Statistical Post Editing System (SPES) Applied to Hindi-Punjabi PB-SMT System

Ajit Kumar and Vishal Goyal

1Multani Mal Modi College, Patiala -147 001, Punjab, India; ajit8671@gmail.com
2Department of Computer Science, Punjabi University, Patiala -147 002, Punjab, India; vishal.pup@gmail.com

Abstract

Post Editing System is an important module for improving the translation accuracy of any Machine Translation System. It is necessary to post-edit the output generated from machine translation system to correct the errors (word errors, syntactic errors, agreement errors, omission, addition and deletion errors, miscellaneous errors). Post Editing can be done extrinsically and intrinsically. Extrinsic Post Editing is done manually with the help of language experts, and is very costly, time consuming and laborious activity, whereas Intrinsic Post Editing refers to automatic post-editing which can be either rule-based or statistical approach-based. The rule-based post editing is language dependent, and statistical post-editing is language independent. The Statistical Post Editing System described in this paper is tested on phrase-based Hindi-Punjabi Statistical Machine Translation System and has been found to improve the translation accuracy from 1.5% to 12% on BLEU score.

Keywords: Extrinsic Post-Editing, Hindi to Punjabi Machine Translation System, Intrinsic Post-Editing, Statistical Machine Translation, Statistical Post-Editing

1. Introduction

The outputs generated from most of the machine translation systems are far from the perfect. To achieve high-quality output, the raw translation needs to be corrected or post-edited by human translators. One way to increase the productivity of the whole process is the development of Automatic Post Editing (APE) systems (Dugast et al. 2007; Simard et al. 2007).

Although existing machine translation systems can lead to impressive accuracy, thus, translated text require human post-editing for making it usable. However, extrinsic post-editing is a costly affair depending on the amount of corrections required in the translated output. Therefore, the intrinsic post-editing is an important task which can lead to high quality machine translation without requiring human intervention.

Our system follow the statistical post-editing design of Simard et al. (2007a), where the output of a first-stage system is used to train a mono-lingual second stage system, that has the potential to correct or otherwise improve on (i.e. post-edit) the output of the first-stage system. In contrast to Simard et al. (2007a), but like Oflarer and El-Kahlout (2007), our experiments use PB-SMT systems throughout both stages. The objective is to investigate in more detail whether and to what extent state-of-the-art PB-SMT technology can be used to post-edit itself, i.e. its own output.

Many of these works propose a combination of Rule-Based Machine Translation (RBMT) and Statistical Machine Translation (SMT) systems, in order to take advantage of the particular capabilities of each system (Chen and Chen, 1997). A possible combination is to automatically post-edit the output of an SMT system using a rule-based system or the other way round. In this work, we have applied statistical post editing to improve the output of SMT system and the resulting system has been tested on Hindi-Punjabi machine translation system.

The SPES that we are proposing, can be integrated with any machine translation system or it can be used as
independent system to improve the quality of translation by taking output of machine translation system as input. The text taken from the Hindi news papers for testing the system covers various domains. Through this work, we evaluated how translation quality can be improved with a post-editing step based on a phrase-based alignment approach. The results presented in this paper are based on the experiments carried out by applying SPES on the output of Hindi-Punjabi PB-SMT system.

More recently, Bechara et al. (2011) design a full PB-SMT pipeline that includes a translation step and a post-editing step. The authors report a significant improvement of 2 BLEU points for a French to English translation task, using a novel context aware approach. This method takes into account the source sentences during the post-editing process through a word-to-word alignment between the source words and the target words generated by the translation system. This latter work is, to the best of our knowledge, the first attempt to combine two PB-SMT systems, one for translating from the source to the target language, and another one for post-editing the first system's output.

The kind of pipeline we are using had already been suggested by previous authors (Isabelle et al., 2007; Oflazer and El-Kahlout, 2007). But in their work they are not targeting to improve the outputs of a general SMT system, in this study we are working to improve the translation quality not for a specific domain but the overall quality of translation. Another recent approach related to our work was presented in (Suzuki, 2011) to select sentences for post-editing.

### 3. System Architecture

The architecture of SPES is similar to the architecture of Statistical Machine Translation System with the difference that, in SPES, Source Language (SL) is the output of Hindi-Punjabi PB-SMT system and Target Language (TL) is the manually corrected translation. A parallel corpus of 90,000 sentences of SL and TL has been created to train the system.

In our experiments first stage PB-SMT system is trained in the usual way using Hindi (H) and Punjabi (P) languages parallel training corpus. This system provides us with the output P’ (SMT output in Punjabi), which is the input data for our second-stage SPES. For the second-stage PB-SMT, 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 SPES a parallel corpus of translated output text P’ from Hindi-Punjabi PB-SMT system and manually corrected output P has been used. The system training is done with this corpus using SRILM and GIZA++ to generate language model and translation model. It has been observed that when SPES is integrated with the existing PB-SMT system the translation time has increased but the accuracy has improved by approximately 12% on BLEU score. Most of the errors corrected manually in the corpus are eliminated in the final output by integrating this post-editing module with the existing system.

In our opinion, the training data required for SPES can be developed easily for training the system, particularly, for those machine translation systems which have been put to use in reality by different organizations like newspapers etc. When such organizations, translate their documents using machine translation system, they have to apply extrinsic post-editing to make these documents usable. So translated text and manually edited text is available with such organizations and can be used to train the SPES. The trained SPES can be attached with the translation system, which they are using, as it reduces the manual editing efforts in further translations.

Statistical approach-based Hindi to Punjabi machine translation system used in the evaluation of proposed SPES has been trained using tradition phrase-based translation model implemented in GIZA++ and 5-gram language model developed using SRILM. The supporting scripts for tokenization, text normalization, and transliteration have been written in Perl to act as preprocessing and post-processing modules with Moses decoder.

The output generated by the baseline Phrase-based Hindi-Punjabi machine translation system, when evaluated using automatic evaluation shows BLEU score from 0.1 to 0.3. Then transliteration module was used for transliterating out-of-vocabulary words and the results after transliteration shows improvement in BLEU score from 0.2 to 0.35, which is still very low score. Then Hindi-Punjabi parallel corpus was augmented with Hindi-Punjabi lexicon consisting of 1.2 lakh words and 70,000 named entities as described in Vogel et al. (2004). The Phrase-based SMT system again trained on this augmented corpus and a considerable improvement on the translation results has been observed.

4. Testing and Evaluation

The system has been tested on the text taken from various daily news papers like bhaskar.com, jagran.com and bbc.com/hindi. The selected text belongs to different domains namely agriculture, business, entertainment, health, politics, science, small stories, sports, and tourism.

Phrase-based Hindi-Punjabi machine translation system has been used to translate Hindi text to Punjabi. The quality of translated text is evaluated manually as well as automatically on BLEU score. The SPES has been applied to post edit the results obtained from the translation system and again evaluated extrinsically with the help of language experts. The improvement in the translation quality is evident from the results. The following example shows the type of errors corrected by using SPE from the output of Hindi-Punjabi PB-SMT system.

Source sentence: फांसी की सज़ा पर चुके व्यक्तियों की मनोदशा आखिर क्या होती है। वो भी जब आपकी सज़ा के लिए आप लंबे समय से इंतजार कर रहे हों।

Phânsi kî sazâ pâ cukê vyaktîyō mî manôdâśâ âkhir kî hûtî hai. Vô bhi jadô âapkî sazâ kî li' âpa lambê samaya sê intajâra kara rahê hôm.

Target output: फांसी की सज़ा पा चुके आपकी सज़ा के लिए आप लंबे समय से इंतजार कर रहे हों।

Phânsi kî sazâ pâ cukê âapkî sazâ kî li’ âpa lambê samaya sê intajâra kara rahê hôm.

Tourist output: फांसी की सज़ा पा चुके आपकी सज़ा के लिए आप लंबे समय से इंतजार कर रहे हों।

Phânsi kî sazâ pâ cukê âapkî sazâ kे li’ âpa lambê samaya sê intajâra kara rahê hôm.

Here, in this example two corrections are made by SPES i.e. तब जब (taba jada) - हुई तब (huîi tada) and रहें हो (rahê hôm) - रहे रहें (rahê rahe). But some errors are still present in the post-edited sentence. The reason behind this
might be that Hindi words आपकी (āpakī) and आप (āpa) are more likely translated as ਤੁਹਾਡੀ (tuhāmī) and ਤੁਸੀਂ (tusīm) in the corpus and otherwise also. So the words which are translated differently in different context are potential source of error in the post-edited text also, and need to be handled separately. This is because; in one context the frequency of translation is higher as compared to other context, and 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 SPES trained on manually corrected corpus, as is evident from the extrinsic and intrinsic evaluation.

4.1 Extrinsic Evaluation

The similar experiments have been performed by taking fifty sentences each from nine different domains namely agriculture, business, entertainment, health, politics, science, short-stories, sports, and tourism. The number of sentences containing errors is manually counted in the translated output. The output generated by the PB-SMT system is post edited using SPES and, again, the number of sentences containing errors in the post-edited text is determined. The counted sentences are those which contain either one or more errors of type wrong words, word agreement error, and syntax errors. Here, $N_{te}$ represents the number of sentences containing one or more errors out of the fifty translated sentences and $N_{pe}$ represents the number sentences containing errors out of fifty post-edited sentences. It has also been observed that the errors in the output are reduced but the number of completely correct sentences is increased by small number. The results obtained are summarized in the Table 1.

| Domain       | $N_{te}$ | $N_{pe}$ | %age improvement |
|--------------|----------|----------|------------------|
| Agriculture  | 26       | 19       | 14.0             |
| Business     | 20       | 16       | 8.0              |
| Entertainment| 22       | 17       | 10.0             |
| Health       | 24       | 18       | 12.0             |
| Politics     | 21       | 17       | 8.0              |
| Science      | 19       | 15       | 8.0              |
| Short-stories| 18       | 15       | 6.0              |
| Sports       | 19       | 16       | 6.0              |
| Tourism      | 20       | 15       | 10.0             |
| Average %age improvement |         |         | 9.1              |

The number of errors at word level are also counted in the output of PB-SMT system as well as after applying SPES. It has been observed that most of the agreement errors and wrong words are eliminated after applying SPES to the output of PB-SMT system, but some context sensitive translation errors are still present in the output of SPES. In this table $NW_t$ is the total number of words in the source document, $NW_{te}$ is the number of word-errors in the output of PB-SMT system, and $NW_{pe}$ is the number of word-errors in the output of SPES as manually counted. The results of word-level improvement are given in Table 2.

4.2 Intrinsic Evaluation

BLEU is one of the metrics to measure the quality of translation (Doddington, 2002; Kishore et al. 2002; Kumar and Goyal, 2011). The output generated from the phrase-based Hindi-Punjabi machine translation system has been post-edited using SPES. Both the outputs are evaluated automatically to obtain the BLEU score. The results are shown in Table 3.

| Domain       | PB-SMT | SPE   | %age Improvement |
|--------------|--------|-------|------------------|
| Agriculture  | 0.7992 | 0.8512| 6.50             |
| Business     | 0.9349 | 0.9567| 2.33             |
| Entertainment| 0.9079 | 0.9214| 1.49             |
| Health       | 0.8312 | 0.9352| 12.51            |
| Politics     | 0.9226 | 0.9533| 3.33             |
| Science      | 0.8564 | 0.9090| 6.14             |
| Short-stories| 0.8080 | 0.8892| 10.05            |
| Sports       | 0.8367 | 0.9293| 11.07            |
| Tourism      | 0.8291 | 0.9198| 10.94            |
| Average %age improvement |         |       | 7.11             |
5. Conclusion

The experimental results of Hindi to Punjabi PB-SMT system using Moses and post-edited with SPES are quite impressive and have shown a lot of improvement. Our automatic post-editing using statistical approach is technically language independent and can theoretically be integrated with any existing machine translation system, although we have not evaluated that yet. The quality of post-edited text, again, depends upon the quality and size of parallel corpus used for the training of SPES. Certainly, if SPES is attached with machine translation system it will reduce the efforts involved in manual editing. Further research is required to handle low frequency context dependent word translation.

6. References

1. Bechara H, Ma Y, van Genabith J. Statistical Post-editing for a Statistical MT System. In Mt Summit XIII; 2011. p. 308–15.
2. Bharti A, Sriram V, Vimshi K, Sangal R, Bendre SM. Algorithm for Aligning Sentences in Bilingual Corpora Using Lexical Information. Proceedings of the International Conference ICON-2002; 2002 Dec. p. 153–61.
3. Brown PF, Lai JCM, Mercer RL. Aligning Sentences in Parallel Corpora. Proceedings of the 29th annual meeting on Association for Computational Linguistics; 1991. p. 169–76.
4. Brown PF, Stephen A, Pietra D, Vincent JR, Mercer L. The Mathematics of Statistical Machine Translation: Parameter Estimation. In Computational Linguistics. 1993 Jun; 19(2):263–311.
5. Vilar D, Leusch G, Banchs RE. Human Evaluation of Machine Translation Through Binary System Comparisons. Proceedings of the Second Workshop on Statistical Machine Translation; 2007. p. 96–103.
6. de Ilarraza AD, Labaka G, Sarasola K. Statistical post-editing: a valuable method in domain adaptation of RBMT System for less-resourced languages. In MATMT; 2008. p. 35–40.
7. Doddington G. Automatic Evaluation of Machine Translation Quality using n-gram Co-occurrence Statistics. Proceedings of the Human Language Technology Conference (HLT). San Diego CA; 2002. p. 138–45.
8. Dugast L, Senellart J, Koehn P. Statistical Post-editing on SYSTRAN's Rule-based Translation System. In WMT; 2007 Jun. p. 220–3.
9. Dugast L, Senellart J, Koehn P. Statistical post-editing and dictionary extraction: SYSTRAN/Edinburgh submissions for ACL-WMT 2009. In WMT; 2009. p. 110–4.
10. Dwivedi SK, Sukhadeve PP. Machine translation system in Indian perspectives. Journal of Computer Science. 2010; 6(10):1111–6.
11. Gale WA, Church KW. A program for aligning sentences in bilingual corpora. Association for Computational Linguistics. 1993 Mar; 19(1):75–102.
12. Goyal V, Lelah GS. Comparative Study of Hindi and Punjabi Language Script. Napalese Linguistics. Journal of the Linguistics Society of Nepal. 2008; 23:67–82.
13. Isabelle P, Goutte C, Simard M. Domain adaptation of MT Systems through automatic Post-editing. In MT Summit XII; 2007 Sep. p. 255–61.
14. Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: A method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL). Philadelphia; 2002 Jul. p. 311–8.
15. Koehn P, Hoang H, Birch A, Other. Moses: Open Source Toolkit for Statistical Machine Translation. In proceeding of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions. 2007 Jun; p. 177–80.
16. Kumar A, Goyal V. Comparative analysis of tools available for developing statistical approach based machine translation system. Proceedings of International Conference on Information Systems for Indian Languages (ICISIL-2011). CCIS. 2011; 139:254–60.
17. Kumar A, Goyal V. Practical approach for developing Hindi-Punjabi parallel corpus. LREC 2012 Workshop on Indian Language and Data: Resources and Evaluation; 2012. p. 65–9.
18. Och FJ, Ney H. A Systematic Comparison of Various Statistical Alignment Models, Computational Linguistics. 2003 Mar; 29(1):19–51.
19. Ollazer K, El-Kahlout ID. Exploring different representational units in English-to-Turkish statistical machine translation. In WMT; 2007 Jun. p. 25–32.
20. Rao D. Machine translation in India: A brief survey. Proceedings of SCALLA (2001) Conference (SCALLA’01). National Centre for Software Technology. Bangalore India; 2001. p. 1–6.
21. Simard M, Goutte C, Isabelle P. Statistical Phrase-based Post-editing. In NAACL-HLT; 2007; p. 508–15.
22. Simard, M, Ueffing N, Isabelle P, Kuhn R. Rule-based Translation with Statistical Phrase-based Post-editing. In WMT; 2007 Jun. p. 203–6.
23. Stolck, A. SRILM - An Extensible Language Modeling Toolkit. In proceeding of the 7th International Conference on Spoken Language Processing (ICSLP 2002); 2002 Nov. p. 257–86.
24. Suzuk, H. Automatic Post-editing Based on SMT and its Selective Application by Sentence-level Automatic Quality Evaluation. In MT Summit XIII; 2011. p. 156–63.
25. Vogel S, Monson C. Augmenting manual dictionaries for statistical machine translation systems. Proceedings of LREC in 2003; 2004. p. 1593–6.