An Incremental Technique of Improving Translation

Aasim Ali
Department of Computer Science
Bahria University
Lahore, Pakistan

Arshad Hussain
Department of Electrical Engineering
University of Central Punjab
Lahore, Pakistan

***Abstract*—Statistical machine translation (SMT) refers to using probabilistic methods of learning translation process primarily from the parallel text. In SMT, the linguistic information such as morphology and syntax can be added to the parallel text for improved results. However, adding such linguistic matter is costly, in terms of time and expert effort. Here, we introduce a technique that can learn better shapes (morphological process) and more appropriate positioning (syntactic realization) of target words, without linguistic annotations. Our method improves result iteratively over multiple passes of translation. Our experiments showed better accuracy of translation, using a well-known scoring tool. There is no language specific step in this technique.**

Keywords—Statistical machine translation; incremental learning algorithm; English; Urdu

I. Introduction

Recent trend in machine translation is mostly towards data-driven methods including Statistical Machine Translation (SMT), which uses parallel text. This approach learns translation through phrase alignments [1] which are based on word alignments. In SMT, the morphological information improves learnability for realizing the correct shape of words, especially for morphologically rich languages like Arabic and Urdu. Similarly, the syntactic information improves positioning of words in the given context, especially when source and target pair has different positions for grammatical relations (Subject, Object, etc.) like English versus Urdu. An intuitive way of algorithmic evaluation of translation output is based on the number of matching sequences and subsequences of words in comparison with human translation. We have used BLEU [2] for an automatic evaluation of progress in translation improvement. A freely available toolkit for training and decoding of SMT systems, Moses [3] is used in our experiments, along with the supportive tools [4] for intermediate tasks like text alignment. Open source tools [5] are used for English (the source side of parallel text), and locally developed morphology analyzer [6] and POS tagger [7] of Urdu (target side of parallel text) are used for morpho-syntactic experiment. The experiments for baseline and proposed technique, both, use plain parallel text.

In the proposed method, the system gradually learns these linguistic elements (shapes and orders of words, etc.) from the surface forms of the target side, without any explicit knowledge, hint and tagging. There is no need of mono-lingual resources either, in addition to the parallel text. We have improved the shapes and arrangements of words on the target side by using the SMT process iteratively, to incrementally learn such information from simply the parallel text itself.

The rest of this paper has been organized in the following sections. Section II gives a review of the existing work on the statistical machine translation and the incremental learning. Section III details the methodology of the proposed algorithm. Section IV describes the data, experimental setup, and results. Section V discusses the proposed technique in the light of obtained results.

II. Literature Review

Statistical machine translation [8], being a machine learning approach towards translation [9], is used in the proposed work. A more detailed and updated record of statistical machine translation may be found in [8]. The proposed work considers linguistic knowledge (morphology, syntax, and word sense) to be "hidden" elements and uses the iterations of machine translation in the form of expectation maximization algorithm [10] without any external knowledge, to reach a better output. Words in our output are better in terms of correctness of shapes, sequences, and senses. The proposed work considers the intermediate translation of source as a pivot language [11], which is then used to improve the model to gradually reach the target language, by utilizing the power of incremental learning [12; 13; 14]. Gradual learning in several iterations reduces the impact of noise and irrelevant attributes [15] for automatically learning the word mappings to generate more correct sentences as output of translation.

The approach of incremental machine translation [16] uses the knowledge of human translator for enhancing the confidence of correct translations, and using that confidence for future translations. The proposed work uses the same idea of enhanced confidence with the help of an automatic tool, BLEU, for evaluation of translated output of one pass to be used as input for translation of next pass. Daybelge and Cicekli [17] have used a similar approach of using BLEU score as a measure of incremental learning and reported improvement in the translation quality using example based machine translation. Quality of translation does not depend only on the syntax and morphology but also on the sense of the source word [18; 19]. Using the translation of phrases observed previously increases the translation correctness when they occur subsequently [20; 21]. This is another view of "incremental" learning in which already observed high probability mappings help improving the mappings of other translation units in subsequent passes of learning. We have successfully experimented and introduced a technique that

This paper is out of my PhD work at NUCES (Lahore Campus) which is partially funded by HEC (Pakistan).

gradually learns the linguistic information from parallel text in several iterations of translation, which is detailed in the next section.

```
1   $x \leftarrow 0$
2   Diff $\leftarrow 0$
3   $i \leftarrow 1$
4   BScore$_0 \leftarrow 0$  // considering that two texts
5                      // (source and target) are disjoint
6   do
7   {
8       Model$_i \leftarrow$ SMT_Learner (TT$_{i-1}$ , TT$_n$)
9       THOT$_i \leftarrow$ SMT_Decoder (Model$_i$ , THOT$_{i-1}$)
10      BScore$_i \leftarrow$ BLEU_Score (THOT$_i$ , THOT$_n$)
11      Diff $\leftarrow$ BScore$_i$ – BScore$_{i-1}$
12      TT$_i \leftarrow$ SMT_Decoder (Model$_i$ , TT$_{i-1}$)
13  } while (Diff > $x$)
```

Fig. 1. Algorithm for Incremental Learning.

III. Algorithm for Incremental Learning

Fig. 1 shows the complete algorithm of incremental learning. The labels and variables used in the following algorithm are defined as: BScore$_i$ means BLEU Score of *i*th iteration; Diff means the difference of two consecutive BScore values to be compared with the Threshold (that is *x*); Model$_i$ is the SMT model learnt in *i*th iteration; when *i*=1 then TT$_{i-1}$ (TT$_0$) means Training Text which is source side, and TT$_i$ (TT$_1$) means translation of source side, same goes for all values of *i*; TT$_n$ means the target side for learning next SMT model; THOT$_0$ is the source side of held-out text, THOT$_i$ denotes the *i*th translated version of the source side of held-out text; and THOT$_n$ denotes target side of held-out text.

`Line 1` and **`2`** initialize the variables *x* and *Diff* to 0. **`Line 3`** initializes the iteration counter *i* to 1. **`Line 4`** initializes the BLEU score variable BScore$_0$ to 0, which means BScore for 0th iteration is 0. This variable will be used to compute the improvement in the translation for comparison with the BScore of i^{th} iteration for measuring the threshold. **`Line 6`** to **`13`** is a loop that will continue for the specific threshold. In this instance of the algorithm, the loop will stop when there is no improvement in the BScore, because the threshold testing variable x is kept 0.

Inside this loop, **`line 8`** updates the SMT Model for i^{th} iteration, between the TT$_{i-1}$ and TT$_n$. When $i = 1$, for first iteration, then TT$_0$ is the original training text on the source side of translation (English in our case, see section 4). When $i > 1$, for subsequent iterations, then TT$_1$, TT$_2$, and so on, are the i^{th} translated versions of source training text; thus termed as translated text. TT$_n$ is always the original training text on the target side of translation (Urdu in our case, see section 4). Hence, in the first iteration we obtain Model$_1$ which is trained SMT model for translation from English into Urdu. In **`line 9`**, the original held-out text on the source side (THOT$_0$) is translated using Model$_1$ thus generating the translated version of held-out text (THOT$_1$), when $i = 1$. When $i > 1$ then every THOT$_i$ is the translated version of THOT$_{i-1}$ using Model$_i$. The **`line 10`** computes the BLEU score between translated version and the target side of the held-out data. **`Line 11`** subtracts the BLEU score of previous iteration (BScore$_{i-1}$) from the BLEU score of current iteration (BScore$_i$). When $i = 1$ then BScore$_0$ is 0 and BScore$_1$ is the BLEU score of first iteration. Thus, TH holds the difference between BScore$_i$ and BScore$_{i-1}$ for every iteration. The processing of **`line 12`** produces the translation of source side of training text which may have to be used in the subsequent iteration if the loop continues to next iteration. TT$_i$ is the translated version of TT$_{i-1}$ using Model$_i$. When $i = 1$ then TT$_1$ is the translation of original source text TT$_0$. When $i > 1$ then every TT$_i$ is the translation of corresponding TT$_{i-1}$. If there is no gain in the BScore then the value of TH remains equal to or less than 0 thus the loop terminates. As the final output of this algorithm we obtain SMT Models from Model$_1$ to Model$_i$, from parallel training data. Our stopping criteria depends on the held-out data; and we use all these models in an incremental way to decode the evaluation data (test data). All these datasets are distinct for our experiment.

IV. Verifying Experiment

The baseline experiment is performed by learning simple phrase based machine translation (PBMT) [22] from plain parallel text. Next, we added POS tags and morphological annotations as factors in the factor based [23] PBMT, to see the improved result. Then we trained using our proposed model to achieve the best result. The data is described below in subsection A, and the experiments are detailed in subsection B, of this section.

A. Data

Text from two books is used in this study. The English and Urdu versions of these books are already aligned at topic level (containing one or more paragraphs). There are 497,354 words in 26,822 sentences on English side and 513,550 words in parallel Urdu translations.

We partitioned our data into three disjoint segments: 75% as training data, 19% as held-out data, and 6% as evaluation data. Plain bi-text is used for baseline and for incremental learning. The lemma, morphological tags and POS on source side (English) are computed using open source tools [5]. Similar tools for the target side (Urdu) of the parallel corpus are developed locally. The finite state transducer [6; 24] is used for morphology, and TNT tagger [7; 25] is used for POS tagging.

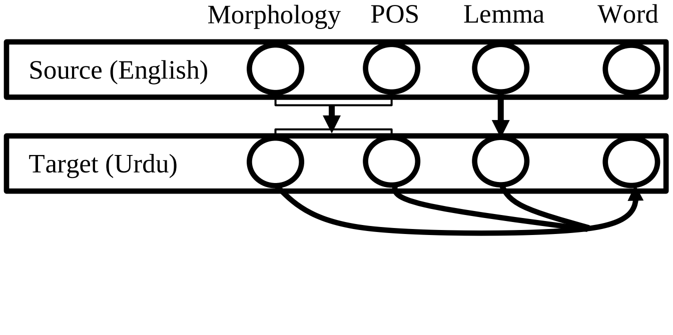

Fig. 2. Mapping of Factors.

Factored translation model is used for incorporating linguistic information at word level. Each such additional information attached to a word is termed as factor. These factors are used in a series of mapping stages or steps. The steps may be of two types: (a) translating the factors on input side to those on the output side, and (b) using the existing factors on output side to generate other factors on the same side for rendering the final shape of the word. Fig. 2 shows the mapping of factors.

The following mapping of input/output factors is used:

1) Lemma of input side is translated into the lemma on the output side.

2) POS and morphological factors on the input side are also translated into the factors on the output side.

3) Surface form on the output side is generated using the translated morpho-syntactic factors on the output side.

B. Experiment and Result

First of all, plain bi-text is used to obtain the baseline resuts. Then the same model is tuned for held-out data using minimum error rate training [26], which improved results from 32.10 to 37.10. Then the morpho-syntactic model of translation is used for which words are annotated with lemma and POS tag factors. This experiment produced the BLEU score of 36.73.

The proposed technique of incremental learning is designed to test if the un-annotated text can itself incrementally take the desired shapes and sequences of words induced by the implicit morpho-syntactic knowledge which is always present in the running text. The proposed algorithm is implemented in the following way:

1) Executed the training model of baseline, i.e. Source-to-Target Model (Model$_1$), on the training set (TT$_0$) itself (to prepare an intermediate train set TT$_1$). The translation of held-out data (THOT$_0$) from Model$_1$ is also saved and termed as THOT$_1$ to be used in the next stage.

2) Used that translated training part (TT$_1$) to pair with the target side (TT$_n$) of the corpora to learn a new model (Model$_2$) to automatically learn the good mappings which were missed in the first pass (while learning the Model$_1$).

3) Used Model$_2$ on THOT$_1$ to obtain the next version of translated held-out data (THOT$_2$), and found the improvement in the BLEU score by comparing between THOT$_2$ and THOT$_n$.

4) Executed Model$_2$ on the TT$_1$ (to prepare another intermediate train set TT$_2$) for next stage of learning.

5) Then used that latest translated training part (TT$_2$) to pair with the target side (TT$_n$) of the corpora to learn another model (Model$_3$) to further learn the good mappings which were missed even in the second pass.

6) Then executed Model$_3$ on THOT$_2$ to obtain THOT$_3$ and found the improvement in the BLEU score by comparing between THOT$_3$ and THOT$_n$.

7) Finally, for the sake of evaluation on a data set which is kept separately (apart from training and held-out data sets), executed Model$_1$ on original source side of the evaluation data (ET$_0$) to obtain ET$_1$. Then executed Model$_2$ on ET$_1$ to obtain ET$_2$. Afterwards executed Model$_3$ on ET$_2$ to obtain ET$_3$, which produced the highest BLEU score from ET$_3$ versus ET$_n$. The detail of this step is shown in Fig. 3.

In Fig. 3, each rectangle represents the process, each parallelogram signifies an input/output of the process, each solid arrow shows the sequence of flow, and each dashed arrow denotes the SMT model used in the process.

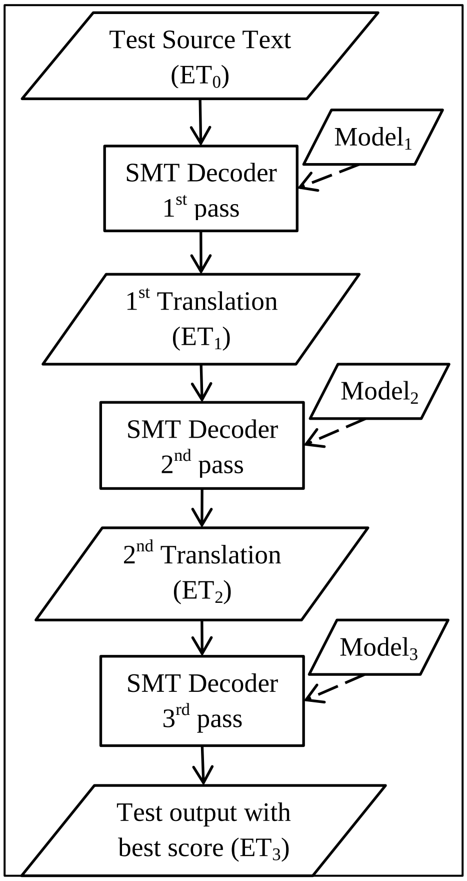

Fig. 3. Application of Three Models Learnt with Incremental Technique.

V. Discussion and Conclusion

The summary of results shown in Table I clearly shows that incremental learning proposed in this paper gives the highest BLEU score. One reason of unprecedentedly high score under proposed technique is the significant overlap of phrases in the data. However, it is also important to keep in mind that gain from this overlap could not be exploited without using the power [12; 13; 14; 15] of incremental learning.

Since this approach involves no language-specific steps therefore it may be applied to any language pair. The technique of exploiting the overlapping in the training set, the held-out set and the evaluation set, may work well for translation of any other text that typically has significant overlap of phrases including user manuals, blogs, specific news genre, and research articles from a specific field. It may also be applied for word sense disambiguation [27] using parallel corpus, instead of using explicit linguistic knowledge to resolve the word sense.

TABLE I. RESULTS OF TRANSLATION OF EVALUATION SET

Experiment	**BLEU**
Translation with Trained Model$_1$	32.10
Translation with Tuned Model$_1$	37.10
Translation with Morpho-Syntactic Model	30.73
Translation with Model$_{1..3}$ obtained from *Incremental Technique*	42.91

REFERENCES

[1] Koehn, P., Och, F. J., & Marcu, D. (2003, May). Statistical phrase-based translation. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1 (pp. 48-54). Association for Computational Linguistics.

[2] Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting on association for computational linguistics (pp. 311-318). Association for Computational Linguistics.

[3] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., ... & Dyer, C. (2007, June). Moses: Open source toolkit for statistical machine translation. In Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions (pp. 177-180). Association for Computational Linguistics.

[4] Och, F. J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. Computational linguistics, 29(1), 19-51.

[5] Carreras, X., Chao, I., Padró, L., & Padró, M. (2004, May). FreeLing: An Open-Source Suite of Language Analyzers. In LREC (pp. 239-242).

[6] Ali, Aasim. (2010). Study of Morphology of Urdu Language, for Its Computational Modeling: Study of Morphological Patterns in Urdu Language, and Partial Implementation of Computational Solution for the Same Using a Finite State Tool. VDM Publishing.

[7] Asif, T., Ali, A. and Malik, M. K. (2015). Developing a POS Tagged Resource of Urdu. Science International, 27(5), 4479-4483.

[8] Koehn, P. (2010). Statistical Machine Translation (1st ed.). Cambridge University Press, New York, NY, USA.

[9] Cardie, C., & Mooney, R. J. (1999). Guest editors' introduction: Machine learning and natural language. Machine Learning, 34(1), 5-9.

[10] Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological), 1-38.

[11] Wu, H., & Wang, H. (2007). Pivot language approach for phrase-based statistical machine translation. Machine Translation, 21(3), 165-181.

[12] Kirby, S., & Hurford, J. (1997). The evolution of incremental learning: language, development and critical periods. Edinburgh Occasional Papers in Linguistics, 97(2), 1-33.

[13] Giraud-Carrier, C. (2000). A note on the utility of incremental learning. AI Communications, 13(4), 215-223.

[14] Solomonoff, R. J. (2002, December). Progress in incremental machine learning. In NIPS Workshop on Universal Learning Algorithms and Optimal Search, Whistler, BC.

[15] Aha, D. W. (1992). Tolerating noisy, irrelevant and novel attributes in instance-based learning algorithms. International Journal of Man-Machine Studies, 36(2), 267-287.

[16] Toselli, A. H., Vidal, E., & Casacuberta, F. (2011). Incremental and Adaptive Learning for Interactive Machine Translation. In Multimodal Interactive Pattern Recognition and Applications (pp. 169-177). Springer London.

[17] Daybelge, T., & Cicekli, I. (2011). A ranking method for example based machine translation results by learning from user feedback. Applied Intelligence, 35(2), 296-321.

[18] Lee, H. A. (2006). Translation selection through machine learning with language resources. In Computer Processing of Oriental Languages. Beyond the Orient: The Research Challenges Ahead (pp. 370-377). Springer Berlin Heidelberg.

[19] Carpuat, M., & Wu, D. (2007, June). Improving Statistical Machine Translation Using Word Sense Disambiguation. In EMNLP-CoNLL (Vol. 7, pp. 61-72).

[20] Bannard, C., & Callison-Burch, C. (2005, June). Paraphrasing with bilingual parallel corpora. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (pp. 597-604). Association for Computational Linguistics.

[21] Callison-Burch, C., Koehn, P., & Osborne, M. (2006, June). Improved statistical machine translation using paraphrases. In Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (pp. 17-24). Association for Computational Linguistics.

[22] Zens, R., Och, F. J., & Ney, H. (2002, September). Phrase-based statistical machine translation. In Annual Conference on Artificial Intelligence (pp. 18-32). Springer, Berlin, Heidelberg.

[23] Koehn, P., & Hoang, H. (2007). Factored translation models. In Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL).

[24] Beesley, K. R., & Karttunen, L. (2003). Finite-state morphology: Xerox tools and techniques. CSLI, Stanford.

[25] Brants, T. (2000, April). TnT: a statistical part-of-speech tagger. In Proceedings of the sixth conference on Applied natural language processing (pp. 224-231). Association for Computational Linguistics.

[26] Och, F. J. (2003, July). Minimum error rate training in statistical machine translation. In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1 (pp. 160-167). Association for Computational Linguistics.

[27] Specia, L., Srinivasan, A., Joshi, S., Ramakrishnan, G., & Nunes, M. D. G. V. (2009). An investigation into feature construction to assist word sense disambiguation. Machine Learning, 76(1), 109-136.