Multilingual Machine Translation approach based Systems and their Analysis : An Indian Perspective

Harjit Singh
Neighbourhood Campus Dehla Seehan, Punjabi University, Patiala, Punjab, India

ABSTRACT

In a massive society like India, there is an excellent demand for translation of documents from one language to a different language. Most of the state government works in there provincial languages, whereas the central government's official documents and reports are in English and Hindi. So there is a need to translate these documents and reports within the individual provincial languages. The realm of AI has been occurring for previous few decades however the promising translation work began within the early Nineteen Nineties thanks to advanced analysis in computer science and Computational Linguistics. India may be a multilingual and philosophical system country with over one billion population and twenty two constitutionally recognized languages that are written in twelve totally different scripts. This necessitates the machine-controlled AI system for English to Indian languages and among Indian languages thus on exchange the knowledge amongst individuals in their native language. Several usable machine translation systems are developed and are underneath development in India and round the world. The paper focuses on totally different approaches utilized in the event of AI Systems and additionally briefly delineates a number of the AI Systems in conjunction with their options, domains and limitations. There are a number of machine translation systems that are developed for translation from English to Indian languages by exploitation totally different approaches.

Keywords: Machine Translation, Indian Languages Conversion, Approaches in Language Conversion, Multilingual Machine Translation.

I. INTRODUCTION

India could be a multilingual country wherever it is said that the oral communication changes at each fifty miles. There are 22 official languages; and close to 2000 dialects are spoken by totally different communities in India. English and Hindi are used for official and most states in India. The state governments in Asian nation preponderantly perform their official add their individual regional language whereas the official work of Union government is disbursed in English and/or Hindi. All the official documents and reports of Union government are revealed in English or Hindi or in each English and Hindi. Several newspapers are revealed in regional languages. Translating these documents manually is incredibly time overwhelming and expensive. Thus there is ought to develop smart machine translation systems to handle of these problems, so as to establish a stronger communication between states and Union governments and exchange of information amongst the individuals states with different regional languages. Indian languages ar divided into 5 language families; viz. Dravidian, Indo-Aryan, Tibeto-Burman, Austro-Asiatic and Andmanese. Several Indian languages are low resource languages create problems within the development of MT systems for Indian languages [1][2].

In India the earliest efforts starts after the middle 80s. In India many Institutes and research organizations have started researching on Machine Translation systems for English to Indian languages and vice-versa [3]. The distinguished Institutes are as follows:

- ✓ National Centre for Software Technology (NCST) Mumbai
- ✓ The research and development projects at Indian Institute of Technology (IIT), Kanpur

- ✓ Centre for Development of Advanced Computing (CDAC), Pune
- ✓ Computer and information Sciences Department, University of Hyderabad
- ✓ Government of India, through its Technology Development in Indian Languages (TDIL) Project
- ✓ Ministry of Communications and Information Technology

The Government of India has determined to relinquish additional thrust to Language Technology for Indian languages throughout VIIIth set up and to start a programme that may emphasize on quality and participation of cognitive content and R & D efforts in Indian languages. Government of India started a National level programme throughout the year 1990-91 on Technology Development for Indian Languages (TDIL) [4]. Alternative establishments like IIT Bombay,IIT Kanpur, IIIT Hyderabad, CDAC Pune, University of Hyderabad, NCST urban center, CDAC Noida, JNU New Delhi, Jadavpur University, Kolkata, etc are taking part in a serious role in developing the machine translation systems. Several machine translation systems have been developed and still the work is going on and more are being developed. The machine translation systems are developed or being developed by using different machine translation approaches. This paper provides an analysis of major machine translation systems in India.

II. Machine Translation at International Level [5]

1. In 1949, Warren Weaver planned the primary plan on the employment of computers in translation by adopting the term computer translation.
2. In 1952, the first conference of machine translation at university under Yehoshua Bar-Hillel.
3. In 1954, the first basic automatic Russian-English translator by researchers at Georgetown University unitedly with IBM that converted sixty Russian sentences. In the same year, Victor Yngve printed the first journal on machine translation.
4. In 1961, the computer linguistics was born thanks to weekly lectures organized by David G. Hays at the Rand Corporation in la. In the same year, first International Conference on machine translation of Languages and Applied Language Analysis of Teddington where linguists and computer scientists participated.
5. In 1964, Creation of committee ALPAC (Automatic Language Processing informatory Committee) to study the views and also the possibilities of machine translation.
6. In 1966, ALPAC revealed its note that its work on artificial intelligence was simply waste of time and money. As a result it created a negative effect on the research in machine translation for many of years.
7. In 1970, the project REVERSO started by Russian researchers. Development of System SYSTRAN1 (Russian-English) by Peter Toma, WHO was member.
8. In 1976, a machine translation system called WEATHER was developed within the project TAUM (machine translation within the university of Montreal) under Alai Colmerauer for the AI weather-forecasts for public. A group of researchers worked on this project.
9. In 1978, a machine translation system called ATLAS2 was developed by the FUJITSU (a Japanese Firm). This translator was able to convert from Korean to Japanese and contrariwise.
10. In 1981, a system called SHARP was developed by a The Japanese firm. It was an Automatic translator DUET (English - Japanese).
11. In 1983, NEC developed a system named as Honyaku device II, used for artificial language approach. It was based on algorithmic program referred to as PIVOT
12. In 1986, a Japanese-English system PENSEE was developed by OKI3. It was rule based system. HICATS (Hitachi laptop power-assisted Translation System) was developed by Hitachi. It was a Japanese-English translation system.
13. In 1993, C-STAR (Consortium for Speech Translation Advanced Research) was started. It was multilingual machine translation (English, German & Japanese) system for the field of tourism.
14. In 1998, Softissimo did the marketing of machine translation system REVERSO.
15. In 2000, a multilingual machine translation system (Japanese-English & Chinese - English) system called ALPH was developed by Japanese Laboratory ATR. It used Example based approach of machine translation.
16. In 2005, Google launched the first internet site (web site) for automatic translation.
17. In 2007, a machine translation systems called MT METIS-II was developed and it was a hybrid system that used the SMT, RBMT, and EBMT artificial intelligence approaches.

III. Machine Translation approaches

A number of machine translation systems across the world have already been developed for the foremost usually used natural languages like English, Japanese, Russian, Spanish, Chinese, Hindi and different Indian languages etc. Following figure depicts the various approaches utilized in developing these systems [5].
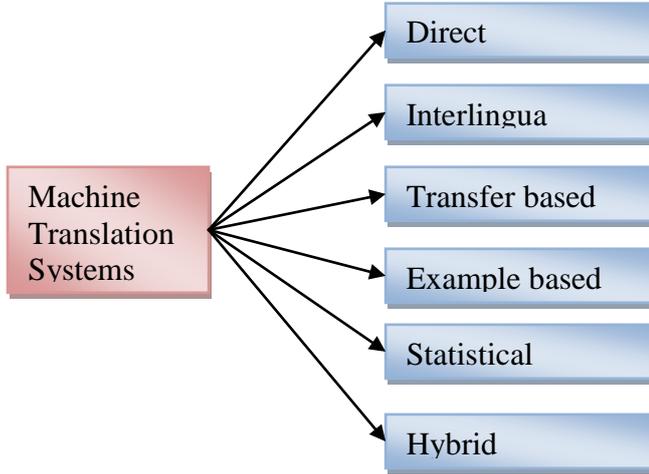

Figure 1. Machine Translation Approaches.

IV. Multilingual Machine Translation approaches in India [6]

In India, research in Machine Translation for Indian Languages is being performed at regional level by a number of Institutes and Research Centres within the country. However there's a requirement for multilingual translation at one place for any Indian language to the other Indian language. Although to a limited extent, the researchers adopted a mixer of following approaches to satisfy the need. Here is a brief introduction with the help of figures:

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

It is an easy approach in which the translators already developed in different languages are combined together to make a multilingual machine translator.
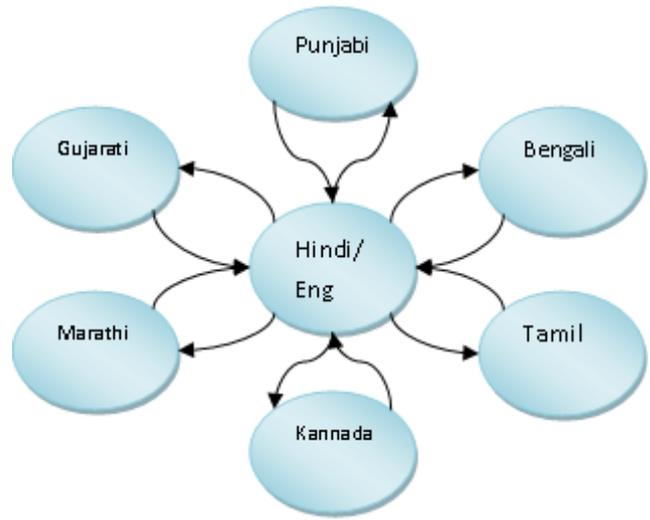

Figure 2. Conversion using Intermediate Language. Reprinted from Harjit Singh, "Multilingual Machine Translation Approaches for Consolidated Translator: An Indian Languages Perspective", IJOES, vol. 21, pp. 33-38, Dec. 2016

B. *Direct Conversion*

It is somewhat complicated approach to directly translate each language to each other language, requires large number of translators.
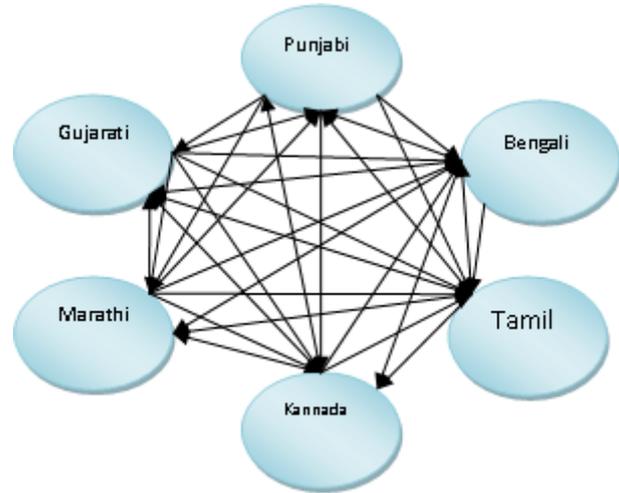

Figure 3. Direct Conversion. Reprinted from Harjit Singh, "Multilingual Machine Translation Approaches for Consolidated Translator: An Indian Languages Perspective", IJOES, vol. 21, pp. 33-38, Dec. 2016

C. *Hybrid Conversion*

Taking the advantages of both direct conversion and conversion through intermediate language, is the hybrid approach.
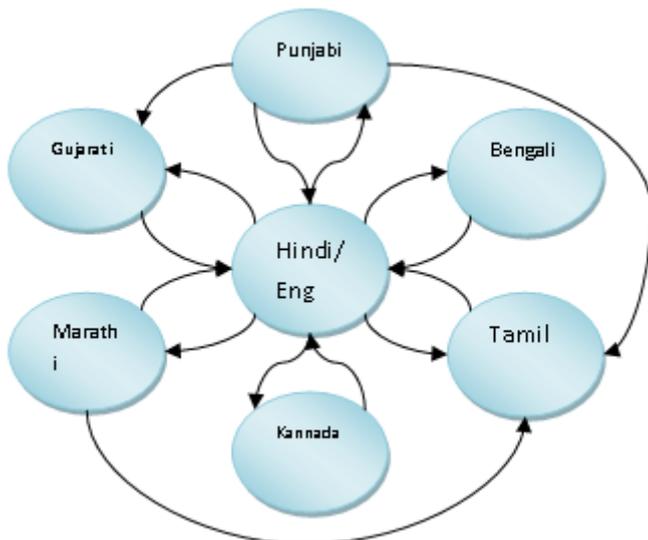
Figure 4. Hybrid Conversion. Reprinted from Harjit Singh, "Multilingual Machine Translation Approaches for Consolidated Translator: An Indian Languages Perspective", IJOES, vol. 21, pp. 33-38, Dec. 2016

V. Indian Machine Translation Systems

Various machine translation systems developed in India using Direct Translation Approach, Interlingua based approach, Hybrid Translation approach and Statistical Machine translation approach are analysed as follows:

a) Direct Machine Translation Systems

Following Machine translation systems are developed based on direct translation approach:

Punjabi to Hindi MT System

This machine translation system based on direct word-to-word translation approach was developed by G. S. Josan and G. S. Lehal. The system is composed of a number of modules such as pre-processing, word-to-word translation with lexicon, morphological analysis, disambiguation, transliteration and at the end post processing. [7][8].

Hindi-to-Punjabi MT System

This machine translation system based on direct word-to-word translation approach was developed by V. Goyal and G. S. Lehal. The system is composed of a number of modules such as pre-processing, word-to-word translation with lexicon, morphological analysis,

disambiguation, transliteration and at the end post processing. [1][9][10][11].

b) Interlingua Machine Translation Systems

Following Machine translation systems are developed based on Interlingua machine translation approach:

ANGLABHARTI

This machine translation system was developed by R. M. K. Sinha, R. Jain, A. Jain for translating English to Indian languages which was based on pseudo-interlingua approach. This interlingua approach adopted in this systems allowed to use the same system for translating the language English to multiple Indian languages and has eliminate the requirement of developing different translation system for translating English to each other Indian language. The English sentences as a source language sentences are analysed only once and then it creates PLIL (Pseudo Lingua for Indian Languages), an intermediate structure. The PLIL is then translated to other Indian language. [13].

AnglaHindi

This machine translation system is derived from AnglaBharti MT System developed by R. M. K. Sinha and A. Jain. It is developed for English to Hindi language translation which is based on pseudo interlingual rule-based approach. It makes use of all modules of AnglaBharti MT System and also uses abstracted example-base for converting commonly encountered verb phrases and noun phrases. [13].

c) Hybrid Machine Translation Systems

Following Machine translation systems are developed based on Hybrid machine translation approach:

Anubharti Technology

This machine translation system is developed using a hybridized example-based machine translation approach by Sinha. Actually it is a combined approach including corpus-based approach, example-based approach and some elementary grammatical analysis. The example-based approach is based on human-learning process in which knowledge from past experiences is stored to be used in future. In the development of Anubharti, EBMT approach is modified to reduce the need of a big

example-base. To modify EBMT, constituents are generalized and then replaced with abstracted form from some example data. Syntactic groups are identified to achieve abstraction. Syntactic category and semantic tags of the source language structure is used for matching of the input sentence with abstracted examples. Hindi language is used by AnuBharti-II as a source language for conversion to any other language. The generalization used in the example-base is much dependent on the target language [14][2].

Lattice Based Lexical Transfer in Bengali Hindi MT Framework

This machine translation system was developed by Sanjay Chatterji, Praveen Sonare, Sudeshna Sarkar, and Anupam Basu. They used a transfer based machine translation approach and proposed a technique for proper lexical conversion from Bengali language to Hindi language called Bengali to Hindi Machine Translation Framework. In this machine translation system, most frequently used Hindi word replace the Bengali word. But considering the context of the words, most frequent translation may be incorrect. The contextual information from a monolingual corpus of Hindi is used to find the better lexical choice from the options presented from dictionary. The system makes use of lattice-based data structure to translate the Bengali sentence to the equivalent Hindi sentence. This translation system is evaluated using the BLEU automatic evaluation tool along with human evaluation process and it is found that the proposed system performs better. [15].

d) Statistical Machine Translation Systems

Following Machine translation systems are developed based on Statistical translation approach:

Shakti

This machine translation system was developed by Bharati, R. Moona, P. Reddy, B. Sankar, D. M. Sharma and R. Sangal and it translates English language to any Indian language with very simple system architecture. It combines two approaches together i.e. linguistic rule based approach and statistical approach. The system is composed of 69 separate modules. From these 9 modules are used for the source language analysis (English) and 24 modules are used for performing bilingual tasks. The remaining thirty six (36) modules

are used for producing target Indian language text [14][12].

English to Indian Languages MT System (E-ILMT)

This machine translation system is developed by a group of 9 institutions namely IIIT Hyderabad, C-DAC Mumbai, C-DAC Pune, IISc Bangalore, IIT Mumbai, IIIT Allahabad, Jadavpur University Kolkata, Amrita University Coimbatore, Utkal University Bangalore and Banasthali Vidyapeeth, Banasthali. The EILMT is a machine translation System for translation from English to Indian Languages developed for Tourism and Healthcare Domains. The project is funded by Department of Information Technology, MCIT Government of India. The system was initially developed using the statistical techniques along with the contemporary tools which include the POS tagger, parser, decoder etc. The objective is to initially develop an English-Hindi machine translation system having the capability of translating of free text as found on the web sites and then gradually develop it for additional Indian languages [16].

VI. CONCLUSION

This paper explored machine translation approaches with the stress on the machine translation development for Indian languages. From this study of existing systems, it is found that about all Indian language machine translation systems are developed on the basis of rule-based, hybrid and applied math approaches.

Most of the Indian machine translation systems are developed for specific domains such as some business enterprise, medical, agriculture, health care, news headlines, government circulars, technical documents and notifications etc.

The rule based approach used in specific domain dependent systems, might fail in the development of general purpose machine translation systems because if may require a huge set of rules which will be difficult to maintain. Also to develop such systems the researchers must make use of applied mathematics and hybrid approaches including the support of linguistic consultants.

Most of the Indian machine translation systems require post correcting to produce more correct results.

VII. REFERENCES

[1] Sitender & Seema Bawa, (2012) "Survey of Indian Machine Translation Systems", International Journal Computer Science and Technolgy, Vol. 3, Issue 1, pp. 286-290, ISSN : 0976-8491 (Online) | ISSN : 2229-4333 (Print)

[2] projects.uptuwatch.com/cs-it/anubharti-an-hybrid-example-based-approach-for-machine-aidedtrapnslation/

[3] Sanjay Kumar Dwivedi, Pramod Premdas Sukhadeve, "Machine Translation System in Indian Perspectives", Journal of Computer Science 6 (10): 1111-1116, 2010, ISSN 1549-3636

[4] Bharati, R. Moona, P. Reddy, B. Sankar, D.M. Sharma & R. Sangal, (2003) "Machine Translation: The Shakti Approach", Pre-Conference Tutorial, ICON-2003.

[5] G. V. Garje, G. K. Kharate, "Survey of Machine Translation Systems in India", International Journal on Natural Language Computing (IJNLC) Vol. 2, No.4, October 2013

[6] Harjit Singh, "Multilingual Machine Translation Approaches for Consolidated Translator: An Indian Languages Perspective", IJOES, vol. 21, pp. 33-38, Dec. 2016

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

[8] Sugata Sanyal & Rajdeep Borgohain, (2013) "Machine Translation Systems in India", Cornel University Library , arxiv.org/ftp/arxiv/papers/1304/1304.7728.pdf

[9] Gurpreet Singh Josan & Jagroop Kaur, (2011) "Punjabi To Hindi Statistical Machine Transliteration", International Journal of Information Technology and Knowledge Management , Volume 4, No. 2, pp. 459-463.

[10] Vishal Goyal & Gurpreet Singh Lehal, (2011) "Hindi to Punjabi Machine Translation System", in proceedings of the ACL-HLT 2011 System Demonstrations, pages 1–6, Portland, Oregon, USA, 21 June 2011.

[11] Ananthakrishnan R, Kavitha M, Jayprasad J Hegde, Chandra Shekhar, Ritesh Shah, Sawani Bade & Sasikumar M, (2006) "MaTra: A Practical Approach to Fully-Automatic Indicative English-Hindi Machine Translation", in proceedings of the first national symposium on Modelling and shallow parsing of Indian languages (MSPIL-06) organized by IIT Bambay, 202.141.152.9/clir/papers/matra_mspil06.pdf

[12] Vishal Goyal & Gurpreet Singh Lehal, (2009) "Advances in Machine Translation Systems", National Open Access Journal, Volume 9, ISSN 1930-2940 http://www.languageinindia.

[13] R.M.K. Sinha & A. Jain, (2002) "AnglaHindi: An English to Hindi Machine-Aided Translation System", International Conference AMTA(Association of Machine Translation in the Americas)

[14] Sanjay Kumar Dwivedi & Pramod Premdas Sukhadeve, (2010) "Machine Translation System in Indian Perspectives", Journal of Computer Science 6 (10): 1082-1087, ISSN 1549-3636, © 2010 Science.

[15] Sanjay Chatterji, Praveen Sonare, Sudeshna Sarkar & Anupam Basu, (2011) "Lattice Based Lexical Transfer in Bengali Hindi Machine Translation Framework", in Proceedings of ICON- 2011: 9th International Conference on Natural Language Processing, Macmillan Publishers, India. Also accessible from ltrc.iiit.ac.in/proceedings/ICON-2011.

[16] R. Ananthakrishnan, Jayprasad Hegde, Pushpak Bhattacharyya, Ritesh Shah & M. Sasikumar, (2008) "Simple Syntactic and Morphological Processing Can Help English-Hindi Statistical Machine Translation", in proceedings of International Joint Conference on NLP (IJCNLP08), Hyderabad, India.