IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-8, Aug 2018 E-ISSN: 2347-2693

Machine Translation In Indian Languages

Deepti Chopra[1], Nisheeth Joshi[2], Iti Mathur[3]

[1,2,3]Dept. of Computer Science, Banasthali Vidyapith, Newai, India

[*]*Corresponding Author: deeptichopra11@yahoo.co.in*

Available online at www.ijcseonline.org

Accepted: 15/Aug/2018, Published: 31/Aug/2018

***Abstract*—** Machine Translation (MT) is one of the tasks of Natural Language Processing. It can be used by intellectuals to attain information from the documents written in different languages. In the following paper, we have discussed problems faced in MT in Indian languages, various approaches of MT, limitations of some of the current existing MT Systems and the work that has been done till now in MT in Indian language perspective. We have also discussed performance metrics that are used for evaluation of MT System.

***Keywords*—**Machine Translation, Rule Based Approach, Example Based Approach

I. INTRODUCTION

Machine Translation may be defined as the task of translation of text from one language to another. There are two kinds of MT namely Metaphrase and Paraphrase. In Metaphrase, there is an exact word for word translation or lexical translation but the translated text may or may not have the similar semantics as the source text. In Paraphrase, translation is not performed at the word level but at the sentence level. Here, the semantics of source text is conserved while translating it into the translated text. Today, there are many Machine Translators available pertaining to Indian languages such as Anusaaraka, Mantra, Punjabi to Hindi MT Systems, Shiv, and Shakti, Anglabharti, Anubaad, Vaasaanubaada, Hinglish MT Systems, Anubharti etc[1]. But, still, these machine translators do not produce translations with very high accuracy.

II. PROBLEMS FACED IN MT IN INDIAN LANGUAGES

Problems faced in MT in Indian languages include the following-

1. Indian languages are free word order languages.

2. They are morphologically and inflectionally rich languages.

3. Named Entity Recognition (NER) can be used to improve MT. But, NER in Indian languages is not an easy task since these languages do not provide capitalization information that helps in performing NER.

4. Many common nouns exist as proper nouns. So, these languages involve a large amount of semantic ambiguity.
5. There is a scarcity of resources pertaining to Indian languages on the web.

Consider the following Source text-
"Rajasthan is India's largest state by area (342,239 square kilometers (132,139 sq mi) or 10.4% of India's total area). It is located on the northwestern side of India, where it comprises most of the wide and inhospitable Thar Desert (also known as the "Rajasthan Desert" and "Great Indian Desert") and shares a border with the Pakistani provinces of Punjab to the northwest and Sindh to the west, along the Sutlej-Indus river valley. Elsewhere it is bordered by five other Indian states: Punjab to the north; Haryana and Uttar Pradesh to the northeast; Madhya Pradesh to the southeast; and Gujarat to the southwest.

This source text in English is translated into Hindi using different machine translators i.e. Google Translator, Bing Translator and Babylon Translator. The translations are shown in figure 1.

Machine Translator	Translated Text
Google Translator	राजस्थान क्षेत्र द्वारा भारत का सबसे बड़ा राज्य है (342,239 वर्ग किलोमीटर (132,139 वर्ग मील) या भारत के कुल क्षेत्रफल का 10.4%)। यह भारत के उत्तर पश्चिमी तरफ स्थित है, जहां इसमें व्यापक और अप्रचलित थार रेगिस्तान (जिसे "राजस्थान रेगिस्तान" और "ग्रेट इंडियन रेगिस्तान" भी कहा जाता है) शामिल है और पंजाब के पाकिस्तानी प्रांतों के साथ सीमा साझा करता है। उत्तर-पश्चिम और सिंध पश्चिम में, सतलज-सिंधु नदी घाटी के साथ। दूसरी जगह यह पांच अन्य भारतीय राज्यों से घिरा हुआ है: पंजाब उत्तर में; हरियाणा और उत्तर प्रदेश पूर्वोत्तर में; मध्यप्रदेश में मध्य प्रदेश; और गुजरात दक्षिणपश्चिम में।
Bing Translator	क्षेत्रफल के अनुसार राजस्थान भारत का सबसे बड़ा राज्य है (३४२,२३९ वर्ग किलोमीटर (१३२,१३९ वर्ग मील) या भारत के कुल क्षेत्रफल का १०.४%) । यह भारत के उत्तर पश्चिमी ओर स्थित है, जहां यह व्यापक और दुर्गम थार रेगिस्तान के सबसे शामिल है (के रूप में भी जाना जाता "राजस्थान डेजर्ट " और "ग्रेट इंडियन डेजर्ट ") और पंजाब के पाकिस्तानी प्रांतों के साथ एक सीमा के शेयर पश्चिमोत्तर और सिन्ध पश्चिम में, सतलुज-सिंधु नदी घाटी के साथ । कहीं यह पांच अंय भारतीय राज्यों की सीमा है: पंजाब को उत्तर; हरियाणा और उत्तर प्रदेश पूर्वोत्तर के लिए; मध्य प्रदेश दक्षिणपूर्व; और दक्षिण पश्चिम के लिए गुजरात ।
Babylon Translator	राजस्थान राज्य क्षेत्र द्वारा भारत के सबसे वंशवृक्ष&#वर्ग से इतर वाहनों (342,239 132,139 वर्गमी.) 10.4% या एमाऐ वर्ग किलोमीटर (इंडिया) कुल क्षेत्र के वंशवृक्ष&#वर्ग से इतर वाहनों यह उत्तर पवश्चमी किनारे पर स्थित है, जिसमें भारत के अधिकांश व् यापक और असहयोगात्मक थार मरूस्थल (34,790 से भी राजस्थान मरुभूमि एंड quot " सिखाता है; और " वंशवृक्ष महान भारतीय वन एवं quot;) और शेयरों के साथ पाकिस्तान के पंजाब प्रांत की सीमा के साथ-साथ पश्चिमी देशों, सिनध तथा पश्चिमोत्तर Sutlej-Indus नदी घाटी

Figure 1 Output of Existing Machine Translators

The translated texts as shown in Figure 1 are not of good quality. Some of the tokens in a translated text appear in English. And, some of the words are transliterated instead of translated. Named Entities are not recognized properly. Babylon Translator has not translated or transliterated 'Sutlej-Indus' present in the source text. Complex sentences present in the source text are not translated properly by Google Translator, Bing Translator, and Babylon Translator.

III. APPROACHES OF MT

Various Approaches of MT include the following-

1. Rule based MT
2. Corpus-based MT
3. Example based MT
4. Statistical MT
5. Context-based MT

TABLE.1- Various Approaches of MT

Approach	Description	Advantage / Disadvantage
Rule based MT (RBMT)	Rule ar s e formed and it involve analysi s s of source and target text atsyntactic, semanti and c morphological level .	Provid good e quality Translations Complex Rules need to be constructed Tedious task Time-consuming
Corpu s based MT	Rules are constructed by paralle analysis of l bitext corpu s	ca Accuracy n be improved by adding exampl more es to the corpus.
Example based MT (EBMT)	It performs translation by analogy. It makes use of Translatio Memorie n s. It involve s three stages: Alignmen t, Matching and Recombinati on. Preprocessing the source text is not compulsory to perform.	Quality of MT can be improved by adding exampl more es to parallel corpus
Statistical MT (SMT)	Statistical models are used to perform MT. Parameters for statistic model al s are constructe analys d by is of parallel corpus.	Qualit y of MT can improve be d by addin g more examples to parallel corpu s

		□ Even if training data has same input, translation may not be produced in output
Context based MT	No Parallel Corpus is used. It make use of Bilingual Dictionary, target language Corpus and Source Langauge	□ Produces good accuracy □ Less tedious and Less time consuming

The hybrid approach involves a combination of above-listed approaches. Quality of MT is expected to improve if a hybrid approach is used to perform MT in Indian languages.

IV. LITERATURE REVIEW

The work that is done in MT pertaining to Indian languages is shown in TABLE II.

TABLE.1- Detailed Description of MT of different language pairs

Authors	About	Detailed Description
Ramanathan, Ananthakrishnan(2008)[2]	English to Hindi MT using SMT approach	Training- 120153 words, Testing- 8557 words. BLEU score (using Baseline Approach) is 12.10. BLEU score (by combining syntactic, morphological and baseline approach) is 15.88
Sinha et. Al (2003)[1]	English to Hindi MT(EBMT+ RBMT+ Post Editing Approach)	Angla MT can produce 90% correct results for sentences to a length of 20 words
Sinha et. Al (2005)[3]	English-Hindi bilingual text to Hindi	Morphological Analyzer is used to detect unknown words and unknown plural words of Hindi and English. This approach has given correct result in 90% of cases.
Ambati et. Al (2007)[4]	English to Hindi MT (Hybrid Approach)	EBMT and SMT are used to perform MT. Parallel corpus (54K English-Hindi sentences) is used. Training- 53K sentences, Testing- 100 sentences. BLEU score is 0.432.
Soni A et. al (2013)[5]	Hindi to English MT	MT output is improved by simplification of the source text. Testing- 100 sentences are taken. BLEU score is 0.805
Goyal, Vishal and Gurpreet Singh Lehal (2009)[6]	Hindi to Punjabi MT	Overall Accuracy- 95.12%. Input is taken from daily news, articles, official language quotes, blog and literature. 95.4% of sentences are found to be intelligible. The accuracy obtained is 87.6%

Josan, Gurpreet Singh and Gurpreet Singh Lehal (2008)[7]	Bilingual Hindi English Text to pure Hindi and pure English	English and Hindi Morphological Analyzers is used to detect English and Hindi words. Plural forms are also identified. The unknown words are considered to be proper nouns. Complex sentences are converted to simplified sentences and finally source text is translated to pure Hindi and pure English. In 90% of cases, this approach has obtained satisfactory results.
Rama, Taraka and Karthik Gali (2009)[8]	Transliteration of English to Hindi using SMT	Accuracy-46.3%. Alignment of English and Hindi letters is done using GIZA++, SRILM toolkit was used for training language toolkit. Mean F-measure obtained using this approach is 0.876.
Poornima C. et. al (2011)[9]	English to Tamil MT	Rule based Text simplification approach is used for enhancing English- Hindi MT. Testing-200 sentences. 115 sentences gave accurate results. The accuracy of the MT system increased by 28% by introduction of Text Simplification Approach in the MT system.
Germann Ulrich (2001)[10]	Tamil to English MT	A statistical machine translation system is built to perform Tamil to English machine translation. A bilingual corpus comprising of Tamil and English sentences is formed consisting of 1300 Tamil-English sentence pairs. Tamil side consisted of 24, 000 tokens.
Islam M et. al (2010) [11]	English to Bangla MT (SMT)	Phrase-based MT is performed in English to Bangla MT. Transliteration Approach is used to deal with the words not present in vocabulary. The accuracy of the Transliteration

		module is 0.18. Preposition handling is also performed. Overall the BLEU score of our system is 11.7. BLEU score obtained for short sentences is 23.3 and 0.63 TER

V. EVALUATION

For evaluation of MT system, three metrics can be used i.e. Precision, Recall, and F-measure.

Precision (P) = Match/System Output

Recall (R) = Match/Human Output

F-Measure(R) = 2*P*R / (P+R)

Here, Precision is calculated by considering the number of matches between the two outputs divided by the total number of system outputs. The recall is calculated by considering the number of matches between the two outputs divided by the total number of human outputs and F-Measure would be the combination of the two. Apart from these metrics, BLEU, METEOR etc. can also be used for evaluation of MT output.

Bilingual Evaluation Understudy (BLEU) - Its value lies between 0 and 1. It indicates how close a Machine translated text is to the expected translated text. Average of BLEU scores of all sentences is taken to get the overall score of the whole corpus.

NIST - Apart from calculating n-gram precision, it also assigns weights to n-gram. A low weight is assigned if n-gram matches exactly with the expected translation otherwise high weight are assigned.

Word Error Rate (WER) - This metrics estimates the number of tokens that differ between Machine translated text and expected translated text.

Meteor - This metrics estimates weighted harmonic mean of unigram precision and recall. It also involves matching of synonyms and lemmatized forms.

LEPOR - This metrics involved collection of different evaluation factors such as precision, recall, sentence length penalty and word order penalty based on n-gram.

VI. CONCLUSION

In this paper we have discussed MT, the problems faced in MT in Indian language context, problems with existing machine translators, approaches of MT and the work that has been done till now in Indian languages in MT. There is a lot of scope in MT in Indian languages. There is a need to develop a machine translator that can provide good translations with high accuracy.

REFERENCES

[1]. Sinha, R. M. K., and A. Jain. "AnglaHindi: an English to Hindi machine-aided translation system." MT Summit IX, New Orleans, USA (2003): 494-497.

[2] Ramanathan, Ananthakrishnan, et al. "Simple Syntactic and Morphological Processing Can Help English-Hindi Statistical Machine Translation." IJCNLP. 2008.

[3]. Sinha, R. Mahesh K., and Anil Thakur. "Machine translation of bi-lingual Hindi-English (hinglish) text." 10th Machine Translation Summit (MT Summit X), Phuket, Thailand (2005): 149-156.

[4]. Ambati, Vamshi, and U. Rohini. "A hybrid approach to example-based machine translation for Indian languages." Proceedings ICON (2007).

[5] Soni A et al. (2013) "Exploring Verb Frames for Sentence Simplification in Hindi" In Proceedings of International Joint Conference on Natural Language Processing, Nagoya, Japan, 14-18 October (2013), pp-1082-1086

[6] Goyal, Vishal, and Gurpreet Singh Lehal. "Evaluation of Hindi to Punjabi machine translation system." arXiv preprint arXiv:0910.1868 (2009).

[7]. Josan, Gurpreet Singh, and Gurpreet Singh Lehal. "A Punjabi to Hindi machine translation system." 22nd International Conference on Computational Linguistics: Demonstration Papers. Association for Computational Linguistics, 2008

[8] Rama, Taraka, and Karthik Gali. "Modeling machine transliteration as a phrase-based statistical machine translation problem." Proceedings of 2009 Named Entities Workshop: Shared Task on Transliteration. Association for Computational Linguistics, 2009.

[9] Poornima, C., et al. "Rule based sentence simplification for English to Tamil machine translation system." International Journal of Computer Applications 25.8 (2011): 38-42.

[10] Germann, Ulrich. "Building a statistical machine translation system from scratch: how much bang for the buck can we expect?." Proceedings of the workshop on Data-driven methods in machine translation-Volume 14. Association for Computational Linguistics, 2001.

[11]. Islam, Md Zahurul, Jörg Tiedemann, and Andreas Eisele."English to Bangla phrase-based machine translation."Proceedings of the 14th Annual conference of the European Association for Machine Translation (2010).

[12]. Joshi, Nisheeth, Hemant Darbari, and Iti Mathur. "Human and Automatic Evaluation of English to Hindi Machine Translation Systems."Advances in Computer Science, Engineering & Applications. Springer Berlin Heidelberg, 2012. 423-432.

Authors Profile

Deepti Chopra has worked as an Assistant Professor in the Department of Computer Science at Banasthali University. She has also worked as a guest faculty at Guru Nanak Dev Institute of Technology, Delhi and Integrated Institute of Technology, Delhi. Her primary area of research is Computational linguistics, Natural Language Processing and Artificial Intelligence. She is also involved in development of MT engines.

Nisheeth Joshi is working as Associate Professor in the Department of Computer Science at Banasthali University. His areas of interest include Computational Linguistic, Natural Language Processing and Artificial Intelligence. Besides this, he is also very actively involved in the development of MT Engines for English to Indian Languages. He is one of the experts empaneled with the TDIL program, Dept. of Information Technology. He has several publications in various journals and conferences and also serves on the program committees and editorial boards of several conferences and journals.

Iti Mathur has worked as Associate Professor in the Department of Computer Science at Banasthali University. Her areas of interest include Computational Linguistic, Soft Computing, Natural Language Processing and Artificial Intelligence. She has several publications in various journals and conferences and also serves on the program committees and editorial boards of several conferences and journals.

