International Journal of Electrical and Computer Engineering (IJECE)
Vol. 6, No. 3, June 2016, pp. 1059 ~ 1071
ISSN: 2088-8708, DOI: 10.11591/ijece.v6i3.8904

Exploration of Corpus Augmentation Approach for English-Hindi Bidirectional Statistical Machine Translation System

K. Jaya, Deepa Gupta
[1]Department of Computer Science and Engineering, Amrita Vishwa Vidyapeetham Amrita School of Engineering, Bangalore Campus, India
[2]Department of Mathematics, Amrita Vishwa Vidyapeetham Amrita School of Engineering, Bangalore Campus, India

Article Info

Article history:

Received Aug 25, 2015
Revised Dec 4, 2015
Accepted Jan 2, 2016

Keyword:

Corpus augmentation
Indian language
Machine translation
Moses SMT
OOV
Statistical machine translation

ABSTRACT

Even though lot of Statistical Machine Translation (SMT) research work is happening for English-Hindi language pair, there is no effort done to standardize the dataset. Each of the research work uses different number of sentences, datasets and parameters during various phases of translation resulting in varied translation output. So comparing these models, understand the result of these models, to get insight into corpus behavior for these models, regenerating the result of these research work becomes tedious. This necessitates the need for standardization of dataset and to identify the common parameter for the development of model. The main contribution of this paper is to introduce an approach to standardize the dataset and to identify the best parameter which in combination gives best performance. It also investigates a novel corpus augmentation approach to improve the translation quality of English-Hindi bidirectional statistical machine translation system. This model works well for the scarce resource without incorporating the external parallel data corpus of the underlying language. This experiment is carried out using Open Source phrase-based toolkit Moses. Indian Languages Corpora Initiative (ILCI) Hindi-English tourism corpus is used. With limited dataset, considerable improvement is achieved using the corpus augmentation approach for the English-Hindi bidirectional SMT system.

Copyright © 2016 Institute of Advanced Engineering and Science.
All rights reserved.

Corresponding Author:

Dr. Deepa Gupta,
Department of Mathematics,
Amrita School of Engineering,
Bangalore, India.
Email: jayakarayil@gmail.com

1. INTRODUCTION

Exponential growth of internet and huge availability of information pose a new challenge to language technology. Generation of knowledge and need to observe the knowledge at the power at which it is dissipated requires one to be well versed with the language in which the knowledge is published. But mastering all languages is impossible. There comes Machine Translation (MT) technique to translate document in any language to document in any other language. Of all the machine translation technologies, Statistical Machine Translation (SMT) [1] considered as an important MT Technique. SMT can be developed independent of the underlying language and are based on bilingual sentence aligned parallel corpus. With increased availability of free, large language corpus, and high speed processor with huge memory SMT has become an important paradigm in machine translation.

India is a multilingual country with Hindi as official language. English being lingua-franca of science, media and technology, is a de-facto medium of educational materials created world over, the importance of English-Hindi Machine translation is obvious. Hindi is morphologically rich language and is

Subject-Object-Verb (SOV) word order language. Difference in word ordering between languages and the morphological rich nature of Hindi proves to be challenge in Statistical Machine Translation. English-Hindi(Eng-Hin) language pair is considered for this experiment and use ILCI Hindi-English tourism corpus (http://tdil-dc.in/).

Even though there are SMT research [2],[3] using ILCI Hindi-English tourism corpus, there is no set standard on splitting the parallel corpus i.e., splitting the parallel corpus into training, development and test dataset, for the translation task. Variations in number of sentences used in translation, varies the translation quality. When models are generated using various SMT techniques, comparison of these models, reproducibility of the result or understanding behavior of the corpus for the various SMT techniques is not possible. So there is a necessary to standardize the corpus for the SMT research. One of the aim of this work is to provide a method to generate dataset for English-Hindi (Eng-Hin) Bidirectional System. This methodology can be used to standardize the dataset and can be adapted for any language pair being considered for translation. This paper also focuses on exploring the various model parameters which gives the best translation quality for generating the baseline for Eng-Hin Bidirectional system using using ILCI Hindi-English tourism corpus. This work is first of its kind for Eng-Hin Bidirectional system. Further contribution of this paper discusses corpus parallel augmentation to improve the translation quality of the Eng-Hin Bidirectional system.

The rest of the paper is organized as follows. Section 2 discusses in brief the machine translation research carried out for Indian languages using various Machine Translation techniques. In Section 3 the statistical machine translation approach is covered in brief. Section 4 gives information about the corpus and experimental setup. Section 5 discusses about the proposed approach and Section 6 about the experiment and result.

2. MT IN INDIA

For a multilingual country like India, development of good machine translation system for the various local languages is necessary for people to communicate and share knowledge without actually mastering the individual language. Considering the importance of MT for India, the Government of India initiated TDIL(Technology Development for Indian Languages) with the intention of creating tool and techniques for machine translation. There are large number of active groups working on MT Some of the active players in MT research are CDAC, IIT Bombay, IISC Bangalore, IIIT Hyderabad, IIT Kanpur, Tamil University, Cochin University, Amrita University. Some of the projects funded by TDIL are – Angalabharthi [4], Anusaarka [5], Anubharathi [6] systems developed by IIT Kanpur, MaTra (http://www.cdacmumbai.in/matra) developed by CDAC, Mumbai, Mantra (http://www.cdac.in/html/aai/mantra.asp) developed by CDAC, Pune, Shiva and Shakthi are the projects jointly developed by IISC Bangalore and IIIT, Hyderabad.

Research work in MT for Indian languages involves various MT techniques like Rule Based, Empirical Based. Rule based machine translation system retrieves language knowledge from dictionary and grammar of the respective language to aid in translation. Rule Based MT (RBMT) is of three categories – Direct, Transfer, Interlingual RBMT. Direct Translation system [7] does a word by word translation using bilingual dictionary. Anusarak, Direct MT, translate between two closely related Indian languages using the principles of paninian grammar. An interactive English-Tamil MT [8] allows user to update the system by adding more words into the lexicon and rules into the rule-base. Interlingua based MT [9],[10] translates the source language to intermediate language and then to target language. Angalabharthi an Interlingua based approach analyses English sentences and creates an intermediate structure called PLIL(Pseudo Lingua for Indian Languages). An English-Hindi interlingua MT system using Universal Natural Language (UNL) as interlingua, which converts source sentence into UNL and from which the target sentence is generated. This system does part of speech disambiguation and some sense disambiguation for postposition markers and pronouns. Mantra (MAchiNe assisted TRAnslation tool) developed by CDAC uses Transfer Based Approach. An English to Kannada MT system [11] is developed at Resource centre for Indian Language Technology Solutions uses transfer based approach, funded by Govt of Karnataka, and is applied to the domain of government circulars.

Empirical Based Machine Translation(EPBMT) uses large amount of data in form of corpora. EPBMT is of two categories – Example Based(EBMT) and Statistical Based. MATREX [12], English to Hindi EBMT System [13], Anubharti, Shiva and Shakti MT System, use Example Based MT technique. Example Based MT translation is by analogy and works well with domain with limited words.

The disadvantage of the varied MT techniques discussed above is its inability to generate a language independent model. This results in Statistical Machine Translation technique gaining momentum in language technology, which provides near to human translation and is language independent. Figure 1 reflects the SMT technique having an edge over other MT techniques in the recent years. Figure 1 is generated based on the number of papers published for Indian languages utilizing various MT technologies. Use of SMT technique for Indian language gained momentum when Moses , open source toolkit for SMT became popular, from year 2005, which can be seen from the Figure 1. A survey of various machine translation methods is carried out in [14]. The major challenge in using SMT for Indian languages is the availability of large parallel corpus. Generating large parallel corpus is expensive and time consuming. So most of the research work concentrate on improving the translation quality using the scare resource. Research work in statistical machine translation for Indian languages were done incorporating preprocessing approaches – Morphology and Dependency Relation [15], incorporation reordering rules in the preprocessing stage [16],[17], POS tagging [18],[19], concept labeling [20], syntactic and morphological information [21]-[24] source side reordering [25],[26] in translating from one local language to other or from English to other local languages. One of the example of SMT system for Indian language is 'Google Translate'(translate.google.com/about/intl/en_ALL/) multilingual service provided by Google. It is based on Statistical Machine Translation MT technique. Google Translate translates source language to intermediate language and then to target language. It uses millions of document during translation.

One another MT technique is Hybrid Machine Translation (HMT), uses multiple MT techniques in translating the language. Sampaark (http://sampark.iiit.ac.in/) Anuvadaksh are a Hybrid MT System funded by TDIL. Anuvadaksh is a hybrid SMT which is an integration of four different MT technologies - Tree-Adjoining-Grammar (TAG) based MT, Statistical based Machine translation, Analyze and Generate rules (Anlagen) based MT, Example based MT. This system translates the text from English to six other Indian languages i.e. Hindi, Urdu, Oriya, Bangla, Marathi, Tamil.

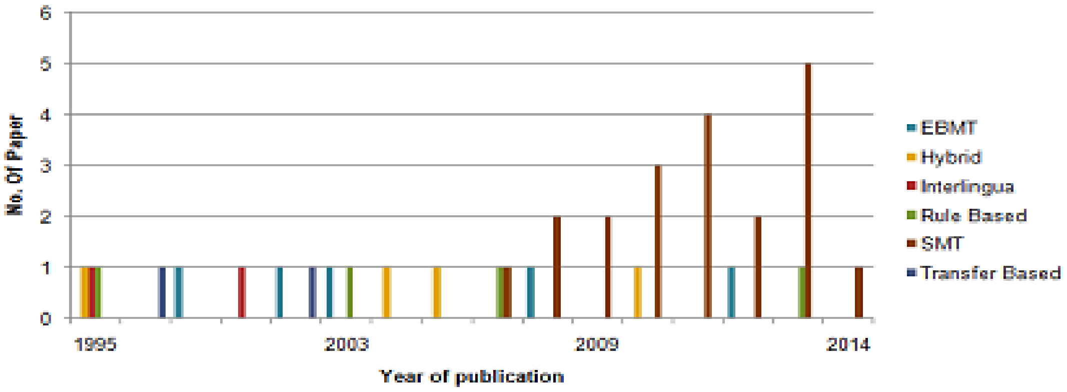

Figure 1. MT Research Trend for Indian Language

The system handles language divergence in a better way. Importance of HMT is with SMT adding more value to the translation with other MT techniques providing support to enhance the translation output. Inspite of SMT popularity, there is little or no effort in terms of standardizing dataset to be used in various stage in generating the translation model or method to create the standardized dataset. Because of this limitation in SMT research, comparison of various SMT model on a common baseline or reproducing the result of these experiments or understanding the corpus behavior when various techniques are used, is not studied. This calls for the need to standardize the data and use this standardize data to create translation model which helps the SMT researchers to be in sync with model behavior and to understand the translation quality better. In this paper, a method is described, which is language independent, can be used to split the corpus as test, train and development dataset. This dataset is used during the various phases of translation In addition to standardizing the dataset, the best parameter to be used for generating the baseline for bidirectional Eng-Hin SMT system is also identified.

Using the standardized dataset, parallel corpus augmentation – preprocessing approach is used to improvise the baseline translation output. This augmentation helps to improve the word alignment and reduce the OOV in the test set resulting in better translation output. In this experiment Moses [27] phrase-based open source toolkit, a complete SMT system is used. The other SMT toolkit available is MARIE (http://www.talp.upc.edu/index.php/technology/tools/machine-translation-tools/75-marie) developed at TALP Research center of the Universitat Politècnica de Catalunya (UPC) by Joseph M. Crego in 2005. Phramer [28], Open Source Phrase-Based SMT, compatible with Pharaoh (2004) written in Java, Joshua

[29], a decoder developed as a research collaboration between Johns Hopkins University and University of Pennsylvania in 2009. Of all these Open Source tool, Moses is a complete SMT system which is widely used in SMT research and has wide development and support community. Before discussing the experimental setup and proposed approach, a brief discussion on basics of SMT approach is provided.

3. BASICS OF STATISTICAL MACHINE TRANSLATION APPROACH

Statistical Machine Translation is a corpus based machine translation system based on Noisy Channel model. When huge parallel corpus as input, it provides near to error-free translation. SMT uses bilingual sentence aligned model, which is dependent of language being translated. Generation of translation through SMT is inexpensive, requires no human intervention in the translation and also can generate language independent model. The goal of SMT is to generate target sentence from the source sentence using the parallel corpus. SMT has three components: Language Model which reflects the fluency of the target language, Translation Model - identifies the correspondence between words and phrases in source and target languages; and Decoder which identifies best target sentence for a given input sentence using the translation and language models.

Thus three components - a language model, a translation model, and a decoder form the core component in Statistical Machine Translation. Figure 2 show the SMT architecture. When an Source language sentence(S) is given as input to the decoder, the corresponding Target language sentence (T) is generated based on the equation.

4. CORPUS CONTOUR & EXPERIMENTAL SETUP

In this experiment, Hindi-English tourism corpus, collected under Indian Languages Corpora Initiative (ILCI) project initiated by the DeitY, Govt. of India, Jawaharlal Nehru University, New Delhi, is used. The corpus statistics is presented in the Table 1.

Table 1. Statistics of ILCI Hindi-English tourism corpus

Data Statistics	English	Hindi
No. Of Sentences	12194	12200
No. of words	35876	36786
Min Sentence length	9	11
Max Sentence Length	79	94
Avg Sentence Length	50	55

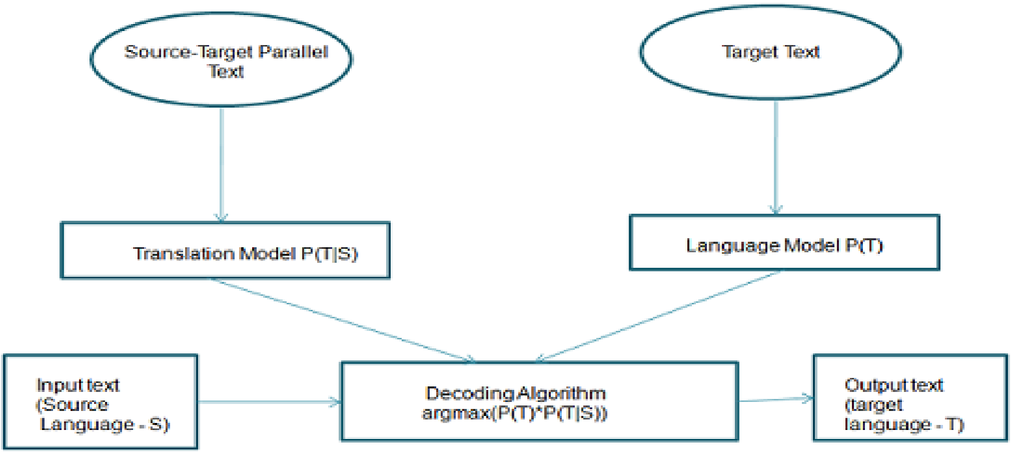
Source-Target Parallel Text
Target Text
Translation Model P(T|S)
Language Model P(T)
Input text (Source Language - S)
Decoding Algorithm argmax(P(T)*P(T|S))
Output text (target language - T)

Figure 2. High Level Design of SMT System

The toolkit used in this experiment - Moses toolkit requires Giza++ [30] open source implementation of the IBM models, for word alignment. Tuning is done by decoding and minimum error rate training (MERT) [31]. KenLM toolkit [32] is used for building language model. BLEU [33] is used for the automatic evaluation of the SMT System. Bleu score is widely used metric which is language independent, inexpensive and correlates highly with human evaluation. This metric gives the precision of n-grams with respect to the reference translation. The score is usually between 0 and 1. Score closer to 1 represents a good translation. The modified n-gram precision score, p_n for each n-gram length by summing over the matches for every hypothesis sentence S in the complete corpus C as:

$$p_n = \frac{\sum_{S \in C} \sum_{ngram \in S} count_{matched}(ngram)}{\sum_{S \in C} \sum_{ngram \in S} count(ngram)}$$

5. PROPOSED APPROACH

There are 6 stages in this experiment and Figure 3 represents the sequence of the approach. They are described in subsequent sections.

5.1. Preprocessing

The parallel corpus even though is published by TDIL, need to be cleaned , to be relevant for the experiment. The first part of the experiment is cleaning the corpus presented in Table 1. In this stage, blank lines, sentences that have no equivalent/valid translation are removed from the parallel corpus. After this step, the basic preprocessing like, case conversion, tokenization is done. Table 2 list the statistics of the cleaned parallel corpus. After corpus clean, parallel corpus is ready to be used in generating translation model.

Table 2. Statistics of parallel corpus after cleaning

Data Statistics	English	Hindi
No. Of Sentences	11700	11700
No. of words	33720	35930

5.2. Generation of Standardized Dataset

After preprocessing stage, the corpus is split as training, test, development data set to be used for the various phases of SMT. The training set is used for generating the translation model, development set is unique and is neither part of training or testing, is needed for SMT model parameter combination and test set is for testing the build model. The translation quality varies with difference in split of the data. One of the main contribution of this paper is to define a method that can be used as a standard method to generate data set. This method uses the Out-Of-Vocabulary (OOV) criteria. Out-Of-Vocabulary is number of unknown words seen in the test set that is not in train set. There are two types of dataset created - 0% Out-Of-Vocabulary (OOV) and Least Out-Of-Vocabulary(LOOV) . For 0% OOV dataset, the test and dev datasets have sentences which are there in the training dataset. There are no new words in 0% OOV test dataset i.e., test set is subset of training set. Translation accuracy measure on 0% OOV (0OOV) dataset provides an insight into the corpus ability to translate the document. In case of LOOV dataset, the test and dev dataset will have some words which are not see in training dataset. Dataset with least number of OOV is considered for LOOV dataset.

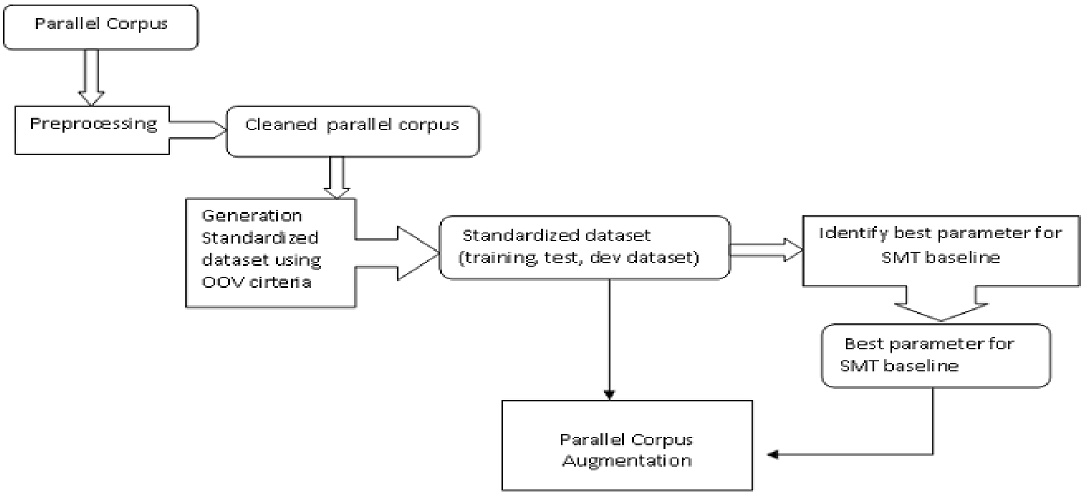
Parallel Corpus
Preprocessing
Cleaned parallel corpus
Generation Standardized dataset using OOV cirteria
Standardized dataset (training, test, dev dataset)
Identify best parameter for SMT baseline
Best parameter for SMT baseline
Parallel Corpus Augmentation

Figure 3. Schematic Diagram of Proposed Approach

For creation of dataset, 5% of parallel corpus is segregated as test set, 15% corpus as dev and the remaining 80% of the corpus as train set. Test and Dev set of LOOV data set have unique sentences with respect to train dataset. The LOOV dataset is constructed to have least OOV in the test set. This is achieved by generating dataset with various combinations of sentences in train,test and dev in such a way that the number of unknown words in test set is minimum.

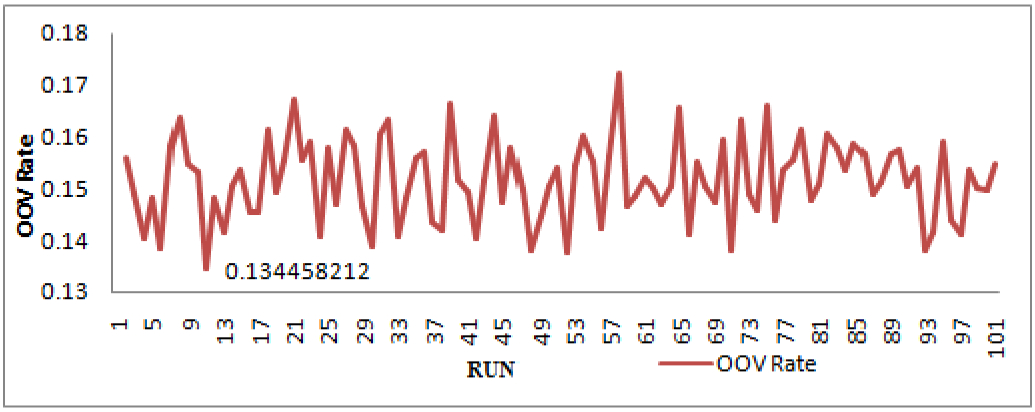

Figure 4. Generation of English Dataset with LOOV

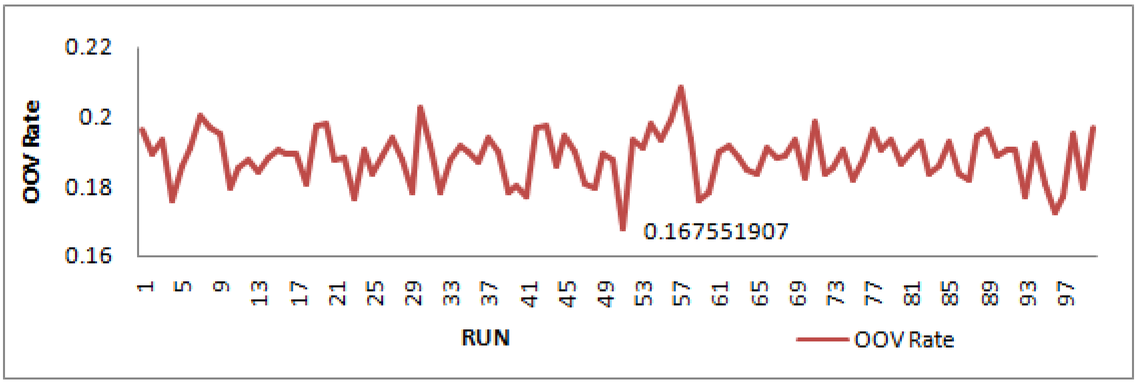

Figure 5. Generation of Hindi Dataset with LOOV

Table 3. Data statistics of Training, Dev and Test dataset

	LOOV Eng-to-Hin Datasets Statistics			
	English		Hindi	
	# Sentences	# words	# Sentences	# words
Train	10200	23967	10200	30208
Dev	500	3649	500	4200
Test	1000	6104	1000	7038
	LOOV Hin-to-Eng Datasets Statistics			
	Hindi		English	
	# Sentences	# words	# Sentences	# words
Train	10200	26163	10200	24926
Dev	500	3749	500	4142
Test	1000	6018	1000	6862

The Figure 4 shows the OOV rate for various English runs. A run with the minimum OOV in the test set is considered as Least OOV dataset for Eng-to-Hin bidirectional system and the corresponding dev and train set are also extracted. The Least OOV dataset is generated for both English and Hindi language. Figure 5 shows OOV rate for various Hindi test dataset which is used for the Eng-Hin SMT system. Table 3 gives data statistics of training, dev and test dataset.

5.3. Identifying Best Model Configuration For Eng-Hin Bidirectional System

Moses comprise of Language Model and Translation Model. These model supports various parameters which on tuning will find the baseline model with best translation quality. Language model has phrase length as one of its parameter. Parameters that translation model supports are translation phrase length, alignment and reordering. Translation phrase length specifies the number of words within a phrase. Word alignment is task of identifying translation relationships among words. There are various alignment heuristics ,supported by Moses. Some of them are intersection, grow-diagonal, union etc. In intersection, the Giza++ alignment is taken and for Union, the union of Giza++ alignment is considered. Reordering identifyies the meaning of the sentences. Various reordering supported by moses are distance, hierarchial, phrase-msd-bidirectional etc. Distance based reordering assigns a penalty to every reordering, and the penarty increases as the reordering distance increases..

One another objective of this experiment is to find a best model which gives a top performance for baseline model. The language pair features, features such as word order of the language, morphological richness of the language influences the model. For Eng-Hin bidirectional SMT system, there is no previous work where there is discussion on the best model for baseline system. Identifying the best model for Eng-Hin bidirectional SMT system is one of the task of this experiment.

Using the dataset 0OOV and LOOV described in the previous section the baseline is generated.In this experiment the language model(lm) phrase length is set to 3. Because of the limitation of the system resource, the lm length is chosen to be 3(default). Experiments with various translation phrase length are carried out. From this experiment it is found that setting translation phrase length to 7 gives top performance for Eng-Hin bidirectional system. Figure 6 represent the Bleu score of the model by varying phrase length. Increase in phrase length ex, phrase length=11 gives better score for the 0OOV dataset but it deteriorates for the LOOV dataset.

Using various alignments and reordering parameters, base model for both 0OOV and LOOV dataset are generated. Thus Figure 7 and Figure 8 shows the Bleu Score for 0OOV dataset and LOOV dataset respectively using various combination of parameters. The Table 4 gives the summary of the best parameter identified in this experiment. From Table 4, for baseline LOOV system dataset Grow-diag-final-and and msd-bidirectional-fe/distance as reordering and alignment parameter gives the best result. 0 OOV baseline smt system output specifies the maximum translation accuracy an SMT system can achieve.

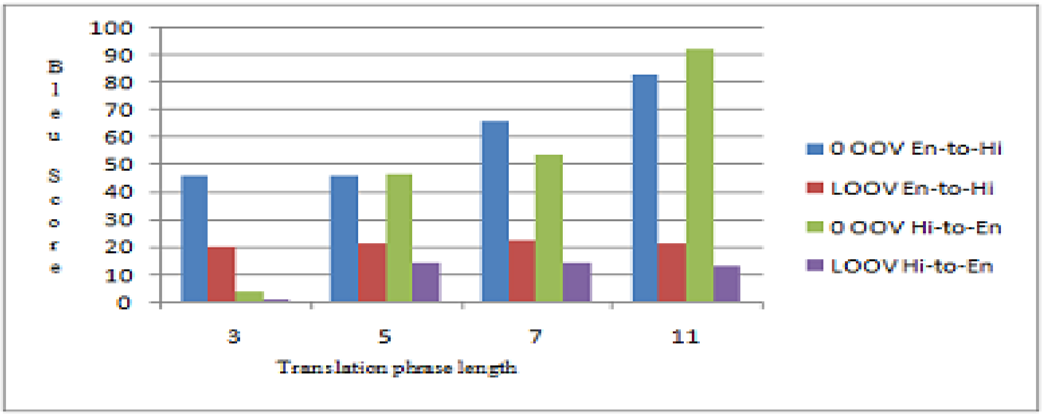

Figure 6. Baseline Bleu Score for various phrase length

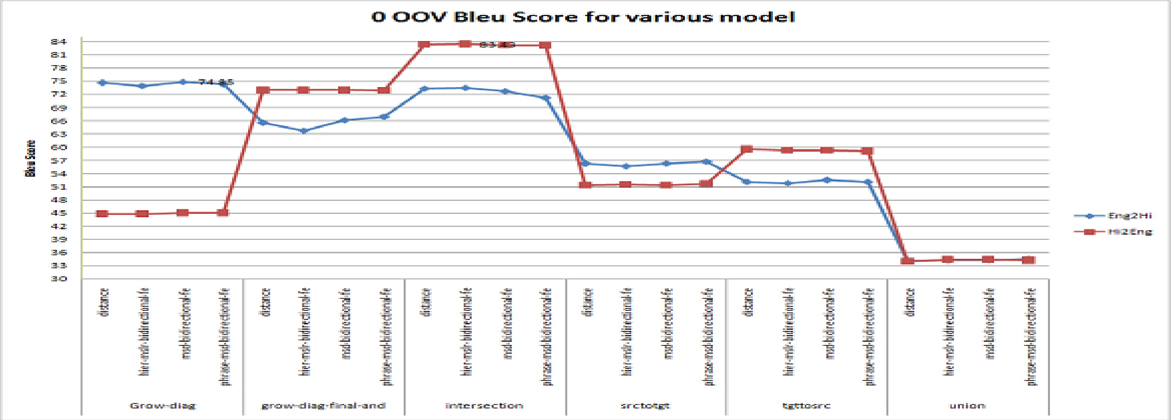

Figure 7. OOV Bleu Score for various Model using various parameters

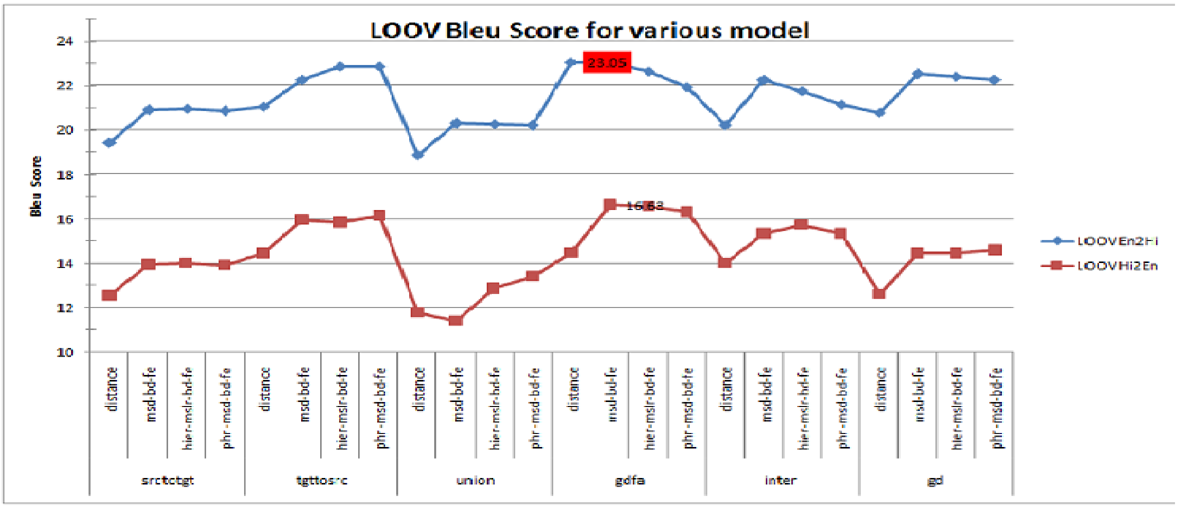

Figure 8. LOOV Bleu Score for various Model using various parameters

Table 4. Summary of the best parameter identified for generating baseline

Language pair	Alignment	Reordering	Bleu Score
Eng-to-Hin(0% OOV)	Grow-diag	msd-bidirectional-fe	74.85
Hin-to-Eng(0% OOV)	Intersection	Hier-mslr-bidirectional-fe	83.23
Eng-to-Hin(least OOV)	Grow-diag-final-and	msd-bidirectional-fe/distance	23.05
Hin-to-Eng(Least OOV)	Grow-diag-final-and	msd-bidirectional-fe	16.03

5.4. ENHANCEMENT TO BASELINE SYSTEM

Using the standardized dataset discussed in section 5.2 and best model specified in Table 3 for the Eng-Hin bidirectional system, corpus augmentation pre processing approach is used to improve the translation. In this stage, the parallel corpus is augmented with additional information extracted from original parallel corpus. For the scarce resource Eng-Hin bidirectional system, this preprocessing approach proves to be a viable solution to improve the translation quality. Corpus augmentation approach for Eng-to-Hin SMT system and Hin-to-Eng SMT system varies, because of difference in morphological richness of the language. To incorporate morphological variations, the corpus augmentation is done differently to get better performance.

For scarce resource like Hindi, corpus augmentation approach work well to improve translation. In this approach, the frequently relevant phrases are extracted from the parallel corpus and augmented by increasing the weight linearly and added to the original parallel corpus to get better alignment which in turn to get top performance. Using the corpus and without using the underlying language knowledge, corpus augmentation approach helps to improve the translation quality. Two types of corpus augmentation is done in this phase – Lexicalized and Lemmatized corpus for Hin-to-Eng Bidirectional SMT system and are explained in subsequent section.

5.4.1. Parallel Corpus Augmentation for Hin-to-Eng SMT System

For Hin-to-Eng SMT system, corpus augmentation is done by augmenting the Lexical and Lemmatized corpus to the original corpus. The resultant corpus is used for generating Hin-to-Eng SMT system. Algorithm to generate the Lexicalized and Lemmatized corpus is as follows.

1. Generate phrase table of phrase length 1.
2. Extract phrase from the phrase table. For each source phrase(h), there is a target translation phrase(e)
3. For each phrase (h),

- Fetch φ(h|e) and φ(e|h) probability, where φ(h|e) is inverse phrase translation probability and φ(e|h) is direct phrase translation probability
- Get phrases(h,e) ,

$$(h,e) = \begin{cases} (h,e)\ \text{where}\ (\text{argmax}(\varphi(h|e)) \cap \text{argmax}(\varphi(e|h))),\ \text{if}\ \text{count}(e) > 1 \\ (h,e)\ \text{if}\ \text{count}(e) = 1 \end{cases}$$

The phrases which are more relevant in the corpus are extracted from the corpus and augmented to the original corpus so as to get better word alignment. Phrase length one , alighment is set as intersection to construct phrase table from the clean parallel corpus . For example the phrase ‘अंग्रेजों’ has multiple translation in phrase table as shown in the example.

अंग्रेजों ||| british ||| 1 0.0130719 0.333333 0.512821

अंग्रेजों ||| built ||| 0.027027 0.0036101 0.166667 0.025641

अंग्रेजों ||| few ||| 0.0714286 0.009434 0.166667 0.025641

अंग्रेजों ||| jia ||| 0.14 0.333333 0.166667 0.025641

अंग्रेजों ||| legacies ||| 0.5 0.25 0.166667 0.025641

There are 4 different scores for each phrase. In this experiment only the first and third score - inverse phrase translation probability φ(h|e) and direct phrase translation probability φ(e|h) respectively are considered. More reliable words are extracted from the phrase table and these are ones with highest direct/inverse translation probabilities.

अंग्रेजों ||| british ||| 1 0.0130719 0.333333 0.512821

So the phrase is considered for the augmentation. The phrases ‘british’ is augmented to the English training corpus and ‘अंग्रेजों’ is added to the Hindi corpus and the weight of this corpus is linearly scaled.

Intention of generating lemmatized corpus is to reduce the number of unknown words. Multiple hindi words map to a English word. For example Hindi noun ‘बिल्ला’,‘बिल्ली’ are mapped to ‘cat’ in English and adjective ‘अच्छा’,’ आचे’ are mapped to ‘good’ in English.There are various inflection in Hindi – Noun, Adjective,Verb. Nouns in Hindi are inflected for gender, number and case. Adjectives need to agree with noun which are inflected for gender, number and case. Verbs are inflected for gender, number, case ,tense and voice. This inflection result is OOV i.e, words that are seen in test set are not seen during training. To handle this morphological variation, the parallel corpus is lemmatized. English lemmatization is done using python web mining module Pattern(http://www.clips.ua.ac.be/pattern) and Hindi lemmatization is done using the hindi shallow parser (http://ltrc.iiit.ac.in/analyzer/hindi/). After lemmatization the more reliable words are extracted from lemmatized corpus using the same procedure used for lexical corpus extraction.

The lexical corpus and lemmatized corpus generated as described above are augmented to the original parallel corpus. The resultant augmented corpus is used to generate the Hin-to-Eng SMT system.

5.4.2. Parallel Corpus Augmentation for Eng-to-Hin SMT System

Corpus Augmentation for Eng-to-Hin SMT System is done differently when compared to Eng-to-Hin SMT System. Translation from English to Hindi is a challenge because of the difference in morphological nature of the language. Corpus augmentation for Eng-to-Hin is done in similar manner as explained in section 5.4.1. In addition to that, morphological variations for Hindi noun and adjective are added to the parallel corpus. For example, 'bad' has translation phrase 'बुरा' in phrase table. Various morphological form 'बुरे', बुरी, of 'बुरा' are generated and augmented to the parallel corpus. Similarly noun morphological variations are added to the augmented list. These variations are augmented to the paralle corpus for the corresponding English phrase. The resultant transformed corpus is used to generate the Eng-to-Hin SMT system.

6. EXPERIMENTAL RESULT ANALYSIS

The proposed parallel corpus augmentation system is compared against one of the best performing open source SMT tools, Mosses. Initially the parameter tuning for system is done and the best parameters were found to be the following.

- language model phrase length = 3
- translation phrase length = 7

The alignment and reordering parameters identified and summarized in the Table 4 is used to build Eng-Hin Bidirectional SMT System using data set listed in Table 3. Table 5 list the result of the corpus augmentation preprocessing approach for the Eng-Hin bidirectional SMT system.

Table 5. Bleu Scores

	Baseline		Augmentation	
	Dev	Test	Dev	Test
Eng-to-Hin system	22.43	23.05	26.52	25.12
Hin-to-Eng system	15.59	16.63	20.12	19.56

From the table it can be observed that the Bleu score has improved by approximately 2 points in case of Eng-to-Hin translation system. For Hin-to-Eng translation system, the score is improved by 2.93 points. The improved translation system identifies the correct word which improves the translation quality. Some of the OOV words not translated in Baseline SMT system output are translated because of the addition of parallel corpus augmentation. Further local reordering is also taken care in the proposed SMT.

An example of Hin-to-Eng parallel corpus augmented system is illustrated in Table 6. The reference translation gives the actually expected translation. From the Table 6, it can be observed that in the Baeline SMT system, some words are not properly translated, viz, 'गहनों' , 'ज़री', 'कशीदाकारी'. While these OOV words in Baseline SMT system output, viz, 'गहनों' , 'ज़री', 'कशीदाकारी' are translated to 'jewellery' , 'brocade' and 'handicraft' in Parallel Corpus Augmented System output. Further the phrase 'उत्पादों हस्तशिल्प' is translated to 'handicraft such a the work of embroidery which is notable translation improvement over the baseline system'

Table 6. Hin-to-Eng sample output

Test Sentence	आगरा संगमरमर पर जड़ाऊ काम, चर्मकार्य, जूते, तथा पीतल का काम, कालीनों, गहनों, ज़री तथा कशीदाकारी के काम जैसे हस्तशिल्प उत्पादों के लिए प्रसिद्ध है ।
Reference Translation	agra is famous for handicrafts, products such as inlay work on marble, leatherwork, footwear, and brass work, carpets, jewellery, zari and embroidery work.
Baseline System output	agra inlay works on marble , चर्मकार्य , shoes , and brass work , carpets , the गहनों , ज़री and कशीदाकारी work as उत्पादों हस्तशिल्प is famous for .
Parallel Corpus Augmented System output	agra inlay work on marble , चर्मकार्य , shoe , and brass work , carpet , jewellery , brocade and handicraft such a the work of embroidery be famous for the .

Table 7. Eng -to- Hin sample output

Test Sentence	travelers anywhere must always carry essential items like a basic first aid kit with medicines for general ailments like fevers, colds and coughs, cuts and scrapes as well as specific medication for allergies, insect and mosquito repellents etc all these are available in the city .
Reference Translation	कहीं भी यात्रीगणों को आवश्यक वस्तुएँ जैसे दवाइयाँ जैसे साधारण बीमारियों जैसे बुखार, ठंड, कफ, कट तथा खुरचों साथ ही साथ एलर्जी, कीट एवं मच्छर निरोधकों आदि के लिए विशिष्ट औषधियों के साथ मूलभूत प्राथमिक चिकित्सा बॉक्स हमेशा लेना ही चाहिए।
Baseline System output	यात्रियों के लिए हमेशा अवश्य कहीं भी ले जाने के लिए आवश्यक वस्‌तुएँ जैसे पेट की एक बुनियादी दवाओं जनरल त्वचा समस्या भी ठीक करती जैसे fevers , colds और coughs scrapes है , और साथ ही साथ विशिष्ट medication allergies के लिए , कीट और इन सभी उठाएंगे हालाँकि मच्छर repellents इत्यादि हैं शहर में उपलब्ध हैं |
Corpus Augmented System output	यात्रियों के लिए हमेशा अवश्य कहीं भी ले जाने के लिए आवश्यक वस्‌तुएँ जैसे पेट की एक बुनियादी दवाओं जनरल त्वचा समस्या भी ठीक करती जैसे बुखार , ठंड और कफ है , और साथ ही साथ विशिष्ट medication allergies के लिए , कीट और इन सभी उठाएंगे हालाँकि मच्छर निरोधकों इत्यादि हैं शहर में उपलब्ध हैं |

Similarly, a sample output for Eng-to-Hin SMT is shown in Table 7. Here again in baseline SMT some words, viz, ‘fevers’ , ‘colds’, ‘coughs, ‘scrapes’, ‘repellents’ are not translated to Hindi. While these OOV words ‘fevers’ , ‘colds’, ‘coughs, ‘scrapes’, ‘repellents’ in Baseline System output are translated into ‘बुखार ‘, ‘कफ’, ‘ठंड’, ‘निरोधकों’ in Corpus Augmented System output.

Thus this reflects that corpus augmentation helps to reduce the OOV words in the Baseline system output. Thus in both translations, from English to Hindi and vice versa, it is noted that the proposed system exhibits an improved translation performance compared to the baseline open source SMT.

7. CONCLUSION

In this paper, standardization of the corpus and a method to develop standardized dataset is discussed in. The best parameter to generate top performance baseline for bi-directional Eng-Hin SMT is also identified. A pre-processing approach – corpus augmentation, for scarce resource, implemented to improve the translation quality. For using ILCI Hindi-English tourism corpus, this baseline score can be used as benchmark and any translation improvement on this standardized parallel corpus will helps us understand translation model better. English follows a Subject-Verb-Object (SVO) word orderwhereas Hindi follows Subject-Object-Verb (SOV) word order. This word order difference is a challenge in the machine translation which when handled will improve the Bleu score of the translation.

The proposed SMT system presents an improved translation quality with 2 point gain in Eng-Hin system and a 2.93 point gain in Hin-Eng system. Reordering of language will be the future work for the English-Hindi Bidirectional system to improve the system further.

REFERENCES

[1] P. Koehn, *et al.*, “Statistical Phrase-based Translation,” *Human Language Technology Conference. Edmonton, Canada*, pp. 127- 133, 2003.

[2] A. Ramanathan, A, *et al.*, “Simple Syntactic and Morphological Processing Can Help English-Hindi SMT,” *IJCNLP,* 2008.

[3] S. Venkatapathy, *et al.*, “A Discriminative Approach for Dependency Based Statistical Machine Translation,” *SSST,* 2010.

[4] R. M. K. Sinha, *et al.*, “ANGLABHARTI: A Multi-lingual Machine Aided Translation Project on Translation from English to Hindi,” *IEEE International Conference on Systems, Man and Cybernetics*. Vancouver, Canada, pp. 1609-1614, 1995.

[5] Bharati A., *et al.*, “Anusaaraka: Machine Translation in Stages,” *Vivek: A Quarterly in Artificial Intelligence*, vol/issue: 10(3), pp. 22-25, 1997.

[6] Sinha R. M. K., “An Engineering Perspective of Machine Translation: AnglaBharti-II and AnuBharti-II Architectures,” *International Symposium on Machine Translation, NLP and TranslationSupport System (iSTRANS-2004),* 2004.

[7] Goyal V. and Lehal G. S., “Web Based Hindi to Punjabi Machine Translation System,” *Journal of Emerging Technologies in Web Intelligence*, vol. 2, pp. 148-151, 2010.

[8] V. Renganathan, “An interactive approach to development of English tamil machine translation system on the web,” *The international Tamil Internet 2002 Conference and Exhibition (TI2002)*, 2002.

[9] S. Dave, *et al.*, “Interlingua Based English Hindi Machine Translation and Language Divergence,” *Journal of Machine Translation*, pp. 17, 2002.

[10] K. Vijayanand, *et al.*, “Vaasaanubaada Automatic MachineTranslation Of Bilingual Bengali - Assamese News Texts,” *Language Engineering Conference, University of Hyderabad, India.*

[11] Kumar G. B. and Murthy K. N., “UCSG Shallow Parser,” *Proceedings of CICLing-2006, Lecture Notes in Computer Science*, vol. 3878, pp. 156-167, 2006.

[12] A. K. Srivastava, *et al.*, “The MATREX (Machine Translation using Example): The DCU Machine Translation System for ICON 2008,” *ICON-2008: 6th International Conference on Natural Language Processing, Macmillan Publishers, India,* 2008.

[13] R. A. Sinhal, *et al.*, “A Pure Example Based approach for English to Hindi Sentence MT Systems,” *I.J. Modern Education and Computer Science,* pp. 51-59, 2012.

[14] Peng L., “A Survey of Machine Translation Methods,” *TELKOMNIKA Indonesian Journal of Electrical Engineering*, vol/issue: 11(12), pp. 7125-7130, 2013.

[15] T. D. Singh and S. Bandyopadhyay, “Manipuri-English Bidirectional SMT Systems using Morphology and Dependency Relations,” *SSST-4, Fourth Workshop on Syntax and Structure in Statistical Translation. Beijing, COLING*, pp. 83-91, 2010.

[16] R. N. Patel, *et al.*, “Reordering rules for English-Hindi SMT,” *2nd Workshop on Hybrid Approaches to Translation, Bulgaria*, 2013.

[17] A. Ramanathan, *et al.*, “Clause-Based Reordering Constraints to Improve Statistical Machine Translation,” *5th International Joint Conference on Natural Language Processing. Chiang Mai, Thailand,* pp. 1351–1355, 2011.

[18] A. Dalal, *et al.*, “Building Feature Rich POS Tagger for Morphologically Rich Languages: Experience in Hindi,” *ICON,* 2007.

[19] J. Ameta, *et al.*, “Improving the quality of Gujarati-Hindi Machine Translation through part-of-speech tagging and stemmer-assisted transliteration,” *International Journal on Natural Language Computing*, vol/issue: 2(3), 2013.

[20] R. Harshawardhan, *et al.*, “Phrase based English - Tamil Translation System by Concept Labeling using Translation Memory,” *International Journal of Computer Applications*, vol/issue: 20(3), pp. 1-6, 2011.

[21] C. Rahul, *et al.*, “Rule-based Reordering and Morphological Processing For English-Malayalam SMT,” *International Conference on Advances in Computing, Control, and Telecommunication Technologies,* pp. 458-460.

[22] K. Visweswariah, *et al.*, “A word reordering model for improved machine translation,” *Conference on Empirical Methods in Natural Language Processing. Edinburgh, United Kingdom*, July 27-31, 2011.

[23] K. Toutanova, *et al.*, “Applying Morphology Generation Models to Machine Translation,” *ACL,* pp. 514—522, 2008.

[24] A. Ramanathan, *et al.*, “Case markers and morphology: addressing the crux of the fluency problem in English-Hindi SMT,” *Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint conference on Natural Language Processing of the AFNLP. Suntec, Singapore*, 2009.

[25] M. A. Kumar, *et al.*, “Improving the Performance of English-Tamil Statistical Machine Translation System using Source-Side Pre-Processing,” *Int. Conf. on Advances in Computer Science, AETACS*, 2013.

[26] N. Durrani, *et al*., “Integrating an unsupervised transliteration model into statistical machine translation,” *EACL*, pp. 148, 2014.

[27] P. Koehn, *et al.,* “Moses: Open Source Toolkit for Statistical Machine Translation,” *Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic*, June 2007.

[28] M. Olteanu, “Phramer - An open source statistical phrase-based translator, *Workshop on Statistical Machine Translation. New York City,* pp. 146–149, 2006.

[29] Z. Li, *et al.*, “Demonstration of Joshua: An Open Source Toolkit for Parsing-based Machine Translation,” *47th Annual Meeting of the Association for Computational Linguistics, Software Demonstrations,* pp. 25-28.

[30] F. J. Och and H. Ney, “A Systematic Comparison of Various Statistical Alignment Models,” *Computational Linguistics*, vol/issue: 29(1), pp. 19-51, 2003.

[31] F. J. Och, “Minimum error rate training in statistical machine translation,” *41st Annual Meeting of the ACL,* pp. 160–167, 2003.

[32] K. Heafield, “KenLM: Faster and smaller language model queries,” *6th Workshop on Statistical Machine Translation, Edinburgh, UK,* 2011.

[33] K. Papineni, *et al.*, “BLEU: a method for automatic evaluation of machine translation,” *40th Annual Meeting on Association for Computational Linguistics. Philadelphia, Pennsylvania,* July 07-12, 2002.

BIOGRAPHIES OF AUTHORS

K. Jaya, B.Tech. , received the B.Tech. degree in Computer Science and Engineering from Pondicherry Engineering College, Pondicherry, India in 1997. She is currently pursuing M.Tech. in Computer Science and Engineering from Amrita School of Engineering, Amirta Vishwa Vidyapeetam, Bangalore. Her research interest includes Machine Learning, Natural Language Processing, and Statistical Machine Translation.

Deepa Gupta received the bachelor degree in Mathematics from Delhi University (India) with honors in 1997, the Master in Mathematics from Indian Institute of Technology Delhi (IITD), India in 1999, and the Ph.D. in Natural Language Processing from Department of Mathematics and computer application from IIT Delhi in 2005. She worked as a Postdoc researcher at FBK-IRST (Center for Scientific and Technological Research), Trento, Italy from February 2005 to March 2007. She worked as an Assistant Professor at IIIT Bangalore, India from April 2007 to December 2007.. And from 2009 to July 2014, she is associated with Amrita Vishwa Vidyapeetam as Assistant Professor in the department of Mathematics. From 2014 she is working as an associate professor in the same institute. Her research interest include Natural Language Processing, Statistical Machine Translation, Data Mining in particular HealthCare and Robotics.