi, Bodo , Dogri , Gujarati, Hindi, Kannada, Kashmiri, Konkani, Maithili, Malayalam, Bengali, Manipuri, , Nepali, Oriya, Punjabi, Sanskrit, Santali, Sindhi, Tamil, Telugu , Assamese, and Urdu[1].Most of the official and administrative work are done in English but, English speaking people are very less in number. Sometimes the government offices also do not know the regional language of the state due to which they face lot of problem in communication with the common public of the state. They need human translator for the translation of document .The efficiency of human translator is less and there is always chance of error during translation. Due to this limitation the automated machine translation system can play important role in language translation process.

The machine translation system work starts in the decade of 90's in India and it finds its application in various areas like in administrative work, State Assemblies and Parliament ,Education and News paper industry and Advertisement industry. There are various institutions like IIT Kanpur, IIT Bombay ,IIIT Hyderabad, University of Hyderabad, NCST Mumbai, The Technology Development in Indian Languages (TDIL), and CDAC Pune who are playing important role in developing the machine translation systems [2]. Many Machine Translation systems have been developed in India which has used different approaches for translating between source and target language.

2. Approaches for Machine Translation

The Machine Translation approaches can be broadly classified into following types: Direct Machine Translation, Rule Based Machine Translation, Corpus Based Machine Translation. The approaches for MT system has been given in Figure1.

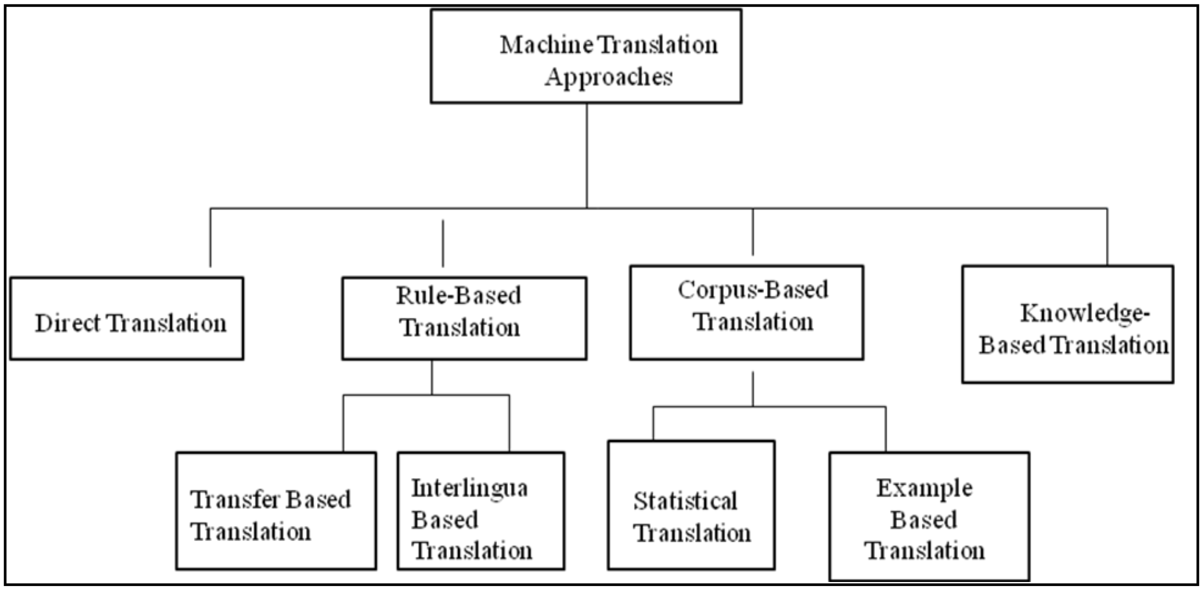
Machine Translation Approaches
Direct Translation
Rule-Based Translation
Corpus-Based Translation
Knowledge-Based Translation
Transfer Based Translation
Interlingua Based Translation
Statistical Translation
Example Based Translation

Figure1:Various Machine Translation Approaches

Direct Machine Translation

Direct MT technique was developed during 1950s to make use of newly invented computers for MT. It is based on a straightforward and easily implementable technique, keeping in view less processing power of computers available at that time. A direct translation system carries out word-by-word translation with the help of bilingual dictionary. As such, it is also known as dictionary driven machine translation approach. It involves a parser, which performs preliminary analysis of the source language sentence to produce its parts of speech information. This information is processed by a rule base to transform the source language sentence into a target language sentence. These rules include bilingual dictionary rules and rules to re-order the words. The direct machine translation system with parser and rule-base is also known as Transformer.

Rule-Based Machine Translation

The rule-based MT is used to remove major shortcomings of direct machine translation system. It parses the source text and produces an intermediate representation, which may be a parse tree or some abstract representation. The target language text is generated from the intermediate representation. These systems rely on the specification of rules for morphology, syntax, lexical selection, semantic analysis, transfer and generation process. Due to the extensive use of rule-base, these systems are known as rule-based systems. These systems are further divided as transfer-based machine translation and interlingua based machine translation.

Interlingua based MT is inspired by Chomsky's findings that regardless of varying surface syntactic structures, languages share a common deep structure. In interlingua-based MT approach, the source language text is converted into a language independent meaning representation called Interlingua. Interlingua based MT system, involves two stages in the translation process, including the analysis stage: to deeply analyze the source sentence for producing a language independent representation; and the synthesis stage: the target language is generated from the interlingua.

Corpus-Based Machine Translation

Corpus-based MT systems have become popular in recent years. These are fully automatic systems that require significantly less human labor than traditional rule-based approaches. However, they require sentence aligned parallel text for the language pair. The corpus-based approach is further divided into statistical and example based machine translation approaches.

Statistical machine translation (SMT) uses statistical models for translation whose parameters are derived from the analysis of bilingual text corpora. It does not make use of linguistic rules. SMT was introduced by Warren Weaver in 1949. SMT was re-introduced in 1991 by researchers at IBM. The essence of this method is first to align phrases, word groups and individual words of the parallel texts, and then calculate the probabilities that any one word in a sentence of one language corresponds to a word or words in the translated sentence with which it is aligned in other language. SMT has given more acceptable results by picking the word(s) that has the highest probability of occupying its current position, given the surrounding words

The *Example based Machine Translation* (EBMT) approach was suggested by Makoto Nagao in 1984. The EBMT approach requires a bilingual corpus with parallel texts. This approach works on the principle of translation by analogy. This principle is encoded in EBMT through example translations. An EBMT system has two main modules, namely, retrieval and adaptation. The retrieval module is used to retrieve translation examples from example-base or translation memory for a given input and adaptation is used to carry out necessary modifications in the retrieved example pair to generate translation of target language sentence.

Knowledge-Based Machine Translation

The important process in knowledge-based translation is to capture as much linguistic knowledge as possible from the source language sentences and store this into the translation system's knowledge base. For this, the system makes the use of source and target language dictionaries; source and target language structures and rules; word meanings in different contexts and language constructs; domain specific terminology; previously translated words, phrases, sentences, paragraphs; ontological and lexical knowledge; language style and cultural differences etc. By capturing all these knowledge sources, the system produces a high quality output. It is implemented on the Interlingua architecture, but differs from interlingua technique by the depth with which it analyzes the source language and its reliance on explicit knowledge of the world. The only problem of KBMT is that it is quite expensive to produce such a system because it requires a large amount of knowledge.

3. Indian Machine Translation Systems

The various types of Machine Translation systems for Indian languages with their source and target language are given in Table 1.

SNo.	MT SYSTEM	YEAR	SOURCE LANGAUGE	TARGET LANGUAGE	DESCRIPTION
I	DIRECT MACHINE TRANSLATION				
a.	Anusaaraka[3]	1995	Telugu, Kannada, Bengali, Punjabi and Marathi	Hindi	It uses Paninian grammar and matches related words between source and target language. Developed in IIIT Hyderabad.

b.	Punjabi to Hindi MT[4] System	2007	Punjabi	Hindi	It is based on direct word-to-word MT approach. Developed by Punjabi University, Patiala.
c.	Hindi-to-Punjabi MT[5] System	2009	Hindi	Punjabi	It is based on Direct word to word translation, consist of Morphological analysis, word sense disambiguation, post processing and Transliteration module.
II	**Transfer-Based MT Systems**				
a.	Mantra MT[6]	1997	English	Hindi	It uses XTAG based super tagger and dependency analyzer for performing analysis of the input English text.
b.	Shakti[7]	2003	English	Indian languages	It combines linguistic rule-based approach with statistical approach. The system consists of 69 modules
c.	Telugu-Tamil MT System[8]	2004	Telugu	Tamil	It uses the Telugu Morphological analyzer and Tamil generator for translation.
III	**Interlingua Machine Translation Systems**				
a.	ANGLABHARTI[9]	2001	English	Indian Languages	It is developed by pseudo-interlingua approach.

b.	UNL-based English-Hindi MT System[10]	2001	English, Hindi	Hindi, Bengali, Marathi	It uses Universal Networking Language (UNL) as the Interlingua structure. Developed by IIT Mumbai.
	Hybrid Machine Translation Systems				
a.	Anubharti Technology[10]	1995	Hindi	Indian Languages	It is a combination of example-based, corpus-based approaches
b.	Bengali to Hindi MT System[11]	2009	Bengali	Hindi	It uses a combined approach of SMT with a lexical transfer based system (RBMT)
	Example Based Machine Translation Systems				
a.	ANUBAAD[12]	2004	English	Bengali	It is specific to English Headlines translation Example-base, Tagged example-base and Phrasal example-base are separately maintained
b.	Shiva and Shakti MT System[10]	2003	English	Hindi, Marathi and Telugu	Uses combination of Example-based, rule based and statistical approaches

Table 1: Various Indian Machine Translation Systems

4. Tools used in Machine Translation

For Rule-based systems following tools can be used:

a. Apertium is a open-source rule-based machine translation platform.
b. Matxin is a open-source rule-based machine translation system for Basque.
c. OpenLogos, a open-source version of the historical Logos machine translation system.

For Statistical machine translation systems

a. Moses, a statistical machine translation system.
b. Marie is an n-gram-based statistical machine translation decoder.
c. Joshua is an open source decoder for SMT models based on synchronous context free grammars
d. Phramer, an open-source statistical phrase-based machine translation decoder
e. GREAT, a decoder based on stochastic finite-state transducers, which includes a training toolkit.
f. Giza++ is a tool to train translation models for statistical machine translation

5. Conclusion

This paper discussed about various approaches to machine translation system and gave a brief survey of various Indian machine translation system and tools available for their implementation. There are many languages called low resource languages for which machine translation has not been made.Further research work can be done to include many of these languages.

6. References

[1] Bandyopadhyay, S., 2004. *Use of machine translation in India*. AAMT J., 36: 25-31.

[2] Goyal.V. and Lehal. S .G. 2009. *Advances in Machine Translation Systems Language in India*, Vol. 9, No. 11, 2009, pp. 138-150.

[3] Bharti. A., Chaitanya. V, Kulkarni. P.A & Sangal. R. 2001. *ANUSAARAKA: overcoming the language barrier in India, published in Anuvad: approaches to Translation*
.

[4] Josan. S. G & Lehal. S.G. 2008 . *Punjabi to Hindi Machine Translation System*, in proceedings of COLING-2008: Companion volume: Posters and Demonstrations, Manchester, UK, pp. 157-160.

[5] Goyal. V & Lehal. G. S. 2010. *Web Based Hindi to Punjabi Machine Translation System*, International Journal of Emerging Technologies in Web Intelligence, Vol. 2, no. 2, pp. 148-151, ACADEMY PUBLISHER.

[6] Naskar. S & Bandyopadhyay. S. 2005.*Use of Machine Translation in India: Current status, AAMT* Journal, pp 25-31.

[7] Bharati, Sankar. B, Sharma D.M & Sangal.R. 2003. *Machine Translation: The Shakti Approach*, Pre-Conference Tutorial, ICON-2003.

[8] Parameswari. K, & Christopher. M. 2012. *Development of Telugu-Tamil Bidirectional Machine Translation System: A special focus on case divergence*, in proceedings of 11th International Tamil Internet conference, pp 180-191.

[9] Sinha.R.M.K. and Sivaraman.K. 1995. *ANGLABHARTI: A multilingual machine aided translation project on translation from English to Hindi*. IEEE Explore, DOI: 10.1109/ICSMC.1995.538002,pp.1609-1614.

[10] Rao.D. 2001. Machine Translation in India: A brief survey. In proceedings of SCALLA Conference, Banglaore, India.

[11] Chatterji. S,Roy. D, Sarkar. S & Basu. A. 2009. *A Hybrid Approach for Bengali to Hindi Machine Translation*, In proceedings of ICON-2009, 7th International Conference on Natural Language Processing, pp. 83-91.

[12] Bandyopadhyay. S. 2004. *ANUBAAD - The Translator from English to Indian Languages*, in proceedings of the VIIth State Science and Technology Congress. Calcutta. India. pp. 43-51