Preordering based Low Resource Translation Using Pretrained Multilingual Model

Jingyi Zhu[1] Takuya Tamura[1] Fuzhu Zhu[1] Xiaotian Wang[1] Taiki Sakai[1] Takehito Utsuro[1] Masaaki Nagata[2]
[1]Deg. Prog. Sys.&Inf. Eng., Grad. Sch. Sci.&Tech., University of Tsukuba
[2]NTT Communication Science Laboratories, NTT Corporation, Japan

Abstract

In this paper, we propose to apply the pretrained seq2seq model to conduct preordering process and translation. We use the manual word alignment data to make preorder training data, and we compared the performance of different kinds of mT5 and mBART in preordering. For the translation procedure, we choose mBART as our baseline model. We evaluated our approach on the Asian Language Treebank dataset (total of 20,000 parallel data), with directions in Japanese and English, and in-house parallel data (total of 3,000 parallel data) of Japanese and Chinese from NTT corporation. For the results, our proposed approach exceeds the baseline on the translation direction of Zh-Ja pairs, and is close to level with the baseline on Ja-En pairs.

1 Introduction

In recent years, more and more researches have been conducted on sequence-to-sequence (seq2seq) models based on pretraining [14, 6, 5]. Since the introduction of Transformer [13], the quality of translation has been greatly improved. However, in the task of low resource translation, due to the dataset's size limitation, this kind of parameter randomization model often performs poorly [16]. In order to meet this challenge, many researches have proposed large-scale pretraining models, which have been widely used in several tasks in NLP [1, 12].

In this paper, We propose applying the pretrained seq2seq model to both preordering and translation. We discussed different sizes of mT5s [14] and mBART [6], to verify the performance that those models could make in preordering when using manual word alignment data. For the translation process, we choose mBART as our baseline translation model. Translation results with the original source language as input and generated preordering as input is verified respectively. In the validation of the ALT Japanese-English dataset [9], the result of preordering as input is similar to that of the baseline. On the other hand, the result of using preordering as input is higher than the baseline in the verification of the in-house Chinese-Japanese parallel dataset provided by NTT.

2 Ralated Work

Kawara et al.[4] discussed the influence of word order on the NMT model and concluded that it is important to maintain the consistency between an input source word order and the output target word orders, to improve the translation accuracy.

Murthy et al.[7] proposed a transfer learning approach for NMT, that trains an NMT model on an assisting language-target language pair, and improves the translation quality in extremely low-resource scenarios.

Nevertheless, both methods above rely on separately pretraining a translation model using a large-scale parallel corpus, and handle the preordering based on the syntax tree.

Zhu et al.[16] discussed a framework that focuses on the translation task limited to a small-scale corpus using preordering and highly accurate word alignment in low-resource translation. In their work, they used an SMT model as a solution for translation and received a better result compared with Transformer. But they did not explore the use of the seq2seq large-scale pretrained model.

Our work focuses on the low-resource translation task and uses the large-scale pretrained multilingual model for fine-tuning not only the preordering but also the translation procedure.

3 Using Seq2seq Models for Preordering and Translation

3.1 Seq2seq Models

Seq2seq can be described as a sequence input of the source sequence $S = \{s_1, s_2, \ldots, s_k\}$ to a sequence output of the target $T = \{t_1, t_2, \ldots, t_m\}$, where $s_i(i = 1, \ldots, k)$ and $t_j(j = 1, \ldots, m)$ represent the tokens in the source sequence and the target sequence, respectively.

Seq2seq models are basically composed of an encoder and a decoder [11, 2]. The encoder does a high-dimensional vector conversion of the input sequence, and the decoder maps the high-dimensional vectors into the output dictionary from the encoder's output. This process has applications in tasks such as machine summarization [10], question-answering systems [15], and machine translation [11]. Therefore, we also tried to use the seq2seq model to conduct preordering and translation.

3.2 Seq2seq Models for Preordering

3.2.1 Preordering Process

The preordering process transforms the orders of the tokens in a source sequence to those of the tokens in its target sequence before translation is performed. An example of transferring a Japanese sentence is shown in Figure 1.

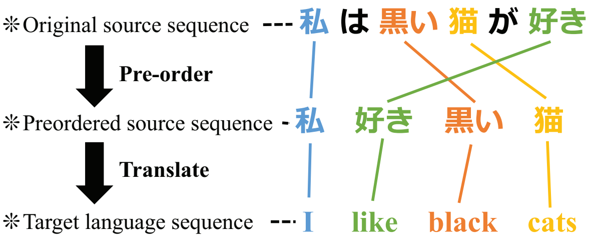
※Original source sequence --- 私 は 黒い 猫 が 好き
Pre-order
※Preordered source sequence - 私 好き 黒い 猫
Translate
※Target language sequence --- I like black cats

Figure 1 Transform the word order of the source Japanese language to the target English language before translation.

For the preordering procedure, we use mT5 [14] and mBART [6], which now are kinds of state-of-the-art seq2seq models. Both of them have similar self-attention-based encoder-decoder structures, with a slight difference in the task of pretraining. The output of mT5 and mBART are mapped to the entire dictionary rather than the input dictionary.

3.2.2 Reordered Training Data

We followed our previous work [16] for conducting the training data for preordering as in Figure 1.

For the model input, we use the original source sequence. On the output side, we simply ignored those NULL-aligned tokens, because they were not aligned with any tokens on the target side. Specifically, according to the word alignment, the Japanese sentence ” 私 (I) は 黒い (black) 猫 (cat) が 好き (like) ” can be easily preordered into the English order of ” 私 (I) 好き (like) 黒い (black) 猫 (cat)” with the alignments of (私-*I*), (黒い-*black*), (猫-*cat*) and (好き-*like*) based on the word alignment. Therefore, we ignore ” は” and ” が” in the preordered sequence, because they were not aligned to any tokens. Thus, after removing ” は” and ” が” in the output side of preordered sequence, the training pair is ” 私 (I) は 黒い (black) 猫 (cat) が 好き (like)” and ” 私 (I) 好き (like) 黒い (black) 猫 (cat) ”. We use such training pairs to train order transformation seq2seq neural networks.

3.3 Seq2seq Model for Translation

For the translation process, we use mBART as the base translation model. To compare different input results, we tried multiple input patterns, which are original input, preordered input, tagged input, concatenated input, and mixed input. Original input uses the original source language sequence as input, and outputs the target language sequence. We see this pattern as the seq2seq translation baseline.

- **Original input**: Original source sequence ⇒ Target language sequence

Preordered input uses the preordered source language sequence as input, and outputs the target language sequence.

- **Preordered input**: Preordered source sequence ⇒ Target language sequence

Tagged input uses both original and preordered source language sequences as input but carries the sequence type tag at the head of the sequence (for example, using [ord] and [pre] to represent the original sequence and preordered sequence). Note that the size of the training set is two of the baseline, because of using original input and preordered input separately.

- **Tagged input**: [ord] Original source sequence ⇒ Target language sequence
  [pre] Preordered source sequence ⇒ Target language sequence

Concatenated input merges the original and preordered source language sequence into one sequence but split by a learnable symbol.

- **Concatenated input**: Original source sequence [SEP] Preordered source sequence ⇒ Target language sequence

Mixed input is a combination of the three types of input above, with the addition of a learning process for preordering. To enable the model to distinguish the expected output from the different inputs, we put a type tag before each input similar to **Tagged input**.

- **Mixed input**: [ord2pre] Original source sequence ⇒ Preordered source sequence
  [ord2tgt] Original source sequence ⇒ Target language sequence
  [pre2tgt] Preordered source sequence ⇒ Target language sequence
  [concat2tgt] Original source sequence [SEP] Preordered source sequence ⇒ Target language sequence

4 Experiments

4.1 Dataset

We use ALT [1] Japanese-English and in-house Chinese-Japanese parallel data as our base dataset in our seq2seq experiments[2]. For word alignment, we choose to use the word alignment data based on human annotations. The data is split into the training, validation, and test parts. Each of them in ALT data includes parallel sequence pairs of 18K, 1K, and 1K. The in-house data includes parallel sequence pairs of 2K, 0.5K, and 0.5K.

4.2 Preordering Setting

Training data for seq2seq preordering is made by manual word alignment as described in section 3.2. We compare preordering results using RIBES [3] between mT5-small, mT5-base, mT5-large and mBART-large[3]. Every model is ensured to be trained for 40,000 steps, with a training batch size of 16, and a learning rate of 3e-5. Furthermore, due to the mT5-large obtained the best preordering result with the condition of a batch size of 16, we trained another mT5-large with a batch size of 32[4]. We generate the preordered sequence using the model of the maximum BLEU score which is evaluated on the validation parts.

Table 1 RIBES result of seq2seq model trained by manual word alignments of transferring Japanese order into English order and opposite.

Model	Training Batch Size	Ja⇒En	En⇒Ja
mT5-small	16	0.876	0.872
mT5-base	16	0.895	0.889
mT5-large	16	0.901	0.905
mT5-large	32	0.904	0.909
mBART-large	16	0.883	0.894

4.3 Translation Setting

We trained mBART models for translation using Fairseq [5], while for each input pattern, every model is trained for 40,000 steps, with a max input length of 1024 and a learning rate of 3e-5 (the same number of update steps and learning rate with the preordering process). For the generation process, we use the model with minimum label-smoothed cross-entropy loss on the validation set to generate the target translation. For preordering inputs, we use sequences generated by mT5-large, which is trained with a batch size of 32.

Table 2 BLEU score of different models when applying the preorder input pattern for translating Ja-En pairs.

Preordering Model	Training Batch Size	Ja⇒En	En⇒Ja
Oracle	-	34.82	34.27
mT5-small	16	21.44	26.06
mT5-base	16	24.38	27.68
mT5-large	16	24.83	28.48
mT5-large	32	25.22	28.34
mBART-large	16	23.28	27.28

5 Result

5.1 Preordering Performance

Table 1 shows the RIBES result between different seq2seq models. As we can see, the RIBES score of mT5-large models have broken through 0.9, regardless of

[1] https://www2.nict.go.jp/astrec-att/member/mutiyama/ALT/

[2] In ALT, for language pairs other than Ja-En, there is no supervised word alignment data available.

[3] All pretrained seq2seq models are downloaded from the public Huggingface library.

[4] We also tried training mT5-large with a batch size of 64, but for preordering process, the result is lower than training with a batch size of 32.

[5] https://github.com/facebookresearch/fairseq

Table 3 BLEU scores between the different Model input types and input order. Oracle rearranges the test set according to the manual word alignment data, which is the most perfect data. Rows with ***Tagged*** as model input represent the mixed input pattern result, while the results in parentheses represent the BLEU score of tagging only on the original order and preordered sequence. † represents the significant difference ($p < 0.05$) with baseline using mT5-large.

Model Input	Input	Preorder Model	Ja-En	En-Ja	Ja-Zh	Zh-Ja
Normal	Original Order (Baseline)	-	25.74	29.32	14.11	17.93
	Preordered	Oracle	34.82	34.27	17.19	22.73
		mT5-large	25.22	28.34	14.25	18.46†
	Concatenation	Oracle	35.53	35.74	17.77	22.87
		mT5-large	25.62	29.61	**14.71**†	19.22†
Tagged	Original Order	-	25.78(25.81)	28.94(29.21)	13.96(14.16)	18.80†(17.95)
	Preordered	Oracle	33.67(33.64)	33.48(33.61)	15.86(16.13)	20.81(20.86)
		mT5-large	25.13(25.51)	28.18(28.45)	13.68(13.98)	18.33(17.83)
	Concatenation	Oracle	35.01	35.07	16.14	22.03
		mT5-large	**25.92**	**29.66**	14.35	**19.41**†

Table 4 RIBES score when transferring the original source sequence to preordered source sequence using mT5-large.

	Ja⇒En	En⇒Ja	Ja⇒Zh	Zh⇒Ja
RIBES	0.904	0.909	0.927	0.919
Unigram Precision	0.91	0.92	0.89	0.83
Normalized Kendall's Tau	0.93	0.94	0.96	0.97
Brevity Penalty	0.96	0.95	0.93	0.95

whether the model is trained with a batch size of 16 or 32. Table 4 contains the RIBES score of transferring the original source sequence to preordered source sequence using mT5-large. Meanwhile, thanks to the pretraining task, the preordering result fully demonstrates the feasibility of using the seq2seq model in the preordering task.

5.2 Translation Performance

As shown in Table 3, the concatenated inputs acquire the highest BLEU score [8] compared to other input patterns. For comparison experiments, we conducted experiments using **oracle**. **Oracle** means that the test set is also preordered using manual word alignment data instead of being generated by the seq2seq model. Based on the results, **oracle** outperformed the baseline by a wide margin. Although this result is impractical, it still shows the possibility of applying our method to the seq2seq model. Table 2 shows the BLEU score of different models when applying the preorder input pattern for translating Ja-En pairs. It is apparent that the final translation result corresponds with the performance of preordering. The higher the quality of the preordering, the better the final translation will gain.

In addition, in the results of preordering using mT5-large, the translation quality of the concatenated input is better than that of the other inputs and baseline. By combining the original and preorder inputs into one sequence, the models could learn more about their relative positions.

For Chinese and Japanese data, the proposed approach is better than the baseline, which we believe is due to the size of the dataset. For the pretrained seq2seq model, the 18,000 training data of ALT is sufficient for fine-tuning with the original source input. However, under the condition of 2,000 parallel data, the model could learn more translation rules through the preordered input.

6 Conclusion

In this paper, we proposed to utilize seq2seq multilingual pretrained models for the process of preordering and translation. We used mT5-large to generate preordering sequences and mBART to translate. In our experiments, we estimated our approach on ALT Ja-En pairs and in-house Zh-Ja pairs. For the results, Our method is effective in the experiments using Chinese and Japanese parallel data and is mostly equal to the baseline in Japanese and English pairs. In our future work, we will try the method of data expansion on the pretraining model, and we will also transplant our method to mT5s to evaluate the translation accuracy.

References

[1] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proc. NAACL**, pp. 4171–4186, 2019.

[2] S. Hochreiter and J. Schmidhuber. Long short-term memory. **Neural Computation**, Vol. 9, No. 8, p. 1735–1780, 1997.

[3] H. Isozaki, T. Hirao, K. Duh, K. Sudoh, and H. Tsukada. Automatic evaluation of translation quality for distant language pairs. In **Proc. EMNLP**, pp. 944–952, 2010.

[4] Y. Kawara, C. Chu, and Y. Arase. Recursive neural network-based preordering for statistical machine translation and its analysis. **Journal of Natural Language Processing**, Vol. 26, No. 1, pp. 155–178, 2019.

[5] Z. Lin, X. Pan, M. Wang, X. Qiu, J. Feng, H. Zhou, and L. Li. Pre-training multilingual neural machine translation by leveraging alignment information. In **Proc. EMNLP**, pp. 2649–2663, 2020.

[6] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer. Multilingual denoising pre-training for neural machine translation. **Transactions of the ACL**, pp. 726–742, 2020.

[7] R. Murthy, A. Kunchukuttan, and P. Bhattacharyya. Addressing word-order divergence in multilingual neural machine translation for extremely low resource languages. In **Proc. NAACL**, pp. 3868–3873, 2019.

[8] K. Papineni, S. Roukos, T. Ward, and W. Zhu. Bleu: a method for automatic evaluation of machine translation. In **Proc. 40th ACL**, pp. 311–318, 2002.

[9] H. Riza, M. Purwoadi, Gunarso, T. Uliniansyah, A. Ti, S. Aljunied, L. Mai, V. Thang, N. Thai, V. Chea, R. Sun, S. Sam, S. Seng, K. Soe, K. Nwet, M. Utiyama, and C. Ding. Introduction of the Asian language treebank. In **Proc. Oriental COCOSDA**, pp. 1–6, 2016.

[10] T. Shi, Y. Keneshloo, N. Ramakrishnan, and Chandan K. Reddy. Neural abstractive text summarization with sequence-to-sequence models. Vol. 2, 2021.

[11] I. Sutskever, O. Vinyals, and Q. Le. Sequence to sequence learning with neural networks. In **NIPS**, Vol. 27, 2014.

[12] B. Tom, M. Benjamin, and et al. Nick, R. Language models are few-shot learners. In **Advances in NIPS**, Vol. 33, pp. 1877–1901, 2020.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, U. Kaiser, and I. Polosukhin. Attention is all you need. In **Proc. 30th NIPS**, pp. 5998–6008, 2017.

[14] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In **Proc. 15th NAACL**, pp. 483–498, 2021.

[15] J. Yin, X. Jiang, Z. Lu, L. Shang, H. Li, and X. Li. Neural generative question answering. In **Proceedings of the Workshop on Human-Computer Question Answering**, pp. 36–42, 2016.

[16] J. Zhu, Y. Wei, T. Tamura, T. Utsuro, and N Masaaki. A framework for low resource language translation based on smt and highly accurate word alignment. In **The Association for Natural Language Processing**, pp. 1312–1316, 2022.

A Appendix

Table 5 shows our actual transformation results. In the first result, although most of the preordered tokens are arranged in relatively correct places, those have meaningful tokens such as “金曜日 夜” and “停車場” are not placed correctly. However, in the second result, the generated preordered sequence is exactly the same as the reference preordered sequence.

Table 5 Results for preordering generated by mT5-large.

Japanese original sequence	彼 は 金曜日 の 夜 に サウス メルボルン の 停車場 から トラム を 盗ん だ こと でも 訴え られ て いる 。
English target sequence	He is also accused of stealing a tram on Friday night , from South Melbourne depot .
Reference preordered sequence	彼 て いる でも 訴え られ でも 盗ん だ こと トラム に 金曜日 夜 から サウス メルボルン 停車場 。
Generated preordered sequence	彼 られ て いる でも 訴え こと 盗ん だ トラム から 停車場 の サウス メルボルン に 金曜日 夜 。
Japanese original sequence	知事 の ジョン ・ コーザイン 氏 は 、 幹 細胞 研究 に 2 億 7000 万 ドル を 提供 する 法案 を 発表 し た 。
English target sequence	Governor Jon Corzine announced a bill that would provide $ 270 million to stem cell research .
Reference preordered sequence	知事 ジョン コーザイン 発表 し た 法案 提供 する ドル 2 億 7000 万 に 幹 細胞 研究 。
Generated preordered sequence	知事 ジョン コーザイン 発表 し た 法案 提供 する ドル 2 億 7000 万 に 幹 細胞 研究 。