Study on Machine translation approaches for Indian languages and their challenges

Sindhu D.V
Department of information science
R.VCollege of engineering
Sindhu.489@gmail.com

Sagar B M
Department of information science
R.V College of engineering
Sagar.bm@gmail.com

Abstract - This survey mainly focuses on the developments of machine translation for the Indian languages. The survey throws a light on rule-based approach, empirical based approach and hybrid based approaches for machine translation. Every approach has its own advantages and disadvantages. Machine Translation (MT) is a process which translates from one language to another language. Due to rapid globalisation there is an increased data over the web machine translation plays a very important role to reduce the language barrier between different regions. In a country like India with 22 official languages shows a high attention for the translation. This paper focuses on the different MT systems for Indian languages and also their challenges.

Keywords: Source language, Target language, Machine Translation, computational linguistics

I. INTORDUCTION

Machine Translation is the branch of computational linguistics deals with translating from source language (SL) to target language (TL) with or without assistance of human. With advanced researchin the field of computational linguistics changes the way of dealing with the data as a means of effective communication in the society. In a multilingual country like India the demand for translation tool as a means of exchange of information between the regions.

Machine Translation systems are designed either for two languages (source language and target language) or for more than two languages called as multilingual translation and also the direction of the system may be either bidirectional or unidirectional. Many Indian systems are basically unidirectional.

The research in machine translation started with the idea of using dictionaries to overcome the language barriers between different regions during 17th century.InIndia the research started bit late just more than two decades ago. Machine Translation for Indian languages is bit hard compared to other language because of their morphological richness and they process thousands of dialects.

India has 25 years of history in computational linguistics which has so many ups and downs. But during the last decade with increased computational resources India has marked a significant growth in the field of language computing and Machine Translation. Researchers in the field of computational linguistics has chosen translation as their main focus of research and in parallel many industries like Microsoft and Google supported the language computing in a big way. Active research in the field of MTmade a big shift in language technology which made language computing tools to enhance the products and services in high growth markets such as health care, mobile communication, online retail, call centres, media and publishing.

II. APPROACHES TO MACHINE TRANSLATION

. MT systems classified according to their working methodology. The various approaches of machine translation are shown in fig 1.Under this classification three main paradigms can be found they are Human translation with machine support, Machine Translation with human support and fully automated translation. Fully automated translation can be defined as translation process where there is no human intervention. Fully automated translation mainly categorised in to three main paradigms they are rule-based, empirical based and hybrid based. In the Rule based approach researchers provide set of rules to derive the translation on the other hand empirical based system generates the rules from the corpusand whereas hybrid system uses both the rules defined by the experts and by the system. Rule based approach mainly categorised in to three main paradigms they are direct translation, transfer based and Interlingua. Empirical based MT broadly classified in to two paradigms they are example based and statistical based.

The following section discusses different approaches in fully automated translation.

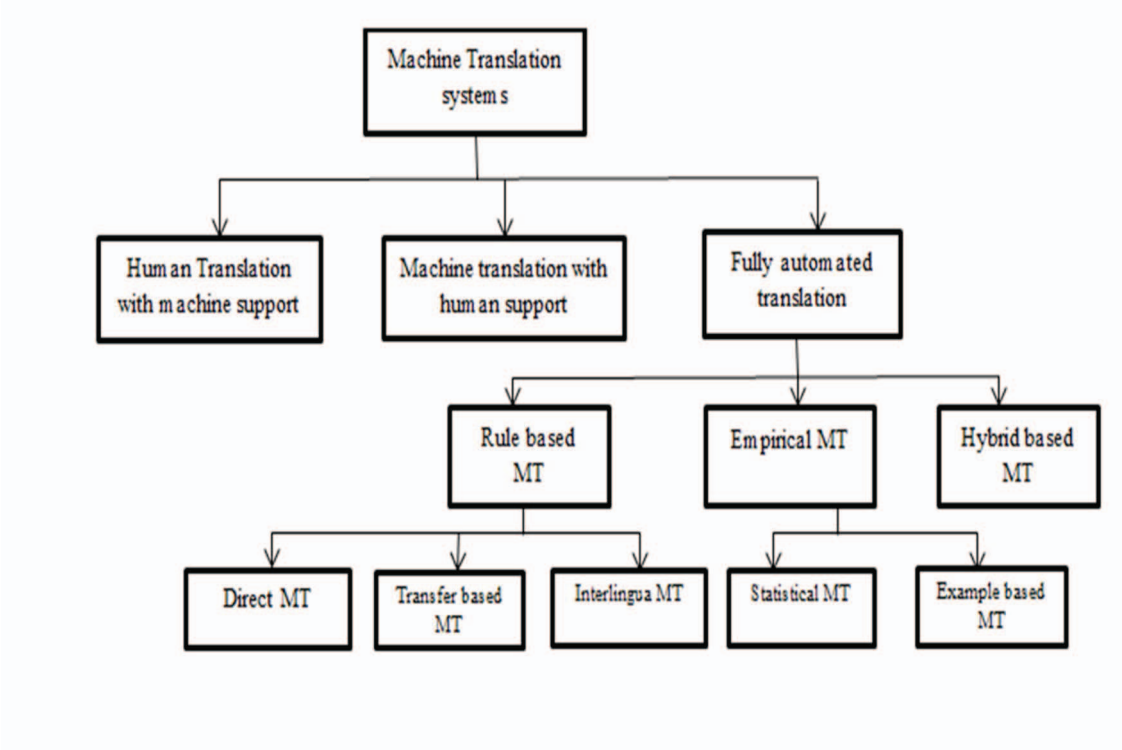

Figure 1: Approaches to MT

A. Rule based MT

In Rule based machine translation systems human linguisticexperts define the rules and these rules are applied at the three different stages of the MT system i.e. analysis, transfer and generation. Rule based MT is one of the oldestapproaches in the field of machine translation. The quality of machine translation depends on the how effectively the rules are defined. Rule based Machine translation in turn classified in to three different paradigms they are direct MT, transfer-based MT and Interlingua MT.

a. Direct based MT

Direct based translation approach is the first and least used approach. MT systems using this approach are capable translating from word to word direct translation systems are basically bilingual and uni-directional. Direct translation approach needs only few resources they are bilingual dictionary and little bit of syntactic and semantic analysis.

This section discusses Direct based MT systems for various indian languages. First Indian direct based system Anusaaraka was developed in 1995 for English-Hindi pair. The initial data was sufficient for bootstrapping process for better performance of the system the data needs to be enhanced qualitatively and quantitatively many times for the general use[1]. Later Anusaaraka was extended to Telugu, Kannada, Bengali, Punjabi and Marathi to Hindi.Later Hindi to Punjabi translation was developed in the year of 2010 for translating webpages and emails.

ANGLABHARTI MT system supports multilingual machine translation with human support works for English to Indian languages, mainly Hindi, which works on a pattern directed approach [2].Punjabi to Hindi MT system was developed at Punjabi University, Patiala which reports 92% accuracy because system is able to handle the word sense disambiguation, transliteration, and post-processing.

Challenges in Direct MT

- Main limitation of the Direct MT is lack of qualitative and quantitative bilingual dictionaries.
- Since the direct translation mainly works on word to word translation there will be a chance of mistranslation at the levels of lexical, word order and word sense.
- At the linguistic level the internal structure analysis of the source text is missing,

particularly syntactic analysis of the source text.

b. Transfer Based MT

To overcome the issues and drawbacks of direct translation the transfer based approach is developed. This approach involves three stages. First stage is analysis, where source text is analysed syntactically and in the second stage is transfer the syntactic structure of source text is transferred into syntactic structure of target text and third stage is generation where target text is generated from the syntactic structure of target text [3].

This section discusses Transfer MT systems for various indian languages. MaTra system was developed in 2008 by CDAC, Mumbai which uses transfer based approach using a frame like structured representation which resolves the disambiguation and provides the correct translation [5].

MaTra was developed for general purpose but mainly concentrated on translating news articles, annual reports and technical phrases.Mantra MT system was developed by CDAC, Bangaloreaim of this MT system was to translate the gazette notifications regarding government appointments and proceedings of the parliament between English and Indian languages and vice-versa.

UCSG-based MAT system is based on a transfer approach and developed for translating the government budget orders from English to Kannada the system works at sentence level and requires post-editing [4].

English –Hindi MAT was developed by Jadavpur University, Kolkata it is a transfer based domain specific MT system works for news articles[13]

CALTS in collaboration with IIIT, Hyderabad, Osmania Universityand Telugu University developed for the two language pairs they are English to Telugu and Telugu to Tamil. The system used bilingual dictionary with 42,000 entries and used already existing analyser generator and word sense disambiguate. It handles the source sentences with variable complexity [3].

Challenges in Transfer-based MT

- Defining rules and applying at all the stages is difficult.
- It is hard to build the reusable synthesis and analysis modules.
- It is difficult to build simple transfer module with minimal rules.
- It is hard for deriving rules for the big data with various ambiguities and dialects.

c. Interlingua based MT system

In Interlingua based MT system will take source text and converts into the intermediate representation called as Interlingua and then Interlingua is converted in to target text. One of the main advantages of the Interlingua is it plays a valuable role if the number of target languages increases because Interlingua is language independent.

This section discusses Interlingua based MT systems for various Indian languages. UNL (Universal Networking Language)-based MT for English-Hindi and Marathi developed by IIT, Mumbai uses the Interlingua approach. The source text information is represented in sentences later converted in to hypergraph [13]. The document knowledge can be represented in three columns as conceptual knowledge, attributes and lexical knowledge. Translation system using pictorial knowledge representation was developed in 2010. Here source language images and animations are converted in to pictorially grounded language which serves as Interlingua and later converted in to target language [6].

Challenges in Interlingua based MT system

- It is very hard to define an Interlingua even for closely related languages
- It is difficult to extract the information from source text to build the Interlingua.
- Defining universal Interlingua is the challenging task.

B. Empirical MT System

Empirical MT system is the emerging system which uses the large amount of parallel corpora. The empirical system uses the sentence aligned source and target text. Empirical system is broadly classified in two paradigms they are statistical and example based systems. Statistical machine translation works on the statistical modelling for the word order of the target

language text and source target word pair. The example based MT systemworks on analogical reasoning between the source and target language text. Example based system relies on large parallel aligned corpora and uses the set of source text and its equivalent target text as knowledge data base and generates the translation according to the example data base.

a. Statistical Machine Translation

Statistical MT [SMT] system is data oriented frame work for translating from source language to the target language based on the statistical models extracted from large amount of aligned bilingual corpora. Supervised and un-supervised machine learning algorithms are used for building the statistical models. The statistical models consist of statistical information such as co-relation between the SL and TL and well-formed sentences. During the translation statistical models helped to find the best translation for the source text.

English – Hindi statistical MT was developed by IBM research lab at New Delhi using the statistical based approach and this machine was designed for general purpose. A statistical MT system developed for Sinhala-Tamil language pair in 2011this system also works for English and Sinhala pair and this system proved that SMT gives good accuracy even for the languages which become to different families. Rule-based word Reordering and Morphological Processing for English-Malayalam SMT used the statistical based approach and proved that statistical based approach can increase the accuracy by using word ordering and morphological information of source and target languages [10].SMT was developed for English-Telugu system called "enTel" using Johns Hopkins University Open Source Architecture (JOSHUA). Parallel corpus was developed by using the CIIL Mysore Corpus andCharles Philip Browndeveloped English to Telugu Dictionary and that is used.

Challenges in Statistical MT system

- It's very hard to create the parallel corpus with the limited resources.
- The results are unexpected depends on the quality of resources available.
- Statistical MT system doesn't work well for the languages with different word order.
- Creating corpus for the languages with limited resources is time consuming task.

b. Example – Based MT system.

In the example based MT [EBMT] system, examples of few source language sentences and their equivalent target language sentences are stored in a database. During the run time, example based translation use bilingual translated text as its database. This database is stored in the translation memory. During the translation process the bilingual corpus is used as the reference for translation if there occurs a same sentence we can directly withdraw from memory and if a similar sentence is there then we can learn from bilingual corpus database and give the appropriate translation. The advantage of the example based translation the translation memory saves the user efforts for re translating the sentence and this saves the processor time and also the user time.

This section discusses Statistical MT systems for various Indian languages. SHIVA MT system was developed by Indian Institute of Science, Bangalore, and IIIT, Hyderabad, along with Carnegie Mellon University on example based approach and this was released for the user trails and feedbacks. From user feedbacks the performance of the system was developed [11].

English to Hindi, Kannada and Tamil and for Kannada – Tamil pair EBMT was developed in 2006 by balajapally. This system used sentence dictionary, bilingual dictionary and phonetic dictionary build using the parallel corpora and phonetic mappings are used to produce the translation. Corpus of 75,000 most commonly used sentences was used to build the system [12].

VAASAANUBAADA is EBMT system developed for the Bengali-Assamese language pair. The system uses aligned bilingual corpus created manually by using real examples. The performance of the systemcan be improved by handling the long sentences at the pre-processing stage only [3].

Challenges in EBMT System

- EBMT system avoids the burden of deriving the rules manually but to produce the exhaustive rules the database should contain exclusive examples which covers all the grammatical features.
- Identifying the feasible examples for database is time consuming task.

- Computational efficiency is the main problem with EBMT for large databases.

C. *Hybrid Machine Translation System*

Hybrid MT system takes the advantage both the rule-based and statistical based methodologies and by combining both the approaches new hybrid approach is derived to overcome the drawbacks of both systems , which has proven to give the better performance than other systems. At present many organisations use hybrid based approach which works on both statistics and user defined rules. In few cases the translations are performed using a rule-based approach and later output by using statistical information or vice versa also can be done [2]. Hybrid based MT is more flexible when compared to other systems in increasing the performance.

This section discusses Hybrid based MT systems for various Indian languages. METIS-II MT system is a Hybrid system this system uses the bilingual dictionary and avoids the use of parallel corpora and uses monolingual corpus in the target language. Rule based paradigm is provided by Oepen, which integrates the statistics within RBMT [11].

ANUBHARATI is an hybrid based system derived by combining the advantages of pattern based and example based approach is derived combining the some of the features of both the methods. This system can be used for Indian languages with minor modifications in the modules[6].

SivajiBandyopadhyay, has developed a hybrid-based MT system for English to Bengali language pair was developed in 2004 at Jadavpur University, Kolkata,. The present version of the system works at the sentence level [13]..

VamshiAmbatihas developed a hybrid approach for English to Indian languages and system uses minimal linguistic resources. Dictionary is built manually and also by using SMT tool from the example data base contains both source and target language sentences [8].

Challenges in Hybrid MT System

- The accuracy of the translation depends on the quality of bilingual corpus. Building bilingual corpus with high level of similarity is very tedious and costly process.
- Developing good quality of training corpus is hard job.
- Hybrid MT fails for the language pairs with minimal resources.

CONCLUSION

Machine Translation is as old as computers but in recent years it has become the active research in Natural Language Processing. Machine Translation is hard problem for Indian languages because of their complex language structure.

From the survey for Indian languages it is found that many Indian systems are working on statistical and hybrid based systems. The reason is rule based system have failed to give good accuracy because Indian languages are morphologically rich and process hundreds of dialects. The paper also describes the different approaches to the Machine Translation and their challenges. The survey shows that there is no full-fledged approach, though the severities in some of the approaches are very less. Taking advantage of best features of few systems a new system can be derived, which helps to overcome the challenges of many MT systems.

REFERENCES

1. Amba P. Kulkarni ,Design and Architecture of ‘Anusaaraka’-An Approach to Machine Translation in Research Gate April 2003 volume.1.
2. Sinha, R. M. K., Sivaraman, K., Agrawal, A., Jain, R., Srivastava, R. & Jain, A. ANGLABHARTI: a multilingual machine aided translation project on translation from English to Indian languages. IEEE International Conference on: Systems, Man and Cybernetics, 1995. Intelligent Systems for the 21st Century, 1609-1614.
3. M. D. OkporMachine Translation Approaches: Issues and ChallengesJCSI International Journal of Computer Science Issues,September 2014 Vol. 11, Issue 5, No 2
4. Murthy. K, “MAT: A Machine Assisted Translation System”, In Proceedings of Symposium on Translation Support System (STRANS-2002), IIT Kanpur. pp. 134-139.
5. CDAC Mumbai, “MaTra: an English to Hindi Machine Translation System”, a report by CDAC Mumbai formerly NCST 2008.
6. Antony P.J “Machine Translation Approaches And Survey For Indian Languages, Computational Linguistics And Chinese Language Processing , Vol.18 ,No.1 , March 2013 ,pp.47-78.
7. Bharati, R. Moona, P. Reddy, B. Sankar, D.M. Sharma & R. Sangal, “Machine Translation: The Shakti Approach”, Pre-Conference Tutorial, ICON-2003.
8. R.M.K. Sinha & A. Jain, “AnglaHindi: An English to Hindi Machine-Aided Translation System”, International

Conference AMTA(Association of Machine Translation in the Americas) 2002.

9. AksharBharti, ChaitanyaVineet, Amba P. Kulkarni& Rajiv Sangal,"ANUSAARAKA: Overcoming the language barrier in India", published in Anuvad: approaches to Translation , 2001.
10. Rahul C, Dinunath K ,RemyaRavivardhan ,K.P SomanRule Based Reordering and Morphological Processing for English-Malayalam Statistical Machine TranslationAdvances in Computing, Control, & Telecommunication Technologies,. ACT,2009.
11. Ambati, V &Rohini, U. A Hybrid Approach to EBMT for Indian Languages.ICON 2007.
12. Balajapally, P., Bandaru, P., Ganapathiraju, M., Balakrishnan, N., & Reddy, R. Multilingual Book Reader: Transliteration, Word-to-Word Translation and Full-text Translation. In VAVA 2006.
13. Dave, S., Parikh, J., & Bhattacharya, P. Interlingua-based English-Hindi Machine Translation and Language Divergence. Journal of Machine Translation, 2001, 251-304.
14. Dwivedi, S. K., &Sukhadeve, P. P. MachineTranslation System in Indian Perspectives. Journal of Computer Science, 2010, 1111-1116.
15. Goyal, V., &Lehal, G. S. Evaluation of Hindi to Punjabi Machine Translation System. IJCSI International Journal of Computer Science,2009, 4(1), 36-39