Investigating the potential of post-ordering SMT output to improve translation quality

Pratik Mehta, Anoop Kunchukuttan, Pushpak Bhattacharyya
Center for Indian Language Technology
Department of Computer Science and Engineering
Indian Institute of Technology Bombay
{pratikm,anoopk,pb}@cse.iitb.ac.in

Abstract

Post-ordering of Statistical Machine Translation (SMT) output to correct word order errors could be a promising area of research to overcome structural divergence between language pairs. This is especially true when it is difficult to incorporate rich linguistic features into the baseline decoder. In this paper, we propose an algorithm for generating oracle reorderings of MT output. We use the oracle reorderings to empirically quantify an upper bound on improvement in translation quality through post-ordering techniques. In our study encompassing multiple language pairs, we show that significant improvement in translation quality can be obtained by applying reordering transformations on the output of the SMT system. This presents a strong case for investing effort in exploring the post-ordering problem.

1 Introduction

Word order divergence is a central problem in Statistical Machine Translation (SMT) and major stumbling block to generating comprehensible translations. Many solutions for reordering have been proposed to bridge this divergence. Word order divergence is generally handled within the core SMT model or using source-side reordering as a pre-processing step. In the core SMT system, word order can be tackled using a variety of models of varying complexity: word-level alignment models (Brown et al., 1993), lexicalized reordering models (Tillmann, 2004; Galley and Manning, 2008), hierarchical SMT (Chiang, 2005), syntax based SMT (Yamada and Knight, 2002), *etc*. SMT models like phrase-based SMT are not good at bridging word order divergence (Marie et al., 2014). In these cases, source-side reordering is used as a pre-processing step to convert the source sentence to target language word order (Collins et al., 2005; Ramanathan et al., 2008). The best performing approaches generally rely on parse information on the source side to generate the correct word order. However, it has proved to be a very difficult problem which is far from being solved, especially when parse information is not forthcoming. The computational complexity of searching through a large space of potential reorderings and the need for incorporating higher level linguistic information are the primary challenges in tackling the reordering problem.

While there is active research in preordering and in-decoder approaches, there has been little work on the problem of post-ordering of SMT output. We define the word-order post-ordering as follows:

Given the output of an MT system, permute the words of the output to generate a better word order.

The following example shows how simply reordering the words in the SMT output can improve translation quality:

Source:

ये बीते पांच सालों में ग्रीस को मिलने वाला तीसरा राहत पैकेज है

Translation (Google Translate[1]):

`They meet Greece in the last five years is the third bailout package`

Post-ordered Translation:

`They is the third bailout package in the last five years meet Greece`

The following are a few reasons why post-ordering may be an interesting direction to explore:

- SMT decoders have to search through a very large search space to find the best translations. Hence, the translation models are generally simple and use a limited number of fea-

[1] `https://translate.google.com/`

tures so as to make decoding computationally tractable. Generating correct word order generally requires richer models which can look at long distance dependencies. The post-ordering stage is a good stage in the SMT pipeline where richer models can be applied to the best translation candidates to correct word order errors.

- If the target language is resource rich, we can use resources of the target language. For instance, chunkers, constituency/dependency parsers, *etc.* may be available for the target language. This is the case for translation from many Indian language to English. Hence, our experiments in the paper have focussed on Indian language to English translation. However, an important limitation of this approach could be the inability of these tools to perform with high accuracy in the face of errors in the baseline translation output.

- Post-ordering can take advantage of human-postediting of MT output. The post-edited output can be useful to learn post-ordering models that are customized to the baseline SMT system.

- Even if human post-edited output is not available, aligning the baseline output with the reference translation can help construct oracle reorderings. The parallel corpus comprising the baseline output and their oracle reorderings could be used to learn post-ordering models customized to the baseline SMT system.

Before embarking of designing post-ordering methods, it would be prudent to estimate if post-ordering methods can actually improve translation quality. In this paper, we study the viability of post-ordering *i.e. can significant improvements in translation quality be obtained by simply permuting the underlying MT system output's word order?*. We try to answer this question by estimating an upper bound on the translation accuracy after post-ordering. For this, we propose to compute oracle reorderings of the translation output by comparing it with the reference translation. Our experiments using this approach show significant improvement in translation quality over baselines, as measured by both automatic and manual evaluation metrics. This puts forward a good case for exploring post-ordering methods for machine translation.

The following is an outline of the paper. In Section 2, we describe related work. In the remainder of the paper, we estimate an upper bound on the potential gains in translation accuracy by post-ordering. Section 3 describes our method for computing oracle reorderings from the translation output, which is used to estimate the upper bound. Section 4 describes our experimental setup and 5 presents the results and discusses the observations. Section 6 concludes the paper and points out future work.

2 Related Work

Oracle translations have been used by many researchers for diagnosing translation output. Auli et al. (2009) and Wisniewski and Yvon (2013) have used oracle translations to do reference reachability analysis and identify model and search errors. Wisniewski and Yvon (2013) have used the oracle translations to conduct various kinds of failure analysis and study effect of various search parameters. Dreyer et al. (2007), Li and Khudanpur (2009) and Wisniewski and Yvon (2013) use oracle translations to understand the limitations of various reordering constraints imposed on translation decoders. In the same spirit, we try to estimate an upper bound on the benefits of post-ordering the baseline SMT output.

Though we do not tackle the problem of post-ordering in this work, we summarize the existing work on post-ordering for SMT. There has been work on post-editing of machine translation output. The method described in (Simard et al., 2007) is most commonly used. It involves automatically post-editing the output of an MT system using another phrase-based MT system trained on parallel data constructed from previously decoded output (e) and corresponding references (e'). Béchara et al. (2011) improvizes on this approach by appending source words (f) to the output part of the parallel data (e), creating a new language (e'#f) and retaining source context. Marie et al. (2014) use a second-pass decoder to improve translation quality. However, none of these works have focussed on word order and the effect on the word order has not been explicitly evaluated.

Post-ordering as a problem has been introduced by Sudoh et al. (2011). However, it is not a

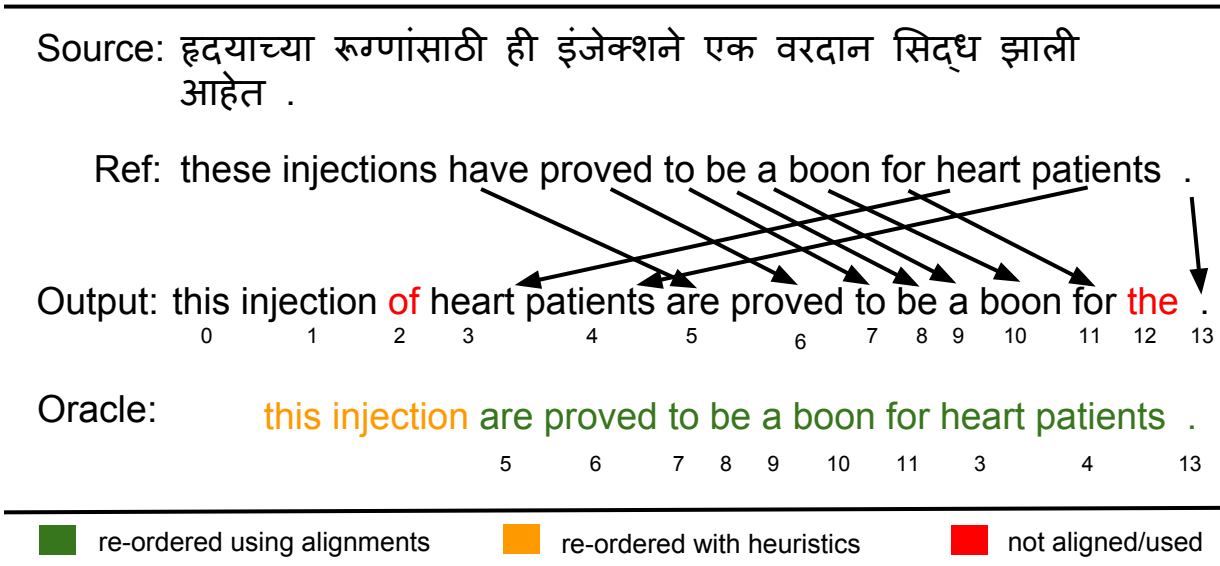

Figure 1: Construction of Oracle reordering

post-ordering system in the sense in which we have defined it. Theirs is actually a two stage translation system that decomposes the translation problem into lexical transfer and reordering sub-problems. Goto et al. (2012) and Goto et al. (2013) also propose post-ordering systems with the same architecture, but different reordering methods in the second stage. The motivation in these post-ordering methods is not to improve upon the word order. Rather, lexical mappings are learnt in the first stage after reordering the target text to match the source order, thus necessitating the second reordering stage.

3 Generating Oracle Translations

To estimate an upper bound on the improvement in translation accuracy possible with post-ordering, we generate *oracle* reorderings of the baseline SMT output hypothesis. An *oracle* reordering is the best possible word order of the hypothesis, in terms of fluency and syntactic correctness. We propose the following algorithm for computing the oracle reordering.

1. Obtain word alignments between the hypothesis and reference using the monolingual aligner algorithm in METEOR (Denkowski and Lavie, 2014).

2. Construct a new sentence by rearranging aligned words from the hypothesis using the word-order from the reference.

3. Use additional heuristics to include as many unaligned words from the hypothesis into the *oracle* reordering as possible. In our experiments, the words in the hypothesis that were not aligned by METEOR but found a stem-match in the reference were inserted in the oracle sentence.

The resulting oracle hypothesis is a permutation of a subset of words in the original MT decoding step, such that they reflect the word order in the reference.

4 Experimental Setup

We studied different SMT systems from 10 Indian languages to English for quantifying the potential improvement in translation accuracy. The experiments were carried across 10 Indian languages included in the multilingual ILCI corpus (Jha, 2010), which contains nearly 50,000 parallel sentences . For each language pair, the corpus was split into 46,277 sentences for training, 500 sentences for tuning and 2000 sentences for testing. We trained phrase-based and hierarchical phrase-based systems on this data.

The phrase-based systems were trained using the Moses SMT toolkit (Koehn et al., 2007) with the *grow-diag-final-and heuristic* for phrase extraction and the *msd-bidirectional-fe* model for lexicalized reordering. The trained models were tuned with Minimum Error Rate Training (MERT) (Och, 2003) with default parameters. We trained

Score	Adequacy	Fluency
1	No meaning	Incomprehensible
2	Little meaning	Disfluent
3	Much meaning	Non-native fluency
4	Most meaning	Good fluency
5	All meaning	Flawless fluency

Table 1: Description of scores for manual evaluation

5-gram language models on the training-set using Kneser-Ney smoothing with SRILM (Stolcke and others, 2002). The hierarchical systems were also trained with Moses using default parameters.

For three phrase-based SMT systems with Hindi, Marathi and Malayalam respectively as source and English as target language, qualitative analysis was performed through manual evaluation of output sentences by native speakers of each of the source languages. Given the source and gold reference, the evaluators were asked to rate the adequacy and fluency of a system's output and oracle sentences on a scale of 1 to 5, as described in (Koehn and Monz, 2006) (see Table 1). The weighted average of the scores over all sentences was then calculated as:

$$\text{average\_score} = \sum_{s=1}^{5} s.f(s) \qquad (1)$$

where, s: the score ranging from 1 to 5
$f(s)$: frequency of occurrence of score s

5 Results & Discussion

Table 2 shows the results in case-insensitive BLEU (Papineni et al., 2002). There was significant improvement in oracle reordering over the baseline SMT systems. This trend was consistent across all studied language-pairs and in both phrase-based and hierarchical SMT systems. We see that the oracle sentences were often shorter than the translation hypotheses because words that were not aligned with the reference translations nor accounted for by the heuristics were left out. For fair evaluation, we removed these *outlier* words from the original translations to create *pruned* hypotheses containing the same bag of words as their corresponding oracle sentences and compute BLEU scores. The average improvement in oracle BLEU scores over all language pairs was 59.5% for phrase-based systems, and 60.45% for hierarchical systems. Table 4 shows examples illustrating the oracle reorderings from the Hindi-English phrase-based SMT experiment.

Lang-pair	Model	Original	Pruned	Oracle
hin-eng	PBSMT	22.77	23.12	36.74
	HPBMT	23.87	24.5	37.36
mar-eng	PBSMT	12.6	11.23	18.07
	HPBMT	14.65	13.95	21.97
ben-eng	PBSMT	16.24	15.67	24.38
	HPBMT	14.69	14.01	24.59
guj-eng	PBSMT	17.66	17.03	26.32
	HPBMT	15.96	15.17	24.59
kon-eng	PBSMT	15.56	14.74	22.7
	HPBMT	13.67	12.77	20.85
pan-eng	PBSMT	19.98	19.87	32.72
	HPBMT	20.12	20.00	30.64
urd-eng	PBSMT	17.31	17.17	29.89
	HPBMT	19.05	18.58	29.52
tam-eng	PBSMT	10.54	8.9	14.77
	HPBMT	10.3	9.0	15.4
tel-eng	PBSMT	12.63	11.42	17.96
	HPBMT	11.9	10.66	17.42
mal-eng	PBSMT	8.3	6.03	10.32
	HPBMT	8.46	6.48	11.24

Table 2: Experimental results (BLEU)
ISO-639-2 language codes are shown

In the manual evaluation task, the evaluators frequently rated oracle sentences as being more fluent than the decoded sentences. Average fluency scores across all evaluated sentences improved by a margin of 0.45 in Hindi (16.4%), 0.15 in Malayalam (6.8%) and 0.11 in Marathi (3.9%), as seen in Table 3. The small drop in adequacy of oracle was expected because of imperfect alignments between the disfluent output and their references, which affected the construction of complete oracle sentences. We suspect that this loss in adequacy must also have affected the perception of fluency in cases where the oracle sentence was significantly shorter than the original output. With better alignment aided by transliteration and more sophisticated heuristics, construction of more adequate oracle sentences would be possible. However, this will only serve to reinforce belief in post-ordering as a beneficial exercise - something our results already show.

Lang-pair	Test	Adequacy	Fluency
hin-eng	Original	3.29	2.74
	Oracle	3.19	3.19
mar-eng	Original	3.12	2.83
	Oracle	2.96	2.94
mal-eng	Original	1.92	2.19
	Oracle	1.40	2.34

Table 3: Manual evaluation scores

6 Conclusion

We see that the BLEU score of oracle reorderings show substantial improvements of upto 60.45% over the baseline output. Manual evaluation of MT output also shows 20% improvement in fluency of translations. These improvements were obtained by simply reordering the output of the baseline SMT systems. Our study thus establishes the potential for further research in post-ordering of machine translation output to provide significant gains in translation quality. The post-ordered parallel translation corpus obtained by oracle alignment may be used for learning post-ordering models.

Acknowledgements

We thank the Technology Development for Indian Languages (TDIL) Programme and the Department of Electronics & Information Technology, Govt. of India for providing the ILCI corpus. We thank the National Knowledge Network for making available computational facilities on the Garuda Cloud for performing computationally intensive tasks. Thanks to Aditya Joshi, Joe Cheri Ross and Anupam Ghosh for helping us in evaluation of the MT outputs and for their feedback on translation quality.

References

Michael Auli, Adam Lopez, Hieu Hoang, and Philipp Koehn. 2009. A systematic analysis of translation model search spaces. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, StatMT '09, pages 224--232, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hanna Béchara, Yanjun Ma, and Josef van Genabith. 2011. Statistical post-editing for a statistical mt system.

Peter F Brown, Vincent J Della Pietra, Stephen A Della Pietra, and Robert L Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 263--270. Association for Computational Linguistics.

Michael Collins, Philipp Koehn, and Ivona Kučerová. 2005. Clause restructuring for statistical machine translation. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 531--540. Association for Computational Linguistics.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*.

Markus Dreyer, Keith Hall, and Sanjeev Khudanpur. 2007. Comparing reordering constraints for smt using efficient bleu oracle computation. In *Proceedings of the NAACL-HLT 2007/AMTA Workshop on Syntax and Structure in Statistical Translation*, pages 103--110. Association for Computational Linguistics.

Michel Galley and Christopher D Manning. 2008. A simple and effective hierarchical phrase reordering model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 848--856. Association for Computational Linguistics.

Isao Goto, Masao Utiyama, and Eiichiro Sumita. 2012. Post-ordering by parsing for japanese-english statistical machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, ACL '12, pages 311--316, Stroudsburg, PA, USA. Association for Computational Linguistics.

Isao Goto, Masao Utiyama, and Eiichiro Sumita. 2013. Post-ordering by parsing with itg for japanese-english statistical machine translation. 12(4):17:1--17:22, oct.

Girish Nath Jha. 2010. The tdil program and the indian language corpora initiative (ilci).

Philipp Koehn and Christof Monz. 2006. Manual and automatic evaluation of machine translation between european languages. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 102--121. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177--180. Association for Computational Linguistics.

Source	मैंने अमेजन नदी के बियाबान एवं नदी तटों की सुंदरता के बारे में पढ़ा व सुना है ।
Reference	I have read and heard about the beauty of the wilderness and river banks of the Amazon river .
Original hypothesis	I Amazon and river banks of the river wilderness about the beauty of read and have heard.
Pruned hypothesis	I Amazon and river banks of the river wilderness about the beauty of read and have heard .
Oracle	I have read and heard about the beauty of wilderness and river banks of the Amazon river .

Source	इस अध्ययन में 5800 परिवारों के 14 हजार बच्चों को शामिल किया गया था ।
Reference	In this study 14 thousand children of 5800 families were included .
Original hypothesis	In this study 5800 families of 14 thousand children had been included.
Pruned hypothesis	In this study 5800 families of 14 thousand children had included .
Oracle	In this study 14 thousand children of 5800 families had included .

Table 4: Oracle post-ordering examples for Hindi-English

Zhifei Li and Sanjeev Khudanpur. 2009. Efficient extraction of oracle-best translations from hypergraphs. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 9--12. Association for Computational Linguistics.

Benjamin Marie, Lingua et Machina, and Aurélien Max. 2014. Confidence-based rewriting of machine translation output. In *Conference on Empirical Methods in Natural Language Processing*, pages 1261--1272.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 160--167. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311--318. Association for Computational Linguistics.

Ananthakrishnan Ramanathan, Jayprasad Hegde, Ritesh M Shah, Pushpak Bhattacharyya, and M Sasikumar. 2008. Simple syntactic and morphological processing can help english-hindi statistical machine translation. In *IJCNLP*, pages 513--520.

Michel Simard, Nicola Ueffing, Pierre Isabelle, and Roland Kuhn. 2007. Rule-based translation with statistical phrase-based post-editing.

Andreas Stolcke et al. 2002. Srilm-an extensible language modeling toolkit.

Katsuhito Sudoh, Xianchao Wu, Kevin Duh, Hajime Tsukada, and Masaaki Nagata. 2011. Post-ordering in statistical machine translation. In *Proc. MT Summit*.

Christoph Tillmann. 2004. A unigram orientation model for statistical machine translation. In *Proceedings of HLT-NAACL 2004: Short Papers*, pages 101--104. Association for Computational Linguistics.

Guillaume Wisniewski and François Yvon. 2013. Oracle decoding as a new way to analyze phrase-based machine translation. *Machine translation*, 27(2):115--138.

Kenji Yamada and Kevin Knight. 2002. A decoder for syntax-based statistical mt. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 303--310. Association for Computational Linguistics.