Vol. (1) No. (1) March 2018

JK Research Journal in Mathematics and Computer Sciences

Experimental Evaluation of Machine Translation: Source to Target Language

*Kumar Sourabh[1] Rakesh Kumar[2]

[1]Government Gandhi Memorial Science College Jammu, J&K- India
[2]Government Degree College Billawar, J&K -India

Abstract: Machine Translation has gained popularity over the years and has become one of the promising areas of research in computer science. Due to a consistent growth of internet users across the world information is now more versatile and dynamic available in almost all popular spoken languages throughout the world. From Indian perspective importance of machine translation become very obvious because Hindi is a language that is widely used across India and whole world. Many initiatives have been taken to facilitate Indian users so that information may be accessed in Hindi by converting it from one language to other. In this paper we have evaluated the performance of the various online MT softwares. This paper deals with evaluation of the translated text from source as an English and target as Hindi language. We have designed an experimental setup for which data is tested for various outputs and tested for accuracy produced by softwares under consideration. The paper is divided into three major sections starting with introduction, machine translation in Indian perspective followed by research methodology and experiments and results.
Keywords: *Machine Translation; Corpus; Rule based; Statistical; Interlingua; Accuracy.*

Introduction

Machine translation is a sub-field of computational linguistics that investigates the use of software to translate text or speech from one language to another. As internet is now flooded with multilingual information for global community, research and development giving space to Machine translation to grow at a rapid rate. There are 22 official languages; and approximately 2000 dialects are spoken by different communities in India. English and Hindi are used for official work in most states of India. The state governments in India predominantly carry out their official work in their respective regional language whereas the official work of Union government is carried out in English and/or Hindi. All the official documents and reports of Union government are published in English or Hindi or in both English and Hindi. Many newspapers are also published in regional languages. Translating these documents manually is very time consuming and costly. G V Garje et.al in his paper suggests that there is need to develop good machine translation systems to establish a better communication between states and Union governments and exchange of information amongst the people of different states with different regional languages. [1] Indians use

Kumar9211.sourabh@gmail.com
*Kumar Sourabh

22 official languages and 11 written script forms and among all the languages Hindi language is spoken by the major population of India. About 5% of population understands English as their second language. Hindi is spoken about 30% of the population [2].

With the web content being written in different languages of the world, it has become important to have tools that can convert information from the documents written in different languages to the favorable language. A Delhi based research organization - Juxt Consult - says that 44 % of existing Internet users in India prefers Hindi to English, if made available. Similarly 25% existing Internet users prefer other regional languages. [3]

Asia is the largest and the most culturally and linguistically diverse continent. It covers 39 million square kilometers, about 60% of land area of the world, and has an estimated 3.8 billion population, which is approximately 60% of the world's population. There are more than 50 countries and roughly 2200 languages spoken in Asia. Hindi is the third most widely-spoken language in the world (after English and Mandarin): an estimated 500-600 million people speak this language.

In the context of above information need for translation of language particularly from English to Hindi becomes obvious.

Machine Translation Initiatives in India

The Department of Information Technology under Ministry of Communication and Information Technology is also putting the efforts for proliferation of Language Technology in India. Other Indian government ministries, departments and agencies such as the Ministry of Human Resource, DRDO (Defense Research and Development Organization), Department of Atomic Energy, All India Council of Technical Education, UGC (Union Grants Commission), are also involved directly and indirectly in research and development of Language Technology. All these agencies help develop important areas of research and provide funds for research, to development agencies. In our previous work [4] initiatives taken in India for development of MT systems have been discussed in which organizations like TDIL (Technology Development for Indian languages), C-DAC, IIT Hydrabad, IIT Mumbai and Amrita Vishwa Vidyapeetham are the key role players which resulted into development of MT systems. Other institutions University of Hyderabad, NCST Mumbai JNU New Delhi etc are also playing a major role in developing the MT systems in India. Many MT systems have been developed and are being developed. The MT systems have been developed using

different machine translation approaches. In the next section of this paper we have discussed in brief various types of MT.

Approaches towards MT

The following diagram reflects the various approached followed in MT. Some of the approached have been discussed in this section briefly particularly with the example of the outcomes of research from Indian perspective.

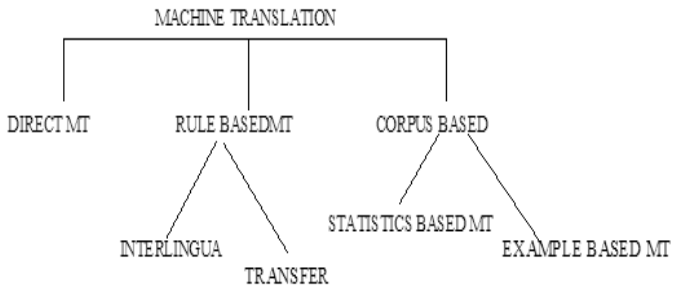

Fig1.

Direct machine translation: Direct translation is made at the word level. Machine translation systems that use this approach are capable of translating a language, called source language (SL) directly to another language, called target language without passing through an intermediary representation. The analysis of SL texts is oriented to only one TL. Direct translation systems are basically bilingual and uni-directional. Direct translation approach needs only a little syntactic and semantic analysis [5]. Some examples of Direct MT includes

a) Anusaaraka systems among Indian
b) Punjabi to Hindi MT System
c) Web based Hindi-to-Punjabi MT
d) Hindi-to-Punjabi MT System (2009, 2011)

Transfer-Based Machine Translation: is a type of machine translation (MT). It is currently one of the most widely used methods of machine translation. In contrast to the simpler direct model of MT, transfer MT breaks translation into three steps: analysis of the source language text to determine its grammatical structure, transfer of the resulting structure to a structure suitable for generating text in the target language, and finally generation of this text. Transfer-based MT systems are thus capable of using knowledge of the source and target languages. [6]

Some examples of transfer based MT Systems include:

Mantra MT (1997), MANTRA MT(1999), An English–Hindi Translation System (2002), MAT (2002), Shakti (2003), English-Telugu MT System (2004), Telugu-Tamil MT System (2004), The MaTra System (2004, 2006), Sampark System: Automated Translation among Indian Languages (2009) **[1]**

Corpus based MT: Since 1989 corpus based approach has emerged as widely explored area because of the level of accuracy achieved [8]. Statistical and example based approaches are corpus based approaches.

Statistical machine translation (SMT) is a machine translation paradigm where translations are generated on the basis of statistical models whose parameters are derived from the analysis of bilingual text corpora. The statistical approach contrasts with the rule-based approaches to machine translation as well as with example-based machine translation [7]

a) Shakti (2003)
b) English to Indian Languages MT System (E-ILMT) (2006)

Example based MT Example-based machine translation (EBMT) is a method of machine translation often characterized by its use of a bilingual corpus with parallel texts as its main knowledge base at run-time. It is essentially a translation by analogy and can be viewed as an implementation of a case-based reasoning approach to machine learning.

a) ANUBAAD (2000, 2004)
b) VAASAANUBAADA (2002)
c) Shiva and Shakti MT System (2003)
d) ANGLABHARTI-II (2004)

Objective and research methodology

Our objective in this paper is to test various machine translation tools available online for accuracy especially for translation of text from English to Hindi Language. The data to be translated into Hindi is divided into three categories keeping in view the time and space framework.

A. Simple Single line text (Easy)
B. Complex single line text (Medium)
C. Simple and complex multiline text (Hard)

The data will be tested upon online free machine translation tools which offer services to translate text from English to Hindi for evaluation of accuracy being produced by the tools. It is important to mention here that text containing *named entities and ambiguous words* will not be tested as this still remains a challenge in context with MT. In the later section of this paper we will conclude with the evaluation of the results. There are several online software tools which provide facility to convert text from one language to another. In this paper we have conducted tests upon following online tools selected (based on popularity of use) to conduct experiments which provide services for conversion of text especially from English to Hindi. First three out of which are internationally used and remaining one is developed and used in India.

1. SDL Free translation
2. MICROSOFT BING
3. Google Translation
4. Indian Language Machine Translations TDIL

Experiments and results As mentioned in section above data to be tested is divided into three categories. It is pertinent to mention that only selected lines of text are displayed in this section which produced a mixture of good and bad output or say average case analysis.

For first and second category we used one hundred lines of simple as well as complex text sample and for third category we used about fifty test sample cases. Evaluation of translated text is based on the score gained in terms of Syntax generated corresponding to the target language, semantic score (grammar) and accuracy (overall meaningful arrangement of the translated text). In this section we present two out outcomes out of one hundred obtained in each category.

Experimental Result Evaluation

In this section we provided simple single line Input for *example: 1. you are a very good human being 2. Input Example: Today and tomorrow is holiday in our state.* The expected output for two inputs is आप बहुत अच्छे इंसान हैं and आज हमारे राज्य में छुट्टी है. Likewise a total of one hundred inputs are provided and outputs are checked for a few parameters for which result is summarized in table 1. It is significant to mention here that to calculate syntax score major keywords are considered excluding articles conjunctions etc whereas to calculate semantic score whole line is

considered to be evaluated with a maximum score of five based on human observation e.g. in case of syntactic score evaluation only works like "you" "very" "good" "human" "being" is considered which counts to five and a count of five has been set as a standard for all hundred samples for category one and seven for category two excluding "are" and "a" "to" etc and for the semantic score whole line is considered including articles ,conjunctions and punctuations.

Also to evaluate accuracy we have opted for Relevance feedback method.

Mean Syntactic score (MSyS)= ∑ Qn/n

Results obtained after translation for simple single line of text.

Example Sample two out of one hundred

Example 1: "Schools remained closed due to extreme weather conditions" for 7-8 Keywords

Example 2:"There is no scholarly consensus over what precisely constitutes a religion" for 8-9 keywords

Example 3:"Feminism has created what is undoubtedly the weakest generation of women in world history 9-10 The results obtained for a sample run of one hundred lines of text are shown in table 1. The sample text for third category of test is as follows. We have selected a sample text with the output produced by only two selected tools out of four, as a single case of experiment out of fifty likewise cases.

"*Deep in the glorious vision, she sees a dazzling lightning sail, some stars and some black holes shining with a constant brightness and blinking with a rhythm. There comes a blue white world out of black hole.*"

Translated by Google: गौरवशाली दृष्टि में दीप, वह एक चमकदार बिजली पाल, कुछ सितारों और कुछ ब्लैक होल देखती है जो निरंतर चमक के साथ चमकता है और ताल के साथ चमकती है। ब्लैक होल के बाहर नीली सफेद दुनिया आती है

Translated by TDIL tool गौरवपूर्ण कल्पना में गहरी , वह एक चकाचौंधपूर्ण तड़ित कुछ स्टार और कुछ काल कोठरियाँ जलयात्रा करते हुए देखती है जो एक स्थिर चमक से चमका रही हैं और एक लय से आँख झपका रही है । वहाँ काल कोठरी का एक एक नीला सफेद आउत आता है

Observation and Discussion

It has been observed that for first case i.e. simple single line of text containing 5 to 6 keywords the performance by all the translating tools is good where as with respect to semantic or grammatical arrangement in the output varies. As it can be observed in table 1 and fig1 below

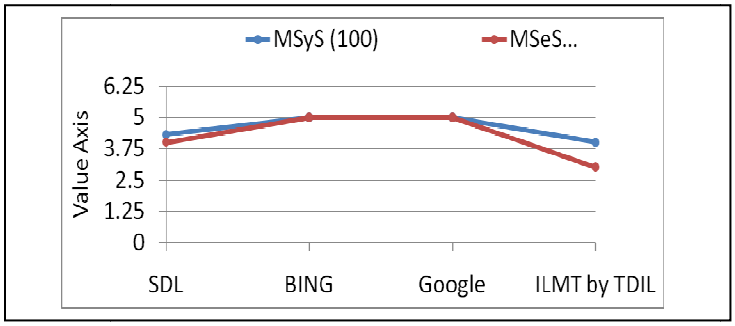

Fig 2

that SDL and ILMT perform equally with low scores while BING and GOOGLE produce output with almost equal high score. However Bing and Google outperform when it comes to the accuracy (exact expected output with respect to meaning). In second case when translators are supplied with complex lines as input containing keywords in the range of seven to eight, eight to nine and nine to ten a different range of outputs can be observed in both parameters of evaluation i.e. syntactic as well as semantic. It can be observed that in first case i.e. for seven to eight keywords syntactic expected output fall in case of SDL and ILMT as compared to BING and Google and in second and third case and depreciation is observed for output produced by ITML in comparison with rest of the MS tools.

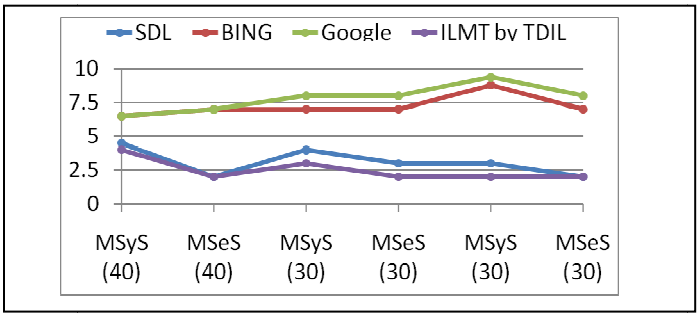

Fig 3

It can be observed from the above fig 3 that for semantic evaluation as the range of keywords increase the expected output produced by MS tools falls down in following order. ILMT, SDL, Bing, Google and the order suggests that Google outperforms rest of the tools.

Whereas accuracy is concerned Google produces more accurate translation as compared to rest of the MT tools. The results produced by Google are more understandable as compared to others

Lastly when input supplied with complex multiline texts, it can be clearly seen that none of the MT tools performs and does not produce the expected output. The one of the translation samples is provided above for readers for reference the translation output is provided being produced by Google and ITML. It can be observed from the table 2 and fig 4

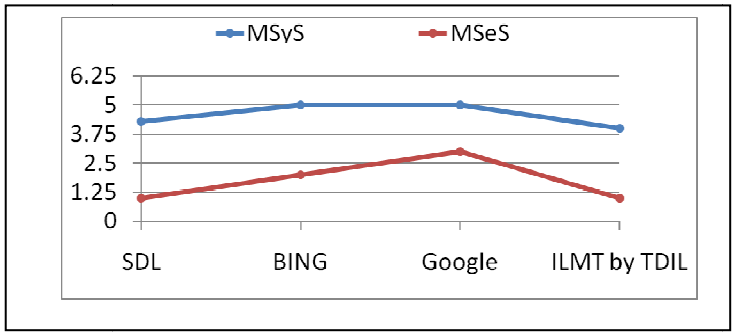

Fig 4

that accuracy offered by IMTL is VERY LOW by SDL and BING is LOW and by Google is MEDIUM. It is also important to mention here that not all the MT tools are capable of conversion of English keywords to Hindi Keywords.

Conclusion

Due to an increase in interest of online users for the information in their native language especially in India need for MT is now obvious. There is information in form of news, articles, essays, poems literature etc. in different languages and to facilitate the users with information in their own language various government and private organizations have come up with an initiative to promote machine translation. In this paper we have made an attempt to manually evaluate the outputs produced by various MT tools available online particularly those tools which provide online service to translate text from English to Hindi language. It has been observed that for simple line of text almost all tools provide an output with almost high accuracy where as for complex line of inputs accuracy drops except in case of Google and in case of multiline inputs none of the MT tool is able to produce output with 100% accuracy. Still a lot of investigation is required for moving on to conclusion an alternate text input test is still required which will be carried out in future. Also data sets need to be increased with wide variety of inputs like named entity and ambiguous words etc to carry out further tests for accuracy.

References

[1] G v garje and g k kharate "survey of machine translation systems in india" international journal on natural language computing (ijnlc) vol. 2, no.4, october2013

[2] Praveen kumar, shrikant kashyap, ankush mittal indian institute of technology, roorkee, india. "a query answering system for e-learning hindi documents" south asian language review vol. xiii, nos 1&2, january-june,2003

[3] S.k. dwivedi and parul rastogi babasaheb bhimrao ambedkar university (a central university), lucknow, uttar pradesh, india "an entropy based method for removing web query ambiguity in hindi language" journal of computer science 4 (9): 762-767, 2008 issn 1549-3636 © 2008 science publications

[4] Kumar sourabh "an extensive literature review on clir and mt activities in india ". international journal of scientific & engineering research volume 4, issue 2, february-2013 1 issn 2229-5518

[5] M. d. okpor department of computer science, delta state polytechnic, ozoro delta state, nigeria "machine translation approaches: issues and challenges" ijcsi international journal of computer science issues, vol. 11, issue 5, no 2, september 2014 issn (print): 1694-0814 | issn (online): 1694-0784 www.ijcsi.org

[6] Jurafsky, daniel; martin, james h. (2009). speech and language processing. pearson. pp. 906–908.
[7] Philipp koehn (2009). statistical machine translation. cambridge university press. p. 27. isbn 0521874157

	SIMPLE INPUT TEXT			COMPLEX INPUT TEXT						
Keywords	Four to five			Seven to Eight		Eight to Nine		Nine to Ten		
TRANSLATOR	MSyS (100)	MSeS (100)	Accuracy (RFM)	MSyS (40)	MSeS (40)	MSyS (30)	MSeS (30)	MSyS (30)	MSeS (30)	Accuracy (RFM)
SDL	4.3	4.0	Low	4.5	2.0	4.0	3.0	3.0	2.0	Low
BING	5.0	5.0	High	6.5	7.0	7.0	7.0	8.8	7.0	Medium
Google	5.0	5.0	High	6.5	7.0	8.0	8.0	9.4	8.0	High
ILMT by TDIL	4.0	3.0	Very low	4.0	2.0	3.0	2.0	2.0	2.0	Very low

Table 1 Result summarization for first two category simple and complex input text

TRANSLATOR	MSyS	MSeS	Ac. (RFM)
SDL	4.3	1	Low
BING	5.0	2	Low
Google	5.0	3	Medium
ILMT by TDIL	4.0	1	Very low

Table 2 Result Summarization for third category multiline input text