Volume 4, Issue 8, August 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A Review on a Web Based Punjabi to Hindi Statistical Machine Translation System

Amarpreet Kaur*
M.tech Student CSE Department
GZS PTU Campus Bathinda, India

Er. Jyoti Rani
Assistant Professor, CSE Department
GZS PTU Campus Bathinda, India

Abstract---In machine translation (MT) system, one natural language is automatically translated into another language with the help of computers. Machine translation is a key application in the field of natural language processing. In existing system, the direct and rule-based approaches are used to translate one natural language into another one. But, Statistical Machine Translation (SMT) approach can also be used for this purpose to translate Punjabi language into Hindi language. SMT is a simple and most widely used approach. In SMT, every sentence in the target language is a translation of the source language with some probability and the best translation which the system will attain in the form of sentence, will be of high probability.

Keywords- Machine translation (MT), Statistical Machine Translation (SMT), Punjabi, Hindi, and Natural Language Processing, transliteration and translation.

I. INTRODUCTION

The main aim of Machine Translation (MT) system is to translate Punjabi text into Hindi text using various approaches i.e., Direct approach, Rule Based approach and Statistical approach. This paper represents two regional languages such as Punjabi and Hindi. Punjabi is the language which is most widely used in Pakistan and 11th most widely spoken in India. Punjabi is the 4th spoken language in England and 3rd most spoken in Canada. Similarly, Hindi language is also a part of India. It is a official language of India. Hindi language is also known as national language of India which is spoken by Indians in each state to communicate with each other. Punjabi and Hindi are closely related languages in terms of syntax and vocabulary. The people who can't understand Punjabi, those one can translate Punjabi text into Hindi text and understand it. So this translation system is made to remove this kind of barrier between communications of people living over the world. This research work is based on AnmolLipi and KrutiDev scripts. These are non standard fonts which need to be converted into Unicode first. Conversion is done through Font Converter that convert non standard fonts into Unicode form. Besides translation, transliteration is also performed as well. Some words such as name, city name, country name does not need to be translated because such words has same pronunciation in both Punjabi and Hindi language.

This research work introduces the concept of Machine translation system, various approaches and the main activities in MT. A Machine Translation System is categorized into the following approaches: Direct Machine translation, Rule Based Machine translation and Corpus Based Machine translation i.e. Statistical Machine translation.

First, ***Direct Machine Translation*** (DMT) system is a simple form of machine translation system. In DMT, a word to word translation of the input text is performed and the result is obtained in the form of output text. In DMT, a language which is called a source language (Punjabi) is given as input and the output is received which is called a target language. E.g,

"ਅਸੀਂ ਦੇਸ਼ ਵਿਚ ਸ਼ਾਂਤੀ ਦੀ ਮੰਗ ਕਰਾਂਗੇ"

Can be translate in Hindi as:

"हम देश में शांती की मांग करेंगे"

In the above example, the Punjabi words ਅਸੀਂ, ਵਿਚ, ਦੀ, ਮੰਗ, ਕਰਾਂਗੇ are stored in the database in the form of source language and the Hindi words हम, में, की, मांग, करेंगे are stored with respect to these words. Then these words are accessed from the database according to the requirements of the sentence. Direct Machine Translation is a unidirectional approach and access only one language pair at a time.

Second, ***Rule Based Machine Translation*** (RBMT) is also known as Knowledge Based Machine Translation system. It is a system which is based on linguistic information related to source and target languages and retrieves this information from dictionaries (bilingual) and grammars which includes semantic and syntactic information of each language. RBMT system generates output text from this information. E.g.,

"ਹਸਪਤਾਲ ਚ ਮੋਬਾਈਲ ਵਰਤਣ ਤੇ ਰੋਕ"

Can be translate in Hindi as:

"हस्पताल में मोबाईल के उपयोग पर रोक"

This example shows that we need to add some words to convey the relation in Hindi. These words are known as preposition and postposition. Such type of words is necessary in Hindi.

Third, ***Statistical Machine Translation*** (SMT) is a new approach which is based on statistical models and in this approach; a word is translated to one of a number of possibilities based on the probability. The whole process is performed by dividing sentences into N-grams. N-gram is a contiguous sequence of n items from a given text. The items can be phonemes, letters, and words. An N-gram of size 1 is known as a unigram; size 2 is a bigram; size 3 is a trigram. Larger sizes are represented by the value of n i.e. four-gram, five-gram and so on. Statistical system will analyze the position of N-grams in relation to one another within sentences. E.g, suppose we want to translate a name into through Statistical Machine Translation, then below process is followed:

In Unigram, ਅਮਰਪ੍ਰੀਤ

"ਅ can be translate in Hindi as अ"

In Bigram,

"ਅਮ can be translate in Hindi as अम"

In trigram,

"ਅਮਰ can be translate in Hindi as अमर"

In four-gram,

"ਅਮਰਪ can be translate in Hindi as अमरप"

In five-gram,

"ਅਮਰਪ੍ can be translate in Hindi as अमरप्z"

In six-gram,

"ਅਮਰਪ੍ਰੀ can be translate in Hindi as अमरप्री"

In seven-gram,

"ਅਮਰਪ੍ਰੀਤ can be translate in Hindi as अमरप्रीत"

II. EXISTING WORK

Machine Translation projects related Indian languages are given below:

English to Hindi Statistical Machine Translation system has been developed by Nakul Sharma, Parteek Bhatia at Thapar University, Patiala. This system is based on Statistical Machine Translation. Statistical Machine Translation consists of Language Model, Translation Model and Decoder. Language Model computes the probability of target language sentences. Translation Model computes the probability of target sentences given the source sentence and the decoder maximize the probability of translated text of target language. The translation of 90 sentences was evaluated using human evaluation method. On the parameters of fluency and adequacy a geometric average of 2.693 and 2.93 was calculated.

Direct Machine Translation System from Punjabi to Hindi for Newspapers headlines Domain has been developed by the Sumita Rani at Guru Kashi University, Talwandi Sabo Bathinda. The Direct Machine Translation system is based on the utilization of syntactic and vocabulary similarities between more or few related natural languages. The similarity between Punjabi and Hindi languages is due to their parent language Sanskrit. Accurate sentences are calculated and then their percentage is found out which is come out to be 97%. The percentage of error rating is found out from the total sentences which is come out to be 3%.

Punjabi to Hindi Statistical Machine Transliteration has been developed by Gurpreet Singh Josan and Jagroop Kaur. In Transliteration, system converts an input string to a string in target alphabet based on the phonetics of the original word. This system shows 87.72% accuracy rate.

A Review on Machine Transliteration of related languages: Punjabi to Hindi has been developed by Sumita Rani and Dr. Vijay Laxmi. This system maps the source language into target language. This system is based on character to character transliteration approach. Most of the characters have same matching part in both languages. But there are some characters exist in Hindi which are double sounds and no such characters are available for Punjabi. The major inaccuracies in the transliteration are due to poor word selection.

III. PROBLEMS IN TRANSLATION

The Punjabi letters ੳ and ੲ have no mapping in Hindi. Similarly, there are letters in Hindi that have no mapping in Punjabi *e.g*. ऋ. These letters will never be mapped in Punjabi to Hindi transliteration using a direct mapping method. Some letters have more than one representation in Hindi, *e.g.,* स may be mapped to श or ष. Another problem is the use of conjunct consonant forms in Hindi. In Hindi, a syllable may consist of a vowel, a consonant followed by vowel, or a consonant cluster followed by a vowel. The last form *i.e.,* when two or more consonants are used within a word with no intervening vowel sound, is known as a conjunct consonant. Use of conjunct consonants is limited in Punjabi. Only three letters can be used as conjuncts *i.e.*, ਹ, ਰ, and ਵ. Their representation is also unique. It is not a trivial task to find out which combinations of letters in Punjabi will take conjunct consonant form in Hindi. For example, why the word ਿਨਊ(new) in Punjabi takes the conjunct consonant form in Hindi न्यू, is not clear. Also, the mapping of nasal consonants is not clear. Nasal consonants in initial place in a conjunct may be expressed using the anusvara over the previous vowel, rather than as a half-glyph attached to the following consonant. The anusvara is written above the headstroke, at the right-hand end of the preceding character. In the list below, both spellings are correct and equivalent, although anusvara is preferred in the case of the first two: रंग = रङ्ग, िहदी = िहन्दी, लंबा = लम्बा. Anusvara is still applied when previous character has its own vowel sign. If the vowel sign is [aa], the anusvara appears over the [aa], *e.g.* आंदोलन. Similarly, the position of the character "र" is not specific. It can be used as consonant, subscripted consonant, or can take a position above a consonant or diacritics, and it is typically displayed as a small mark above the right shoulder of the last letter in the syllable. It is not clear how the character ਰ in Punjabi will map to which form in Hindi, *e.g.*, the three cases where mapping of ਰ is not clear are: from ਤਰਕਸ਼ [tarkash] to तरकश, from ਵਰਤ [varat] to वर्त, and from ਬਰਤਨ[bartan] to बतर्न.

Table I
Comparison Of Approaches Used In Punjabi To Hindi Translation

Approach Used	**Overall Accuracy**
Direct Mapping	67%
Rule Based approach	81%
Statistical Approach	91%

Hence maximum accuracy for translation comes to be 91% which needs further improvements. Hybrid approach can also be used to improve the accuracy for Punjabi to Hindi translation system.

IV. CONCLUSIONS

The accuracy of the translation achieved by our system justifies the suggestion that word-for-word translation for machine translation system for language pair of Punjabi-Hindi. The major inaccuracies in the translation system are due to poor word choice for confusing words and some corrections regarding preposition and post positions. The lack of information about Hindi language is sometimes causes an unnecessary translation error. We can conclude that this study is beneficial to remove the language barrier between two closely related language pair Punjabi-Hindi. A hybrid approach is required to translate the given Punjabi text into its Hindi text equivalent.

REFERENCES

[1] Gurpreet Singh Josan and Gurpreet Singh Lehal, *"Direct Approach for Machine Translation from Punjabi to Hindi"* CSI Journal of Computing | Vol. 1, No.1, 2012.

[2] Gurpreet Singh Josan1& Jagroop Kaur, *"Punjabi to Hindi statistical machine transliteration"* International Journal of Information Technology and Knowledge Management July-December 2011, Volume 4, No. 2, pp. 459-463.

[3] Vishal Goyal and Gurpreet Singh Lehal, *"Hindi to Punjabi machine translation system"* Proceedings of the ACL-HLT 2011 System Demonstrations, pages 1–6, Portland, Oregon, USA, 21 June 2011. Association for Computational Linguistics.

[4] Vishal Goyal and Gurpreet Singh Lehal, *"Web based Hindi to Punjabi machine translation system"* journal of emerging technologies in web intelligence, vol. 2, no. 2, may 2010.

[5] Gurpreet Singh Josan and Gurpreet Singh Lehal, *"A Punjabi to Hindi Machine Transliteration System"* Computational Linguistics and Chinese Language Processing Vol. 15, No. 2, June 2010, pp. 77-102 The Association for Computational Linguistics and Chinese Language Processing.

[6] Vishal Goyal and Gurpreet Singh Lehal *"Evaluation of Hindi to Punjabi Machine Translation System"* IJCSI International Journal of Computer Science Issues, Vol. 4, No. 1, 2009.

[7] Gurpreet Singh Josan and Gurpreet Singh Lehal, *"Evaluation of Direct Machine Translation System For Punjabi To Hindi"*

[8] Vishal Goyal and Gurpreet Singh Lehal, *"Hindi-Punjabi Machine Transliteration System (For Machine Translation System)"*

