A Review on Machine Translation in Indian Languages

Deepti Chopra
Department of Computer Science
Banasthali University
Newai, India
deeptichopra11@yahoo.co.in

Nisheeth Joshi
Department of Computer Science
Banasthali University
Newai, India
jnisheeth@banasthali.in

Iti Mathur
Department of Computer Science
Banasthali University
Newai, India
mathur.iti@rediffmail.com

***Abstract*—Machine translation (MT) is considered an important task that can be used to attain information from documents written in different languages. In the current paper, we discuss different approaches of MT, the problems faced in MT in Indian languages, the limitations of some of the current existing MT systems, and present a review of the work that has been done until now in MT in Indian language perspective.**

Keywords-machine translation; rule based approach; example based approach; natural language processing; bilingual evaluation understudy

I. INTRODUCTION

Machine translation may be defined as the task of conversion of text from one language, called source language, to another language, called target language. There are two kinds of MT, metaphrase and paraphrase. In metaphrase, an exact word to word translation takes place but the translated text may or may not have the similar semantics as the source text. In paraphrase, translation is not performed at word level but at sentence level. Here, the semantics of source text are conserved while translating into translated text. MT is one application of natural language processing. MT involves five major goals:

1. Morphological analysis is the process of generation of all possible roots from word level information.
2. Part-of-speech tagging is the process of assigning part-of-speech tags to every word in a given sentence.
3. Chunking is the process of identification of phrases such as noun phrase (NP), adjective phrase (JJP), verb phrase (VP) etc. in a given sentence.
4. Parsing is the process of generation of a parse tree with the help of the information obtained from part-of-speech tagging and chunking.
5. Word sense disambiguation is the process of identification of meaning of a word in a particular sentence when a given word has multiple meanings.

II. PROBLEMS FACED IN MT IN INDIAN LANGUAGES

Problems faced in MT in Indian languages include:

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 scarcity of resources pertaining to Indian languages on web.

Today, there are many available machine translators pertaining Indian languages but still these machine translators do not produce translations with very high accuracy. Consider the following Source text: "Jammu and Kashmir, India's one of the most picturesque state lies on the peaks of Himalayan Ranges with varying topography and culture. Jammu was the stronghold of Hindu Dogra kings and abounds with popular temples and secluded forest retreats. Kashmir's capital city, Srinagar offers delightful holidays on the lakes with their shikaras and houseboats". This source text in English is translated into Hindi using different machine translators. The translations are shown in Figure 1. It is undeniable that the translated texts obtained from existing machine translators are not of good quality. Some of the words in the translated text appear in English, and some of the words are transliterated instead of translated. So, there is a need to develop machine translator that can produce good translations.

III. APPROACHES OF MT

Various approaches of MT are depicted in Figure 2. These include the following:

- Direct machine translation
- Rule based machine translation
- Corpus based machine translation, including statistical machine translation and example based machine translation.

The description, advantages and disadvantages of the

different approaches are shown in Table I. Hybrid approach involves a combination of the above listed approaches. MT quality is expected to improve if hybrid approach is used to perform MT in Indian languages.

MACHINE TRANSLATOR	TRANSLATED TEXT
Google Translator	जम्मू-कश्मीर, भारत सबसे खूबसूरत राज्यों में से एक है हिमालयी रेंजों की चोटी पर विभिन्न भूगोल और संस्कृति के साथ है। जम्मू हिंदू डोगरा राजाओं का गढ़ था और लोकप्रिय मंदिरों और एकांत वन वापसी के साथ है। कश्मीर की राजधानी शहर, श्रीनगर झीलों पर अपने शिकारों और हाउसबोटों के साथ रमणीय छुट्टियां प्रदान करता है
Bing Translator	जंमू और कश्मीर, भारत की सबसे खूबसूरत राज्य की एक अलग स्थलाकृति और संस्कृति के साथ हिमालय पर्वतमाला की चोटियों पर निहित है । जंमू हिंदू डोगरा राजाओं के गढ़ था और लोकप्रिय मंदिरों और एकांत वन retreat के साथ लाजिमी है । कश्मीर की राजधानी शहर, श्रीनगर अपने shikaras और हाउसबोटों के साथ झीलों पर रमणीय छुट्टियां प्रदान करता है ।
Babylon Translator	जम्मू और ९श्मीर भारत की सर्वाधिक मनोरम राज्य निहित हिमालय पर्वतश्रृंखला की चोटियों पर भिन्न-भिन्न स्थलाकृति तथा संस्कृति है। जम्मू के गढ में हिंदू राजाओं के लोकप्रिय मंदिरों और एकांत सार्वनामिक डोगरा वन विहारों है। कश्मीर की राजधानी श्रीनगर के प्रस्ताव पर अपनी ३०ीलों को स्वच्छ और सभी जरूरतेंशिकारों अवकाश। हाऊसबोट
Anglabharti	जम्मू - कश्मीर इन्दिआस कोई एक अधिकतम चित्रात्मक अवस्था झूठ पर शिखर का/की/के हिमालायन रान्गेस के साथ परिवर्तित करनेवाला स्थलाकृति विज्ञान और संस्कृति । जम्मू हिंदू डोगरा राजाओं का गड्ढ था और प्रसिद्ध मंदिर और एकान्त जंगल की वापसी से प्रचुर होता है ।काश्मिर्स राजधानी शहर का श्रीनगर उनके शिकारास और शिकारों के साथ झीलों पर रुचिकर अवकाश प्रस्ताव रखता है ।

Fig. 1. Output of existing machine translators

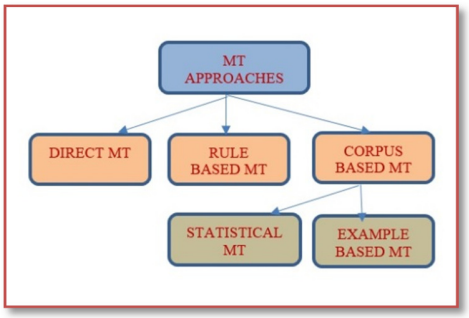

Fig. 2. Classification of approaches of MT

IV. LITERATURE REVIEW

The work done in MT pertaining Indian languages is shown in Table II.

V. EVALUATION

For evaluation of MT system, automatic evaluation metrics or human evaluation metrics may be used.

A. Automatic Evaluation Metrics

1) 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. 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 automatic evaluation metrics, BLEU, METEOR etc. can also be used for evaluation of MT output.

TABLE I. MT APPROACHES

Approach	Description	Advantages/Disadvantages
Direct MT	No parallel corpus is used. It makes use of bilingual dictionary, target language and source language corpus.	-Produces good accuracy. -Less tedious and less time consuming
Rule Based MT	Rules are constructed. Involves analysis of source and target text at syntactic, semantic and morphological level.	-Good quality translation. -Complex rules are needed to be constructed. -It involves tedious tasks and is time consuming.
Corpus based MT	Rules are constructed by analysis of parallel corpus.	Accuracy can be improved by adding more examples to corpus.
Statistical MT	Statistical models are used to perform MT.	MT quality can be improved by adding more examples to parallel corpora.
Example based MT	It makes use of translation memories. It performs translation by analogy.	MT quality can be improved by adding more examples to parallel corpora.

2) 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 whole corpus overall score.

3) 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 is assigned.

4) Word Error Rate (WER)

Estimates the number of tokens that differ between machine translated text and expected translated text.

5) METEOR

Estimates weighted harmonic mean of unigram precision and recall. It also involves matching of synonyms and lemmatized forms.

6) LEPOR

Involves a collection of different evaluation factors such as precision, recall, sentence length penalty, and word order penalty based on n-gram.

TABLE II. DESCRIPTION OF MTs OF DIFFERENT LANGUAGE PAIRS

Reference	About	Detailed Description
[1]	English to Hindi MT	Hybrid approach. EBMT and SMT is used to perform MT. Parallel corpus (54K English-Hindi sentences) is used. Training: 53K sentences. Testing: 100 random sentences. BLEU score:0.432.
[3]	English to Assamese	Phrase based MT+Transliteration. Parallel Corpus: 14,371 sentences of English and Assamese. Testing: 500 sentences. Wordnet of Assamese is used to improve MT output
[9]	Hindi to Punjabi MT	Overall accuracy: 95.12%. Input taken from daily news, articles, official language quotes, blogs and literature. 95.4% sentences are found to be intelligible. Accuracy obtained: 87.6%
[10]	Bilingual Hindi English text to pure Hindi and pure English	English and Hindi morphological analyzers are used. Plural forms are identified. Unknown words are considered to be proper nouns. Complex sentences are converted to simplified sentences and source text is translated to pure Hindi and pure English. In 90% cases this approach has obtained satisfactory results.
[13]	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
[14]	English to Hindi MT	EBMT+RBMT+Post editing approach. Can produce 90% correct results for sentences upto length of 20 words.
[15]	English-Hindi bilingual text to Hindi	Morphological analyzer is used to detect unknown words and unknown plural words of Hindi and English. Correct results: 90%.
[19]	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. Mean F-Measure obtained: 0.876.
[20]	English to Tamil MT	Rule based text simplification approach is used for enhancing English- Hindi MT. Testing: 200 sentences. Accurate results in 115 sentences. Accuracy of MT system increased by 28% by introduction of text simplification approach.
[21]	Hindi to English MT	MT output is improved by simplification of source text. Testing: 100 sentences. BLEU score: 0.805
[22]	Tamil to English MT	Statistical machine translation system, performs Tamil to English MT. 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.
[23]	English to Bangla MT (SMT)	Phrase based MT is performed for English to Bangla MT. Transliteration approach is used to deal with the words not present in vocabulary. Accuracy of transliteration module: 0.18. Preposition handling is also performed. Overall BLEU score: 11.7. BLEU score obtained for short sentences is 23.3 and 0.63 TER
[24]	Hindi to English MT	Testing: 100 sentences taken from Hindi Treebank. Source text simplification is used to improve MT. BLEU score: 4.45.

B. Human Evaluation Metrics

For human evaluation, authors in [10] used some linguistic features that include:

- Translation of gender and number of noun(s)
- Translation of voice in the sentence.
- Translation of tense in the sentence
- Identification of the proper noun(s)
- Use of adjectives and adverbs corresponding to nouns and verbs
- Selection of proper words/synonyms (lexical choice).
- Sequence of phrases and clauses in the translation.
- Use of punctuation marks in the translation.
- Fluency of translated text and translator's proficiency.
- Maintaining semantics of source sentence in the translation.
- Evaluating the translation of source sentence (with respect to syntax and intended meaning).

In order to access the translation quality, a five-point scale is used, which is shown in Table III. Similarly, adequacy and fluency score may be calculated using five-point scales as represented in Tables IV and V.

TABLE III. FIVE- POINT SCALE TO ACCESS TRANSLATION QUALITY

Score	Meaning
4	Ideal
3	Perfect
2	Acceptable
1	Partially acceptable
0	Not acceptable

TABLE IV. FIVE POINT SCALE TO ACCESS ADEQUACY

Score	Meaning
5	Complete Information
4	Most Information
3	Much Information
2	Little Information
1	None

TABLE V. FIVE POINT SCALE TO ACCESS FLUENCY

Score	Meaning
5	Ideal
4	Good
3	Non Native
2	Disfluent
1	Incomprehensible

VI. CONCLUSION

In this paper we discussed about MT, 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 MT regarding Indian languages. As we have seen, the quality of existing MT systems is not good, so there is a need to develop machine translators that can provide

good translation with high accuracy. We have discussed about automatic evaluation metrics and human evaluation metrics that can be used to access the translation quality.

REFERENCES

[1] V. Ambati, U. Rohini, "A hybrid approach to example based machine translation for Indian languages", 5th International Conference on Natural Language, Hyderabad, India, January, 2007

[2] B. Babych, A. Hartley, "Improving machine translation quality with automatic named entity recognition", 7th International EAMT Workshop on MT and other Language Technology Tools, Improving MT through other Language Technology Tools: Resources and Tools for Building MT, Budapest, Hungary, April 13, 2003

[3] A. K. Barman, J. Sarmah, S. K. Sarma, "Assamese WordNet based Quality Enhancement of Bilingual Machine Translation System", 7th Global WordNet Conference, Tartu, Estonia, January 25-29, 2014

[4] J. G. Carbonell, S. Klein, D. Miller, M. Steinbaum, T. Grassiany, J. Frey, "Context-based machine translation", 7th Conference of the Association for Machine Translation in the Americas, Cambridge, USA, August, 2006

[5] G. V. Garje, G. K. Kharate, "Survey of machine translation systems in India", International Journal on Natural Language Computing, Vol. 2, No. 4, pp. 47-65, 2013

[6] A. Hassan, H. Fahmy, H. Hassan, "Improving named entity translation by exploiting comparable and parallel corpora", AMML07, 2007

[7] J. Hutchins, "Towards a definition of example-based machine translation", MT Summit X, Workshop on Example-Based Machine Translation, Phuket, Thailand, September 16, 2005

[8] L. Jiang, M. Zhou, L. F. Chien, C. Niu, "Named Entity Translation with Web Mining and Transliteration", IJCAI-07, Hyderabad, India, pp. 1629-1634, January 6-12, 2007

[9] N. Joshi, H. Darbari, I. Mathur, "Human and Automatic Evaluation of English to Hindi Machine Translation Systems", in: Advances in Computer Science, Engineering & Applications, pp. 423-432, Springer Berlin Heidelberg, 2012

[10] N. Joshi, I. Mathur, H. Darbari, A. Kumar, "HEval: Yet another human evaluation metric", International Journal on Natural Language Computing, Vol. 2, No. 5, pp. 21-36, 2013

[11] N. Joshi, Implications of Linguistic Feature Based Evaluation in Improving Machine Translation Quality: A case of English to Hindi Machine Translation, PhD Thesis, Banasthali University, India, 2014

[12] S. Nirenburg, C. Domashnev, D. J. Grannes, "Two approaches to matching in example-based machine translation", 5th International Conference on Theoretical and Methodological Issues in Machine Translation, Kyoto, Japan, 1993

[13] A. Ramanathan, P. Bhattacharyya, J. Hegde, R. M. Shah, M. Sasikumar, "Simple Syntactic and Morphological Processing Can Help English-Hindi Statistical Machine Translation", Proceedings of the Third International Joint Conference on Natural Language Processing, Vol. I, pp. 513-520, 2008

[14] R. M. K. Sinha, A. Jain, "AnglaHindi: an English to Hindi machine-aided translation system", MT Summit IX, New Orleans, USA, September 23-27, 2003

[15] R. M. K. Sinha, A. Thakur, "Machine translation of bi-lingual hindi-english (hinglish) text", MT Summit X, Workshop on Example-Based Machine Translation, Phuket, Thailand, September 16, 2005

[16] V. Goyal, G. S. Lehal, "Evaluation of Hindi to Punjabi machine translation system", International Journal of Computer Science Issues, Vol. 4, No. 1, pp. 36-39, 2009

[17] V. Goyal, G. S. Lehal, "Web based Hindi to Punjabi machine translation system", Journal of Emerging Technologies in Web Intelligence, Vol. 2, No. 2, pp. 148-151, 2010

[18] G. S. Josan, G. S. Lehal, "A Punjabi to Hindi machine translation system", 22nd International Conference on Computational Linguistics: Demonstration Papers, Manchester, UK, August 18-22, 2008

[19] T. Rama, K. Gali, "Modeling machine transliteration as a phrase based statistical machine translation problem", 2009 Named Entities Workshop: Shared Task on Transliteration, Suntec, Singapore, August 7, 2009

[20] C. Poornima, V. Dhanalakshmi, M. A. Kumar, K. P. Soman, "Rule based sentence simplification for english to tamil machine translation system", International Journal of Computer Applications, Vol. 25, No.8, pp. 38-42, 2011

[21] A. Soni, S. Jain, D. M. Sharma, "Exploring Verb Frames for Sentence Simplification in Hindi", International Joint Conference on Natural Language Processing, Nagoya, Japan, October 14-18, 2013

[22] U. Germann, "Building a statistical machine translation system from scratch: how much bang for the buck can we expect?", Workshop on Data-driven Machine Translation, Toulouse, France, July 7, 2001

[23] M. Z. Islam, J. Tiedemann, A. Eisele, "English to Bangla phrase-based machine translation", 14th Annual Conference of the European Association for Machine Translation, St Raphael, France, May, 2010

[24] K. Mishra, A. Soni, R. Sharma, D. M. Sharma, "Exploring the effects of Sentence Simplification on Hindi to English Machine Translation System", Workshop on Automatic Text Simplification: Methods and Applications in the Multilingual Society, Dublin, Ireland, August 24, 2014