Hybrid Machine Translation for Indian Languages – A Comparative Study

Literature Review

Jyoti Sheoran
Computer Science Department
University of Victoria, Victoria, Canada
jsheoran@uvic.ca

Abstract **--- The author proposes a comparative study of hybrid Machine Translation approach for Indian language pairs – both related and un-related language pairs. India being a multi-lingual country has huge requirement for government documents, news and text to be translated in other official languages. However, current Machine translation systems for Indian languages lack accuracy and cannot be used as a fully automated translation systems. Hybrid Machine Translation approaches show some improvement in the accuracy but these are biased towards a specific language pair. Further research is necessary to test their accuracy for Indian language pairs. This paper provides a brief overview of the Machine Translation approaches and the strategy for conducting a comparative study of hybrid Machine Translation system for Indian languages.**

Keywords—Machine Translation, RBMT, SMT, hybrid MT, Indian languages, UNL, AnglaBharti, AnglaHindi

I. INTRODUCTION

Machine Translation (MT) can be defined as the translation from one natural language (source language) to another natural language (target language) using machines. There is increasing demand for translation of government documents, manuals, news, websites, e-mails, etc. because of increase in cross-culture communication. As most of these translation tasks are repetitive, there is a need for using a machine to do the translation. Moreover, there is shortage of qualified human translators for many technical fields. The need for MT systems is even more in multi-lingual nations like India. India has 22 official languages and more than 100 regional languages and dialects. While the central government work is bilingual - English and Hindi, the state governments are free to use any regional language. Hence, majority of state governments use their regional language for government documents, education and businesses. There is huge requirement for translation of documents for effective communication. Indian tourism and media industry is also interested in using a machine translation system to reach broader audience.

Indian languages can be classified into 4 main families – Indo-Aryan family (Hindi, Bangla, Punjabi, Marathi, etc.), Dravidian family (Tamil, Telugu, Kannada, Malayalam, etc.), Austro-Asian family and Tibetan-Burmese family. Languages that belong to same family are called related languages and are require less effort to translate as compared to un-related languages. This is because related language are similar in grammatical structure. Many MT systems have been proposed for Indian languages and are widely used today. Most of the proposed systems are for English-Hindi translations for example, MANTRA, Angla-Hindi, Anuvadak, etc. There are also a few MT systems for languages belonging to same family and for un-related language pairs. AnglaBharti is a Machine-aided Translation system proposed to translate English to Indian language families. Shakti is another Machine-aided Translation system to translate English text to Hindi, Telugu or Marathi. But there is no full-fledged MT system that can translate between any of the 22 Indian languages. Also, these systems lack accuracy and require post-processing by human translators.

Machine Translation process involves decoding of the meaning of the source language text and then re-encoding the meaning in the target language. This process requires in-depth knowledge of the grammar, semantics, syntax, idioms, etc. of both source and target languages. Many models have been proposed for Machine Translation. But the translation done by these MT models is still not reliable. Today's MT systems require post-processing by human translators and hence cannot be used as fully automated systems. In the recent past, some hybrid MT approaches have been proposed that combines two or more individual MT models. The evaluation of hybrid MT approaches for translating English text to other languages shows significant improvement in accuracy. English sentences have Subject+Verb+Object structure, but Indian languages have Subject+Object+Verb structure. Indian languages are morphologically richer than English. Further research is required to study the accuracy of hybrid MT systems for Indian language pairs. In this survey we will study the approaches for MT systems and their applicability for Indian languages.

II. MT Approaches

There are two generations of MT systems – the direct systems and the indirect systems. First generation systems known as direct systems used word-by-word or phrase-by-phrase translation. The source language text is first analyzed to the minimal extent to apply transformational rules, and then source language words are replaced with target language words using a bilingual dictionary. Finally words are re-arranged according to the target language. The resulted target language sentence is poor in grammar. These systems are difficult to extend to other languages as rules are written in one direction and are language specific. SYSTRAN is a direct system described in Hutchins and Somers [1]. It was used by Xerox to translate technical manuals.

The second-generation MT systems use linguistics of text for translation. These systems are called indirect systems or Rule-based MT systems (RBMT). The source language text is analyzed and transformed into a logical form to generate target text. Indirect systems can be further divided into two systems – Interlingua based systems and Transfer based systems. **Interlingua based** systems use an intermediate language for the transfer of source language text to the target language text. These systems are easy to extend to other languages as only analysis and generation modules are required to add a new language. Universal Networking Language (UNL) [2] project is an example of Interlingua based systems. UNL was proposed by UNU to access, transfer and process information on the Internet. It has been used to develop translation between UNL to Hindi, Bengali and Marathi. But there are no evaluations done to test or compare the accuracy of this system. **Transfer-based** systems create an abstract level of source text. This abstract level is then converted into the corresponding abstract level in target language. The final text is generated from the target language's abstract level. These systems use independent grammar of both source and target language. These systems are easy to extend to other languages and have good grammatical structure. However, the target language text has poor lexical selection. Also, these systems fail to parse any grammatically incorrect sentence. AnglaBharti [3] is an example of pattern-directed rule-based system for English to Indian Languages. The input English text is analyzed and a pseudo-target (intermediate) text is generated. This intermediate text is then translated to target language. The pseudo-target text can be used for any language belonging to same family. The results lack accuracy and require post-editing by human translators.

Example-based machine translation (EBMT) [4] translates by matching source language text with stored example translations. It uses bilingual corpus of translation pairs. This MT model has limited use, as it requires huge bilingually aligned corpus. ANUBAAD is a MT system that uses example-based approach to translate English news headlines to Bengali. ANUBAAD was also used to develop English-Hindi language pair, but it only works for simple sentences. Shakti is another EBMT system developed for English to Hindi, Marathi and Telugu. Shakti is an interactive Machine-aided Translation system that allows users to replace words or phrases. Both these systems lack accuracy and require post-processing by humans. Also, there is lack of large bilingual corpus for most of the Indian languages. Hence, these systems cannot be extended to other Indian languages.

There are also some **empirical approaches** for Machine Translation. These apply statistical or pattern matching techniques. Statistical Machine Translation (SMT) [5] uses probabilistic analysis for tasks like word-sense disambiguation, or lexical structure disambiguation. It uses bilingual text corpora for probabilistic analysis. The target language text has good lexical structure but lack in grammar. The quality of target language text depends on the availability of large bilingual text corpora. SMT are widely used today. Google Translator is a SMT based system. It provides translation between 7 Indian languages. SMT systems require post-processing by human translators. These systems are less feasible for Indian languages due to lack of availability of large bilingual corpora.

III. Hybrid MT Approaches

In the past decade, significant research has been carried out in knowledge-based approaches (RBMT) and data-driven approaches (SMT and example-based). The study by Callison [6] showed that both RBMT and SMT systems reach comparable translation quality but the accuracy is only about 50%. These systems require post-processing of target text by human translators and hence cannot be used as fully automated MT systems. However, both these systems have complementary properties. RBMT systems are good in grammar but lack lexical selection. SMT systems are good in lexical selection but lack grammar. Hence it is required to come up with hybrid solutions combining both SMT and RBMT systems in order to increase the accuracy of the MT systems.

Thurmair [7] provides a classification of Hybrid MT system architectures. Parallel-coupling which is a combination of multiple individual MT systems in parallel performs poorly than individual MT systems. Moreover, it is impractical to use in practice because of huge resource requirement. Serial coupling involves feeding the output of one MT system to the other MT system. These systems also do not provide any substantial increase in accuracy. The output is worse when SMT is used in the last, as SMT system don't know how to deal with syntactic structure of text. If output of SMT system is fed to RBMT systems, then there is problem failure at parsing stage in RBMT as SMT systems give grammatically incorrect text. Hence, simple combination of these two MT systems is not enough to get a significant improvement in accuracy. These two systems must be coupled in truly hybrid manner, i.e. combining the individual components of RBMT and SMT systems. Many hybrid MT approaches have been proposed, but their evaluation is done on only one or two language pairs. Hybrid MT approaches tend to perform better in a particular language pair and perform badly in other language pairs. Hence, further research is required to study if these models can be extended to other language pairs.

Statistical pre-editing of RBMT: This approach involves pre-editing the source language text using statistical techniques

before feeding it to the RBMT system. It uses bilingual text corpus. AnglaHindi [8] is an example of such an approach that uses generalized example-base and statistical techniques for pre-editing of RBMT system. It shows better accuracy than RBMT systems. However, it has parse failure issue when it encounters grammatically incorrect sentences. Further research is required to extend this approach to other Indian languages as it currently translates only English-Hindi texts.

RBMT pre-editing of SMT: This approach [9] uses RBMT system at the parsing phase of SMT. The bilingual dictionary is used to translate words. The results are better for both in-domain and out-of-domain. The SMT decoder runs last which leads to grammatical errors in the target text. The evaluations were done on English-German and English-Spanish language pairs. There are no evaluations available for Indian language pairs.

LiSTEX [10] combines statistical term extraction techniques with RBMT techniques. Additional linguistic layers are added to ensure that the extracted terms are tailored for RBMT system. The source and reference texts are tokenized and passed to RBMT engine to generate trees. Then extraction of terms is performed using both statistical data and RBMT trees. Next step involves linguistic filtering and transformation to generate the final term pairs. After this step, a final linguistic based selection of terms is performed and then RBMT system is used for target text generation. The evaluations were done on both related (English-German) and un-related (English-Spanish) language pairs, which showed 40-50% better translation quality and 80% accuracy. However, the results also showed 10-20% worst translation quality. Further research is required to improve the translation quality. There is no research done to evaluate the performance of LiSTEX for Indian language pairs.

CONCLUSION

Some hybrid MT models show promising results in terms of accuracy and translation quality. But these models are tested on only one or two language pairs. Most of the hybrid approaches are evaluated for English language paired with some other related or un-related language. English language is significantly different from Indian languages in terms of grammatical structure and morphology. Also, there is deficiency of large amount of parallel corpus for Indian languages. Hybrid systems tend to be biased towards a specific language pair or a specific evaluation metrics. Hence evaluation done on a Hybrid approach using one particular language pair cannot confirm that it will give same level of accuracy for all families of Indian languages. At present, we do not have a hybrid model that provides better accuracy and can also be extended to other language pairs. Also, there is not much research done on the formal evaluation of the performance of new hybrid models for Indian languages. A comparative study of the hybrid model with Indian languages (both related and un-related) is required in order to test the accuracy of hybrid MT models on Indian languages.

The promising results from LiSTEX MT approach provide clear motivation to compare its the accuracy for Indian language pairs. The comparative study should involve both related and un-related language pairs. My proposal is to conduct a comparative study of LiSTEX MT approach for Indian languages by selecting 3 sets of related and un-related language pair from the Indian language families. The results will be evaluated using 3 evaluation metrics – BLEU [11], Translation Error Rate [12] and NIST. Human evaluation will also be done, as automated evaluation metrics are not perfect. The results will be compared with current individual MT systems. For this, any public MT systems, for example Google Translator could be used. This study will not only test the feasibility of LiSTEX MT approach for Indian language, but the results could also help in extending this system for any language pair. A future research could be to propose improvement in the LiSTEX MT approach to further improve accuracy and translation quality.

REFERENCES

[1] Hutchins, W.J. and Somers, H.L. (1992) An introduction to machine translation. London: Academic Press.

[2] Dave S, Parikh J, Bhattacharyya P 2001 Interlingua based English Hindi machine translation and language divergence. *J. Mach. Trans. (JMT)* 16(4):251–304.

[3] Renu Jain, R.M.K. Sinha and A. Jain. 1995. A Pattern Directed Hybrid Approach to Machine Translation through Examples, In Proc. Symposium on Natural Language Processing, SNLP'95 , Bangkok, Thailand, pp 325-335.

[4] Somers H., Example-based Machine Translation, Machine Translation, 14(2), 113-157,1999.

[5] Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Comput. Linguist.* 16, 2 (June 1990), 79-85.

[6] Callison-Burch, Chris, Philipp Koehn, Christof Monz, and Omar F. Zaidan. "Findings of the 2011 workshop on statistical machine translation." In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pp. 22-64. Association for Computational Linguistics, 2011.

[7] Thurmair. 2009. Comparing different architectures of hybrid machine translation systems. In *Proc MT Summit XII*.

[8] Sinha, R.M.K. and A. Jain, 2003. Anglahindi: An English to Hindi machine-aided translation system. Proceeding of the 9th MT Summit, (MTS'03), MT- Archive, New Orleans, Sept. 23-27, USA., pp: 494-497.

[9] Eisele, Andreas, Christian Federmann, Hans Uszkoreit, Hervé Saint-Amand, Martin Kay, Michael Jellinghaus, Sabine Hunsicker, Teresa Herrmann, and Yu Chen. "Hybrid machine translation architectures within and beyond the EuroMatrix project." In *Proceedings of the 12th annual conference of the European Association for Machine Translation (EAMT 2008)*, pp. 27-34. 2008.

[10] Wolf, Petra, Ulrike Bernardi, Christian Federmann, and Sabine Hunsicker. "From Statistical Term Extraction to Hybrid Machine Translation." In *15th International Conference of the European Association for Machine Translation*, p. 225. 2011.

[11] Papineni, S. Roukos, T. Ward, and W. Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meet- ing of the Association for Computational Linguistics*, pages 311–318.

[12] Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. "A study of translation edit rate with targeted human annotation." In *Proceedings of association for machine translation in the Americas*, pp. 223-231. 2006.