Machine Translation of Natural Language using different Approaches: ETSTS (English to Sanskrit Translator and Synthesizer)

Sarita G. Rathod
Information Technology
VESIT, Chembur
Mumbai, India

ABSTRACT
In our work, we integrate a traditional approach of machine translation which translates source language sentence into equivalent target language sentence. We proposed an English (source language) to Sanskrit (target language) machine translator based on Rule based approach and Example based approach. We further compare the performance of these approaches for different category of sentences as like small, large, and extra large.

We also develop a GUI for making it user friendly and provide the speech output of the target sentence with the help of speech synthesizer a plug in module.

Keywords
Rule based machine translation, Example based machine translation, Parser, Bilingual Dictionary, Formant Synthesizer, Machine translation, Evaluation parameter.

1. INTRODUCTION
Machine translation (MT) is the area of information technology and applied linguistics dealing with the translation of human languages. MT is automated translation. It is the process by which computer software is used to translate a text from one natural language (such as English) to another (such as Spanish). The ideal aim of machine translation systems is to produce the best possible translation without human assistance. Basically every machine translation system requires programs for translation and automated dictionaries and grammars to support translation. To process any translation, human or automated, the meaning of a text in the original (source) language must be fully restored in the target language, i.e. the translation.

The Sanskrit language is basically the language of the ancient India and considered as the mother language from which all other Indian languages evolved. In this time Sanskrit language is a dead language. But it is recognized in the Indian constitution of 1950 because Sanskrit is related and associated with the religion and literature of India. Since Sanskrit is an ancient language and is no more in vogue as an easily understandable language, many communities are constantly working towards the translation of Sanskrit texts to popular languages. Presently, scholars of Sanskrit are doing this translation work manually and they have expressed a need for some software to do this work for them. The project that is modeled here is an attempt to ease the work of scholars and help accelerate the translation efforts. It shall also make it possible for the common people, who are not familiar with Sanskrit, to translate texts and understand them.

The most popular language in the world today is English. So, a translator which can translate English sentences into Sanskrit will prove very useful. The proposed software is an effort to develop this for use in a restricted domain. As the preceding section tells us, there is a need for a translator to translate sentences from English to Sanskrit. The project built here presents a model for this purpose. Broadly, speaking first we perform analysis and tokenization of English sentences and then translate them into Sanskrit using RBMT or EBMT. Then we compare the output of both the techniques for same set of sentences.

Next, the machine translation system is evaluated by using different parameter like precision, recall, F-measure, Meteor, Bleu. According to the results (parameter values) we can define the accuracy and performance of the machine translator. The RBMT and EBMT based translators are compared based on these metrics.

2. LITERATURE SURVEY
The work already has done in the machine translation by different people. There are various methods for machine translation. These are classified according to their possession of knowledge, these groups are as Rule Based Machine Translation and Corpus Based Machine, along with their theoretical foundation (statistical or example-driven) for achieving translation process, Statistical Machine Translation and Example Based Machine Translation. And the combination of any of these three makes the Hybrid Machine Translation. Sitender, Seema Bawa [2] had given survey for Indian machine translation, according to this paper the work done on various Indian machine translation systems either developed or under the development. Some systems are of general domain, but most of the systems have their own particular domains like parliamentary documents translation, news readings, children stories, web based information retrieval etc. Sudip Naskar and Sivaji Bandyopadhyay [5] had given a survey of current status of the machine translation systems that have been developed in India for translation from English to Indian languages and among Indian languages reveals that the MT softwares are used in field testing or are available as web translation service. These systems are also used for teaching machine translation to the students and researchers. Most of these systems are in the English-Hindi or Indian language-Indian language domain. The translation domains are mostly government documents/reports and news stories. In Rule Based machine translation R.M.K. Sinha and A. Jain [1] had given a system overview of English to Hindi Machine Aided Translation System. It is known as AnglaHindi. ANGLAHINDI accepts unconstrained text. The

text may be made up of headings, parenthesized texts, text under quote marks, currencies etc. The performance of the system has been evaluated by human translators. The system generates approximately 90% acceptable translation in case of simple, compound and complex sentences up to a length of 20 words. Current version of AnglaHindi is not tuned to any specific domain of application or topic. However, it has a user friendly interface which allows hierarchical structuring of the lexical database leading to preferences on lexical choice. Khaled Shaalan [4] had given the Rule-based Approach of machine translation for English to Arabic Natural Language Processing and the rule based tools for Arabic natural language. It has given the morphological analyzers and generator and syntactic analyzer and generators. Sandeep Warhade [6] had given a design of Phrase-based decoder for English-to-Sanskrit translation .It describes the Phrase-Based Statistical Machine Translation Decoder for English as source and Sanskrit as target language. Their goal is to improve the translation quality by enhancing the translation table and by preprocessing the source language text research. They discuss the major design objective for the decoder, its performance relative to other SMT decoders.

In Example Based machine translation, Vimal Mishra and R. B. Mishra [3] had given Example Based English to Sanskrit Machine Translation. In this paper, a comparative view of EBMT and RBMT is presented on the basis of some specific features. This paper describes the various research efforts on Example based machine translation and shows the various approaches and problems of EBMT. Rajpal Singh, Dr. Gurpreet Singh Josan [7], this paper presents example-based machine translation architecture using translation memory that integrates the use of examples for flexible, idiomatic translations with the use of linguistic rules for broad coverage and grammatical accuracy.

In Evaluation parameter of machine translation, BLEU is one of the parameter. According to Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu [8]**,** Human evaluations of machine translation are extensive but expensive. They propose a method of automatic machine translation evaluation that is quick, inexpensive, and language-independent, that correlates highly with human evaluation, and that has little marginal cost per run.

3. PROPOSED SCHEME

In this proposed scheme the first objective is to develop translator module which translates source language (English) to target language (Sanskrit) using the two different approaches of machine translation i.e. Rule based machine translation and example based machine translation and

4. DESIGN, IMPLEMENTATION AND RESULT

Design of the system is divided into different parts. First of all we have the detail design of the complete ETSTS System. After that we have given the detail design and working of RBMT and EBMT techniques.

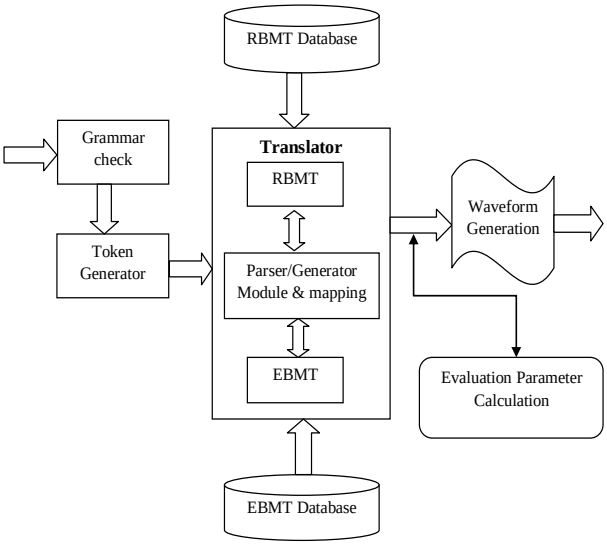

Fig 1: Basic Block diagram of ETSTS

- Text Input: This is the first phase in the machine translation process and is the first module in any MT system. In ETSTS system it is a source English sentence.
- Grammar/Spell Check: This module is use for checking the grammar and spelling mistakes in the input source sentence and modifies it as per requirement of sentence.
- Token Generator: This module splits the given sentence into chunks of strings delimited by spaces. These strings may be simple words or compound words coalesced by the rule of Sandhi.
- Translator: In this system, the translation of source language to the target language has done by two different techniques i.e. RBMT and EBMT. It selects any one technique at a time. This module performs the actual translation. The input to this module is the parse tree which is generated by using the Stanford parser with proper tagging for that purpose here we have the parser/generator module to get the parse of each word. It then generates appropriate equivalents in English for the morphological details of each word and ultimately presents the sentence in the correct order.
- Parser Generator Module: This module contains a set of transducers built for individual Sanskrit words and transforms strings to partial words, which are used by the EBMT/RBMT module. It also gives the parse of the words, which are used by the sentence former to give the output in a structurally correct sentence.
- RBMT/EBMT: These are the two different techniques used for this English to Sanskrit translation. These two techniques have different way to generate the output and it is explain in detail in the following sections. Both are using the different database for generating the output. This module passes their output to the Speech Generation module for further procedure.
- RBMT/EBMT Database: This module provides the database to the system**.** For RBMT it has the lexical

Bi-lingual database i.e. the dictionary of source and target language. After tokenization and tagging of the input sentence into chunk of words, it checks for the meaning of each word from this dictionary which is exactly match and give the target sentence. As like this, in EBMT also it has the database of reference examples stored which is use to generate the target sentence. After tokenization and tagging of words it is divided into different phrases and searches the meaning of it in Sanskrit from the example database.

- Text Output: This module gives the output in the Devnagari script of Sanskrit text.
- Evaluation parameter calculation: This module gives the values of the different parameters which are used to evaluate the any machine translation system. Here we have implemented the five different standard evaluation parameters as Unigram Precision, Unigram Recall, F-Measure, Meteor Metric and Bleu parameter. These parameters are implemented as comparing the reference Sanskrit sentence with Candidate Sanskrit sentence, by using their given formulas. If the value of any parameter is near to one then it is near to the right output and if the value is near to zero then it is near to the wrong output.
- Waveform Generation: The input to this is from the translator module of the system, i.e. target sentence in Sanskrit. Afterwards this text will analyzed linguistically and providing proper phonetic alphabets it is converted into the speech waveform and gives the voice output.

The objectives described in section II of paper may be implemented in numerously many ways. Out of that two has been employed for the present work.

In both the approaches a combination of databases along with logic can implement using some structured or objects oriented programming language. An appropriate programming language in the form of Java has been taken to implement the translator algorithm. The details of the design are given in above. The methodologies used in this work are different to the other which is described in the beginning of the section in sense of the morphological details as well as the lexicon. The main idea behind dictionary based Machine Translation is that input sentence can be transformed into output sentence by carrying out the simplest possible parse, replacing source word with their target language equivalents as specified in a dictionary, and then roughly re-arranging their order considering rules of the target language. And for that we used different functions for converting English sentence into Sanskrit sentence. And in Example based machine translation the input sentence can be converted into small phrase of source language, then in the matching phase the phrases are converted into target language phrase using example database and through the alignment and recombination it gives complete long target output sentence. The comparison of results of both the techniques has been done based on the different parameters of MT and the time required for the execution of both the approaches separately.

4.1 Evaluation parameter

Here we evaluate the performance of our ETSTS system using different MT evaluation methods. These are the five standard evaluation parameter of the Machine Translation System. The parameter are given as follows [6],

- Unigram Precision: Precision is fraction of correct instances among those that algorithm believes to belong to relevant subset and is calculated as,

$$P = m/Wt \quad (1)$$

Where P is Unigram Precision, m is number of unigram matches and Wt is the number of unigram in candidate translation.

- Unigram Recall: Recall is fraction of correct instances among all instances that actually belong to relevant subset and can be calculated as,

$$R = m/Wr \quad (2)$$

Where R is Unigram Recall, m is number of unigram matches and Wr is the number of unigram in reference translation.

- F-Measure: It is an MT evaluation metric developed at the New York University. The F-measure is defined as the harmonic mean of precision and the recall as,

$$\text{F-Measure} = 2*\text{Precision}*\text{Recall} / (\text{Precision}*\text{Recall}) \quad (3)$$

- METEOR (Metric for Evaluation of Translation with Explicit Ordering): It is an MT evaluation metric which is developed at Carnegie Mellon University. The Meteor metric is based on the weighted harmonic mean of unigram precision and unigram recall. Main idea is to Combine Recall and Precision as weighted score components. Fmean is calculated by combining the recall and precision via a harmonic-mean that places equal weight on precision and recall as follows,

$$Fmean = 2PR / (P+R) \quad (4)$$

This measure is for congruity with respect to single words but for considering longer n-gram matches, a penalty p is calculated for the alignment as,

$$P = 0.5(C/um)3 \quad (5)$$

Where c is the number of chunks, and um is the number of unigrams that have been mapped. The more mappings there are, that are not adjacent in the reference and the candidate sentence, the higher the penalty will be. Final Meteor-score (M-score) can be calculated as,

$$M = Fmean(1-P) \quad (6)$$

- BLEU (Bilingual Evaluation Understudy): It is an IBM-developed [14] metric and uses modified n-gram precision to compare the candidate translation against reference translations.

$$P_n = \frac{\sum_{C\in\{Candidates\}}\sum_{n-gram\in C} Countclip(n-gram)}{\sum_{C'\in\{Candidates\}}\sum_{n-gram'\in C'} Count(n-gram')} \quad (7)$$

Where Pn is Corpus-based N-gram Precision, C is the set of candidate translation sentences and C' is the set of reference sentences. Countclip is count of n-gram match found in both candidate and reference, Count is count of n-gram found only in candidate, the formula for calculating brevity penalty is,

$$BP = min\left(1, e^{(1-r/c)}\right) \quad (8)$$

Where BP is brevity penalty, r is length of reference and c is length of candidate. Then Bleu score is calculated as,

$$Bleu = BP * exp\left(\sum_{n=1}^{N} \frac{1}{n} \log\left(Pn\right)\right) \quad (9)$$

4.2 Expected output screen

The expected output screens are as follows, figure.1 represents the grammar and spell checking for the input sentence. We can make the correction and get proper correct sentence. The following figure.2 represents the output of the translator using rule based approach. We can get the speech output of the text which has been translated into Sanskrit using RBMT.

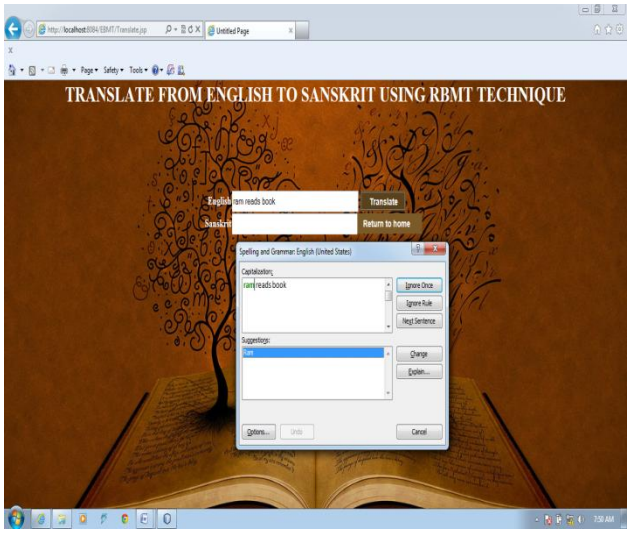

Fig 2: Grammar and Spell Check

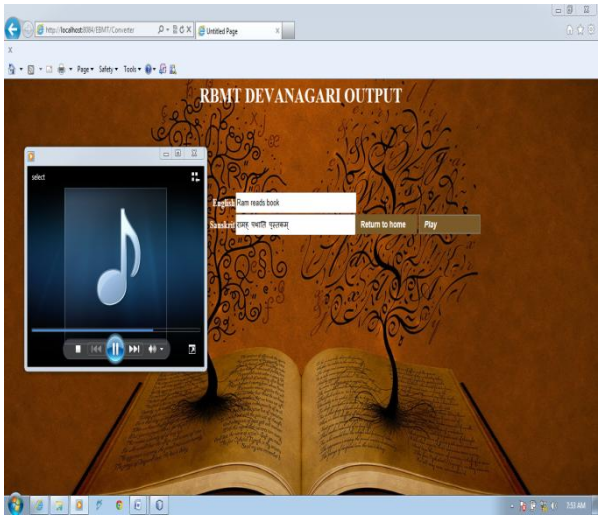

Fig 3: Output Using RBMT

The figure.3 represents the parse tree for the input source sentence, which gives part of speech for each word in sentence and represents the translator output using Example based machine translation approach. Figure 4 also represents the comparison of two approaches for the same input sentence based on the different evaluation parameter values of MT. The important parameters are METEOR and BLEU, where the values of this are in between 0 and 1. If it is near to 0 then it cannot be a good translation and if it is near to 1 then it can be better translation.

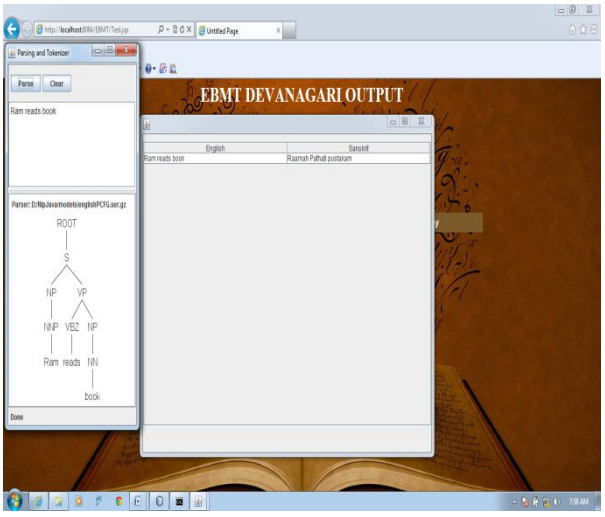

Fig 4: parse tree

To check robustness of the system, ETSTS system took samples of hundred sentences of various types, as the sentences of simple types in active and passive voice. We have considered sentences from all the three tenses i.e. present, past and future. It is our belief that this methodology can be adopted for translation of similar languages. The sentences are divided into three categories that are small, large and extra-large to find out the accuracy of the system as per the parameter values. Category wise results are shown in the following tables with parameter values. Each category has the ten different examples to evaluate.

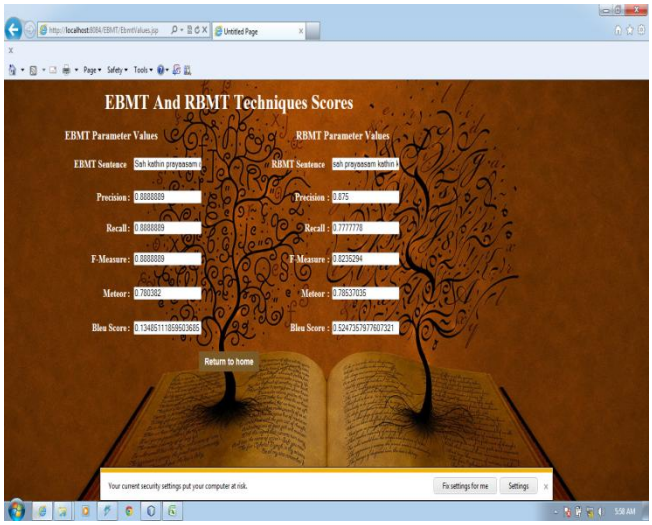

Fig 5: Comparison by parameter values for same sentence

- **Category 1: Small Sentences-** This category will have the small sentence of maximum 3 words in it.
- **Category 2: Large Sentences-** This category will have the large sentence of more than 3 and less than equal 5 words in it.
- **Category 3: Extra large-** This category will have the large sentence of more than 5 and less than equal 8 words in it.

We observed the performance of the system category wise and got the results for randomly selected different sentences. The comparative score of different MT evaluation methods such as BLEU (Bilingual Evaluation Understudy), unigram Precision (P), unigram Recall (R), F-measure (F) and METEOR (M) for randomly selected English sentences of all three categories of sentence by RBMT and EBMT in percentage has given in following table 1,

Table 1. Values of parameter for three categories in percentage

Category	Precision		Recall		Fmeasure		Meteor		Bleu	
	RB	EB	RB	EB	RB	EB	RB	EB	RB	EB
Small	77	70	77	82	76	75	75	79	93	95
Large	63	77	61	81	68	79	68	79	81	93
Extra Large	70	76	69	77	69	76	69	73	86	87

For small sentence category the Precision score of RBMT is 7% more than EBMT, Recall score of EBMT is 5% more than RBMT, F-Measure score RBMT is 1% more than EBMT, METEOR score of EBMT is 4% more than RBMT and BLEU score of EBMT is 2% more than RBMT.

For Large sentence category the Precision score of EBMT is 14% more than RBMT, Recall score of EBMT is 20% more than RBMT, F-Measure score of EBMT is 11% more than RBMT, METEOR score of EBMT is also 11% more than RBMT and BLEU score of EBMT is 12% more than RBMT.

For extra large sentence category the Precision score of EBMT is 6% more than RBMT, Recall score of EBMT is 8% more than RBMT, F-Measure score of EBMT is 7% more than RBMT, METEOR score of EBMT is 4% more than RBMT and BLEU score of EBMT is 1% more than RBMT. As per the overall result of randomly selected sentence of all categories we find that EBMT performs better than RBMT for all the three categories of sentences. We can also observed the comparison of two approaches in graphical view as follows,

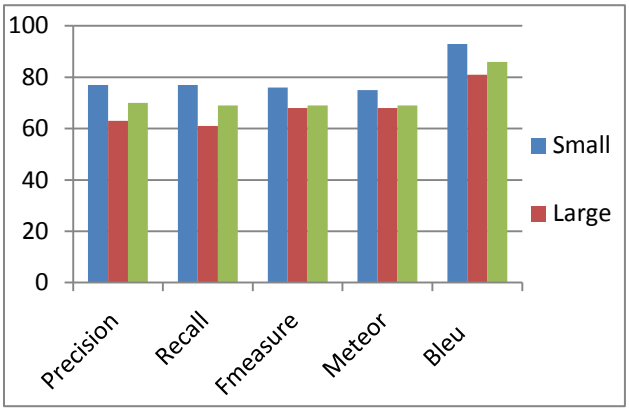

Fig 6: Comparison of parameter values for all three types of sentence using RB

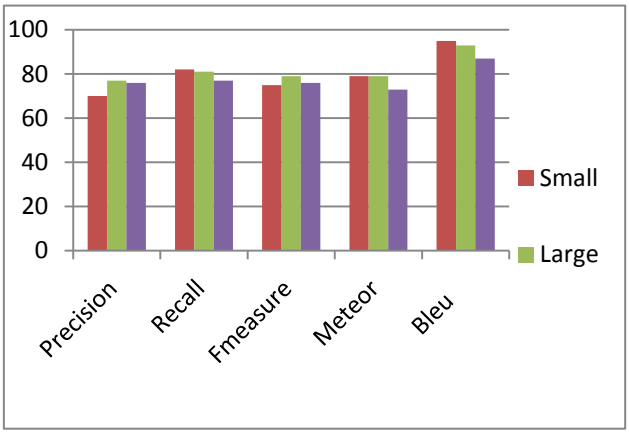

Fig 7: Comparison of parameter values for all three types of sentence using EB

5. CONCLUSION

In our work, the complete framework for Rule Based Translation and example based translation is outlined. The chosen language pair is English and Sanskrit, as a source and target language respectively. The system (ETSTS) supports both English and Sanskrit grammar such as noun, verb adjective etc. This system can be extended to translate various types of literature in English to Sanskrit. Based on the

observations above, several experiments with system were conducted. Our system handles English sentences of types: (i) simple subject, object and verb; (ii) subject, object, adverb and verb; (iii) subject, object, adjective and verb; (iv) subject, object, preposition and verb.

In the proposed work we have compared both the methodologies and obtained the comparative score of different evaluation methods like unigram Precision (P), unigram Recall (R), F-measure (F), METEOR (M) and BLEU (Bilingual Evaluation Understudy) for the sentences in per small, large, extra large categories. So from the results we conclude that Example Based Approach of Machine Translation gives the better performance than the Rule Based Approach for all the three category of sentences. We also presented the translated target sentence output into the speech output using speech synthesizer. So from the performance comparison point of view we can say that the EBMT technique of machine translation gives improvement 10-12% with respect to RBMT technique.

6. REFERENCES

[1] R.M.K. Sinha, A. Jain "AnglaHindi: English to Hindi Machine-Aided Translation System". International Journal of Computer application Vol. 3, No. 3, Oct 2008.

[2] Sitender, Seema Bawa "Survey of Indian Machine Translation Systems". IJCST Vol. 3, Issue 1, Jan. - March 2012 ISSN : 0976-8491 (Online) | ISSN : 2229-4333 (Print).

[3] Mishara Vimal, Mishara RB. "Study of Example Based English to Sanskrit Machine Translation" department of computer engineering .Institute of technology, Banaras Hindu University, Varanasi India.

[4] Khaled Shaalan "Rule-based Approach in Arabic Natural Language Processing" International Journal on Information and Communication Technologies, Vol. 3, No. 3, June 2010.

[5] Sudip Naskar, Sivaji Bandyopadhyay, "Use of Machine Translation in India: Current Status".

[6] Mr. Sandeep Warhade and Mr.Prakash R.Devale "Design of phrase-based decoder for english-to-sanskrit translation" Journal of Global Research in Computer Science, Vol 3 (1), January 2012, 35-38.

[7] Rajpal Singh, Dr.Gurpreet Singh Josan, "An Approach to Example -Based Machine Translator using Translation Memory", ISSN 2250-2459, Volume 2, Issue 5, May 2012

[8] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu "BLEU: a Method for Automatic Evaluation of Machine Translation" Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 311-318.

[9] A Study towards Design of an English to Sanskrit Machine Translation System Sanskrit Computational Linguistics: First and Second International Symposia Rocquencourt, France, October 29-31, 2007 Providence, RI, USA, May 15-17, 2008.

[10] Shahnawaz, R. B. Mishra "*A Neural Network based Approach for English to Hindi Machine Translation*", International Journal of Computer Applications (0975 – 8887) Volume 53– No.18, September 2012.

[11] "Speech and Language Processing- an introduction to Natural Language Processing" by Daniel Jurafsky and James Martin, reprint 2000.

[12] "*Natural Language Processing*" by Aksar Bharati, Vineet Chaitanya, Rajeev Sangal. Prentice Hall of India, New Delhi, June 1994.

[13] "*A Higher Sanskrit Grammar*" by M. R. Kale, Delhi M. Banarassidas Publisher, 1961.