Multilingual Machine Translation Approaches for Consolidated Translator: An Indian Languages Perspective

Harjit Singh
Punjabi University Neighbourhood Campus,
Dehla Seehan (Sangrur)

Abstract— In India, there is an increasing need of consolidated translator to translate any regional language to any other regional language as per the requirement. India is a multilingual country and a variety of languages is used by people to communicate in spoken or written forms. Languages are also the base for dividing the country into states. So, Indian literature is available in various languages. The demand for translation from one regional language to another has increased in the recent years because of the increase in information exchange among different states of India. Machine Translation can be a useful tool to provide to fulfill the gap between languages. Basically Machine Translation is a field which correlates computer science and linguistics and provides human computer interaction in a natural language instead of a computer language. The research work in Machine Translation requires deep knowledge of linguistics, statistics and computer science. So it can be categorized as a multidisciplinary research area. Machine Translation can play a very useful role in language conversions such as from Punjabi to Hindi, Gujarati to Punjabi, and Hindi to Gujarati etc. It seems simple for a person who knows the source language and target language but it is very complex form computers to do the same job perfectly. A broad classification of machine translation approaches provides two paradigms i.e. Rule-Based Approach and Corpus-Based Approach. Machine Translation research for Indian Languages is being done at regional level in the country. But there is a need for multilingual conversion at one place for any Indian language to any other Indian language. Using existing research there are three approaches that can be adopted to fulfill the need. These are Conversion using Hindi/English as Intermediate Language, Direct Conversion from Source language to Target language and the mixture of these two approaches as Hybrid Conversion approach which can take the benefits of both. A consolidated multilingual translator can be very useful for Government, businesses and public to access information from different regions of country under a single platform.

Keywords—*Machine Translation, Indian Languages Conversion, Approaches in Language Conversion, Rule-Based Translation, Corpus-Based Translation.*

INTRODUCTION

In modern India, there is an increasing need of consolidated translator to translate any regional language to any other regional language as per the requirement. India is a multilingual country and variety of languages is used by people to communicate in spoken or written forms. Languages are also the base for dividing the country into states. Almost each state has its own regional language and that language is used by state government and some state governments have made it mandatory to use their regional language as official language of the state. The demand for translation from one regional language to another has increased in the recent years because of the increase in information exchange among different states of India.

Machine translation concept had its origin in the 17th century, when Rene Descartes proposed an idea of Universal Language. Some highlights from the history of machine translation are:

1929: Rene Descartes proposed an idea of Universal Language.
1946: A.D. Booth proposed the idea using computers for natural languages translation.
1949: Warren Weaver presented a Memorandum on Translation.
1951: Yehosha Bar-Hillel started his research on Machine Translation at MIT.

1954: A Georgetown Machine Translation research team demonstrated its system.
1955: Machine Translation research programs initiated in Japan and Russia.
1956: First Machine Translation conference held in London.
1962: Association for Machine Translation and Computational Linguistics formed in U.S.
1964: Automatic Language Processing Advisory Committee (ALPAC) formed by National Academy of Sciences.
1966: ALPAC report disclosed that 10 year research in MT failed to fulfill expectations. Research in MT slowed down.
1996: SYSTRAN offered translation of small text free through Web.
1997: AltaVista Babelfish scored 5,00,00 requests per day.
2003: Franz-Josef Och won DARPA's Machine Translation speed competition.
2007: Open source statistical Machine Translation engine "MOSES" developed.
2008: Mobile SMS translation service in Japan.
2009: Speech to speech translation for English, Chinese and Japanese for Mobile phones.
2012: Google announcement for translating enough text to fill 1 million books in one day.

In India, Hindi is considered as the national language but most of the official and business documents are prepared in English. Hindi is the spoken language and understood by large group of the population. Most of the states use their local language as official language. So in government and legal sector, the translations from one language to another may be required in some cases. In business sector also, the language translations are required according to the targeted audience. Some newspapers are published in multiple languages to target the particular audience. Doing the things manual is very time consuming and cumbersome task, so automation is the best alternative with the help of Machine Translation.

Digitizing Indian literature is a huge challenge because of the variety of languages in which the literature is available. To overcome the language barriers, Machine Translation can be very useful tool for language conversion.

Historical Review

In seventeenth century some philosophers (Leibniz, Descartes and others) put some proposals to relate words between languages. But these proposals are theoretical and no actual machine development was did. During mid-1930s, a patent for bilingual dictionary was applied by Georges Artsrouni. A Russian philosopher, Peter Troyanskii also came forward with a bilingual dictionary and a method that deals with grammar between languages. Alan Turing in 1950 published the famous paper "Computing Machinery and Intelligence" in which he proposed a criterion of intelligence that is now called "Turing Test". An overview of Historical developments is:

A. *Georgetown Experiment*

Developed in 1950 by Georgetown and IBM and was able to do automatic translation of more than 60 Russian sentences into English language.

B. *STUDENT*

Developed by Daniel Bobrow in 1964 and was able to solve high school algebra problems.

C. *ELIZA*

Developed by Joseph Weizenbaum in 1964 and was a simulation of a Rogerian psychotherapist. It was able to rephrase her response with a few grammar rules.

D. *SHRDLU*

It was developed in 1970 by Terry Winograd and was able to manipulate blocks of different colors. It was able to receive instructions like “Pick up the green box” or “where is yellow block”. It was able to answer the questions such as “What does the red box contain”. SHRDLU was the system that combined syntax, semantics and reasoning about the real world though natural language understanding. The system was able to handle limited number of sentences and those sentences should be about blocks.

E. *LUNAR*

The natural language database interface system was LUNAR produced in 1972 with ATNs and Woods’ Procedural Semantics. It was introduced at Second Annual Lunar Science Conference in 1971. The name LUNAR was taken from the database used by the system. Its performance was moderately inspiring.

F. *LIFER/LADDER*

It was a very impressive system developed in 1978. It was a natural language interface to database about US Navy ships. The semantic grammar was used by the system, so it was very much coupled to its domain. The system used the semantic grammar for various user-friendly features such as the ability to add new dictionary entries, to process incomplete input and to define paraphrases. These features made the system very inspiring.

G. *Jabberwacky*

It was developed in 1982 by Rollo Carpenter as a chatterbot. Its aim was to simulate human chatting in an entertaining manner.

H. *Racter*

It was developed in 1983 by William Chamberlain and Thomas Etter as a chatterbot that generated English language prose at random.

I. *Watson*

It was developed in 2006 by IBM and is a question answering system that defeated the best human players in 2011.

MACHINE TRANSLATION – SIMPLIFIED VIEW

It seems simple for a person who knows the source language and target language but it is very complex form computers to do the same job perfectly. Simple word by word substitution cannot produce good translations. Now the complexity is to program a machine (computer) in such a way that it should understand the source language text like a human and generate the target language text according to the meaning of source text.

The core concept of machine translation is to input a source language text, use some methodology to output the target language text with same meaning as original text. A broad classification provides two paradigms:

A. *Rule-Based Approach*

It is the classical approach of Machine Translation also called Knowledge Based Machine Translation. These systems get linguistics information from bilingual dictionaries and grammars of both the source and target languages which are used for translation process. These dictionaries and grammars provide semantic, syntactic and morphological regularities of both languages. The translation process applies a set of linguistic rules to generate the target language text from source language text.

The source text (input) is given to Morphological Analyzer for morphological processing of text which is sent to POS Tagger followed by Lexical Selection. The Structure Transfer is the next step to transfer the structure of source language to target language structure. Morphological Generator provides its text to Post Generator for final touch which provides target text (output). Rule based translation can be Direct, Transfer Based or Interlingua translation.

Rule Based Direct translation is applied at word level. Source language words are directly converted to target language words without any intermediate representation. It is a word by word translation method with some of the grammatical adjustments. These are basically unidirectional translations.

Rule Based Interlingua translation uses an intermediate representation of source text to generate target text from intermediate representation in the second phase. The intermediate representation is actually the representation of meaning of source text and it is independent of any language (language neutral text). So the approach is very useful approach for multilingual translation systems.

Rule Based Transfer translation is somewhat similar to Rule Based Interlingua translation and uses an intermediate representation of source language text to generate target language text. But it partially depends on the source and target languages and partially depends on structural differences between languages involved. This method provides high quality translations.

B. *Corpus-Based Approach*

This approach uses a bilingual Corpus to obtain knowledge for translation. A corpus is a huge amount of raw data that provides knowledge to the system to work. It is a huge collection of text and its parallel translations. The approach uses two methodologies, which are Statistical Based and Example Based.

In Statistical Based Corpus approach, statistical models are used for translation and the parameters are retrieved by analyzing the bilingual corpora. This method initially assumes every sentence in a language as the translation of every sentence of another language. Then the probability is assigned to every assumed translation of sentence and the sentence with highest probability is taken as final translated sentence. The search space is reduced by using heuristics and other methods.

In Example Based Corpus approach, translation by analogy is the basic idea. The system is given a set of source language sentences and corresponding target language translations of each sentence using a point to point mapping between the two. These sets become the base examples to get knowledge which is used to translate similar type of source language sentences. So, the system is first trained to make it capable for further translations.

INDIAN INTER LANGUAGES CONVERSION APPROACHES

Machine Translation research for Indian Languages is being done at regional level in the country. But there is a need for multilingual conversion at one place for any Indian language to any other Indian language. There are three approaches that can be adopted to fulfill the need:

A. *Conversion using Hindi/English as Intermediate Language*

This approach is easy to adopt since the research on regional languages to Hindi/English and vice versa is already being done by researchers in each region or state. All these researches need to be clubbed together to fulfill the need of Government, business and public. Good thing in this approach is that, most of these translators are already available and research is going on to improve their accuracy. So the work done by the researchers can be reused to make if more functional and useful.

As shown in (Fig. 1), it requires only 2xn translators where n is the number of languages we want to integrate. For example, if there are 20 Indian languages we want to integrate in our multilingual translator then 20 translators are required to convert any source language to intermediate language and 20 translators are required to convert intermediate language to any target language. That is, 2x20=40 translators are to be integrated together in multilingual translator.

B. *Direct Conversion*

The approach discussed above is easy to adopt but it has several disadvantages. First is that, it will consume double time than directly converting source language to target language. Secondly, in such type of conversions, percentage of accuracy may reduce to some or large extent. It is because; the intermediate language used may not be fully compatible with source as well as target language. It means that every language has a specific set of words (vocabulary) and when we convert, we may have to substitute a word that may not be the exact meaning of source word but is near to that one. It is obvious and it may convey the right message in that context but when such word is again converted to target language, it may produce a meaningless result. So the things may become more complicated. (Fig. 2) explains this approach.

Alternative but much complicated approach is to convert every source language to every target language directly using separate translator. But it requires large number of translators. It there are n languages then n-1 translators is required for every separate language. That is, nx(n-1) number of translators are needed. For example, if there are 20 languages then 20x19=380 translators are required.

C. *Hybrid Conversion*

Both direct conversion and conversion through intermediate language has their own advantages and disadvantages. Taking the benefits of both these approaches is the hybrid approach. It is an idea based on the fact that a number of translators developed by regional language researchers are available which directly translate their regional language to

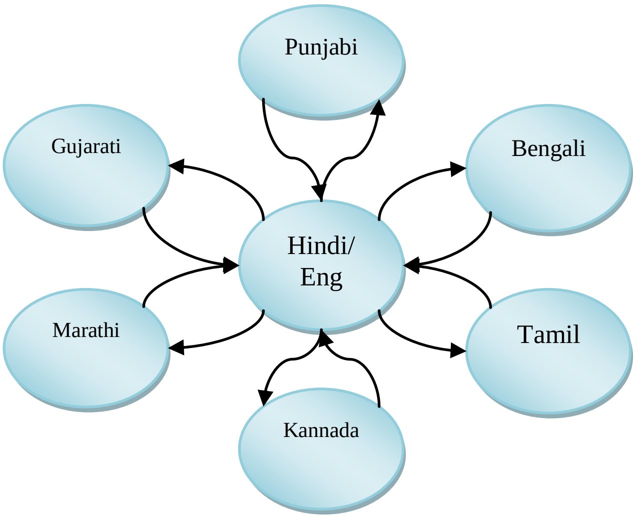

Fig. 1: Hindi/English as Intermediate Language

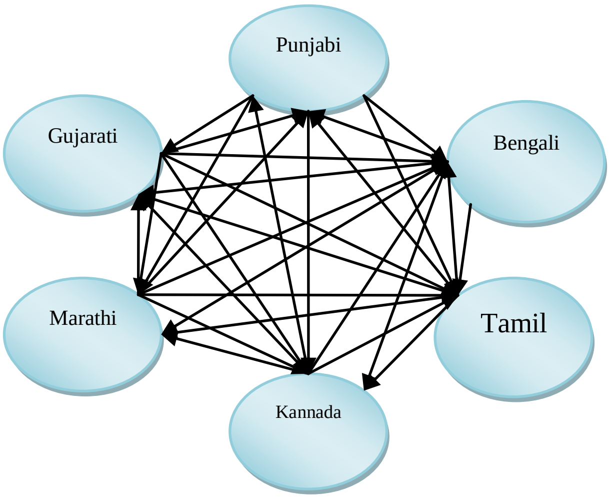

Fig. 2: Direct Conversion

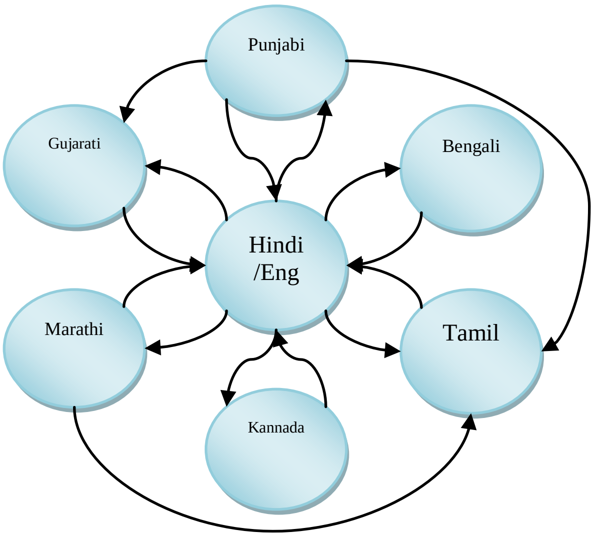

Fig. 3: Hybrid Conversion

multiple other languages and vice versa. So, these translators can be reused as it is in the consolidated translator. It will improve performance when those particular language translations will be required. In other cases for which direct translations do not exist, the translation though intermediate language is suitable. It will reduce the total number of translators required for multilingual translations.

In figure (Fig.3), it is assumed that Direct Translators are available for Punjabi to Gujarati, Punjabi to Tamil and Marathi to Tamil. So, those are used as it is to fasten the translation process and for other target languages Hindi/English is used as intermediate for translation process.

CONCLUSION

Machine Translation can play a great role in Indian Language conversions. The research work in language conversion is being done at regional level. Government sector, business sector and even public face difficulties to access information from different regions of country. A multilingual translator can be very useful to fulfill the need.

One easy approach is to use Hindi/English as an intermediate step for conversion since the research on regional languages to Hindi/English and vice versa is already being done by researchers in each region or state. It requires only 2xn translators where n is the number of languages we want to integrate. But this approach will consume double time than directly converting source language to target language. Secondly, in such type of conversions, percentage of accuracy may reduce to some or large extent. Converting every source language to every target language directly using separate translator is another alternative. If there are n languages nx(n-1) number of translators are needed. The third alternative is to use the mixture of both the approaches in such a way that if a direct translator is available it should be preferred otherwise conversion using intermediate language can be used. It can provide benefit of both the approaches.

REFERENCES

[1] Abhimanyu Chopra, Abhinav Prashar, Chandresh Sain, Natural Language Processing, International Journal Of Technology Enhancements And Emerging Engineering Research, Vol 1, Issue 4

[2] Prof. Langote Manojkumar S, Miss Kulkarni Sweta, Miss Mansuri Shabnam, Miss Pawar Ankita and Miss Bhoknal Kishor, Role of NLP in Indian Regional Languages, IBMRD's Journal of Management and Research Volume-3, Issue-2, September 2014

[3] Gore Lata and Patil Nishigandha, English to Hindi-Translation System,Proceedings of Symposium on translation systems strans (2002).

[4] http://www.slideshare.net/jhonrehmat/natural language processing.

[5] Natural Language Processing,www.myreaders.info /html/artificial_intelligence.html.

[6] Natural Language Processing-Computer science and engineering, www.cse.unt.edu/~rada/CSCE5290/Lectures/Intro.ppt

[7] NLP, https://www.coursera.org/course/nlp

[8] NLP, research.microsoft.com/en-us/groups/nlp/

[9] Dash, N S and B B Chaudhuri. “Why do we need to develop corpora in Indian languages”, International Conference on SCALLA, Banglore, 2001

[10] Murthy, B K and W R. Despande. Language technology in India: past, present, and the future. In the Proceedings of the SAARC Conference on extending the use of Multilingual and Multimedia Information Technology (EMMIT’98). Pune, India

[11] Anoop Kunchukuttan, Abhijit Mishra, Rajen Chatterjee, Ritesh Shah and Shata-Anuvadak: Tackling Multiway Translation of Indian Languages, LREC 2014, Rekjyavik, Iceland, 26-31 May, 2014

[12] R M K Sinha, “ Machine Translation : An Indian Perspective “ , Proceedings of the Language Engineering Conference (LEC’02)

[13] Vishal Goyal and Gurpreet Singh Lehal. "Web Based Hindi to PunjabiMachine Translation System", Journal of Emerging Technologies in Web Intelligence, Vol. 2, No. 2, May 2010, pg(s):148-151.

[14] M.D. Okpor, “Machine Translation Approaches: Issues and Challenges”, IJCSI Vol. 11, Issue 5, No. 2, September 2014.