International Journal of Advanced Research and Development
ISSN: 2455-4030
Impact Factor: RJIF 5.24
www.advancedjournal.com
Volume 2; Issue 6; November 2017; Page No. 407-411

Phrased based T2 model: A review of *Google translate*, *Bing translator* & *Anusaaraka*

Saroj Kumar Jha, Piyush Pratap Singh, Vijay Kumar Kaul
Research Scholar, Dept. of Computational Linguistics, School of Language, Mahatma Gandhi International Hindi University, Wardha, Maharashtra, India

Abstract
This paper is an observation of few attempts under the major area of NLP from the translation perspective especially based on *Google Translation*, *Bing Translation*, and *Anusaarka*, the MT systems for translating Source Language (SL) English to Target Language (TL) Hindi with all the regular/irregular phrases. After the parametric evaluation of these phrase based examples, it's found that all the systems are having some of basic problems at several stages and the result is bit far from desired output. The following outputs of all the translation systems are under evaluation and whatever translation methodology is used by them till date, has some drawbacks. PBMT (Phrase Based Teaching and Testing Modal) is proved useful in solving these issues. Therefore, this approach can be used for phrase to phrase accurate translation, which will enhance the quality and robustness of statistical as well as Rule-Based system.

Keywords: PBMT, NLP, TL, SL, MT, T2

Introduction
Since Machine Translation came under NLP in current research trend, it is directed to one of the most important applications of computational linguistics that uses the computer software and web to translate text from one Source Language (SL) to Target Language (TL). In 21st century the awareness regarding automatic MT system among the people is an astonished work by machine and soon India is going to be huge market for NLP probably in Automated Machine Translation field. There are such priority basis work is under development. The government is funding several projects to develop a quick and fast localize language based MT system where the administrative and government notices can be frequently translated in all the Indian languages whether in text or speech need to be quick translated in the text or speech of another languages without delay. This field has involved several research scholars and renowned NLP contributors from various fields who are working with multiple methodologies and approach used frequently in MT System. We consistently effort to prepare the effective algorithm and apply to get the current MT output and compare it that how is it closer o desired output. Simultaneously, the following mechanism is still under progress to get the closest one. Among all the approaches of MT, one of the renowned and most applicable approaches is called PBMT (Phrased Based Machine Translation) approach for Machine translation which aims to skip the restriction of word-based translation by translating whole sequences or the given phrase of sentence, where lengths of the sentence may differ at many level. This modal is in use from quite some time in different modified form like PBM (Phrase Based Method) and HPBM (Hierarchical Phrased Based Method). In fact, after the certain process in MT System from beginning, it is forced to use (PB) method. So this approach is taken for testing and implementing PB approach with it. There are also some other approaches which are used for MT system as follows.

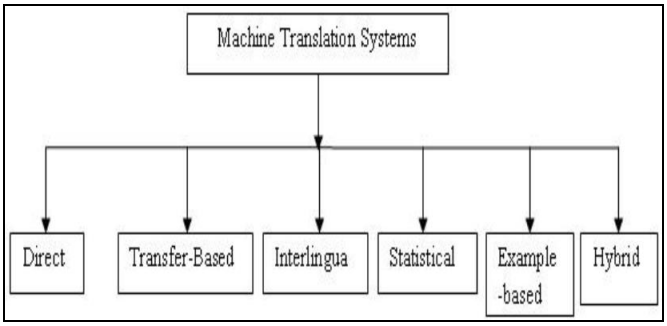


Fig 1

These are the approaches called *Direct*, *Transferred*, *Interlingua*, *Statistical*, *Example Based and Hybrid*, give the sufficient output with accuracy, but still under the development stage.
This effort is to map some of the assumptions which can be used with Phrase-Based methods in statistical MT system, such as *Google* or *Bing* or Rule-based MT system like *Anusaaraka*, which can help to improve the robustness of application.
For testing, we have taken the variety of simple sentences and other like Negative, Interrogative, Interrogative-Positive and Interrogative-Negative, and Wh-questions. The aim of taking such varieties is to identify the translation quality, word sequence translation and multiple sentences in various contexts. These tests will also certify the algorithm & design used by systems.

Related Work
1. According to (P. Koehn, F.J. Och, and D. Marcu. 2003) [1]

the particular issue elected by several renowned Linguist, who are giving their individual and combine efforts in part-wise stage of MT like: Chunking, Tagging, Parsing, and other significant parts of Machine Translation simultaneously carrying multiple approaches along with translation. This approach is good at removal of translation error caused due to local reordering, translation of short idioms, insertions and deletions.

2. W. John Hutchins (2005) [2] gives a brief description of the various approaches and major machine translation developments in India and shows comparison of different MT systems in India on the basis of approaches.
3. Durgesh, D Rao (1998) [3], has presented the Machine translation as the study of designing systems that translate from one human language into another. They introduce the main concepts, issues and techniques involved in machine translation, and look at some applications.
4. Vishal Goyal *et. al*. (2009) [4], has discussed the machine translation systems for non-Indian languages and second part discusses the machine translation systems for Indian languages.
5. Taraka Rama, Karthik Gali, "Modeling Machine Transliteration as a Phrase Based Statistical Machine Translation Problem". The discussion is on the phrase based translation model where the words are added as a phrase and after that the translation is taken.
6. Bibek Behera, Pushpak Bhattacharyya (2013) [6] have discussed the processing of phrase based model for translation and the challenges are encountered while translation.
7. Daniel Marcu and William Wong. 2002 [7] have discussed in this paper that how the phrased based model can be added through probability and make the tree through HMM modelling.
8. T. Rama, A.K. Singh, and S. Kolachina (2009) [8]. "Modelling of the letter to phoneme conversion as a phrase based statistical machine translation problem with minimum error rate training". As per the discussion was taken on the basis of conversion of latter in the phoneme where the possibility of error rate would be minimum in received output.
9. Latha R. Nair & David Peter S., (2012) [9]. "Machine Translation Systems for Indian Languages" In this paper the author proposed the design and development of MT for Indian languages with peculiar approaches with maximum coverage of Indian Languages.
10. Akshar Bharti, Chaitanya Vineet, Amba P. Kulkarni & Rajiv Sangal, (1997) [10]. The Anusaaraka team have propose the translation system for Indian languages covering the four major Indian languages like Hindi, Urdu, Telugu and Tamil. The methodology of the system is based on Hybrid approach and result is satisfactory whereas the further research and development is going on.

Review (*Google, Bing & Anusaaraka*)

Google Translate is a free multilingual statistical machine translation system developed by *Google* to translate text, speech, images, or real-time video from one language into another. For some languages, it can pronounce translated text, highlight corresponding words and phrases in the source and target text, and act as a simple dictionary for single-word input. It does not apply grammatical rules, since its algorithms are based on statistical analysis rather than traditional rule-based analysis. The system's original creator, *Franz Josef Och*, has criticized the effectiveness of rule-based algorithms in favour of statistical approaches. The system doesn't translate from one language to another (L1 → L2). Instead, it often translates the source language first into English and then into the target language (L1 → EN → L2). For example, translating *vous* from French to Russian gives *vous* → *you* → *ты* OR *Вы/вы*. If Google was using an unambiguous, artificial language as the intermediary, it would be *vous* → *you* → *Вы/вы* OR *tu* → *thou* → *ты*. Such a suffixing of words disambiguates their different meanings. Hence, publishing in English, using unambiguous words, providing context, using expressions such as "you all" often make a better one-step translation. *Google Translate*, like other automatic translation tools, has its limitations as well.

Bing Translator (previously Live Search Translator and Windows Live Translator) is a user facing translation portal provided by Microsoft as part of its Bing services to translate texts or entire web pages into different languages. All translation pairs are powered by the *Microsoft Translator*, a statistical machine translation platform and web service, developed by Microsoft Research, as its backend translation software. It can translate phrases entered by the user or acquire a link to a web page and translate it entirely and also helpful for translating an entire web page, or when user selects "Translate this page" in Bing search results, the Bilingual Viewer is shown, which allows users to browse the original web page text and translation in parallel, supported by synchronized highlights, scrolling, and navigation. Four Bilingual Viewer layouts are available: *Side by side*, *Top and bottom*, *Original with hover translation* and *Translation with hover original*. Website owners can add a translation widget to their website for translating it into other languages supported by Bing Translator; this is done by inserting an HTML code snippet on the web page. The widget supports:

- Any-to-any language translation pairs.
- Automatically detect the language of the text or website being translated.
- Ability to easily reverse the translation direction.
- The user can play back a spoken version of the translation through text-to-speech (not supported in every language).

Anusaaraka is an English – Hindi language accessing software. With insights from *Panini's Ashtadhyayi* (Grammar rules), *Anusaaraka* is a machine translation tool being developed by the *Chinmaya International Foundation (CIF)*, International Institute of Information Technology, Hyderabad (IIIT-H) and University of Hyderabad (Department of Sanskrit Studies). It's all about the fusion of traditional Indian *Shastras* and advanced modern technologies. It allows users to access text in any Indian language after translation from the source language (i.e. English or any other regional Indian language). In today's Information Age large volumes of information are available in English – whether it is information for competitive exams or general reading. However, many people whose primary language is Hindi or a regional Indian language are unable to access information in English.

Anusaaraka aims to bridge this language barrier by allowing a user to enter an English text into *Anusaaraka* and read it in an Indian language of their choice.

Current MT Output

The parallel outputs of three current MTs *(Google, Bing & Anusaaraka)*, discussed in the conference, have been taken for analysis. The entire analysis covers the sequence of Word → Phrase → Sentence translated from source to target. First section (F_1) deals with *Word-Word* (W-W) appropriate mapping, second one, (F_2) *Phrase to Phrase* (P-P) mapping and last one deal with (F_3) *Sentence-Sentence* (S-S) mapping. Here some Acronyms are used like: Word, Phrase and Sentence (WPS) to understand the points. As we all know that absolute translation doesn't consider only Morphology, Syntax, but Semantics level too. Therefore, it is need to be identified that, is the meaning of (WPS) in source text appropriately translated into target? As per (W-W) translation are mostly founds nearby 90% accurate and 10% probe with homographs and Irrelevant database provided to system. But the problem mostly occurs with (P-P) and (S-S) part. Some of the translated samples of both statistical and rule-based MT system have been discussed here.

Let's begin with the Google Translate text, some of the examples are somehow correct but some are not even close to relevant translation because of the complexities of the text. There are few criteria followed to represent the problems related to items are enlisted below:

Table 1: Translation Features & Tags

S. No	Items	True	False
1.	Pronominal	Pro-T	Pro-F
2.	Verbal	VM-T	VM-F
3.	Prepositional	Prep-T	Prep-F
4.	Idiomatic Expression	IE-T	IE-F
5.	Honorific	Honf.-T	Honf.-F
6.	Word Sequence	Seq-T	Seq-F
7.	Extra Additions	EA-T	EA-F
8.	Gender	Gen-T	Gen-F

Above the following table represents the True (T) and False (F) section in translated text by multiple MT systems.

A few output of *Google Translate* can draw your attention towards the ill-formedness and grammatical errors in them:

[1] Ram's brother is a good manager.
राम के भाई एक अच्छा प्रबंधक है। (Honf.-F)

[2] That red flat belongs to Mohan.
यही कारण है कि लाल फ्लैट मोहन के अंतर्गत आता है। (Pro-F, VM-F)

[3] He cuts an apple.
उन्होंने कहा कि एक सेब कटौती। (Pro-F, VM-F)

[4] He said that he is a good boy.
उन्होंने कहा कि वह एक अच्छा लड़का है कि कहा। (Pro-F, EA-T)

[5] Ram cuts the tree with an axe.
राम एक कुल्हाड़ी के साथ पेड़ काटता है। (Prep-F)

[6] Is he coming from school?
वह स्कूल से आ रहा है? (Intro-F)

[7] He was not going to market with you.
वह आप के साथ बाजार में नहीं जा रहा था? (Prep-F)

[8] Why is he driving the car?
यही कारण है कि वह कार चला रहा है? (Pro-F, Intro-F)

Bing translator also translated some of the text alike:

[1] Ram's brother is a good manager.
राम का भाई एक अच्छे प्रबंधक है। (Prep-F)

[2] That red flat belongs to Mohan.
उस फ्लैट लाल मोहन के अंतर्गत आता है। (Pro-F, Seq-F)

[3] That man had an iron rod.
आदमी एक लोहे की छड़ कि था। (Pro-F)

[4] He wore a chain made of gold.
वह सोने का बना एक चेन पहनी थी। (Gen-F)

[5] Ram cuts the tree with an axe.
राम के साथ एक कुल्हाड़ी पेड़ काटता है। (Seq-F, Prep-F)

[6] He has a cow.
वह एक गाय है। (Pro-F, VM-F)

[7] Drinking is harmful for health.
पीने के स्वास्थ्य के लिए हानिकारक है। (IE-F)

[8] You are a big cheese of our family.
आप हमारे परिवार का एक बड़ा पनीर रहे हैं। (Honf-F, IE-F)

[9] My brother was a bone of contention in my house.
मेरा भाई मेरे घर में विवाद का एक हड्डी थी। (IE-F, Gen-F)

[10] It's a dead letter now.
अब यह एक मरा पत्र है। (IE-F)

Anusaaraka translation is somehow close to accuracy due to Rule-Based algorithms but at some stages, it's also giving incorrect output like:

[1] That red flat belongs to Mohan.
लाल सपाट वह मोहन का है. (Pro-F,)

[2] That man had an iron rod.
उस आदमी का लोहा छडी थी. (Prep-F)

[3] He wore a chain made of gold.
उसने सोने से बनाई हुई माला अनुमति दी. (VM-F)

[4] Ram cuts the tree with an axe.
राम एक कुल्हाडी के साथ पेड काटता है. (Prep-F)

[5] He kept eating and drinking in whole journey.
उसने पूरी यात्रा में खाना और शराब पीती हुई जारी रखा. (Pro-F, VM-F)

[6] The girl in blue dress had a cup of coffee.
नीले लिबास में लडकी का कॉफी का एक प्याला था. (VM-F)

[7] What kinds of boy you want?
आप लडके के क्या प्रकार चाहते हैं? (Wh-F)

[8] Is someone special in your life?
क्या कोई आपके जीवन में विशेष है? (Seq-F)

[9] You are a big cheese our family.
आप हमारा परिवार एक बडी चीज हैं. (IE-F)

[10] It's now a dead letter.
यह मृत पत्र अब है. (IE-F)

The above English to Hindi Translation output has been translated from the current translation sites of *(Google, Bing and Anusaaraka)* i.e.
www.googletranslate.com
www.bing.com/translator
www.Anusaaraka.iiit.ac.in/drupal/node/32

PB Approach

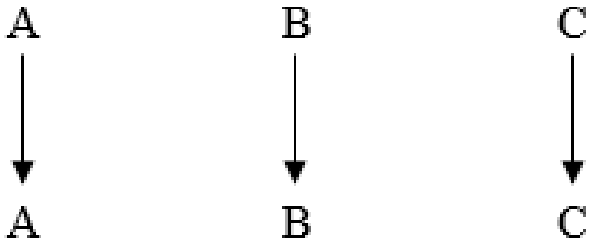
Phrased Based method is unique method found in current trends of translation. A sequence of words called phrase which can be either simple or idiomatic equally translated from one language to another. This method is mostly used for phrase alignment and translation from SL to TL. As this paper accounts for the Teaching and Testing (T&T) modal of MTs like *(Google, Bing and Anusaaraka)*, we have first followed the testing process which is previously defined and later we tried to find the use of this method with every system. As per our test we found that *Anusaaraka* is the one who support this method and rest are not. The analysis part says that it can be applied with the Statistical MT modal *(Google Translate & Bing Translator)* too.

Algorithm

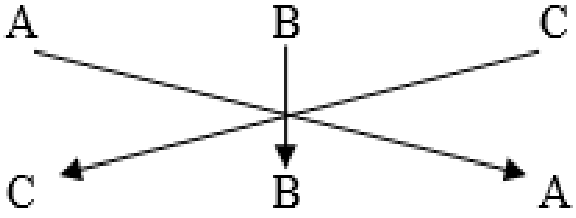
Google and Bing are statistical based modal for Machine Translation from one language to another. Google follows the (W-W) translation where as Bing follows the training data based translation. If we need to use the PB Method among any of the system, we need to change the system translation algorithm. By and large, PB method will mostly work in the two ways, First word to word (W-W) translation and later phrase alignment, Second direct phrase to phrase (P-P) straight translation. If any of these systems will configure in such a way, the translation mechanism will be possible. Some patterns of Phrased Based translation also map the word to word concept. Such patterns are identified as Reordering Phrase movement in these three logics.

- Straight Movement
- Cross Movement
- Zigzag Movement

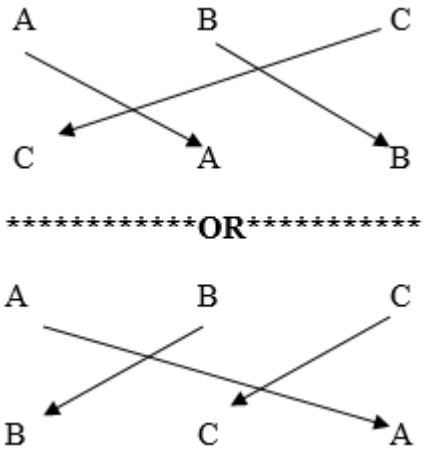
Straight Movement: in this movement ABC are determined as phrases in a sentence and it is translated as one to one straight manner.



Cross Movement: this movement takes the constituents/ phrases translated in one to one cross movement.



Zigzag Movement: in this movement constituents /phrases are translated in zigzag manner.



These movements state the order of translated words or phrase in aligned corpus. Apart from these three, there can be some other ways of movement too.

System Block Diagram

The expected block diagram represents the translation steps through this diagram.

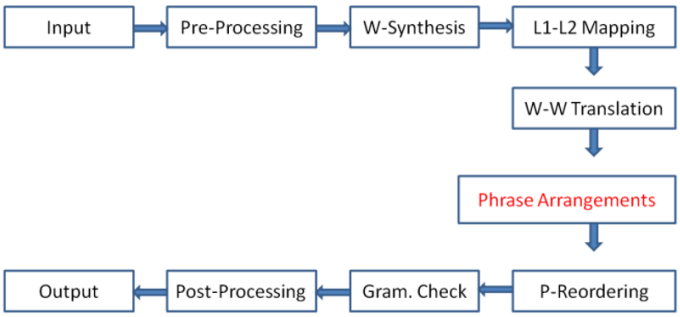


Fig 2

Result & Discussion

This testing analysis and discussion elaborated the detail picture of implementation for PB methodology in statistical system such as Google and Bing would be quite productive for output. A few findings have mentioned in points below:

- Minimise the restriction of word to word translation.
- Phrase can be translated quickly.
- Grammatical reordering of phrases could be possible on the basis of training data.
- Achievement of desired output would be expected.
- Idiomatic expression will be taken from available data source, provided to system.

References

1. Koehn Philipp, Franz J. Och, and Daniel Marcu. Statistical Phrase-Based Translation, 2003. doi:10.21236/ada461156.
2. Hutchins John W. Current commercial machine translation systems and computer-based translation tools: system types and their uses, International Journal of Translation. 2005; 17(1-2):5-38.
3. Rao Durgesh D. Machine translation. Resonance. 1998; 3(7):61-70. doi:10.1007/bf02837314.

4. Goyal V, Lehal GS. Web Based Hindi to Punjabi Machine Translation System. Journal of Emerging Technologies in Web Intelligence. 2010; 2(2). doi:10.4304/jetwi.2.2.148-151
5. Taraka Rama, Karthik Gali. Modeling Machine Transliteration as a Phrase Based Statistical Machine Translation Problem", in Proceedings of the 2009 Named Entities Workshop, ACL-IJCNLP. 2009, 124-127.
6. Bibek Behera, Pushpak Bhattacharyya. Automated Grammar Correction Using Hierarchical Phrase-Based Statistical Machine Translation. In Proceedings of IJCNLP, Nagoya, Japan, 2013.
7. Daniel Marcu and William Wong. A phrasebased, joint probability model for statistical machine translation. In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2002, 133-139.
8. Rama T, Singh AK, Kolachina S. Modeling letter to phoneme conversion as a phrase based statistical machine translation problem with minimum error rate training. In The NAACL Student Research Workshop, Boulder, Colorado, 2009.
9. Latha Nair R, David Peter S. Machine Translation Systems for Indian Languages, International Journal of Computer Applications. 2012; 39(1): 0975-8887.
10. Akshar Bharti, Chaitanya Vineet, Amba Kulkarni P, Rajiv Sangal. ANUSAARAKA: Machine Translation in stages, Vivek, a quarterly in Artificial Intelligence. 1997; 10(3):22-25.
11. Shachi Dave, Jignashu Parikh, Pushpak Bhattacharyya. Interlingua-based English-Hindi Machine Translation and Language Divergence. Journal of Machine Translation. 2002, 251-304.
12. Mahesh R, Sinha K, Anil Thakur. Machine Translation of Bi-lingual Hindi-English (Hinglish) Text, in proceedings of 10th Machine Translation Summit organized by Asia-Pacific Association for Machine Translation (AAMT), Phuket, Thailand, 2005.
13. Richard Zens, Hermann Ney. Improvements in phrase-based statistical machine translation. In Proceedings of HLT-NAACL. 2004, 257-264.
14. Ying Zhang, Stephan Vogel, Alex Waibel. Interpreting BLEU/NIST scores: How much improvement do we need to have a better system? In Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC). 2004, 2051-2054.
15. Imamura K. Application of translation knowledge acquired by hierarchical phrase alignment for pattern-based mt, In Proceedings of TMI, 2002.