Pivot Translation for Zero-resource Language Pairs Based on a Multilingual Pretrained Model

Kenji Imamura kenji.imamura@nict.go.jp
Masao Utiyama mutiyama@nict.go.jp
Eiichiro Sumita eiichiro.sumita@nict.go.jp
National Institute of Information and Communications Technology, Hikari-dai, Seika-cho, Soraku-gun, Kyoto, 619-0289, Japan

Abstract

A multilingual translation model enables a single model to handle multiple languages. However, the translation qualities of unlearned language pairs (i.e., zero-shot translation qualities) are still poor. By contrast, pivot translation translates source texts into target ones via a pivot language such as English, thus enabling machine translation without parallel texts between the source and target languages.

In this paper, we perform pivot translation using a multilingual model and compare it with direct translation. We improve the translation quality without using parallel texts of direct translation by fine-tuning the model with machine-translated pseudo-translations. We also discuss what type of parallel texts are suitable for effectively improving the translation quality in multilingual pivot translation.

1 Introduction

Multilingual neural network models are models in which multiple languages are learned in a single model, and are useful for machine translation and cross-lingual language processing. Multilingual models utilize resources of similar languages (e.g., those in the same language family) and thus provide a relatively high degree of accuracy for even low-resource languages.[1] Machine translation is performed according to a combination of a source language and a target language, and therefore, language-specific models require a model for each possible combination of languages. By contrast, a multilingual model can handle all combinations of source and target languages and is therefore easier to manage. The potential usefulness of the multilingual model has led to the development of several encoder–decoder models pretrained using parallel corpora.

For example, a multilingual translation model pretrained with the OPUS-100 corpus (Zhang et al., 2020)[2] has been developed. This is a multilingual model that translates between English and any of 100 languages (i.e., an English-centric model). The M2M-100 model (Fan

[1] Under high-resource conditions, language-specific models are generally more accurate than multilingual models. This is called the curse of multilinguality.

[2] `https://github.com/bzhangGo/zero/tree/master/docs/multilingual_laln_lalt#pretrained-multilingual-models-many-to-many`

Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track, pages 348–359
September 4–8, 2023, Macau SAR, China.

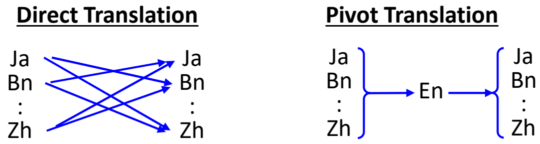

Figure 1: Direct translation and pivot translation.

et al., 2020)[3] also handles 100 languages. It is pretrained in 2,200 directions by adding parallel corpora that do not include English. The NLLB-200 models (Team et al., 2022)[4], which are extended from the M2M-100 model, handle around 200 languages and pretrained from parallel corpora of over 2,600 language pairs. Although the mBART models (Liu et al., 2020; Lewis et al., 2020) are also encoder–decoder pretrained models, they are trained using monolingual corpora only.

If we handle 100 languages for translation, this results in a total of 9,900 translation directions. Even if a multilingual translation model is used, translation quality for language pairs not trained by parallel corpora (called the zero-shot translation (Johnson et al., 2017)) is likely to be insufficient for practical use.

Pivot translation (Utiyama and Isahara, 2007; Cohn and Lapata, 2007) is a known method of achieving moderate quality translation in language pairs for which it is difficult to obtain parallel corpora. The method uses a pivot language between the source and target languages. Source texts are first translated into the pivot language, and then the pivot texts are translated into the target language (Figure 1). English is often used as the pivot language given the benefit of its rich set of parallel corpora. Although pivot translation is often used in statistical machine translation, it is also applicable to neural machine translation. By applying a multilingual pretrained model to the pivot translation, a single model can achieve a practical level of translation, even between languages without parallel corpora (zero-resource language pairs; between non-English languages in most cases).

In this paper, we apply pivot translation based on a multilingual pretrained model to zero-resource language pairs. This study aims to clarify the following points.

Q1 Comparison of the translation quality of pivot and direct translation. If parallel corpora exist, which has better translation quality? In creating new parallel corpora for zero-resource language pairs, should we prefer pivot or direct translation?

Q2 Pivot translation is performed in two stages: translation from the source language to the pivot (first stage), and then translation from the pivot to the target language (second stage). We can use different models in each stage. In regard to improvement, which model should be addressed first, the former or the latter?

Q3 Using pivot translation, we can create pseudo-parallel corpora (i.e., synthetic parallel corpora) between all language pairs of the supported languages if we have monolingual corpora. When we use pseudo-parallel corpora to fine-tune a multilingual model, what level of translation quality can be achieved with respect to manually created parallel corpora?

Q4 When generating pseudo-parallel corpora, which monolingual corpus should be used as the original language, the source, pivot, or target language?

[3] `https://github.com/facebookresearch/fairseq/tree/main/examples/m2m_100`

[4] For example, `https://huggingface.co/facebook/nllb-200-3.3B`

XX Language	Family[5]	Script[5]	CC-100 (monolingual)	OPUS-100 (En-XX)	CCAligned (En-XX)
English (En)	Indo-European	Latin	1,858 M	-	-
Japanese (Ja)	Japonic	Chinese and Kana	393 M	1.0 M	15.0 M
Bengali (Bn)	Indo-European	Bengali-Assamese	54 M	1.0 M	3.5 M
Indonesian (Id)	Austronesian	Latin	969 M	1.0 M	15.7 M
Khmer (Km)	Austroasiatic	Khmer	6.6 M	0.1 M	0.4 M
Lao (Lo)	Kra-Dai	Lao	2.6 M	-	0.2 M
Malay (Ms)	Austronesian	Latin	66 M	1.0 M	5.4 M
Myanmar (My)	Sino-Tibetan	Burmese	2.0 M	0.02 M	0.3 M
Thai (Th)	Kra-Dai	Thai	295 M	1.0 M	10.7 M
Tagalog (Tl)	Austronesian	Latin	27 M	-	6.6 M
Vietnamese (Vi)	Austroasiatic	Latin	939 M	1.0 M	12.4 M
Chinese (Zh)	Sino-Tibetan	Simplified Chinese	169 M	1.0 M	15.2 M

Table 1: Training corpus sizes of the languages used in this paper for the basic model. The values indicate the number of sentences.

Hereafter, Section 2 describes the English-centric multilingual pretrained model used in this study. Section 3 investigates the above questions through experiments.

2 Multilingual Pretrained Model Used in This Study

For this study, we newly trained an English-centric model to focus on translating zero-resource language pairs. We call this the "basic model." Specifically, this model corresponds to the 103 languages covered by the CC-100 corpus (Conneau et al., 2020; Wenzek et al., 2020), and the OPUS-100 corpus (Aharoni et al., 2019; Tiedemann, 2012) or the CCAligned v1 corpus (El-Kishky et al., 2020). CC-100 is a monolingual corpus, and OPUS-100 and CCAligned are parallel corpora. All corpora are based on Web crawl data. Table 1 shows the corpus sizes used for training the basic model (only the languages used in this paper). The number of sentences in CC-100 is for monolingual sentences. OPUS-100 and CCAligned are the number of parallel sentences between English (En) and one of the languages other than English (XX languages).

2.1 Procedure for Building the Basic Model

We built the basic model using the following procedure.

1. Following the method of Wang et al. (2020), the word embeddings of the mBART-50 model were extended to the 109 languages covered by the CC-100 corpus. The extended embeddings were randomly initialized.

2. All corpora were tokenized by SentencePiece (Kudo and Richardson, 2018) using the model attached to mBART-50 (250K subwords). Then, denoising training was additionally performed on the above extended model using the CC-100 corpus. This is the same as the training of the mBART-50.

3. The model was trained using parallel sentences from/to English in the OPUS-100 and CCAligned corpora. Because the corpus sizes for each language pair are substantially different, we applied temperature sampling (Arivazhagan et al., 2019) in the training (inverse

[5] https://en.wikipedia.org/

XX Language	En → XX BLEU	En → XX ChrF2	XX → En BLEU	XX → En ChrF2
Japanese (Ja)	26.0	36.0	26.2	57.2
Bengali (Bn)	9.5	44.5	28.2	56.8
Indonesian (Id)	41.8	67.5	43.0	67.6
Khmer (Km)	52.7 †	47.6	27.0	55.2
Lao (Lo)	27.7 †	24.9	6.3	27.7
Malay (Ms)	44.1	69.1	44.5	68.3
Myanmar (My)	40.9 †	36.2	19.3	49.1
Thai (Th)	53.0 †	48.2	26.9	56.4
Tagalog (Tl)	30.7	59.1	39.2	63.3
Vietnamese (Vi)	39.7	57.8	36.0	61.9
Chinese (Zh)	35.0	31.0	24.8	56.2
Average Score (11 Languages)	36.5	47.4	29.2	56.3
(FYI) Average Score of M2M-100	28.5	40.9	26.0	52.0

Table 2: Translation quality between English (En) and foreign languages (XX) in the basic model. † mark indicates the BLEU scores tokenized into characters because sacreBLEU cannot tokenize the languages. The language-dependent default tokenizers of sacreBLEU were used for other languages.

temperature coefficient $1/T = 0.7$). Namely, we down-sampled training sentences in the language pairs of the large corpora, and up-sampled them in the language pairs of the small corpora.

The basic model has the same structure as the mBART-50 model except for the word embedding table. Thus, the encoder and decoder consist of 12 layers each, 1,024 hidden dimensions, 4,096 FFN dimensions, 16 heads, and 250K word embeddings. Note that the source and target language IDs must be given during translation because the mBART-50 requires the source and target language tags.

2.2 Translation Quality between English and Foreign Languages in the Basic Model

Table 2 lists the quality of translation between English and the selected languages targeted in this study using the basic model.

The Asian Language Treebank (ALT) Parallel Corpus (Riza et al., 2016), which is used in the experiments described in the next section, was translated by the basic model using the direct translation, and the translation qualities were evaluated by sacreBLEU (Post, 2018). Note that several languages are not supported by the tokenizers in sacreBLEU. We used sacreBLEU to evaluate translations in such languages using the character tokenization († marks in Table 2). In addition, we also report the ChrF scores (Popović, 2015) ($\beta = 2$; notated as ChrF2), which are independent of the tokenizers.

For reference, the results of the M2M-100 model (Fan et al., 2020) evaluated on the same test set are also listed at the bottom of the table. The results indicate that translation quality of the basic model is, in the limited languages, better than that of the M2M-100 model on average.

3 Translation Experiments

In this study, we conducted translation experiments between Japanese (Ja) and languages other than English (XX).

	#Sentences			
Corpus	Training	Dev.	Test	Remarks
ALT	18,088	1,000	1,018	
ASPEC-JC	669,923	2,090	2,107	
ASPEC-JE	670,000	-	-	English only, selected from 3M sentences.

Table 3: Corpus size for fine-tuning

3.1 Experimental Settings

3.1.1 Corpora

In our experiments, the following parallel corpora were used to compare a zero-resource condition with a condition when direct parallel corpora exist. The corpus sizes are shown in Table 3. The sizes of the training sets indicate those after removing translations with significantly different lengths.

In the low-resource experiments, we used the ALT Parallel Corpus (Riza et al., 2016)[6]. This is a multilingual corpus that covers English (En), Japanese (Ja), Bengali (Bn), Indonesian (Id), Khmer (Km), Lao (Lo), Malay (Ms), Myanmar (My), Thai (Th), Tagalog (Tl), Vietnamese (Vi), and Simplified Chinese (Zh). This corpus contains translations from the same English WikiNews texts. Therefore, translations are also provided between languages other than English. Hence, translation experiments were conducted between Japanese and languages other than English.

In the mid-resource experiments, we used the Asian Scientific Paper Excerpt Corpus (ASPEC) (Nakazawa et al., 2016)[7], which is based on scientific paper abstracts. The ASPEC-JC corpus is a parallel corpus of Japanese and Chinese, and it does not have English counterparts. We mainly use it to evaluate the effectiveness of pseudo-translations. To generate pseudo-translations from English texts in the same domain, we also used the English part of the ASPEC-JE corpus, which is a Japanese–English parallel corpus. To match the size with ASPEC-JC, we selected 670K sentences from the entire corpus.

These corpora were tokenized by the SentencePiece model attached with the mBART-50 model, in the same way as the basic model.

3.1.2 Comparison of Methods/Systems

In this study, we compare the direct and pivot translations (Figure 1). The multilingual pre-trained model described in Section 2 is called the basic model. We compare the translation results of the basic model with those of models fine-tuned on the parallel corpora (Figure 2). Fine-tuning was performed using the parallel corpora described in Section 3.1.1.

- **+Direct Parallel Model:**
  The model fine-tuned using the direct parallel corpora of Japanese and the XX languages. In this case, we used the direct translation method because the corpora do not go through the pivot.[8]

- **+XX → En Model:**
  The model fine-tuned using the parallel corpora from the XX languages to English. The

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

[7] https://jipsti.jst.go.jp/aspec/

[8] Translation qualities among languages that have not been fine-tuned are significantly degraded due to catastrophic forgetting. Therefore, pivot translation cannot be applied.

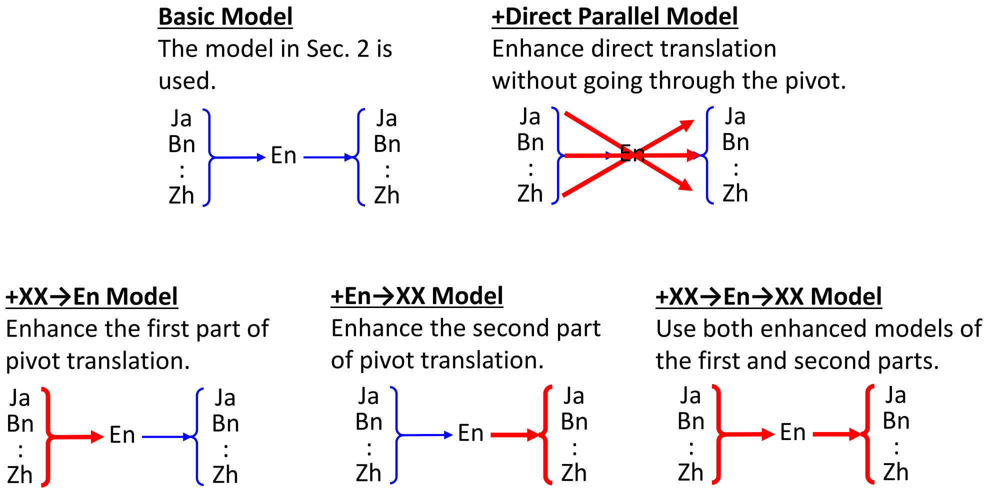
Basic Model
The model in Sec. 2 is used.
Ja
Bn
:
Zh
En
Ja
Bn
:
Zh
+Direct Parallel Model
Enhance direct translation without going through the pivot.
Ja
Bn
:
Zh
En
Ja
Bn
:
Zh
+XX→En Model
Enhance the first part of pivot translation.
Ja
Bn
:
Zh
En
Ja
Bn
:
Zh
+En→XX Model
Enhance the second part of pivot translation.
Ja
Bn
:
Zh
En
Ja
Bn
:
Zh
+XX→En→XX Model
Use both enhanced models of the first and second parts.
Ja
Bn
:
Zh
En
Ja
Bn
:
Zh

Figure 2: Types of fine-tuned models.

first stage of pivot translation was enhanced. The basic model was used in the second stage of pivot translation.

- **+En → XX Model:**
The model fine-tuned using the parallel corpora from English to the XX languages. The second stage of pivot translation was enhanced. The basic model was used in the first stage of pivot translation.

- **+XX → En → XX Model:**
The +XX → En and +En→XX models were used for the first and second stages of pivot translation, respectively.

3.1.3 Pseudo-translations

Using pivot translation, we can perform machine translation even for zero-resource language pairs. Therefore, we can construct direct parallel corpora by machine translation from monolingual corpora. In this study, we also compare the cases fine-tuned by manual translation and pseudo-translations.

All pseudo-translations were generated by the basic model. Although word sampling (Imamura et al., 2018; Edunov et al., 2018) improves translation quality during back-translation (Sennrich et al., 2016) because of increasing the translation diversity, we must switch the translation methods depending on the direction. For the sake of simplicity, we used one-best translations for pseudo-translations in our experiments.[9]

3.1.4 Other Settings

The hyperparameters used during fine-tuning and testing are listed in Table 4. We used one-best translations in both stages of pivot translation.

We used BLEU (Papineni et al., 2002) and ChrF (Popović, 2015) for the evaluation using sacreBLEU (Post, 2018).

[9] In back-translation (target-to-source), final translation quality improves when we increase translation diversity using word sampling. By contrast, one-best translation is preferred in sequence-level knowledge distillation (source-to-target) (Kim and Rush, 2016; Kim et al., 2019)). To apply this distinction, it is necessary to switch the generating method depending on the translation direction.

Type	Value
Fine-tuning	Temperature sampling (Arivazhagan et al., 2019): $1/T = 0.7$, Loss: label_smoothed_cross_entropy=0.1, Dropout: 0.3, Warmup: around one epoch, LR: 0.00008, inverse_sqrt, Early stopping: ten epochs, Batch size: 8K tokens, Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.99$, $\epsilon = 10^{-6}$)
Test (Inference)	Beam width: 10 One-best translation

Table 4: List of hyperparameters.

3.2 Experimental Results

Tables 5 and 6 present the results for the ALT and ASPEC-JC corpora, respectively. The results of the ALT corpus show the average scores on all XX languages. No. 2 is the baseline result using pivot translation.

3.2.1 Pivot Translation vs. Direct Translation

Regardless of language direction, the BLEU scores of the direct translation using the basic model (No. 1) in the ALT and ASPEC-JC corpora are extremely low. This is caused by zero-shot translation. However, when we use pivot translation with the same model (No. 2), the BLEU scores improve by over 12 points. Namely, moderate transition can be obtained even for the zero-resource language pairs.

By contrast, when we fine-tune the model using the direct parallel corpus (No. 6a), the BLEU scores improved by more than 3 points and 18 points in the ALT and ASPEC-JC corpora, respectively, when compared with the results of No. 2. These scores were not the highest in the results of the ALT corpus. We assume that this is because the number of parallel sentences was small (18K sentences), and the improvement obtained using pivot translation with the English-centric model surpassed these results. By contrast, the BLEU scores for the ASPEC-JC corpus were the highest. If we can acquire a corpus of medium size, we should prepare a direct parallel corpus to improve the translation quality.

3.2.2 First and Second Stages of Pivot Translation

No. 3a and 4a in Table 2 are the results of the fine-tuned basic models using the parallel corpora between English and XX languages. No. 3a and 4a enhanced the first and second stages of pivot translation, respectively.

Compared with the baseline results, the results of No. 2 and 3a are not significantly different. By contrast, the BLEU score of No. 4a improved the score of No. 2 by over 4 points. In other words, the translation quality was efficiently improved when we enhanced the second stage of pivot translation. This is because the basic model was trained using corpora of Web crawl data, and their domain was different from that of the test set. Therefore, the quality of the output sentences was improved by the domain adaptation of the second stage of pivot translation.

			Ja → XX		XX → Ja	
		Translation	BLEU	ChrF2	BLEU	ChrF2
No.	Model	Method	(Avr.)	(Avr.)	(Avr.)	(Avr.)
1	Basic Model	Direct	0.5	6.3	0.1	0.8
2	Basic Model	Pivot	27.0	40.4	17.3	26.9
3a	+XX → En (Manual)	Pivot	26.8	40.0	18.8	28.3
4a	+En → XX (Manual)	Pivot	**33.1**	**45.6**	21.4	30.2
5a	+XX → En → XX (Manual)	Pivot	32.9	45.4	**23.1**	**32.0**
6a	+Direct Parallel (Manual)	Direct	32.2	44.8	21.2	30.3

Table 5: Translation results for the ALT corpus. The bold values indicate the highest score among the models and methods.

		Translation	Ja → Zh		Zh → Ja	
No.	Model	Method	BLEU	ChrF2	BLEU	ChrF2
1	Basic Model	Direct	0.0	0.0	0.1	0.2
2	Basic Model	Pivot	19.4	17.6	12.0	21.8
3b	+XX → En (Pseudo)	Pivot	19.6	17.7	12.4	22.2
4b	+En → XX (Pseudo)	Pivot	26.8	23.2	19.2	27.8
5b	+XX → En → XX (Pseudo)	Pivot	27.2	23.6	19.8	28.4
6b	+Direct Parallel (Pseudo)	Direct	31.0	26.4	24.5	32.9
6a	+Direct Parallel (Manual)	Direct	**37.6**	**32.0**	**33.4**	**41.6**

Table 6: Translation results for the ASPEC-JC corpus. The bold values indicate the highest score among the models and methods.

3.2.3 Data Augmentation Generated by Machine Translation

The results for No. 3a, 4a, 5a, and 6a in Tables 5 and 6 are the results of the fine-tuned models with the manually created parallel corpora. The results for No. 3b, 4b, 5b, and 6b in the tables are the results of the fine-tuned models with the pseudo-parallel corpora generated by machine translation.

When we enhance the first stage of the pivot translation using the pseudo-translations (No. 3b), the translation quality rarely changed from that of the baseline (No. 2).

By contrast, the translation qualities were significantly improved when we enhanced the second stage of pivot translation (No. 4b) or fine-tuned using the direct parallel corpus (No. 6b) despite using pseudo-data, even though the quality scores did not reach those of manual translation (No. 6a). However, the pseudo-translations can be generated from monolingual corpora. If we actively use pseudo-translations, the translation quality can be improved even for zero-resource language pairs.

3.2.4 Original Language of Pseudo-Translations

When we generate pseudo-translations from monolingual corpora, either the source, target, or pivot language can be used as the original language. In the experiments described in this section, we created pseudo-translations from various original languages and fine-tuned the basic model. When the source or target language was used as the original one, the ASPEC-JC training set was used. When the pivot language was used as the original one, the English part of ASPEC-JE

		Translation	Original	Ja → Zh		Zh → Ja	
No.	Model	Method	Language	BLEU	ChrF2	BLEU	ChrF2
3b	+ XX → En	Pivot	Source	19.6	17.7	12.4	22.2
3c			Pivot	20.1	17.9	12.8	23.1
4b	+ En → XX	Pivot	Target	26.8	23.2	19.2	27.8
4c			Pivot	19.2	17.4	11.6	21.5
5b	+ XX → En → XX	Pivot	Source & Target	27.2	23.6	19.8	28.4
5c			Pivot	19.8	17.5	12.3	22.9
6b	+ Direct Parallel	Direct	Target	31.0	26.4	24.5	32.9
6c			Pivot	21.7	19.2	15.0	24.9
6d			Source	20.4	18.4	13.1	23.0

Table 7: Translation qualities when pseudo-translations with different original languages are used. The underlined values indicate the highest score of the same model/translation method.

was used.

Table 7 presents the translation quality results obtained on ASPEC-JC and is summarized as follows.

- In the +XX → En model, the pseudo-translations generated from the pivot language (i.e., the translations from the pivot to the source language) had a higher translation quality.
- In the +En → XX model, the quality of the pseudo-translations generated from the target language (i.e., translations from the target to the pivot language) was significantly higher.
- In the +XX → En → XX model, the quality of the pseudo-translations generated from the source and target languages (i.e., translations from the target to the pivot and from the source to the pivot language) was significantly higher.
- In the +Direct Parallel model, the translation quality was highest in the order of the target, pivot, and source languages.

For all these results, the translation qualities were high when we fine-tuned the model with the pseudo-translations translated in the direction opposite that to be tested. Even in pivot translation, the monolingual corpora of the target languages should be collected, if possible.

4 Conclusions

Using the pivot translation, we can translate texts even for zero-resource language pairs. Moreover, we can improve the translation quality without changing the zero-resource condition because we can generate pseudo-parallel corpora from monolingual corpora.

In this study, we applied pivot translation to zero-resource language pairs using a multilingual pretrained model. The answers to the questions studied in this work are summarized as follows.

A1 Comparing pivot translation with direct translation, the quality of pivot translation is higher than that of direct translation when the parallel corpus size is very small. When the corpus size is large, the quality of direct translation increases. If we can acquire a corpus of medium size, we should prepare a direct parallel corpus to improve the translation quality.

A2 Comparing the first and second stages of pivot translation, it is better to enhance the second stage to improve quality.

A3 It is possible to improve translation quality using pseudo-translations generated by pivot translation.

A4 When generating pseudo-translations, it is better to generate them from monolingual corpora of the target language.

The fact that zero-resource language pairs can be translated is helpful when we extend our machine translation to new languages. For example, we can check the quality of a newly created parallel corpus by back-translation, or we can post-edit pseudo-translations to create direct parallel corpora.

We plan to extend multilinguality while appropriately using direct and pivot translation.

Acknowledgment

Part of this work was conducted under the commissioned research program 'Research and Development of Advanced Multilingual Translation Technology' in the 'R&D Project for Information and Communications Technology (JPMI00316)' of the Ministry of Internal Affairs and Communications (MIC), Japan.

References

Aharoni, R., Johnson, M., and Firat, O. (2019). Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Arivazhagan, N., Bapna, A., Firat, O., Lepikhin, D., Johnson, M., Krikun, M., Chen, M. X., Cao, Y., Foster, G., Cherry, C., Macherey, W., Chen, Z., and Wu, Y. (2019). Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv e-print*, 1907.05019.

Cohn, T. and Lapata, M. (2007). Machine translation by triangulation: Making effective use of multi-parallel corpora. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 728–735, Prague, Czech Republic. Association for Computational Linguistics.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Edunov, S., Ott, M., Auli, M., and Grangier, D. (2018). Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

El-Kishky, A., Chaudhary, V., Guzmán, F., and Koehn, P. (2020). CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5960–5969, Online. Association for Computational Linguistics.

Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., Baines, M., Celebi, O., Wenzek, G., Chaudhary, V., Goyal, N., Birch, T., Liptchinsky, V., Edunov, S., Grave, E., Auli, M., and Joulin, A. (2020). Beyond english-centric multilingual machine translation. *arXiv e-print*, 2010.11125.

Imamura, K., Fujita, A., and Sumita, E. (2018). Enhancement of encoder and attention using target monolingual corpora in neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 55–63, Melbourne, Australia. Association for Computational Linguistics.

Johnson, M., Schuster, M., Le, Q., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., and Dean, J. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5(0):339–351.

Kim, Y. and Rush, A. M. (2016). Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas. Association for Computational Linguistics.

Kim, Y. J., Junczys-Dowmunt, M., Hassan, H., Fikri Aji, A., Heafield, K., Grundkiewicz, R., and Bogoychev, N. (2019). From research to production and back: Ludicrously fast neural machine translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 280–288, Hong Kong. Association for Computational Linguistics.

Kudo, T. and Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Nakazawa, T., Yaguchi, M., Uchimoto, K., Utiyama, M., Sumita, E., Kurohashi, S., and Isahara, H. (2016). ASPEC: Asian scientific paper excerpt corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2204–2208, Portorož, Slovenia. European Language Resources Association (ELRA).

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Popović, M. (2015). chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Post, M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Riza, H., Michael Purwoadi, G., Uliniansyah, T., Ti, A. A., Aljunied, S. M., Mai, L. C., Thang, V. T., Thai, N. P., Chea, V., Sun, R., Sam, S., Seng, S., Soe, K. M., Nwet, K. T., Utiyama, M., and Ding, C. (2016). Introduction of the asian language treebank. In *Oriental COCOSDA*.

Sennrich, R., Haddow, B., and Birch, A. (2016). Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Team, N., Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., Sun, A., Wang, S., Wenzek, G., Youngblood, A., Akula, B., Barrault, L., Gonzalez, G. M., Hansanti, P., Hoffman, J., Jarrett, S., Sadagopan, K. R., Rowe, D., Spruit, S., Tran, C., Andrews, P., Ayan, N. F., Bhosale, S., Edunov, S., Fan, A., Gao, C., Goswami, V., Guzmán, F., Koehn, P., Mourachko, A., Ropers, C., Saleem, S., Schwenk, H., and Wang, J. (2022). No language left behind: Scaling human-centered machine translation. *arXiv e-print*, 2207.04672.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In Chair), N. C. C., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Utiyama, M. and Isahara, H. (2007). A comparison of pivot methods for phrase-based statistical machine translation. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 484–491, Rochester, New York.

Wang, Z., K, K., Mayhew, S., and Roth, D. (2020). Extending multilingual BERT to low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2649–2656, Online. Association for Computational Linguistics.

Wenzek, G., Lachaux, M.-A., Conneau, A., Chaudhary, V., Guzmán, F., Joulin, A., and Grave, E. (2020). CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Zhang, B., Williams, P., Titov, I., and Sennrich, R. (2020). Improving massively multilingual neural machine translation and zero-shot translation. *arXiv e-print*, 2004.11867.