How Good are Transformers in Reordering?

Ch Ram Anirudh and K Narayana Murthy

University of Hyderabad,
Hyderabad, India
ramanirudh@uohyd.ac.in, knmuh@yahoo.com

Abstract. Translation requires transfer of lexical items (words / phrases) from Source Language to Target Language and also reordering of the transferred lexical items as appropriate for the target language. Whatever be the approach used, quality of translation depends on both the quality of lexical transfer and quality of reordering. In this paper, we explore how good the state-of-the-art sequence-to-sequence Transformer model is in reordering. Reordering models are tested for sequence to sequence mapping from an Intermediate Language (which uses the words of the target language arranged in the source language order) to target language. We build models using the *samanantar* English-Kannada parallel corpus. BLEU, TER and RIBES scores show significant improvement after reordering. We have also tested our models on the Machine Translation task as a whole. Compared to the default lexicalized reordering models used in Statistical Machine Translation, our transformer based reordering models have shown better performance.

Keywords: Machine Translation, Natural Language Processing, Transformers, Reordering

1 Introduction

Translation requires transfer of lexical items (words and phrases) from Source Language (SL) to Target Language (TL) and also reordering of the transferred lexical items as appropriate for the target language. Any approach to Machine Translation (MT) has to do these implicitly or explicitly. Traditional rule based MT (RBMT) [7] - transfer based methods in particular, perform transfer at morphological and syntactic levels explicitly. Statistical MT (SMT) [4, 13], a more modern data driven approach, builds a lexical table and a phrase table (translation model) from a parallel corpus, which are used for lexical transfer. SMT also learns a reordering model from the parallel corpus. The translation model and reordering model are combined with language model in a log-linear function to generate translations during a process called decoding. Neural MT [1, 24] implicitly learns to translate SL texts into TL texts in appropriate word order. Whatever be the method, quality of translation depends on both the quality of lexical transfer and quality of reordering.

Modern approaches to MT recognize the fact that words may have multiple meanings and connotations, and disambiguation is best done by considering the sentential context in which they occur. That is perhaps why lexical transfer and reordering are not considered as independent tasks. However, lexical items are much smaller and structurally much simpler than sentences. Also, the number of possible lexical items is much smaller when compared to all possible sentences, which is open ended and unbounded. Therefore, conceptually, lexical transfer is a much simpler task compared to sequence to sequence operations such as reordering. Further, the morphological complexity (including inflection, derivation, phonetic conflation, compounding etc.) vary widely across languages. Degree and nature of lexical ambiguities are also not the same in all languages. Word order is very rigid in some languages, while it is relatively free in other languages. Instead of the same-size-fits-all approach, exploring lexical transfer and reordering individually can give us new insights and open up new doors for research. In particular, here we focus on the question of how good the latest models are in the reordering task.

Suppose the MT task starts by transferring SL words to TL words arranged in SL word order. We call this *intermediate language* (IL). Next, the words in IL text are reordered using a reordering model. See fig. 1.

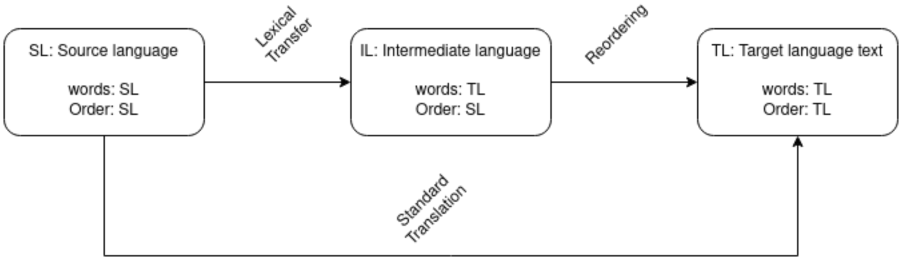

Fig. 1. Typical workflow of MT in explicit lexical transfer followed by reordering compared to standard MT models.

In this paper, we explore how good the state-of-the-art sequence-to-sequence Transformer model [24] is in reordering. The transformer reordering model is built using parallel corpus of IL sentences and corresponding TL sentences. Such a training data is prepared using word-aligned parallel corpus trained using IBM models of SMT [4]. Word alignments from SL to TL help us in arranging TL sentences in SL word order. We then train IL-TL transformer reordering model. We test the reordering model for English and Kannada languages. English belongs to Indo-Germanic language family and Kannada belongs to Dravidian language family. These languages differ widely in morphological and syntactic structures making them good candidates for testing reordering. Results on reordering test sets have shown a substantial improvement in BLEU score from 51.5 before reordering to 85.1 after reordering using transformer reordering model for English. It improved from 47 to 77 for Kannada. Reordering model is tested on MT task also, by applying it after generating monotone (SL ordered) SMT decoder output. Promising results were shown for English to Kannada.

2 Related Work

Simplest reordering in phrase-based MT [13] is a distortion probability distribution model that models the jump of phrase from one position to another. Tillman [23] proposed a lexicalized reordering model that predicts whether next phrase should be oriented to the right (monotone), left (swap) or a different position (discontinuous) relative to the previous translated phrase. This is the most commonly used baseline model in SMT. Li et al. [15] used a neural reordering model comprising of recursive autoencoders to learn the orientations. In all these methods, reordering model is combined with translation model and language model during decoding. Here we rather train a sequence-to-sequence reordering model, which is used to reorder after translating words/phrases.

Very few works in MT literature have had a separate reordering model. Bangalore and Riccardi [2] developed stochastic finite state models for lexical choice and reordering separately. The lexical choice model outputs a sequence of TL words for a given SL sentence. A stochastic finite state transducer learnt from a parallel corpus of source-ordered TL words with their corresponding TL sentences, is used to convert the TL words from lexical choice to a set of reordering rules. Sudoh et al. [22] do a translation from Japanese to English by training two SMT models: first, to translate Japanese to English in Head-Final English (HFE) ordering which is a feature native to Japanese. Second SMT model learns to reorder HFE to appropriate English. The English portion of the parallel corpus is reordered according a set of rules for getting HFE. Our work is similar, where we use word-alignments from IBM models instead of linguistic rules to get a parallel corpus of IL-TL sentences. More information on reordering models for MT can be found in a comprehensive survey by Bisazza and Federico [3].

Here we use transformers to build reordering models. Transformers are the backbone of many state-of-the-art NLP models including BERT [5], GPT [18], BART [14], etc. They have consistently outperformed the previous models in various tasks like machine translation, text summarization and question answering on standard data sets [5,24]. In applications of computer vision, transformers have shown remarkable results [6, 10]. Transformer learns the representations of input and output sequences, completely on a self-attention mechanism, without using any recurrence mechanism in its architecture. This mitigates the problems of long-distance relationship and catastrophic forgetting as with earlier RNN based networks. The authors of the original paper further claim that self-attention mechanism implicitly models the structure of sentence, making the transformer model more interpretable. For more details on transformer network, see Vaswani et al. [24].

3 Experiments

3.1 Dataset and Preprocessing

All the experiments in this work are done on English-Kannada (en-kn) language pair. English and Kannada belong to distinct language families. Kannada is agglutinative, morphologically rich and a free word order language but with a default *subject-object-verb* order. English is a relatively more isolating language with *subject-verb-object* word order.

English-Kannada parallel corpus from *samanantar* [19] is used for all the experiments. *samanantar* is a repository of large-scale parallel corpora for Indic languages, collected mostly by web-scraping and aligning similar sentences in web scraped texts. There are 4014931 translation pairs with around 27.7M words in English and 36.9M words in Kannada.

We have preprocessed the dataset for our experiments using open source tools. English side of the dataset is truecased and tokenized using the recaser and tokenizer respectively, provided with Moses SMT toolkit [12]. Kannada text is tokenized using Indic NLP library[1]. The tokenized parallel corpus is then cleaned to remove very long sentences ($>$ 80 words) and empty lines on source and target sides, using corpus cleaning script provided with Moses.

The dataset required for training our reordering model is actually a parallel corpus where one side is IL and the other side is TL. We take the en and kn sides from the cleaned *samanantar* corpus and create two parallel corpora: $IL_{en} - en$ and $IL_{kn} - kn$. IL sentences are generated using alignments learnt by training SMT models. The corpora obtained thus are split into training, development and test sets. 500 sentences are randomly picked for development set and 3000 sentences are randomly picked for test set. Remaining sentences are used for training. Subword tokenization is learnt for the two datasets using byte-pair-encoding (BPE) technique with 32000 symbols [20]. The datasets are then BPE tokenized before proceeding for training.

3.2 Training

We train two phrase based SMT systems: en-kn and kn-en using Moses. These models are useful for obtaining IL and testing the reordering models later. Cleaned *samanantar* parallel corpus is used for training. 5-gram language models with Kneser-Ney smoothing are built with *kenlm* provided with Moses. Monolingual datasets for building LM are obtained from Kakwani et al. [9]. We use the option `grow-diag-final-and` for learning alignments and `msd-bidirectional-fe` for learning lexicalized reordering.

We build two transformer reordering models: $IL_{en} - en$ and $IL_{kn} - kn$. OpenNMT [11] is used for training. Hyperparameter settings for the transformer network are similar to Vaswani et al. [24], except the word-embeddings. Vocabulary and word-embeddings are shared across input and output, since the words in

[1] https://anoopkunchukuttan.github.io/indicnlplibrary/

IL and TL are same. Encoder and decoder stacks consist of 6 layers each. The model dimension d is 512. Batch size is 2048, validation batch size is 128. Training steps are 200000 and validation steps are 10000. Attention heads are 8 and attention dropout is 0.1. Optimization is done using Adam optimizer. Training is done on 2 NVIDIA GeForce RTX 2080 GPUs. The training takes around 12 hours for $IL_{en} - en$ and 14 hours for $IL_{kn} - kn$.

3.3 Evaluation

Evaluation is done using BLEU [17], TER [21] and RIBES [8] scores. BLEU (BiLingual Evaluation Understudy) uses a modified precision based score that counts n-gram matches between the hypothesis and reference translations. TER (Translation Edit Rate) is based on edit distance between hypothesis and reference which includes a shift operation alongside insertion, deletion and substitution. RIBES (Rank-based Intuitive Bilingual Evaluation Score) directly measures the reordering between hypothesis and reference translation using rank correlation coefficients. The word ranks (indices) in reference are compared to those of the corresponding matching words in the hypothesis to compute the correlation coefficient. For testing our task, RIBES and TER are more relevant.

We evaluate our reordering model in two settings. In one, we report BLEU, TER and RIBES scores before and after reordering on the test sets we created from IL-TL parallel corpus. We call these reordering test sets. Evaluation on these shows whether transformer is good at reordering. The sentences in these test sets represent an ideal situation where lexical transfer is at its best. To check how they perform in a more realistic setup of MT, we also test the models by using them to reorder the monotone-ordered[2] SMT decoder output. We compare the BLEU, TER and RIBES scores of this reordered output with the decoder output generated using the baseline lexicalized reordering in SMT. It may be noted that while lexicalized reordering is combined with translation and language models, we do transformer reordering separately after generating monotone ordered decoder output which resembles our IL. For these, we use test sets provided by Workshop on Asian Translation (WAT) 2021 [16].

3.4 Results

The results are reported in tables 1 and 2. Table 1 shows BLEU, TER and RIBES on the test set before and after reordering. All the scores increase substantially after reordering using the transformer model, showing the promise of transformer for reordering. Table 2 shows MT evaluation results on test set provided with WAT2021 benchmarks dataset. For English to Kannada MT, reordering separately using transformer model is better than lexicalized reordering. This is evident from TER and RIBES scores. For Kannada to English MT, lexicalized reordering gives better results compared to separate transformer model. However, when we look at the counts of edit operations in TER computation

[2] use option -dl 0 in Moses during decoding to generate monotone ordered output

(table 3), we see that shifts are lesser with transformer reordering compared to insertions, deletions and substitutions for both language pairs, indicating the efficacy of transformer reordering model.

Table 1. BLEU and TER scores on reordering test set before and after reordering using transformer model

Language	Metric	Before Reordering	After Reordering
Kannada	BLEU	47.7	**77.4**
	RIBES	0.693	**0.929**
	TER	0.382	**0.117**
English	BLEU	51.5	**85.1**
	RIBES	0.739	**0.961**
	TER	0.360	**0.095**

Table 2. BLEU and TER scores of SMT output with lexicalized reordering model and transformer reordering model on WAT2021 test sets

Language Pair	Metric	Lexicalized Reordering	Transformer Reordering
Kannada-English	BLEU	19.9	16.0
	RIBES	0.541	0.537
	TER	0.754	0.772
English-Kannada	BLEU	9.2	9.2
	RIBES	0.423	**0.465**
	TER	0.891	**0.867**

Table 3. Edit operations in TER when outputs are compared with reference translations

Language Pair	Edit Ops.	Lexicalized Reordering	Transformer Reordering
Kannada-English	Insertions	3423	2525
	Deletions	5463	7253
	Substitutions	15637	15798
	Shifts	4494	**4160**
English-Kannada	Insertions	2586	2444
	Deletions	3043	3233
	Substitutions	15851	15749
	Shifts	2984	**2388**

4 Conclusions

In this paper we have explored transformers for reordering in MT. We build models using the *samanantar* English-Kannada parallel corpus. Our experiments show that transformers are good at the reordering task. This will hopefully encourage further explorations in this direction and open up new avenues of research in the field of MT.

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural Machine Translation by Jointly Learning to Align and Translate. In: Proceedings of the 3rd International Conference on Learning Representations. San Diego, CA, USA (2015)
2. Bangalore, S., Riccardi G.: Finite-state models for lexical reordering in spoken language translation. In: Proceedings of Sixth International Conference on Spoken Language Processing. Beijing, China (2000)
3. Bisazza, A., Federico, M.: A survey of word reordering in statistical machine translation: Computational models and language phenomena. Comput. linguist. 1;42(2):163-205 (June 2016)
4. Brown, P.F., Pietra, V.J.D., Pietra, S.A.D., Mercer, R.L.: The mathematics of statistical machine translation: Parameter estimation. Comput. linguist. 19(2), 263-311 (June 1993)
5. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the NAACL:HLT, Volume 1 (Long and Short Papers), pages 4171-4186, Minneapolis, Minnesota. ACL (2019)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X.,Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: Proceedings of International Conference on Learning Representations. Online (2021)
7. Hutchins, W.J., Somers, H.L.: An introduction to machine translation. Academic Press, London (1992)
8. Isozaki, H., Hirao, T., Duh, K., Sudoh, K., Tsukada, H.: Automatic evaluation of translation quality for distant language pairs. In: Proceedings of the 2010 Conference on EMNLP. pp. 944-952. Massachusetts, USA (2010)
9. Kakwani, D., Kunchukuttan, A., Golla, S., Gokul N.C., Bhattacharyya, A., Khapra, M.M., Kumar, P.: IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In: Findings of the Association for Computational Linguistics: EMNLP 2020, pages 4948-4961, Online. Association for Computational Linguistics (2020)
10. Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M.: Transformers in Vision: A Survey. ACM Comput. Surv. 54, 10s, Article 200 (January 2022), 41 pages. https://doi.org/10.1145/3505244
11. Klein, G., Kim, Y., Deng, Y., Nguyen, V., Senellart, J., Rush, A.: OpenNMT: Neural machine translation toolkit. In: Proceedings of the 13th Conference of the AMTAS (Volume 1: Research Papers). pp. 177-184. Boston, MA, USA (Mar 2018)
12. 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 ACL. pp. 177-180. ACL, Prague, Czech Republic (June 2007)

13. Koehn, P., Och, F.J., Marcu, D.: Statistical phrase-based translation. In: Proceedings of the 2003 Conference of the NAACL:HLT-Volume 1. pp. 48-54. ACL, Edmonton (2003)
14. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In: Proceedings of the 58th Annual Meeting of the ACL, pages 7871-7880, Online. ACL (2020)
15. Li, P., Liu, Y., Sun, M., Izuha, T., Zhang, D.: A neural reordering model for phrase-based translation. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers (August 2014)
16. Nakazawa, T., Nakayama, H., Ding, C., Dabre, R., Higashiyama, S., Mino, H., Goto, I., Pa, W.P., Kunchukuttan, A., Parida, S., Bojar, O., Chu, C., Eriguchi, A., Abe, K., Oda, Y., Kurohashi, S.: Overview of the 8th Workshop on Asian Translation. In: Proceedings of the 8th Workshop on Asian Translation (WAT2021), pages 1-45, Online. ACL (2021)
17. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the ACL. pp. 311-318. ACL, Philadelphia, Pennsylvania, USA (Jul 2002).
18. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding with unsupervised learning. Technical report, OpenAI (2018)
19. Ramesh, G., Doddapaneni, S., Bheemaraj, A., Jobanputra, M., Raghavan, A.K., Sharma, A., Sahoo, S., Diddee, H., Mahalakshmi J, Kakwani, D., Kumar, N., Pradeep, A., Nagaraj, S., Deepak, K., Raghavan, V., Kunchukuttan, A., Kumar, P., Khapra, M.S.: Samanantar: The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages. Transactions of the ACL, 10:145-162 (2022)
20. Sennrich R., Haddow B., Birch A.: Neural Machine Translation of Rare Words with Subword Units. In: Proceedings of the 54th Annual Meeting of the ACL (Volume 1: Long Papers) (pp. 1715-1725) (2016)
21. Snover, M., Dorr, B., Schwartz, R. Micciulla, L., Makhoul, J.: A study of translation edit rate with targeted human annotation. In: Proceedings of the 7th Conference of the AMTAS. Massachusetts, USA (2006)
22. Sudoh, K., Wu, X., Duh, K., Tsukada, H., Nagata, M.: Syntax-Based Post-Ordering for Efficient Japanese-to-English Translation. ACM Trans. on Asian Lang. Inf. Process. 12, 3, Article 12 (August 2013).
23. Tillmann, C.: A unigram orientation model for statistical machine translation. In: Proceedings of the Joint Conference on HLT-NAACL, pages 101-104. Boston, MA (2004)
24. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Proceedings of Neural Information Processing Systems. pp. 6000-6010. Long Beach, CA, USA (2017)