International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 3, March - 2014

Implementation of Example Based Machine Translation System

Sunny Bhavan Sall
Department of Computer Engineering
SardarVallabhbhai Patel Polytechnic
Mumbai- India

Dr. Rekha Sharma
Department of Computer Engineering
Thakur College of Engineering & Technology
Mumbai ,India

***Abstract*—Automatic translation of text form one language into another is a Machine Translation. Due to globalization, it is the need of today's information technology dominated age to understand the text written into different languages by using computers. However, there are numerous challenges for automatic machine translation due to diversity of language constructs. This paper presents the implementation of Example Based Machine Translation (EBMT) by using Natural Language Processing (NLP) techniques. In this system, user can submit the text in English and it will be translated into Devnagari language, Hindi.**

Keywords—Machine Translation; Example Based Machine Translation; Natural Langauge Processing

I. INTRODUCTION

Example based machine translation (EBMT) is one such response against traditional models of translation. Like Statistical MT, it relies on large corpora and tries somewhat to reject traditional linguistic notions (although this does not restrict them entirely from using the said notions to improve their output). EBMT systems are attractive in that they require a minimum of prior knowledge and are therefore quickly adaptable to many language pairs. We ask that authors follow some simple guidelines. In essence, we ask you to make your paper look exactly like this document. The easiest way to do this is simply to download the template, and replace the content with your own material. Machine translation (MT) research has come a long way since the idea to use computer to automate the translation process and the major approach is Statistical Machine Translation (SMT). An alternative to SMT is Example-based machine translation (EBMT) [1]. The most important common feature between SMT and EBMT is to use a bilingual corpus (translation examples) for the translation of new inputs. Both methods exploit translation knowledge implicitly embedded in translation examples, and make MT system maintenance and improvement much easier compared with Rule-Based Machine Translation.

On the other hand, EBMT is different from SMT in that SMT hesitates to exploit rich linguistic resources such as a bilingual lexicon and parsers. EBMT does not consider such a constraint. SMT basically combines words or phrases (relatively small pieces) with high probability [2]; EBMT tries to use larger translation examples. When EBMT tries to use larger examples, it can better handle examples which are discontinuous as a word-string, but continuous structurally. Accordingly, though it is not inevitable, EBMT can quite naturally handle syntactic information. Besides that, the difference in between EBMT and SMT, EBMT is not the replacement for SMT. SMT is a natural approach when linguistic resources such as parsers and a bilingual lexicon are not available. On the other hand, in case that such linguistic resources are available, it is also natural to see how accurate MT can be achieved using all the available resources. EBMT is a more realistic, transparent, scalable and efficient approach in such cases. The language spoken by the human beings in day to day life is nothing but the natural language. There are many different applications under NLP among which Machine Translation is one of the applications. The work on machine translation began in late 1947. Machine translation deals with translating one natural language to another.

The ideal aim of Machine Translation system is to give the possible correct output without human assistance. The example based machine translation use the former examples as the based for translating source language to target language. The database for the two languages is considered for translation. Example based machine translation is bilingual translation. Example based machine translation use the corpus of two languages, the target language and the source language. We are proposing the design and development of an EBMT system. In this system, the English text entered by the user in the box is to be converted to Hindi without any divergence. The sentence i.e. text entered at the source side will be fragmented and the fragmented text will be matched into the corresponding target text. This will be done by using the data mining and the tree formation of the source text. The output then obtained will be aligned and the sentence having proper structure and the meaning will be generated using the corpus.The Example based machine translation is one of the approaches in machine translation. The concept uses the corpus of two languages and then translates the input text to desired target text by proper matching. The different languages have different language structure of the subject-object-verb (SOV) alignment. The matching is then arranged to give proper meaning in target text language and to form proper structure.

1.1 Various Issues in Machine Translation

All The task of translation is needed in day to day life. Humans can also do the task of translation; but now-a-days there is too much data to be coped with. So the job becomes tedious; therefore there is need for a translator which gives proper results for a text without any human assistance. The text should be translated properly without any divergence in the translation; i.e. the output for translation should be proper and no meaningless translation should be done. The speed of translation can also be increased.

The translation done till now is not accurate , to give results with the divergence in conversion form source language to target language. There are certain drawbacks which does not give translation without human assistance. There is a genuine requirement of having a machine translation system which can overcome the limitations of existing machine translation systems , and provide the translated content with high relevance and precision. EBMT is trying to minimize the human assistance and still give a better translation.

Recently corpus based approaches to machine translation have received wide focus. They are namely Example Based Machine Translation (EBMT) [6] and Statistical Machine Translation (SMT) [7]. A combination of statistical and example-based MT approaches shows some promising perspectives for overcoming the shortcomings of each approach. Efforts have been made in this direction using the alignments from both the methods to improve the translation [8], to improve the alignment in the EBMT using the statistical information computed from SMT methods [9] etc. The results obtained have shown improvement in performance. However, these approaches cannot directly be applied to Indian languages due to the small size of the parallel texts available and sparse linguistic resources. Also some of the assumptions made in some of these approaches like marker hypothesis [10], cannot directly be applied to translate from English to Indian languages since word order in the source and target languages is very different and sequential word orderings between source and target sentences do not exist.

Machine translation of Indian Languages has been pursued mostly on the linguistic side. Hand crafted rules were mainly used for translation, [11], [12]. Rule based approaches were combined with EBMT system to build hybrid systems [13], [14] performs interlingua based machine translation. Input in the source language is converted into UNL, the Universal Networking Language and then converted back from UNL to the target language. Recently, Gangadhariahet,al [15] used linguistic rules are used for ordering the output from a generalized example based machine translation [16]. While, in general in the machine translation literature, hybrid approaches have been proposed for EBMT primarily using statistical information most of which have shown improvement in performance over the pure EBMT system. [17] automatically derived a hierarchical TM from a parallel corpus, comprising a set of transducers encoding a simple grammar. [18] used example-based re-scoring method to validate SMT translation candidates. [19] proposed an example based decoding for statistical machine translation which outperformed the beam search based decoder [20]. Kim et al [9] showed improvement in alignment in EBMT using statistical dictionaries and calculating alignment scores bi-directionally. [8], [21] combined the sub-sentential alignments obtained from the EBMT systems with word and phrase alignments from SMT to make 'Example based Statistical Machine Translation' and 'Statistical Example based Machine Translation'.

The EBMT module shares similarities in structure with three stages : analysis, transfer & generation as shown in the figure 1.The Vauquois Pyramid adapted for EBMT [22].

- Direct
- Transfer
- Interlingual

minimum of prior knowledge and are therefore quickly adaptable to many language pairs. The particular EBMT system that we are examining works in the following way. Given an extensive corpus of aligned source-language and target-language sentences, and a source-language sentence to translate:

1. It identifies exact substrings of the sentence to be translated within the source-language corpus, thereby returning a series of source-language sentences
2. It takes the corresponding sentences in the target-language corpus as the translations of the source-language corpus (this should be the case!)
3. Then for each pair of sentences:
   i. It attempts to align the source- and target-language sentences;
   ii. It retrieves the portion of the target-language sentence marked as aligned with the corpus source-language sentence's substring and returns it as the translation of the input source-language chunk

.

The above system is a specialization of generalized EBMT systems. Other specific systems may operate on parse trees or only on entire sentences. The system requires the following:

1. Sentence-aligned source and target corpora.
2. Source- to target- dictionary
3. Stemmer

The stemmer is necessary because we will typically find only uninflected forms in dictionaries. While it is consulted in the alignment algorithm, it is not consulted in the matching step—as stated before, those matches must be exact.

II. RELATED WORK

A. . Researches on Machine Translation

The history of machine translation research can be traced back to the forties of the 20th century. American scientist W. Weaver and British engineers AD, Booth proposed the idea of machine translation. In 1954, Georgetown University collaborating with IBM completed, for the first time, the Anglo-Russian Machine Translation with the IBM-701 computer. It proved the feasibility of a machine translation. Thus studies on machine translation began. In 1964, in order to evaluate the progress of the research on machine translation, Automatic Language Processing Advisory

Committee (ALPAC), began a two-year comprehensive survey and testing on machine translation. In November 1966, the Commission published ALPAC report. The report provided a comprehensive denial of the feasibility of machine translation, and recommended to stop the financial support of the machine translation project. From the 1970s, with the increasing frequency of the exchange among countries, machine translation is urgently needed by society. Meanwhile, the development of computer science, linguistics research, particularly the increase in computer hardware technology and applications of artificial intelligence in natural language processing promote the recovery of study on machine translation. Machine translation projects began to develop. A variety of practical and experimental systems has been introduced .such as, Weinder system, the EURPOTRA multilingual translation system, TAUM-METEO system. With the universal application of the Internet, the acceleration of the integration process of the world economy, machine translation enters a new development level.

III. PROPOSED WORK

The Example based machine translation is one of the approaches in machine translation. The concept uses the corpus of two languages and then translates the input text to desired target text by proper matching.

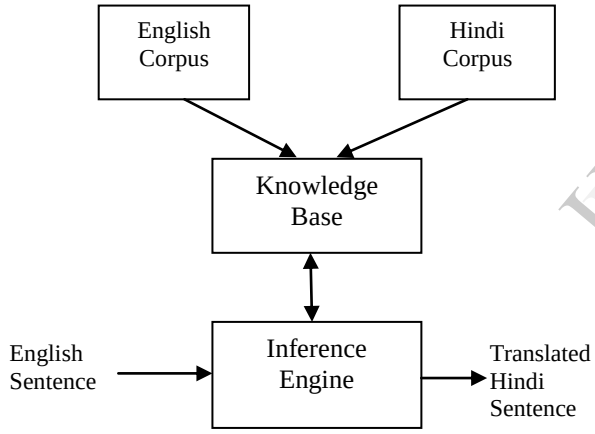

Figure 1 : Example Based Machine Translation

The different languages have different language structure of the subject-object-verb (SOV) alignment. The matching is then arranged to give proper meaning in target text language and to form proper structure. In this paper, we describe the Example Based Machine Translation using Natural Language Processing. The proposed EBMT framework can be used for automatic translation of text by reusing the examples of previous translations. This framework comprises of three phases, matching, alignment and recombination.

A. *Example Based Machine Translation*

- **English Corpus** :We have used 1000 English sentences for forming a corpus. The sentences are the news headlines from reputed newspaper.
- **Hindi Corpus :**It consist of the translated sentences in hindi for each of the English sentences**.**
- **Knowledge Base** : It stores the patterns of how English sentences are translated into Hindi form.
- **Inference Engine :**It is a collection of facts and rules.

  Inference Engine compares the given English sentence with the English sentences stored in the corpus. After finding the best match, it translates it into Hindi according to the Hindi translation present in the Hindi corpus.

B. *Proposed EBMT Implementation*

We have implemented the Example based machine translation system as shown in figure 2.

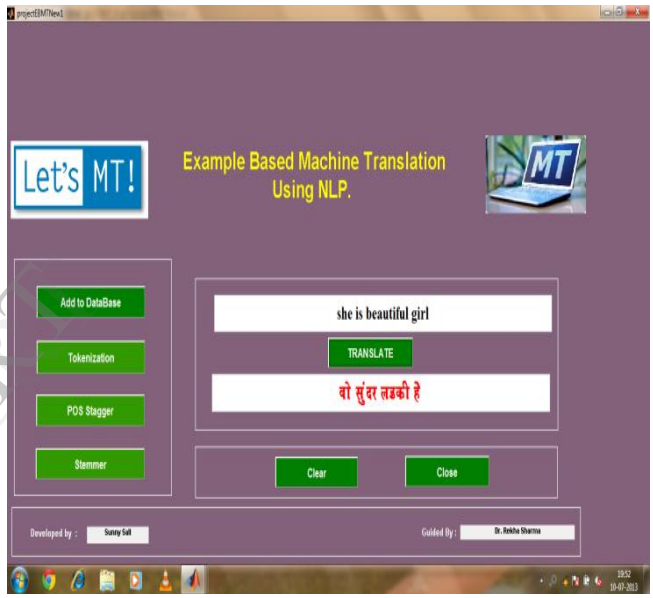

Figure 2 : Proposed EBMT system

Following are some of the examples of Example Based Machine Translation

Example 1

English : India won the match.
Hindi :Hkkjr us eWpftrk

Example 2

English : India is the best
Hindi :HkkjrloZJs"BgS

Example 3

English :Sachin plays well
Hindi :lfpuzvPNk [ksyrkgS

Input

English :Sachin is the best

Translation (Output)

Hindi :lfpuloZJs"BgS

IV. RESULTS AND DISCUSION

In this research work, I have used the corpus consisting of 1000 simple and complex sentences, to carry out the experiments of machine translation. The performance metrics used for evaluation of translation are *unigramprecision*, *unigramrecall*, *F-measure*, *BLUE*, *NIST*, *mWER* and *SSER*. I have tried to compare the results of translation for three techniques : RBMT, SMT and EBMT.

Unigram Precision: As mentioned before, we consider only exact one-to-one matches between words. Precision is calculated as follows:

$$P = \frac{m}{wt}$$

where m is the number of words in the translation that match words in the reference translation, and wt is the number of words in the translation. This may be interpreted as the fraction of the words in the translation that are present in the reference translation.

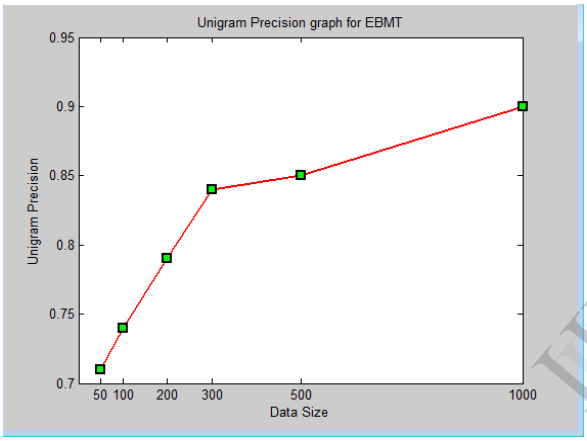

Figure 2 : Unigram Precision

In this graph, corpus size is varied from 0 to 1000 along x-axis and corresponding unigram precision is plotted along y-axis

Unigram Recall: As with precision, only exact one-to-one word matches are considered. Recall is calculated as follows:

$$R = \frac{m}{wr}$$

where m is the number of matching words, and wr is the number of words in the reference translation. This may be interpreted as the fraction of words in the reference that appear in the translation.

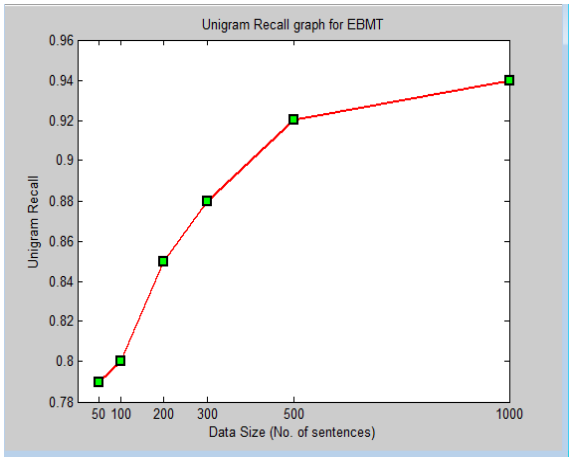

Figure 3 : Unigram Recall

In this graph, corpus size is varies from 0 to 1000 along x-axis and corresponding unigram recall is plotted along y-axis

F-measure : The F-measure of precision and recall is computed as follows:

$$F = \frac{2PR}{P + R}$$

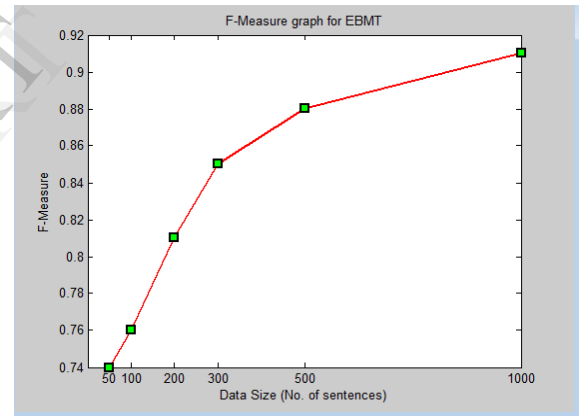

Figure 4 : F-Measure

In this graph, corpus size is varied from 0 to 1000 along x-axis and corresponding F-measure is plotted along y-axis

BLEU (Papineni et al., 2001): This measures the precision of n-grams with respect to the reference translations, with a brevity penalty. A higher BLEU score indicates better translation.

$$\text{BLEU} = BP * \exp(\sum_{n=0}^{n-1} (\text{wn} * \text{logpn})$$

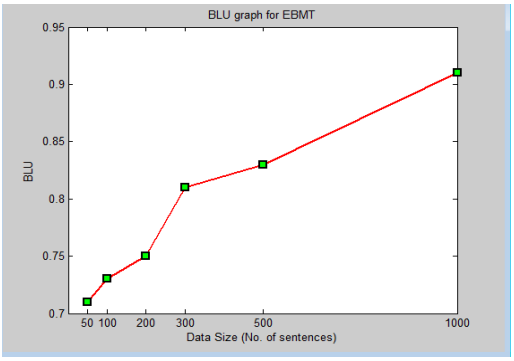

Figure 5 :BLEU

In this graph, corpus size is varied from 0 to 1000 along x-axis and corresponding BLEU is plotted along y-axis

NIST : Translation adequacy is captured by NIST score. A translation using the same words (1-grams) as in the references tends to satisfy *adequacy*.

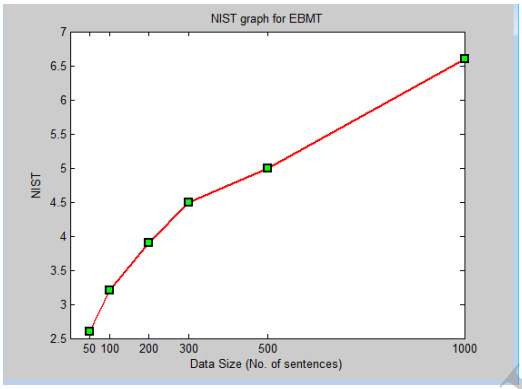

Figure 6 : NIST

$$Info(w1 \dots wn) = \log \frac{\text{No. of occurences of (w1 .... wn} - 1)}{\text{No. of occurences of w1, ... wn}}$$

Performance evaluation of EBMT for NIST is displayed in above graph i.e. figure 6

MWER (multi-reference word error rate) (Nießen et al., 2000): This measures the edit distance with the most similar reference translation. Thus, a lower mWER score is desirable.

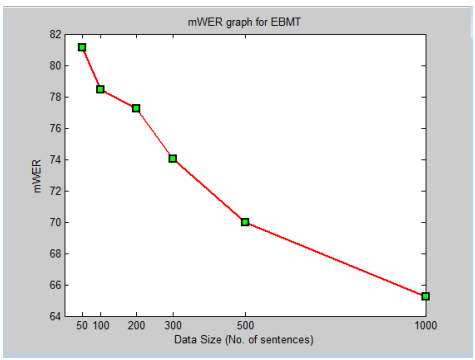

Figure 7 : MWER

Performance evaluation of EBMT for MWER is displayed in above graph i.e. figure 7

SSER (subjective sentence error rate) (Nießen et al., 2000): This is calculated using human judgements. Each sentence was judged by a human evaluator on the following five-point scale, and the SSER was calculated as described in (Nießen et al., 2000).

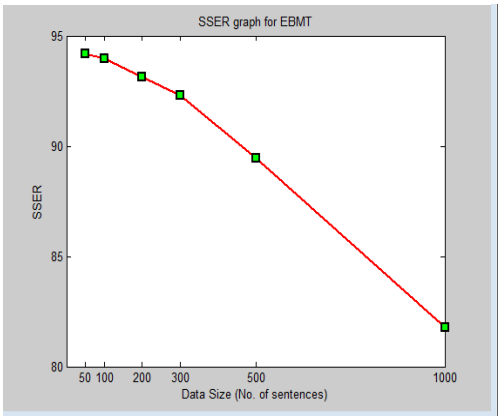

Figure 8 : SSER

0	Nonsense
1	Roughly understandable
2	Understandable
3	Good
4	Perfect

Performance evaluation of EBMT for MWER is displayed in above graph i.e. figure 8

TABLE I. PERFORMANCE EVALUATION OF EBMT

Corpus Size (No. of Sentences)	Unigram Precision	Unigram Recall	F-measure	BLEU	NIST	mWER	SSER
50	0.71	0.79	0.74	0.71	2.6	81.11	94.21
100	0.74	0.80	0.76	0.73	3.2	78.44	93.96
200	0.79	0.85	0.81	0.75	3.9	77.24	93.12
300	0.84	0.88	0.85	0.81	4.5	74.02	92.32
500	0.85	0.92	0.88	0.83	5.0	70.00	89.44
1000	0.90	0.94	0.91	0.91	6.6	65.22	81.77

The performance of EBMT system is evaluated and it is tabulated in Table I. Whereas in Figure 9,10 and 11 the comparison of the unigram precision, unigram recall and F-measure for RBMT, SMT and proposed EMBT system are shown. The unit of corpus is the number of sentences taken for machine translation.

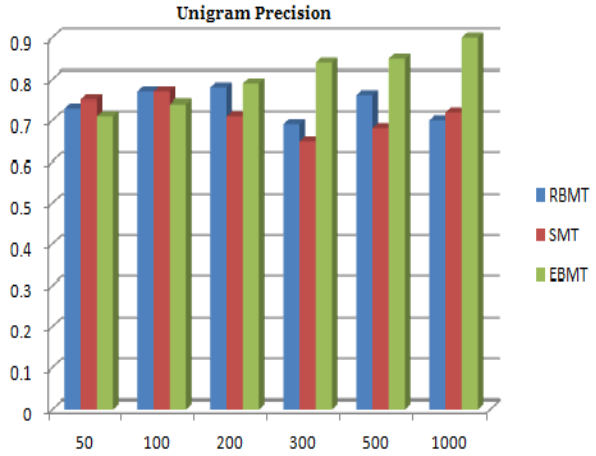

Figure 9 : Comparison of unigram precision

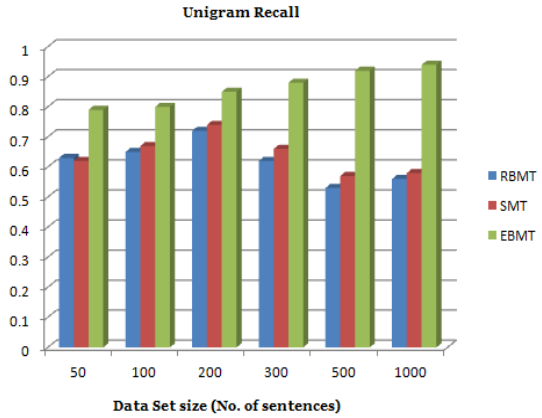

Figure 10 : Comparison of unigram recall

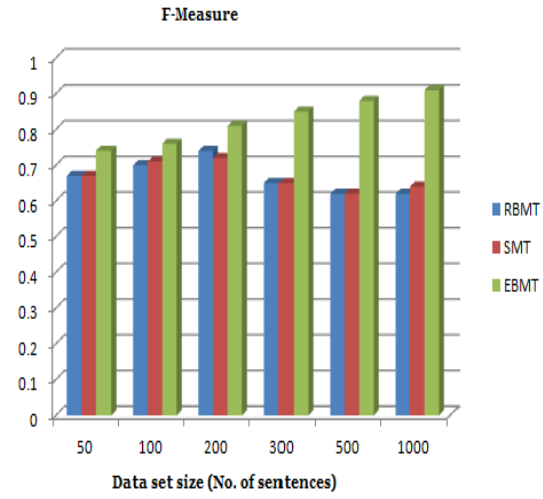

Figure 11 : Comparison of F-measure

From figure 9, 10 and 11 it reveals that the effectiveness and quality of translation can be improved by implementing example based machine translation.

V. CONCLUSION

We proposed a new system, which is scalable, transparent and efficient. The entire system will convert the source language text into target language text using natural language processing. It will use the machine translation technique which is better than the existing tools available in the marketThe algorithm is such that, there is dictionary / corpus / vocabulary of **English** and **Hindi**. The parsing will be proper. The mapping technique will also be used. All the Literals will be separated using partitioning and stemming techniques. The root word will be identified using artificial intelligence and bilingual translation.

We pursue the study of example based machine translation using natural language processing.

VI. ACKNOWLEDGMENT

Our thanks to the experts who have contributed towards development of this research work. We are thankful to Prof. R.R.Sedamkar for his constant guidance and support. We also thankful Dr. B.K.Mishra for providing us environment for research.

VII. REFERENCE

[1] D. Gupta, N. Chatterji, "Identification of divergence for English to Hindi EBMT," , in preceedings of Machine Translation, pp. 141-148, 2010.

[2] E. Sumita, "EBMT Using DP-Matching Between Word Sequences,", in ATR Spoken Language Translation Research Lab., Tokyo Japan, 2001.

[3] Allen J., "Linguistic Aspect of Speech Synthesis,", in Voice Communication Between Humans and Machines, D. B. Roe, National Academy of Science, Washington, Columbia, 1994.

[4] Alexanderson J., Reithinger N., "Insight into the Dialog Processing of Verbomil,", in Proceedings of the Fifth Conferenceon Applied Natural Language Processing, Association for Computational Linguistics, Morgan Kaufmann, San Francisco, California, pp. 33-40, 1997.

[5] Nagao, "A Frame work of a Mechanical Translation Between Japanese and English by Analogy principle,", in Preceedings of the International NATO Symposium on Artificial and Human Intelligence, pp. 173-180, 1984.

[6] R. D. Brown, "EBMT in Pangloss System,", International Workshop on EBMT, 1993.

[7] Groves and Way, "A Memory Based Classification Approach to Marker Based EBMT,", Duplin, Ireland, 2006.

[8] J. D. Kim, ",Chunk Based EBMT", in Annual Conference of European Association for Machine Translation, 2010.

[9] Guogh, "Example Based Machine Translation,", Valetta, Malta, pp.35-42, 2005.

[10] Sinha and A. Jain, "A English to Hindi Machine Aided Translation System,", 2002.

[11] L. K. Bharti, "Constraint Based Hybrid Approach to Parsing Indian Languages,", IIIT Hydrabad, 1997.

[12] V. Jain, "A Smorgasbord of features for Statistical Machine Translation",2003.