DESIDOC Journal of Library & Information Technology, Vol. 37, No. 5, September 2017, pp. 307-312, DOI : 10.14429/djlit.37.11420

Efficient Use of Resources for Statistical Machine Translation

Karunesh Kumar Arora[*] and Shyam S. Agrawal[#]

[*]*Speech and Natural Language Processing, Centre for Development of Advanced Computing, Noida - 201 307, India*
[#]*Kamrah Institute of Information Technology, Gurugram - 122 102, India*
[*]*E-mail: karunesharora@cdac.in*

ABSTRACT

Machine translation has great potential to expand the audience for ever increasing digital collections. Success of data driven machine translation systems is governed by the volume of parallel data on which these systems are being modelled. The languages which do not have such resources in huge quantity, the optimum utilisation of them can only be assured through their quality. Morphologically rich language like Hindi poses further challenge, due to having more number of orthographic inflections for a given word and presence of non-standard word spellings in the corpus. This increases the chances of getting more number of words which are unseen in the training corpus. In this paper, the objective is to reduce redundancy of available corpus and utilise the other resources as well, to make best use of resources. Reduction in number of words unseen to the translation model is achieved through text noise removal, spell normalisation and utilising English WordNet (EWN). The test case presented here is for English-Hindi language pair. The results achieved are promising and set example for other morphological rich languages to optimise the resources to improve the performance of the translation system.

Keywords: Statistical machine translation; Normalisation; WordNet

1. INTRODUCTION

Machine translation technology can play a vital role in any domain where multi-lingual content is used. Library and information science also deal with multi-lingual contents. For accessing multi-lingual contents query translation is the commonly used. The data driven machine learning methods have over-shadowed the traditional approaches of rule based systems. As the name suggests these methods are based on the knowledge mined from the available resources. Statistical machine translation models make use of parallel set of sentences which are translations of each other. Here, the probabilities of translations of different word or phrase pairs are learned from their appearances in the parallel corpus. Bigger the size of this parallel corpus better the chances of learning probabilities reliably. But it goes easy in saying, as practically it has been seen that such large sized parallel resources cease to exist. English-Hindi pair is no exception to it. The building of such resource is time consuming and expensive task.

The phrase pairs which are considered translations of each other are learnt from the corpus. Analysing the corpus, it is observed that various types of noises exist in their orthographic presence. These include different noises in corpus and non-standard use of punctuation symbols and spellings. For making the best use of resources, the orthographic appearance consistency needs to be maintained in training, development and test data. Different noises are filtered and spellings are mapped to single spelling through normalisation process to reduce redundancy.

Further to it, the same word present in lower-case or true case would be treated as two different words in statistical learning e.g. 'Light' and 'light'. Beyond that, the words may be unknown to the translation model as presence of each word or each inflectional form of the word can not be guaranteed. In the paper presented here we have also experimented with different combinations of punctuation symbols and casing. The best combination is decided based on the automatic evaluation measured through BLEU score.

This problem becomes severe when a morphologically rich language is involved in translation model. Morphological rich language encompasses various inflections of a word like 'लड़का, लड़के, लड़कों' and if one of the inflection is not present in the training corpus, it would not get translated and would appear as unknown.

As the training corpus is of limited size, the possibility of a number of untranslated words appearing in the translated sentence can not be overruled. To handle this problem, we used other resource English WordNet which is a lexical database for english language. The methodology is based on the assumption that if a word is not seen in the training corpus, there is possibility that its equivalent synonym form might have appeared in the training corpus. We replace the OOV word with this synonym word. Now this sentence is sent for translation. In the current experiment, this exercise is limited to OOV words of noun and adjective syntactic categories.

Received : 30 May 2017, Revised : 02 August 2017
Accepted : 03 August 2017, Online published : 23 October 2017

The experimental results show that translation performance depends on the quality of the corpus and it needs to be utilised judicially and optimally.

2. LITERATURE REVIEW

A number of research works have been reported for correcting the spelling errors, but pre-processing for machine translation in terms of removing noises and orthographic normalisation has not been reported much. Sproat[1] has also said that "text normalisation is not a problem that has received a great deal of attention, and approaches to it have been mostly ad hoc: to put the issue somewhat bluntly, text normalisation seems to be commonly viewed as a messy chore". Caseli[2] has experimented with casing and punctuation markers and have reported that these changes have significant impact on translation performance. But the experiments have not taken the spell normalisation, and use of other lexical resources like WordNet into consideration. Bojar[3] has looked into the data normalisation issues in phrase based machine translation but have not reported any experiment with punctuation and casing handling. Liu[4] has described some of the techniques for improving the corpus quality by filtering noise and selecting more informative sentences from the training corpus and the development corpus. Their work relates to selecting the best parallel sentence pairs out of a large parallel corpus, while in contrast our paper here talks about ensuring optimised use of small sized corpus. Various related work on pre-processing have shown that datasets require preprocessing based on the purpose it is to be used. Lane[4] used class-based translation and language models in speech-to-speech translation in travel domain and presented performance improvement by using a mechanism to handle out-of-vocabulary words. Singla[5] talks about using English WordNet for reducing data sparsity, but does not handle other preprocessing to reduce redundancy. Further, they are replacing each word with Synset ID and use factored SMT model, while we replace only OOV words with their respective synonym word in desired inflected form. The paper being presented here reports two types of preprocessing – First relates to the noise removal and ensuring consistency in corpus established through punctuation symbols, casing and spell normalisation. The second type of pre-processing is to make use of the other available lexical resources like English WordNet (EWN) to address the problem of Out Of Vocabulary (OOV) words.

3. STATISTICAL MACHINE TRANSLATION

The background papers on this subject[6,7] describe the statistical machine translation as, that if we are given a source language sentence $S = s_1^I = s_1 \ldots s_i \ldots s_I$, which is to be translated into a target language ('English') sentence $T = t_1^J = t_1 \ldots t_j \ldots t_J$.

Statistical machine translation is based on a noisy channel model. It considers T to be the target of a communication channel, and its translation S to be the source of the channel.

System may generate multiple translation sentences options and the problem of translation becomes identifying sentence T which fits as the best translation of the source sentence S. Hence the machine translation task becomes to recover the source from the target.

So, we need to maximise $P(T|S)$. According to the Bayes rule,

$$t^* = \arg\max_t P(t \mid s) = \arg\max_t \frac{P(s \mid t) * P(t)}{P(s)}$$

As, P(S) is constant,

$$t^* = \arg\max_t P(s \mid t) * P(t)$$

Here, $P(s|t)$ represents translation model and $P(t)$ represents language model. Translation model plays the role of ensuring translation faithfulness and language model to ensure the fluency of translated output.

Figure 1 shows the steps of a phrase based SMT system with pre-processing. The bi-lingual and mono-lingual data are fed to the pre-processing as discussed in the following sections and then translation models and language models are trained on the pre-processed data. These trained models are used by the decoder for translating a given source to target language sentence.

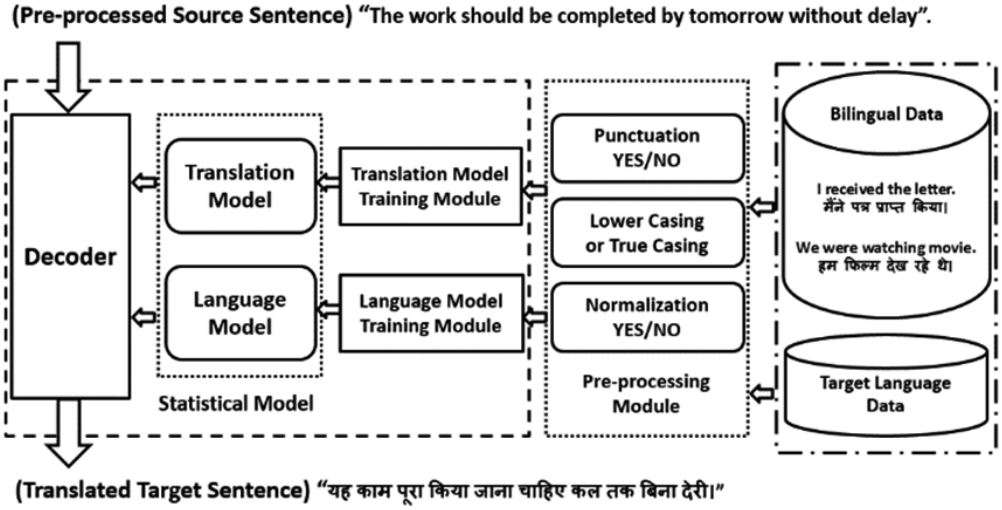

Figure1. Phrase based machine translation with pre-processing.

4. PRE-PROCESSING

The pre-processing described in the paper is limited to casing, punctuation symbols and spell normalisation. These are described in the following sub-sections.

4.1 Casing

Capitalisation is language specific orthographic convention. English uses capitalisation while Hindi does not have this feature. In English, capitalisation is used in the beginning of sentences, to indicate named entities or for proper nouns. This in turn may help to facilitate part-of-speech tagging and named entity recognition (NER). But

capitalisation may counter impact the performance of statistical machine translation, as the occurrence of words with and without capitalisation would be treated as two separate words and this in turn may reduce their number of counts in the corpus. This helps in reducing data sparsity problem. True casing is to keep the word in their natural case and changing the word at the beginning of the sentences to their most frequent form. Lower casing is meant by converting each word to lower cased form irrespective of their position or role in the sentence.

4.2 Punctuations

Punctuation is a mechanism by which one tries to express the emphasis and for clarity of expression. It helps the reader in terms of readability of an expression. Additionally, punctuation becomes very important in respect of conveying the intended meaning of an expression, as the placement of punctuation marks also helps in disambiguation of an expression.

4.3 Spell Normalisation of Hindi Corpus

Hindi language is the official language of India and is the third most spoken language of the world. Spell normalisation is a process by which text is transformed in some way to make it consistent in terms of usage of spellings for the given word throughout the text. The intention behind this activity is to reduce lexical redundancy. Different types of normalisations applied to the corpus for our experiment are described follows:

(a) Same word can be written into multiple ways orthographically e.g. सम्बंध / संबंध / सम्बन्ध / संबन्ध (*sambandh*, 'relation'). These forms are very productive in nature and almost are found in the text. In normalisation process, these are mapped to one single form with the help of rules. The rule for handling these, is that the 'fifth letters' (पंचमाक्षर) of the alphabet sequence for the given class of consonants and *Anuswar* (अनुस्वार) can be used interchangeably. If fifth letter of a class of consonants precedes any of the four remaining letters of the same class, the Anuswar can be used in place of that fifth letter; e.g. कंघा (*kangha*, 'comb'), झंडा (*jhanda*, 'flag'), कंधा (*kandha*, 'shoulder'), कंपन (*kampan*, 'vibration') etc. can be written in place of कङघा, झण्डा, कन्धा, कम्पन, respectively.

(b) Analysing the text it is observed that it is a customary practice to use nasalisation signs *Chandrabindu* (चंद्रबिंदु) or *Anuswar* (चंद्रबिंदु) for marking nasalisation in the word. However, in very few words, the use of nasalisation sign chandrabindu is sometimes necessary to avoid ambiguity in meaning and to mark distinction between words like हंस (*hans* , 'swan') and हँस (*haNs*, 'laugh') etc. But these types of words are very limited, so for the MT experiment purpose we have mapped these to one single form with *Anuswar*.

(c) Words of persio-arabic origin adapted in hindi vocabulary have to be written with *Nukta* (a dot below letter), but these appear with or without *Nukta* in corpus; e.g. ज़मीन / जमीन (*zameen/jameen*, ground/Earth). In the normalisation process all these are mapped to without Nukta form.

(d) Data encoded in unicode may have more than one way of storage for the same words e.g. hindi word पहाड़ (*pahaar*, 'mountain') if written with pre-composed character ड़, will have character storage as प ह ा ड़ , while with combining sequences it will have character sequence of प ह ा ड ◌़ . To normalize this we have mapped both representations to single one, the pre-composed character.

(e) Same words may be represented in more than one way depending on the presence of ZWJ / ZWNJ (zero width joiner / zero width non-joiner). For experiment purpose, in normalisation process these are mapped to single representation by removing the ZWJ / ZWNJ. e.g. भक्ति, भक्‌ति (*bhakti*, 'devotion').

Normalisation of the corpus was done on both English and Hindi corpora. Besides above, it also covered the following:

- Devanagari digits are converted to European digits (०, १, २, ३… to 0,1,2,3…)
- Punctuation symbol semi-colon ';' to comma ',' and sentence end-markers to dot '.' where it appeared as Devanagari danda (।). , as these were inconsistently used in corpus.
- Non-ASCII punctuations are replaced with their ASCII equivalents.

It is also observed that english words which are written in *devanagari* do not appear in a consistent form e.g. साइकल / साइकिल / साईकिल (bicycle). This is because of not having standard dictionary of writing these words. The spell variants which appear due to un-stable orthography are left un-handled in this experiment.

5. CORPUS STATISTICS

The experiments described in this paper were carried out using a corpus of total 43977 pairs of english-hindi (en-hi) parallel sentences with 601924 tokens in english and 615911 tokens in hindi. This corpus contains sentences from the tourism domain (ILCI corpus), grammar books, travel and tourism domain sentences from web and manually translated sentences. The Indian Languages Corpora Initiative (ILCI) project initiated by the MeitY, Govt. of India has a collection parallel corpora for various languages. For our experiment, we have included tourism domain english-hindi corpus of 25 K sentences from ILCI. The corpus contains sentences covering travel conversations and information about different visiting places, the monuments, temples, parks etc.

The corpus has been divided in 3 sets – training, development and test corpus. The size and distribution details are given in the Table 1.

6. PRE-PROCESSING EXPERIMENTS

For arriving at the most suitable settings in pre-processing of the corpus various combinations have been tried out. First, for handling punctuation symbols the two modes used are with retaining or removing the punctuation symbols in both (En-Hi) sides of corpus. Second, for dealing with cases, either lowercasing or true-casing have been tried out. In true-casing, the initial words in each sentence are converted to their most probable casing. This requires True-caser model, which is trained on statistics extracted from the training corpus itself. The model is used for english. The third setting was with or without spell normalisation of the corpus. This is achieved

Table 1. Corpus statistics

	#Sent	#Tokens (En)	#Tokens (Hi)
Total corpus	43977	601924	615911
Test corpus (5 %)	2195	29757	30265
Development corpus (10 %)	4394	60471	61954
Training corpus (75 %)	37388	511701	523 687

through an in-house developed hindi spell normaliser which handles spell normalisation cases as detailed in section 4.

Table 2 given below lists different pre-processing settings. As the last step of pre-processing, to clean-up the parallel corpus, duplicate sentences, empty lines and sentences having more than specified length were removed. Redundant space characters were also removed. For consistent handling of punctuation symbols, spaces have been inserted between words and punctuation symbols.

For the training of the statistical models, standard word alignment GIZA++[8] and language modeling toolkit KenLM[9] were used. For translation, MOSES[10] phrase based SMT decoder has been used. For evaluation, the automatic evaluation metric BLEU[11] scores are calculated to measure translation output. The main parameters of the Moses configuration were 5 iterations of IBM-1 and HMM and 3 iterations of IBM-3 and IBM-4 for GIZA++, the maximum phrase length was set to 7 and the option of reordering set as true. The parameters of phrase-based translation systems are tuned on development set using MERT[12].

Eight experiments were performed with the settings of the above described three features as listed in Tables 2 and 3. The settings of training, test and development corpus were kept similar. The language models (LMs) were built using target language corpus after preprocessing it similar to training corpus for the respective experiment.

The assumption taken was that for better performance of the SMT system the both training and test data should be in sync and should use the consistent forms of the words throughout. Not being so, the words seen by the training corpus may be unseen by the test corpus due to their presence in dissimilar form.

Table 2. Pre-processing settings for experiments

Feature	Value	Feature Description
Punctuation marks (PUNC)	Yes	Punctuation marks are retained for training purpose in parallel corpus.
	No	Punctuation marks are removed in parallel corpus.
True-casing (TCASE)	Yes	The initial words in each sentence are converted to their most probable casing in the source text English.
	No	All words in each sentence are converted to lowercase in the source text English.
Spell normalisation (SNORM)	Yes	Spelling variants are converted to the single form in Hindi text.
	No	Spelling variants are left as they appear in the text

7. REDUCING SPARSITY USING WORDNET

The small size of the training corpus increases chances of words appearing as OOVs. The shortage of parallel corpus is compensated here by utilising other available resources. The experiment presented here makes use of English WordNet, a lexical database of words.

The methodology adopted here is based on the assumption that synonym words mostly have similar translations. So, here we replace an OOV word of particular syntactic category with its most frequent synonym word present in the training model. The presence of synonym word is checked by searching it in the source side training vocabulary list. To maintain the inflectional similarity, the synonym word is converted to desired inflected form as of OOV word.

Table 3. Different Models with data description and BLEU scores

Model trained	PUNC	TCASE	SNORM	BLEU Score
M1	√	√	√	25.47
M2	√	√	×	24.85
M3	√	×	√	25.75
M4	√	×	×	24.99
M5	×	√	√	25.07
M6	×	√	×	24.29
M7	×	×	√	25.17
M8	×	×	×	24.44

The experiment presented here is limited to the OOV words of common noun, adjective and adverb categories. From the syntactic distribution of the OOV words, it is observed that these constitute of 42 per cent of the total OOV words. The proper nouns are open class words and are mostly transliterated. If we leave the proper nouns, we can see that the three categories handled in this paper constitute 76 per cent of total OOV words.

The best setting observed in the pre-processing with respect to casing, punctuation symbols and spell normalisation is used for performing experiment pertaining to use of synonyms for OOV words which are extracted from English WordNet. The Fig. 2 given below shows the flow of the methodology adopted in experiment.

For observing the effect of synonym replacement, we selected a subset of test set which contained at least one word which is OOV. There were 786 such sentences out of test set of 2195 sentences. After that, the set was further reduced which contained sentences wherein synonym of any OOV word (of syntactic category adjective, noun or adverb) is present in training corpus. This sentence set contained 73 sentences.

The BLEU score for this test set was found 11.26. After replacing OOVs with synonym words having presence in training corpus a higher BLEU score of 12.75 was noticed. This shows a gain of +1.49 BLEU points (i.e. 13 per cent) which is quite promising. We also got the translations compared manually and it also resulted in similar observation. Out of this test set of 73 sentences, 47 sentences showed better results while in 26 sentences the replaced words translation was not in context or the replaced word did not get translated.

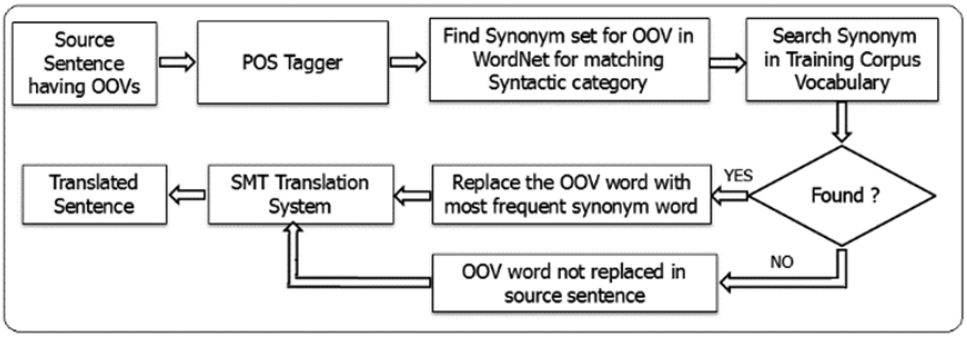

Figure 2. Flow diagram for using synonyms for reducing sparsity.

8. OBSERVATIONS AND DISCUSSION

Table 3 lists the values of BLEU scores[11] for pre-processing experiments. The BLEU scores shown in the table indicate that for English to Hindi translation best results are obtained with keeping the punctuation symbols intact, lower casing the source (English) side and with spell normalized text. The findings with english to hindi translation experiments, are in conformance to the observations of Caseli[2].

Results of english to hindi translation experiments show that the spell normalisation gives improvements in BLEU scores. This can be observed by comparing the BLEU scores of M1-M2, M3-M4, M5-M6 and M7-M8, as for these pairs, the other two features are kept constant. Through experiments with punctuation markers, the BLEU scores indicate that retaining punctuation markers is better than not having punctuation markers (M1-M5, M2-M6, M3-M6 and M4-M8). While with true-casing the reverse phenomenon is observed. The BLEU scores drops down when True-casing is applied. The BLEU scores are higher with training and test corpus in lowercase format.

Out of the three features, it can be observed that spell normalisation process has maximum impact on translation performance improvement. Out of vocabulary words not only lower the performance in terms of BLEU score but also affect the readability and make the translated text difficult to comprehend.

The Table 4 given in *Annexure* lists some of the example source language sentences, its translation, the sentence after replacement of OOV word with its synonym and its translated sentence. The OOV word and their synonym words are shown in italicised form.

9. CONCLUSIONS AND FUTURE WORK

The paper presents some experiments pertaining to pre-processing on training and test corpora and use of other available resources to reduce sparsity. It is observed that for English-Hindi translation best results are obtained with keeping the punctuation symbols intact, lower casing the source (English) side and with spell normalized text. Spell normalisation process influenced the translation to the maximum over the other two preprocessing – casing and punctuation symbols. Participation of punctuation symbols also helps in forming better phrases in the phrase-table and their presence impacted positively.

Similarly, use of other resources for reducing OOVs with their synonyms proved a positive step and it ensures to make maximum out of available resources.

The best combinations in pre-processing will be used as baseline cases in performing future experiments. Future work will include re-ordering source sentences to expected target side word order, and normalisation of named entities in the phrase based SMT.

ACKNOWLEDGEMENTS

We thank Mr Mukund Kumar Roy for helping in programming exercise and for performing manual evaluation of the test set.

REFERENCES

1. Sproat, R.; Black, A.W.; Chen, S.; Kumar, S.; Ostendrof, M. & Richards C. Normalization of non-standard words. *Computer Speech Language,* 2001*,* **15**(3), 287-333.
2. Caseli, H.M. & Nunes, I.A. Statistical machine translation: Little changes big impacts. *In* Proceedings of 7th Brazilian Symposium in Information and Human Language Technology, 2009, pp. 1-9.
3. Bojar, Ondřej; Straňák, Pavel & Zeman, Daniel. Data issues in english-to-hindi machine translation. *In* Proceedings of the Seventh International Language Resources and Evaluation (LREC'2010), 1771–1777, Valletta, Malta, May. ELRA, European Language Resources Association.
4. Lane, I.R. & Waibel, A. Class-based statistical machine translation for field maintainable speech-to-speech translation. *In* Proceedings of International Conference on Speech Communications and Technology, 2008, pp. 2362-2365.
5. Singla, K.; Sachdeva, K.; Yadav, D.; Bangalore, S. & Sharma, D.M. Reducing the impact of data sparsity in statistical machine translation. *In* Proceedings of Workshop on Syntax, Semantics and Structure in Statistical Translation, 2014, pp. 51-56.
6. Brown, P.F.; Pietra, V.J.; Pietra, S.A.D. & Mercer, R.L. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 1993, **19**, 263–311.
7. Och, F. J. & Ney, H. The alignment template approach to statistical machine translation. *Computational Linguistics,* 2004*,* **30**(4), 417-449.
8. Och, F.J. & Ney, H. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 2004, **29**(1), 19–51.
9. Heafield, K. & Ken, L.M. Faster and smaller language model queries. *In* Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation, Edinburgh, Scotland, United Kingdom, 2011, pp. 187–197.
10. Koehn, P.; Hoang, H.; Birch, A.; Callison-Burch, C.; Federico, M.; Bertoldi, N.; Cowan, B.; Shen, W.; Moran, C.; Zens, R.; Dyer, C.; Bojar, O.; Constantin, A. & Herbst,

E. Moses: Open source toolkit for statistical machine translation. *In* Proceedings of the 45th Annual Meeting of the Asso for Computational Linguistics (ACL 2007), Prague, Czech Republic. Association for Computational Linguistics, pp.177–180.

11. Papineni, K.; Roukos, S.; Ward, T. & Zhu, W.J. BLEU: a method for automatic evaluation of machine translation. *In* Proceedings *o*f the 40th Annual meeting of the Asso for Computational Linguistics (ACL 2002), pp. 311–318.
12. Och, F. Minimum Error Rate Training in Statistical Machine Translation. *In* Proceedings of Association of Computational Linguistics, 2003, pp. 160-167.

CONTRIBUTORS

Mr Karunesh Kumar Arora is presently working as Joint Director with Centre for Development of Advanced Computing (CDAC), Noida. He has almost 20 years of experience of working in the field of natural language processing. He has authored 25 research papers and contributed 4 chapters in a book. This paper presents the results and observation of experiments performed in statistical machine translation.

Dr Shyam Sunder Agrawal obtained his PhD from Aligarh Muslim University, India, in 1970. Currently working as Director General of KIIT Group of College, Gurugram. He is having research experience of about 45 years at CEERI, Pilani and subsequently as Emeritus Scientist of CSIR, Advisor to CDAC, Noida. He has published more than 250 research papers. The experiments presented in this paper have been guided by him.

Annexure I

Table 4. Some sample sentences

Improved cases:	
Original sentence:	You never read a lesson of *humbleness* in the school
Translated Sentence:	तुम कभी नहीं पढ़ी *humbleness* का पाठ स्कूल में
OOV replaced with its synonym:	You never read a lesson of *humility* in the school
Translated sentence with synonym:	तुम कभी नहीं पढ़ी नम्रता का पाठ स्कूल में
Original sentence:	This place in a special way, is a refuge place of the permanent and *migrant* birds
Translated Sentence:	यह स्थान विशेष रूप में है , के लिए शरण स्थली स्थायी और *migrant* पक्षी
OOV replaced with its synonym:	This place in a special way, is a refuge place of the permanent and *migratory* birds
Translated sentence with synonym:	यह स्थान विशेष रूप में है , शरण स्थली स्थायी और प्रवासी पक्षियों के लिए
Not improved cases:	
Original sentence:	There is a *boom* of amusement parks here
Translated Sentence:	एक एम्यूजमेंट पाक यहां का *boom*
OOV replaced with its synonym:	There is a *roar* of amusement parks here
Translated sentence with synonym:	एक एम्यूजमेंट पार्क यहां की गर्जना
Original sentence:	Taj Mahal is symbol of Shah Jahan's *everlasting* love for Mumtaz
Translated Sentence:	ताज महल प्रतीक शाहजहाँ के *everlasting* प्रेम मुमताज के लिए
OOV replaced with its synonym:	Taj Mahal is symbol of Shah Jahan's *ageless* love for Mumtaz
Translated sentence with synonym:	ताज महल प्रतीक शाहजहाँ के *ageless* प्रेम मुमताज के लिए