Improving Jejueo-Korean Translation With Cross-Lingual Pretraining Using Japanese and Korean

Francis Zheng, Edison Marrese-Taylor, Yutaka Matsuo
Graduate School of Engineering
The University of Tokyo
{francis, emarrese, matsuo}@weblab.t.u-tokyo.ac.jp

Abstract

Jejueo is a critically endangered language spoken on Jeju Island and is closely related to but mutually unintelligible with Korean. Parallel data between Jejueo and Korean is scarce, and translation between the two languages requires more attention, as current neural machine translation systems typically rely on large amounts of parallel training data. While low-resource machine translation has been shown to benefit from using additional monolingual data during the pretraining process, not as much research has been done on how to select languages other than the source and target languages for use during pretraining. We show that using large amounts of Korean and Japanese data during the pretraining process improves translation by 2.16 BLEU points for translation in the Jejueo → Korean direction and 1.34 BLEU points for translation in the Korean → Jejueo direction compared to the baseline.

1 Introduction

Low-resource machine translation has recently attracted more attention in the field of natural language processing as neural machine translation (NMT) systems typically do not perform well for low-resource languages, where parallel data are lacking (Koehn and Knowles, 2017). Current machine translation systems typically use tens or even hundreds of millions of parallel sentences as training data, but this type of data is only available for a small number of language pairs (Haddow et al., 2022). However, there are many examples of low-resource languages that have many speakers (Haddow et al., 2022), so more attention is needed in the field of machine translation to serve speakers of these languages. Additionally, for the purpose of helping to preserve language and culture and providing equitable access to technology, it is important to improve machine translation for speakers of all languages, even those that have a small number of speakers.

Jejueo (Jeju language, ISO 639-3 language code: *jje*) is a language spoken on Jeju Island, located just south of the Korean Peninsula. It is closely related to but mutually unintelligible with Korean (ISO 639-3 language code: *kor*) (Yang et al., 2020b). It was also classified as a critically endangered language by UNESCO in 2010, meaning that its youngest fluent speakers are grandparents or great-grandparents (Yang et al., 2020b). Despite academic efforts to preserve Jejueo (Yang et al., 2017; Saltzman, 2017; Yang et al., 2020a, 2018), data-driven approaches have not been explored deeply (Park et al., 2020). There are only 5,000 - 10,000 fluent speakers of Jejueo, and most of these speakers are more than 70 years old (Park et al., 2020), so it is hard to acquire Jejueo data themselves, let alone parallel data between Jejueo and Korean. Despite this scarcity of data, translation between Jejueo and Korean is an important task due to their lack of mutual intelligibility.

We propose a method that uses an mBART (Liu et al., 2020) implementation of FAIRSEQ[1] (Ott et al., 2019) and leverages the use of large amounts of linguistically similar languages during pretraining to improve the accuracy of translation between Korean and Jejueo. We show that using large amounts of Japanese and Korean monolingual data during pretraining improves translation by 2.16 BLEU points in the Jejueo → Korean direction and 1.34 BLEU points in the Korean → Jejueo direction over the baseline.

2 Related Work

Park et al. (2020) published a parallel dataset for Korean and Jejueo, described later in Section 3.1.2, and used a Transformer (Vaswani et al., 2017) with six encoder and decoder blocks and eight attention heads for translation in both directions between Korean and Jejueo. The authors used FAIRSEQ (Ott

[1] https://github.com/pytorch/fairseq

Proceedings of the 9th Workshop on Asian Translation, pages 44–50
October 17, 2022.

Table 1: Monolingual Dataset Statistics

Dataset	Description	Size	Tokens
JA	Japanese	6.6 GB	1,638,553,045
KO	Korean	5.7 GB	1,603,938,119
ZH	Chinese (written in traditional characters) data	5.9 GB	2,257,606,300
MIX	A mix of monolingual data from Bulgarian, English, French, Irish, Korean, Latin, Spanish, Sundanese, Vietnamese, and Yoruba	11.5 GB	3,206,224,170

Table 2: JIT Dataset Statistics (Park et al., 2020)

	Total	Train	Dev	Test
Parallel sentences	170,356	160,356	5,000	5,000
Jejueo words	1,421,723	1,298,672	61,448	61,603
Korean words	1,421,836	1,300,489	61,541	61,806
Jejueo word forms	161,200	151,699	17,828	18,029
Korean word forms	110,774	104,874	14,362	14,595

et al., 2019) to run their experiments and SentencePiece [2] (Kudo and Richardson, 2018) for byte-pair encoding (BPE) segmentation. They experimented with different vocabulary sizes and found that a vocabulary size of 4,000 produced the best results, establishing a new baseline for translation between Jejueo and Korean. They achieved BLEU (Papineni et al., 2002) scores of 67.70 for the Jejueo → Korean direction and 43.31 for the Korean → Jejueo direction on the test set of their parallel dataset. Then, they followed an approach by Sennrich et al. (2016), who showed that machine translation models can be improved with monolingual data, and augmented "both the source and target sides of the training set with the same number of randomly sampled Korean sentences from a Wikidump" (Park et al., 2020). This improved their BLEU scores to 67.94 for the Jejueo → Korean direction and 44.19 for the Korean → Jejueo direction on the test set of their parallel dataset.

Zheng et al. (2021) explored the use of large amounts of monolingual data during the pretraining process to improve translation between low-resource languages from the Americas and Spanish. Instead of monolingual data from either the source or target language, languages from all over the world were used in this training process to expose the model to a wide variety of linguistic features, allowing for improvements of BLEU scores that were 1.64 higher and CHRF scores that were 0.0749 higher on average than the baseline for those language pairs.

We build on this work by taking a closer look at how the selection of language for these monolingual data used during the pretraining process affects translation quality in the case of translation between Jejueo and Korean. Our methods are described in the following section.

3 Methods

3.1 Data

We experimented with four sets of monolingual data described in Table 1 and Jejueo-Korean parallel data described in Table 2. Tokenization was performed as described in Section 3.2. Details on the size of and amount of tokens used from each language in the MIX dataset can be found in Table 6 in Appendix A.

3.1.1 Monolingual Data

The monolingual datasets JA, KO, ZH, and MIX were obtained from CC100[3] (Wenzek et al., 2020; Conneau et al., 2020). The Japanese dataset JA was chosen for its similarity in syntax and vocabulary to Korean and Jejueo, and the Korean dataset KO was chosen to provide more data for one side of translation between Korean and Jejueo. The Chinese dataset ZH was selected because both Korean and Jejueo (and Japanese, for that matter) have

[2] https://github.com/google/sentencepiece

[3] http://data.statmt.org/cc-100/

Table 3: Datasets Used in Pretraining

Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
MIX	JA, KO, ZH	JA, KO	KO	JA	ZH

loanwords from Chinese even though Chinese has a vastly different syntax and writing system.

The dataset MIX compiles data from a variety of widely spoken languages across the Americas, Asia, Europe, Africa, and Oceania and was included in hopes of allowing the model to learn from a wider range of language families and linguistic features.

We use these monolingual data as part of our pretraining as this has been shown to improve results with smaller parallel datasets (Conneau and Lample, 2019; Liu et al., 2020; Song et al., 2019). Different combinations of these datasets are used in our pretraining to examine the effect of language similarity on translation accuracy after finetuning.

3.1.2 Parallel Data

Parallel data between Korean and Jejueo are from the Jejueo Interview Transcripts (JIT) dataset[4] (Park et al., 2020). These data were compiled from data from the Center for Jeju Studies, which collected data by interviewing senior Jeju citizens in Jejueo and having these interviews transcribed and translated into Korean by experts (Park et al., 2020).

3.2 Preprocessing

All data were tokenized using a unigram (Kudo, 2018) implementation of SentencePiece (Kudo and Richardson, 2018) in preparation for our multilingual model. We used a vocabulary size of 6,000 and a character coverage of 0.9995 as the languages used have a rich character set, especially the JA, KO, and ZH datasets. Separate SentencePiece models were trained for each combination of datasets shown in Table 3.

All data were then sharded for faster processing. With our SentencePiece model and vocabulary, we used FAIRSEQ (Ott et al., 2019) to build vocabularies and binarize our data.

The Jejueo-Korean parallel training, development, and test sets for finetuning and evaluating our models were the same as those used by the authors of the JIT dataset (Park et al., 2020) and are described in Table 2.

3.3 Pretraining

We pretrained six different models on different combinations (Table 3) of the datasets described in Section 3.1.1 using an mBART (Liu et al., 2020) implementation of FAIRSEQ (Ott et al., 2019). We also included 8.7 MB (160,356 sentences) of Jejueo training data from the JIT dataset as part of the pretraining process for each combination of datasets. Each model was pretrained on 32 NVIDIA V100 GPUs for two hours.

Balancing data across languages

Due to the large variability in size amongst the different datasets used in pretraining, we used an exponential sampling technique used in Conneau and Lample (2019); Liu et al. (2020) to re-sample text according to smoothing parameter α as follows:

$$q_i = \frac{p_i^\alpha}{\sum_{j=1}^N p_j^\alpha} \tag{1}$$

In equation 1, q_i refers to the resampling probability for language i given a multinomial distribution $\{q_i\}_{i=1...N}$ with original sampling probability p_i.

Because we want our model to work well with low-resource languages such as Jejueo, we set the smoothing parameter α to 0.25 (instead of 0.7 as used in mBART (Liu et al., 2020)) to reduce model bias towards the higher proportion of data from high-resource languages.

Hyperparameters

Using FAIRSEQ (Ott et al., 2019), we trained our models using a Transformer (Vaswani et al., 2017) with six encoder and decoder layers with eight attention heads each, a hidden dimension of 512, a feed-forward size of 2048, and a learning rate of 0.0003. Each model was optimized using Adam (Kingma and Ba, 2015) with hyperparameters $\beta = (0.9, 0.98)$ and $\epsilon = 10^{-6}$. For regularization, we used a dropout rate of 0.1 and a weight decay of 0.01.

[4] https://www.kaggle.com/datasets/bryanpark/jit-dataset

3.4 Finetuning

We performed finetuning using the best checkpoints (chosen using loss as a metric) from each of our pretrained models on the Jejueo → Korean translation task and Korean → Jejueo translation task. Using FAIRSEQ (Ott et al., 2019), we finetuned our models using the same hyperparameters used during pretraining, except for the dropout rate, which we changed to 0.5. We found that a higher dropout rate improved the translation output from our models.

3.5 Evaluation

We evaluated translations output by our models with detokenized BLEU (Papineni et al., 2002; Post, 2018) using the SacreBLEU library[5] (Post, 2018) on the test data from the parallel dataset JIT. We also used CHRF (Popović, 2015) to measure performance at the character level.

4 Results and Analysis

Table 4: Jejueo → Korean Results

		BLEU	CHRF
Baseline		67.94	
Model 1	MIX	65.79	0.7664
Model 2	JA, KO, ZH	64.04	0.7542
Model 3	JA, KO	**70.10**	**0.8009**
Model 4	KO	67.61	0.7788
Model 5	JA	66.90	0.7739
Model 6	ZH	62.95	0.7436

Table 5: Korean → Jejueo Results

		BLEU	CHRF
Baseline		44.19	
Model 1	MIX	42.97	0.5665
Model 2	JA, KO, ZH	41.17	0.5553
Model 3	JA, KO	**45.53**	**0.5867**
Model 4	KO	42.58	0.5626
Model 5	JA	42.35	0.5573
Model 6	ZH	42.47	0.5608

We compiled our results in Table 4 and Table 5. The best BLEU scores on the test data achieved by the authors who published the Korean and Jejueo parallel dataset (Park et al., 2020) are displayed as a baseline. To the best of our knowledge, these baseline BLEU scores are the highest published for this dataset, and there are no existing baseline CHRF scores.

Model 3, primarily trained on Japanese and Korean data (in addition to a small amount of Jejueo training data, as described in Section 3.3), performed the best, beating the baseline by 2.16 BLEU points for translation in the Jejueo → Korean direction and 1.34 BLEU points for translation in the Korean → Jejueo direction. Model 4, which made use of only Korean and Jejueo data, performed similarly to the baseline, despite having employed a much larger amount of Korean data. Model 1 and Model 2 performed even worse, which suggests that pretraining using languages that are more different from Korean and Jejueo can be detrimental to model quality. Though Model 1's Korean → Jejueo score is a bit higher than that of Model 4, there is a marked drop in score for the the Korean → Jejueo direction in Model 2 and the Jejueo → Korean direction for both Model 1 and Model 2.

Though Park et al. (2020) did not publish CHRF scores, we calculated CHRF scores to see if a similar trend could still be seen. When using CHRF scores, we can still see that Model 3 performed the best. Additionally, it still holds true that Model 4 performed better than Model 1 and Model 2 in the Jejueo → Korean direction and that Model 1 slightly beats Model 4 in the Korean → Jejueo direction followed by a steeper drop in score for Model 2 in this direction.

The similar trends in CHRF scores and BLEU scores amongst the six models suggest that the selection of languages used in the pretraining stage has a marked effect on model quality. Japanese, Korean, and Jejueo share many similar characteristics, such as having a similar syntax and having a high proportion of vocabulary of Chinese origin. While Chinese shares some vocabulary with Japanese, Korean, and Jejueo, it operates under a vastly different syntax and has a much lower degree of linguistic similarity. As can be seen from the results for Model 2, the addition of the Chinese dataset ZH may have thus hampered model quality. Model 1, which incorporates languages from all over the world, suffers from a similar issue, but the sheer variety of languages used may have helped it perform better than Model 2, as the model was exposed to a larger variety of linguistic features.

Model 4, however, also did not perform as well as Model 3 and achieved close but not higher scores

[5] https://github.com/mjpost/sacrebleu

compared to the baseline as it did not have enough linguistic variety from which to learn. Thus, while it is important to introduce linguistic variety to the model during pretraining, data must be selected carefully such that there is still a relatively high degree of linguistic similarity, perhaps most particularly in terms of syntax.

Model 5 and Model 6 both performed worse than Model 4, which was expected, as Korean is used in translation between Jejueo and Korean and is closely related to Jejueo itself. Model 6's performance displayed a more pronounced drop in translation quality in the Jejueo → Korean direction, performing nearly 5 BLEU points worse than the baseline and more than 7 BLEU points worse than Model 3. This marked difference is also reflected in the CHRF scores. Model 5 performed more similarly to Model 4, which may be due to the linguistic similarity between Korean and Japanese.

Model 4, Model 5, and Model 6 all performed similarly, however, in the Korean → Jejueo direction. Their performance is also similar to that of Model 1 and that of Model 2 in this direction, indicating that only a particular combination of languages can bring about a marked improvement in translation quality. Additionally, the fact that Model 1 and Model 2 achieved similar performance despite having used much more data than Model 4, Model 5, and Model 6 shows that Model 3's higher translation quality may not be due to simply having more data but instead be due to having a more advantageous combination of languages, but this needs more exploration in future work.

It is also worth noting that translation from Jejueo to Korean performs significantly better than translation from Korean to Jejueo. This is likely due to the fact that a single Korean word may have multiple translations in the Jejueo dataset while a single word in Jejueo typically corresponds to just one word in Korean. Thus, translation quality as measured by BLEU and CHRF is higher for translation in the Jejueo → Korean direction. This was also observed in Park et al. (2020)'s baseline translations. Another potential reason for this difference is the fact that Korean data outside of the parallel data were used during the pretraining process, whereas no additional Jejueo data were used, giving the model overwhelmingly more exposure to Korean vocabulary and a relatively small amount of exposure to Jejueo vocabulary. Perhaps more Jejueo data are needed for the model to better learn how different Jejueo words are used in different contexts.

5 Conclusions and Future Work

We have shown how pretraining on a large amount of carefully selected monolingual data can improve the quality of translation between Korean and Jejueo, a low-resource language pair. By using Japanese and Korean data during the pretraining process, our model was exposed to some linguistic diversity beyond Korean and Jejueo from a language of relatively high linguistic similarity, allowing our model to improve translation by 2.16 BLEU points for translation in the Jejueo → Korean direction and 1.34 BLEU points for translation in the Korean → Jejueo direction in comparison to the baseline.

If enough is known linguistically about the source and target languages, it is important to carefully select additional but similar languages to use during the pretraining process. Pretraining with Korean alone and pretraining with other languages of low linguistic similarity generated models that performed worse than the baseline. Syntactic similarity may be of particular importance as Korean, Jejueo, and Japanese all share a similar syntax while differing mostly in vocabulary. Korean, Japanese, and Jejueo are all considered synthetic SOV (subject-object-verb) languages, while Chinese is an analytic SVO language. This drastic difference in syntax may explain how using Chinese during the pretraining process resulted in a marked drop in translation quality.

Japanese, Jejueo, and Korean, however, do share many words that come from Chinese origins. For future work, we are interested in better leveraging these cognates found amongst Korean, Jejueo, and Japanese as shared representations that can be used as additional linguistic information for improving translation quality.

References

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 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.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 7057–7067.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of Low-Resource Machine Translation. *Computational Linguistics*, pages 1–67.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Taku Kudo and John Richardson. 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.

Chunxi Liu, Qiaochu Zhang, Xiaohui Zhang, Kritika Singh, Yatharth Saraf, and Geoffrey Zweig. 2020. Multilingual graphemic hybrid ASR with massive data augmentation. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 46–52, Marseille, France. European Language Resources association.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. 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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Kyubyong Park, Yo Joong Choe, and Jiyeon Ham. 2020. Jejueo datasets for machine translation and speech synthesis. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2615–2621, Marseille, France. European Language Resources Association.

Maja Popović. 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.

Matt Post. 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.

Moira Saltzman. 2017. Jejueo talking dictionary: A collaborative online database for language revitalization. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 122–129, Honolulu. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 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.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: Masked sequence to sequence pre-training for language generation. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 5926–5936. PMLR.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Changyong Yang, William O'Grady, and Sejung Yang. 2017. Toward a linguistically realistic assessment of language vitality: The case of Jejueo.

Changyong Yang, William O'Grady, Sejung Yang, Nanna Haug Hilton, Sang-Gu Kang, and So-Young Kim. 2020a. *Revising the Language Map of Korea*, pages 215–229. Springer International Publishing, Cham.

Changyong Yang, Sejung Yang, and William O'Grady. 2018. Integrating analysis and pedagogy in the revitalization of Jejueo. *Japanese-Korean Linguistics*, 25.

Changyong Yang, Sejung Yang, and William O'Grady. 2020b. *Jejueo: The Language of Korea's Jeju Island*. University of Hawai'i Press.

Francis Zheng, Machel Reid, Edison Marrese-Taylor, and Yutaka Matsuo. 2021. Low-resource machine translation using cross-lingual language model pretraining. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 234–240, Online. Association for Computational Linguistics.

A MIX Dataset Details

Table 6: MIX Dataset Statistics

	Size	Tokens
Bulgarian	2.7 GB	637,886,934
English	1.2 GB	378,524,430
French	1.3 GB	406,561,356
Irish	0.5 GB	121,007,968
Korean	1.6 GB	448,758,999
Latin	1.6 GB	496,141,311
Spanish	1.3 GB	401,305,855
Sudanese	49 MB	15,355,568
Vietnamese	1.2 GB	299,449,330
Yoruba	4.1 MB	1,232,419
Total	11.5 GB	3,206,224,170