IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 3 Issue 4, April 2016.
www.ijiset.com
ISSN 2348 – 7968

Post Editing System for Statistical Machine Translation

Harmanjit Singh[1], Vishal Goyal[2] and Ajit Kumar[3]

[1] University Institute of Computing, Chandigarh University,
Gharuan, Punjab, India

[2] Department of Computer Science, Punjabi University,
Patiala, Punjab, India

[3] Department of Computer Science, Multani Mal Modi College,
Patiala, Punjab, India

Abstract

Hindi is the mother tongue language, official language of India and is 4th most widely spoken in the world. In the 2001 Indian survey, 258 million (258,000,000) people in India reported Hindi to be their native language; whereas Punjabi is the official language of Punjab state of India and the 11th most widely spoken in India. It is also the fourth most vocal language in England and Wales and third most spoken in Canada. Both the language having great impact on Indian officials, journals, articles. The problem starts when the natives of Punjab cannot understand the Hindi language. In order to make it possible very few translation software are available in market but quality of these software are not up to the mark. The idea of this paper is to describe a system that will improve the quality of translation from Hindi to Punjabi. In this paper, describe a post editing module for statistical based Hindi-Punjabi Translation system that will compare the result with previous machine translation system.

Keywords: *Statistical Machine Translation, Post Editing System, Rule Based Approach, Phrase Based Approach, Moses..*

1. Introduction

The present system involves Hindi as a source language and Punjabi as a target language. Both languages are very closely connected languages such as syntax and vocabulary similarities. The both languages belong to the same family but they have lot of differences between them.

Both the language are not mutually comprehensible. Mutual comprehensible of the languages depends on the elements like degree of phonetical, morphological, syntactical and lexical similarities. In written positions, Punjabi and Hindi are not mutually comprehensible but in spoken term, both are mutually comprehensible [1].

The script of Hindi linguistic is Devanagari and the script of Punjabi Linguistic is Gurmukhi[4].Most of the words written in Hindi Language pronounced same in Punjabi e.g. in Hindi the word "बिजली" pronounced same in Punjabi as "ਬਿਜਲੀ"

In spite of these connections maximum of Punjabi knowing individuals still cannot read or understand Hindi language and vice-versa. The dissimilarity is only in words and in pronunciation e.g. in Punjabi the word for boy is ਮੁੰਡਾ and in Hindi it is लड़का. The inflection forms of both these words in Punjabi and Hindi are also similar. There are examples where words are also same but pronunciation is different e.g. ਘਰ and घर [5].

So to make a common platform for both Hindi to Punjabi languages, machine translation system is required to be developed.

Machine Translation (MT) is a sub-field of Artificial Intelligence (AI), which converts the text from one language known as source language into the text of another language known as target language with the help of application software or algorithm without losing the meaning of sentence. The practice of statistical approach to machine translation begins with the use of parallel corpora. The IBM group at Yorktown Heights, New York had got the indication of using Statistical Approach to MT (SMT), based on their insight with speech recognition [6]. Since then SMT has arisen as a foremost research zone in the field of machine translation.

SMT treats machine translation as machine learning problem. This means learning algorithms are applied on previously converted text, known as parallel text, bi-text or parallel corpus. This paper will also explain the methodology adopted for evaluating the system and the outcomes found after evaluation.

2. Related Work

It has been widely acknowledged that when the aim of machine translation is to produce superior translation, then post-editing by human translators is necessary. The time period and effort required for post-editing can be reduced by implementing a number of strategies. For example, previously converted sentences can be leveraged from a translation memory first, thereby reducing the number of words that have to be machine translated and post-edited.

ISSN 2348 – 7968

The first reported results of automatic post-editing of machine translation outputs are where the authors successfully performed statistical post-editing (SPE) of rule-based machine translation outputs. [16] To perform the post editing, they used a phrase-based SMT system in a monolingual setting, trained on the outputs of the RBMT system as the source and the human provided reference translations as the target, to achieve massive translation quality enhancements. The authors also compared the performance of the post-edited rule-based system to right using the SMT system in a bilingual setting, and reported that the SMT system alone performed worse than the post-edited rule-based system. They then tried to post-edit the bilingual SMT system with another monolingual instance of the same SMT system, but concluded that no progress in quality was observed.

The first known positive results in SPE of SMT are reported by Oflazer and El-Kahlour[17] on English to Turkish machine translation. The authors followed a similar approach to Simard et al. [16], training an SMT system to postedit its own output. They use two iterations of post-editing to get an progress of 0.47 BLEU points [18]. The authors used a rather small training set and do not discuss the scalability of their approach.

To the best of our knowledge, the best results reported so far for SPE of SMT are by B´echara et al. [2] on French-to-English translation. The authors start by using a similar approach to Oflazer and El-Kahlout [17], getting a statistically significant improvement of 0.65 BLEU points. They then further improve the performance of their system by addition information from the source side into the post-editing system by concatenating several of the translated words with their source words, eventually reaching an improvement of 2.29 BLEU points. However, similarly to Oflazer and El-Kahlout [17], the training data used are very small, and it is not clear how their method scales on larger training data.

In a study by plitt et. al [12], they evaluated the productivity increase of statistical MT post-editing(Autodesk, a software company, whose products are translated from English into up to twenty languages) as compared to traditional conversion in a two-day test involving twelve participants translating from English to French, Italian, German, and Spanish. The machine translation system they selected for their productivity test was the open-source Moses system, trained solely on our own data without any factored representation.The principal aim of their productivity test was to measure the productivity increase they could expect in production at Autodesk. The test was scheduled to last two days. A small number of sentences, 1.6%, had a duration above five minutes and up to three hours, cumulating to a total 22% of the the period recorded, without there being any explanation such as the complexity of the source text. These sentences were eliminated from the result set.

Apertium: a free/open-source machine translation platform (engine, toolbox and data).

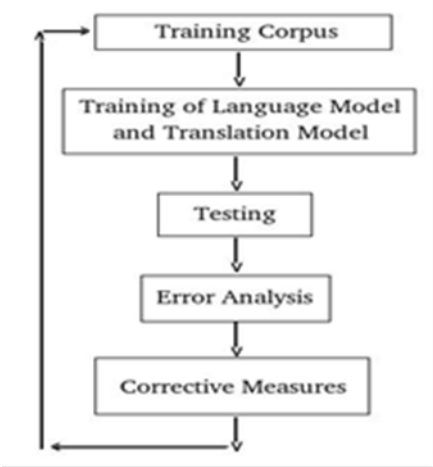
3. Progression Of Statistical Machine Translation

Statistical machine translation facilitates us to automatically build machine translation systems using statistical models capable by text data. The statistical models comprise of translation and language models. The translation model signifies the likely word translations and is trained by bilingual data, which consist of sentence sets in two different languages. The language model encrypts the sentence fluency and is trained by the target language data. The decoder examines for the most likely target word sequence from a large amount of hypotheses using these two models. SMT allows us to construct robust translation systems with low cost in short development cycles if the training data are available.

The Statistical Machine Translation System starts with the collection of training corpus. A training corpus is a large text material written in some language that will take as input to the SMT system. A large corpus of parallel text is processed to align paragraphs in sentence levels. Further, automatic training of MT system is carried out. The output of this phase is an operational part of MT system. Typically, this step is does not requires any human observation. In the next step, the output of machine translation that will not accurate one will go for under observation of error analysis. This phase is called as post editing, in which errors are requires to remove from corpus and depending upon the types of errors the corrective measures are taken and the process is repeated.

After re-input a large corpus in statistical machine translation system, a new result will be generated and will calculated it with earlier result by means of evaluation tools like BLEU.

ISSN 2348 – 7968

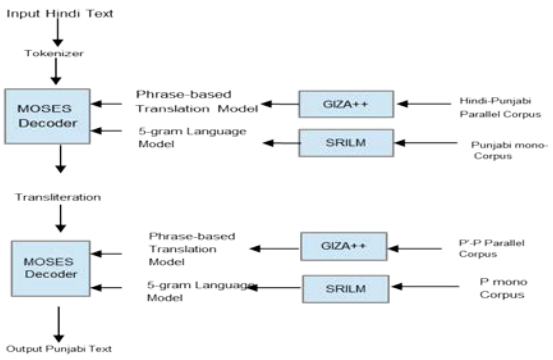


4. Steps In Statistical Machine Translation

Basis stages involved in statistical machine translation systems are: a.) Development, Cleaning and alignment of parallel corpus in the required language pair. B.) Find the probability of p(e) using a language model toolkit like SRILM,IRSTLM (c) Find the probability of p(f|e) using translation modeling toolkit like GIZA,GIZA++. (D) Use the decoder to convert source language sentence to target language using Language model and Translation Model. The resultant translation can be evaluated using evaluation tools like BLEU [7]

5. Methodology

In our research first stage PBSMT system is trained in the normal way using Hindi and Punjabi parallel corpus. This system provides us with the output T, which is the input data for our second-stage SPES. The second-stage PBSMT, which is being called as SPES, P' is manually edited to get correct Punjabi translation (P), and P'-P pair is used for training. To implement the above system a parallel corpus of translated output text from our PB-SMT system and manually corrected output has been created. The system training is done on this corpus using SRILM and GIZA++ to make language model and translation model. Most of the errors which are manually corrected in the corpus are eliminated in the final output after integrating this post-editing module with the existing system.



6. Data

In our experiments we focus on Hindi and Punjabi as these are the languages which are closely related. The present Hindi to Punjabi machine translation has been developed using hybrid approach based on word-for-word translation and rule-based approach [8].

The other Hindi to Punjabi machine translation system 'Sampark' is developed using hybrid approach and its translation quality is not better than the previous system [19]. The parallel corpus developed for the present work is improved with Hindi-Punjabi lexicon of around one lakh words and seventy thousand named entities.

7. Translation of Parallel Corpus

A parallel corpus is a corpus that contains a collection of original texts in language L1 and their conversions into a set of languages L2 ... Ln. In most cases, parallel corpora contain data from only two languages where the texts, passages, sentences, and words are typically linked to each other. [7].The translation of parallel corpus has been accomplished by statistical approach based machine translation system (Developed By: Punjabi University-Patiala) as shown below:

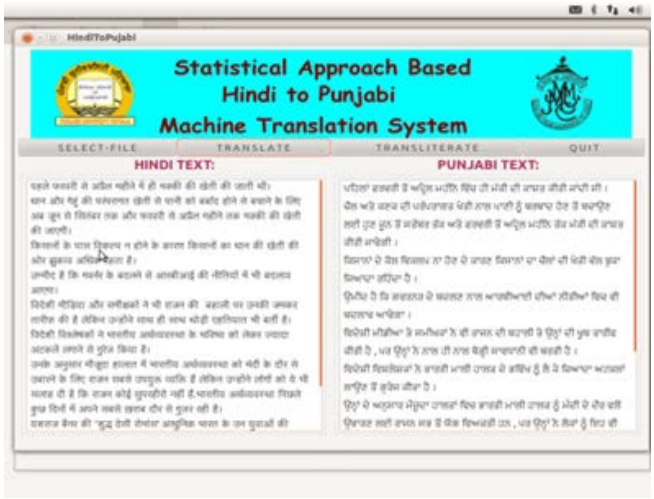


Figure: Translation of Hindi Text into Punjabi Text

As corpus is being available in large text paragraph and in order convert into target language, first we need to convert it into a sentence level. Alignments of parallel corpora at sentence level are prerequisite for many areas of linguistic research. In translation, sentences can be split, merged, deleted, inserted or changed in order. Fundamentally the shorter sentences are aligned with shorter sentences and longer sentences are aligned with longer sentences.[7]

ISSN 2348 – 7968

The existing Hindi to Punjabi machine translation has been developed using hybrid approach based on word-for-word translation and rule-based approach [7].It has been observed that the output from this system contain different kinds of errors like: Incorrect words, Grammatical Errors, Unknown words.

The parallel corpus developed for the present work is augmented with Hindi-Punjabi lexicon of around one lakh words and seventy thousand named entities.

8. Tools of SMT

8.1 SRILM

SRILM is a collection of C++ libraries, executable programs, and helper scripts aimed to allow both production of and experimentation with statistical language models for speech recognition and other applications.

SRILM is easily available for non-commercial purposes. The toolkit supports formation and evaluation of a variety of language model types based on N-gram statistics, as well as several related tasks, such as statistical tagging plus manipulation of N-best lists and word lattices.[9]

8.2 GIZA++

In Linux system environment, it is very common and suitable to use the word alignment generated from GIZA++ for most statistical machine translation (SMT) systems. [11] There are many applications for word alignment in natural language processing, and most of them depend on the quality of word alignment. While GIZA++ can be used on its own, it typically serves as the starting entry point for further machine translation systems, both phrase-based and syntactic. For instance, running GIZA++ is the first stage in training the popular phrase-based translation system Moses [20]

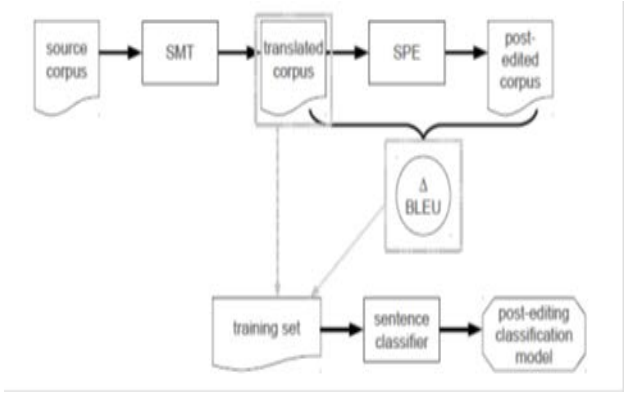
We are using GIZA++ tool to develop translation model from Hindi to Punjabi language. We found out word alignment and translation possibility for words give the impression in the corpus. From the word alignment we come to know that the word direction in Hindi-Punjabi language pair is nearly similar. There are some couples of words that are translated into single word of some word in Hindi are translated into a pair of Punjabi words. The same tool has been used to find phrase-alignment and phrase translation possibilities.

9. Statistical Post-Editing

The main problem occurs after Hindi to Punjabi translation is output that come from translation is not accurate. There are some kind of errors still remain in a corpus that will manually corrected. Translated text and manually edited text can be used for training a system. The trained system can be attached with the translation system, which we are using, as it reduces the manual editing efforts in further translations. This system can be developed by adding post editing module after translation in order to correct those errors.

The design of statistical post-editing system is similar to the architecture of statistical machine translation system with the difference that, in SPES, source language (SL) is the output of machine translation system and target language (TL) is the manually corrected machine translation.

The post-editing of a machine translation output contains generation of a text T" from a translation hypothesis T' of a source text S. When a PBMT system is built on bilingual parallel data, a phrase-based SPE system requires monolingual parallel texts. Current approaches on SPE are based on three-part parallel corpora composed of a source language text, its translation by an MT system and this output manually post-edited. If SPE can correct mistakes made by machine translation systems, it can also be used to acclimatize machine translation outputs to specific domains. [3]



In above figure, the source language part of the in-domain parallel corpus is first translated into the target language by an SMT system. Then, the generated translation hypotheses are affiliated with their translation references in order to form a monolingual parallel corpus and to form a SPE model. When a test corpus is translated and has to be post-edited, we propose approach that to training a post-editing SMT model on the training set decrypted by the first stage SMT model and iterate the approach, post-editing the output of the post-editing system.

ISSN 2348 – 7968

The output of post editing phase evaluated via BLEU score.

10. Evaluation and Result

The Hindi to Punjabi SMT system, accepting Hindi language sentences as input and give Punjabi sentences as training output. The translation of Sixty Five Thousands hindi sentences is done into Punjabi language.

Hindi to Punjabi machine translation system developed by Goyal and Lehal (2010) at Punjabi University, Patiala is available at http://h2p.learnpunjabi.org. This system takes hindi input data and convert it into Punjabi training data. The test data consisting of text take from "Gian Nidhi" corpus is prepared and translated using the above system. The translation quality evaluation is done using BLEU and NIST score.

Phrase-based Hindi-Punjabi machine translation system has been used for translating Hindi input text from "Gian Nidhi" corpus to Punjabi and evaluated manually and also on BLEU score. The Phrase-based post editing system has been applied to post edit the results obtained from the translation system and again evaluated. The improvement in the translation quality is evident from the results. The following example shows the type of errors corrected by using SPE on the output of PB-SMT system and their BLUE Score.

Source Sentence: महाराणा प्रतापसिंह ने गद्दी पर बैठते ही, मेवाड़ के उन अस्त-व्यस्त हिस्सों को, जिन्हें अकबर के हाथों में पड़ने से बचाए रखा गया था, फिर से संगठित करने का प्रयत्न किया ।

English Gloss Source Sentence: mahārāṇā pratāp nē gaddī utē bēṭhdiāṃ hī mēvāṛ dē unhāṃ asat-visat hissīāṃ nūṃ , jinhāṃ nūṃ akabar dē hatthāṃ vicc paiṇ tōṃ bacā liā giā sī , phēr tōṃ saṅgaṭhit karan dā jatan kītā.

Target Output: ਮਹਾਰਾਣਾ ਪ੍ਰਤਾਪ ਨੇ ਗੱਦੀ ਉਤੇ ਬੈਠਦਿਆਂ ਹੀ ਮੇਵਾੜ ਦੇ ਉਨਾਂ ਅਸਤ-ਵਿਅਸਤ ਹਿੱਸਿਆਂ ਨੂੰ , ਜਿਨਾਂ ਨੂੰ ਅਕਬਰ ਦੇ ਹੱਥਾਂ ਵਿੱਚ ਪੈਣ ਤੋਂ ਬਚਾ ਲਿਆ ਗਿਆ ਸੀ , ਫੇਰ ਤੋਂ ਸੰਗਠਿਤ ਕਰਨ ਦਾ ਯਤਨ ਕੀਤਾ ।

English Gloss Target Output: mahārāṇā pratāp nē gaddī utē baiṭhdiāṃ hī mēvāṛ dē unāṃ asat-visat hissiāṃ nūṃ , jināṃ nūṃ akabar dē hatthāṃ vicc paiṇ tōṃ bacā liā giā sī , phēr tōṃ saṅgaṭhit karan dā yatan kītā .

SMT Output: ਮਹਾਰਾਣਾ ਪ੍ਰਤਾਪ ਨੇ ਗੱਦੀ ਉਤੇ ਬੇਠਦਿਆਂ ਹੀ ਮੇਵਾੜ ਦੇ ਉਨ੍ਹਾਂ ਅਸਤ-ਵਿਅਸਤ ਹਿੱਸੀਆਂ ਨੂੰ , ਜਿਨ੍ਹਾਂ ਨੂੰ ਅਕਬਰ ਦੇ ਹੱਥਾਂ ਵਿੱਚ ਪੈਣ ਤੋਂ ਬਚਾ ਲਿਆ ਗਿਆ ਸੀ , ਫੇਰ ਤੋਂ ਸੰਗਠਿਤ ਕਰਨ ਦਾ ਜਤਨ ਕੀਤਾ ।

English Gloss SMT Output: mahārāṇā pratāp nē gaddī utē bēṭhdiāṃ hī mēvāṛ dē unhāṃ asat-visat hissīāṃ nūṃ , jinhāṃ nūṃ akabar dē hatthāṃ vicc paiṇ tōṃ bacā liā giā sī , phēr tōṃ saṅgaṭhit karan dā jatan kītā .

SPE Output: ਮਹਾਰਾਣਾ ਪ੍ਰਤਾਪ ਨੇ ਗੱਦੀ ਉਤੇ ਬੇਠਦਿਆਂ ਹੀ ਮੇਵਾੜ ਦੇ ਉਨ੍ਹਾਂ ਅਸਤ-ਵਿਅਸਤ ਹਿੱਸਿਆਂ ਨੂੰ , ਜਿਨ੍ਹਾਂ ਨੂੰ ਅਕਬਰ ਦੇ ਹੱਥਾਂ ਵਿੱਚ ਪੈਣ ਤੋਂ ਬਚਾ ਲਿਆ ਗਿਆ ਸੀ , ਫੇਰ ਤੋਂ ਸੰਗਠਿਤ ਕਰਨ ਦਾ ਯਤਨ ਕੀਤਾ ।

SPE English Gloss Output: mahārāṇā pratāp nē gaddī utē bēṭhdiāṃ hī mēvāṛ dē unhāṃ asat-visat hissiāṃ nūṃ , jinhāṃ nūṃ akabar dē hatthāṃ vicc paiṇ tōṃ bacā liā giā sī , phēr tōṃ saṅgaṭhit karan dā yatan kītā .

Here in this example two corrections are made by SPE i.e. ਯਤਨ and ਹਿੱਸਿਆਂ with improved BLEU Score 0.8249 which was 0.7846 earlier in SMT system. But one error is still present in the post-edited sentence. So the words which are translated differently in different context are potential source of error in the post-edited text and need to be handled separately. This is because; in one context the frequency of translation is higher as compared to other context. Such words retain their translation corresponding to high frequency context. In spite of this, it has been observed that most of the general grammatical errors are get corrected by the SPE system trained on manually corrected corpus, as is evident from the manual evaluation and improved BLEU score.

We tested a system on ten tests and calculate their BLEU score as follows:

11. Conclusion

It is evident that the results of SMT system are better than existing systems based on rule-based or hybrid approaches when applied to Hindi-Punjabi languages pair. An improvement of average 8.69 BLEU score is achieved by combining SPE with SMT system. Like SMT, the quality of post-edited text again depends upon the quality and size of parallel corpus used for the training of SPE system. Certainly, if SPE is attached with machine

ISSN 2348 – 7968

translation system it will reduce the efforts involved in manual editing.

References

[1]. Rani S., Luxmi V. (2013), "A Review on Machine Transliteration of related languages: Punjabi to Hindi", *IJSETR* Volume 2, Issue 3, March 2013
[2]. B´echara H., Ma Y., Gebabith J. (2011), "Statistical Post-Editing for a Statistical MT System", MT Summit XIII, pages 308–315.
[3]. Rubino R., Huet S., Lef`evre F., Linar`es G (2012), "Statistical Post-Editing of Machine Translation for Domain Adaptation", Proceedings of the 16th EAMT Conference, 28-30 May 2012, Trento, Italy.
[4]. Goyal, Vishal; and Lehal, Gurpreet Singh (2008), "Comparative Study of Hindi and Punjabi Language Script", *Napalese Linguistics, Journal of the Linguistics Society of Nepal*, Volume 23, pp. 67-82
[5]. Joshan G., Lehal G. (2007) , "Evaluation of direct machine translation system from Punjabi to Hindi", *Int J Systemics Cybern Inform*, pp. 76–83
[6]. Brown, Peter F.; Stephen, A.; Della Pietra; Vincent, J. Robert; and Mercer, L (1993), "The Mathematics of Statistical Machine Translation: Parameter Estimation. In Computational Linguistics", 19(2), pp. 263–311.
[7]. Kumar P., Goyal V. (2010), "Development of Hindi-Punjabi parallel corpus using existing Hindi-Punjabi machine translation system and using sentence alignments", *International Journal of Computer Applications* (0975 – 8887) Volume 5– No.9, August 2010.
[8]. Goyal V.,Lehal G.(2010), "Hindi To Punjabi Machine Translation System", Proceedings of the ACL-HLT 2011 System Demonstrations, pages 1–6,Portland, Oregon, USA, 21 June 2011.
[9]. Stolcke A. (2010), "SRILM- An extensible language modeling toolkit", Speech Technology and Research Laboratory", SRI International, Menlo Park, CA, U.S.A
[10]. Riley D., Gildea D.(2010), "Improving the Performance of GIZA++ Using Variational Bayes", NSF grants IIS-0546554 and IIS-0910611
[11]. Tian L., Wong F., Chao S.(2011), "Word Alignment Using GIZA++ on Windows", Proceedings of Thirteenth MT Summit, Xiamen, China
[12]. Plitt M., Masselot F. (2013), "A Productivity Test of Statistical Machine Translation Post-Editing in a Typical Localisation Context", The Prague Bulletin of Mathematical Linguistics, NUMBER 93, JANUARY 2010, pp. 7–16
[13]. Rosa R., Marecek D., Dusek O. (2012), "DEEPFIX: A system for automatic correction of Czech MT outputs", In Proc. of WMT, pages 362– 368. ACL.
[14]. Forcada M., Tyers F., S´anchez G. (2007), "The Apertium machine translation platform: Five years on", Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation, pp. 3-10,Alacant, Spain, November 2009.
[15]. Simard M., Goutte C., Isabelle P.(2013), "Statistical Phrase-based Post-editing", Proceedings of NAACL HLT 2007, pages 508–515,Rochester, NY, April 2007.
[16]. Simard, M.; Goutte, C.; and Isabelle, P. 2007a. Statistical Phrase-based Post-editing. In NAACL-HLT, pp. 508-515.
[17]. Oflazer, K.; and El-Kahlout I.D. 2007. Exploring Different Representational Units in English-to-Turkish Statistical Machine Translation. In WMT, pp. 25-32.
[18]. Kishore, Papineni; Salim, Roukos; Todd, Ward; and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, pp. 311-318.
[19]. Christopher, Mala; and Rao, Uma Maheshwar. 2010. "IL-ILMT SAMPARK: A Hybrid Machine Translation System". Proceedings of 32nd All India Conference of Linguistics (AICL-32), at Lucknow University, Lucknow, 21st to 23rd December, 2010, pp. 69-75.
[20]. Dugast, L.; Senellart, J; and Koehn, P. 2009. Statistical Post-editing and Dictionary Extraction: SYSTRAN/Edinburgh submissions for ACL-WMT2009. In WMT, pp. 110-114.

Harmanjit Singh is a M.Tech and computer faculty in University Institute of Computing in Chandigarh University-Gharuan, Punjab.

Vishal Goyal is a Ph.D. and computer faculty in Department of computer science, Punjabi University-Patiala, Punjab.

Amit Kumar is prior to completing Ph.D. and computer faculty in Department of computer science, Multani Mal Modi College-Patiala, Punjab