Multi-dialect Neural Machine Translation for 48 Low-resource Japanese Dialects

Kaori Abe^{†,††}, Yuichiroh Matsubayashi^{†,††}, Naoaki Okazaki^{†††} and Kentaro Inui^{†,††}

We present a multi-dialect neural machine translation (NMT) model tailored to Japanese. Although the surface forms of Japanese dialects differ from those of standard Japanese, most of the dialects have common fundamental properties, such as word order, and some also use numerous same phonetic correspondence rules. To take advantage of these properties, we integrate multilingual, syllable-level, and fixed-order translation techniques into a general NMT model. Our experimental results demonstrate that this model can outperform a baseline dialect translation model. In addition, we show that visualizing the dialect embeddings learned by the model can facilitate the geographical and typological analyses of the dialects.

Key Words: *Neural Machine Translation, Multilingual, Dialect, Typological Analysis, Dialectometry*

1 Introduction

With the use of automated personal assistants (e.g., Apple’s Siri, Google Assistant, or Microsoft Cortana) and smart speakers (e.g., Amazon Alexa or Google Home) becoming increasingly widespread, the demand to bridge the gap between the standard form of a given language and its dialects has also enlarged. The importance of dealing with dialects is particularly evident in a rapidly aging society, like that in Japan, where older people use them extensively.

To address this issue, we consider a system for machine translation (MT) between Japanese dialects and standard Japanese. If such a system can yield correct dialect-to-standard translations, then other natural language processing systems (e.g., information retrieval or semantic analysis) that adopt standard Japanese as the input could also be applied to dialects. In addition, if a standard-to-dialect translation system becomes available, then smart speakers could respond to the native speakers of a dialect using that dialect. We believe that sympathetic interactions of this type might lead to such systems gaining more widespread acceptance of in the Japanese society.

[†] Tohoku University

^{††} RIKEN

^{†††} Tokyo Institute of Technology

In this paper, we present a multi-dialect neural MT (NMT) system tailored to Japanese. Specifically, we employ *kana*, a Japanese phonetic lettering system, to provide the basic units in the encoder–decoder framework to avoid the followings: ambiguity in converting *kana* to *kanji* (characters in the Japanese writing system), difficulties in identifying word boundaries especially for dialects, and data sparseness problems due to dealing with numerous words originating from different dialects. Because Japanese dialects almost always use the same word order as standard Japanese, we employ *bunsetsu* (a Japanese phrase unit) as a unit of sequences, instead of a sentence, which is more commonly used in NMT.

One issue for Japanese dialects is the lack of training data. To deal with this, we build a unified NMT model covering multiple dialects, inspired by the studies on multilingual NMT (Johnson, Schuster, Le, Krikun, Wu, Chen, Thorat, Viégas, Wattenberg, Corrado, Hughes, and Dean 2016). This approach utilizes *dialect embeddings*, i.e., vector representations of Japanese dialects, to inform the model of the input dialect. An interesting by-product of this approach is that the dialect embeddings that the system learns illustrate the difference between different dialect types from different geographical areas. In addition, we present an example of using these dialect embeddings for dialectometry (Nerbonne and Kretzschmar 2011; Kumagai 2016; Guggilla 2016; Rama and Çöltekin 2016).

Another advantage of adopting a multilingual architecture for multiple related languages is that it can enable gaining knowledge of their lexical and syntactic similarities. For example, Lakew, Mauro, and Federico (2018) reported that including several related languages in supervised training data can improve multilingual NMT. Our results confirm the effectiveness of using closely related languages (i.e., Japanese dialects) in multilingual NMT.

2 Related Work

Dialectal text is scarcely available because dialects are generally spoken, instead of being written. For this reason, many dialect MT researchers study in low-resource settings (Zbib, Malchiodi, Devlin, Stallard, Matsoukas, Schwartz, Makhoul, Zaidan, and Callison-Burch 2012; Scherrer and Ljubešić 2016; Hassan, Elaraby, and Tawfik 2017).

The use of similar dialects has been found to be helpful in learning translation models for particular dialects. Several previous studies have investigated the characteristics of translation models of closely related dialects (Meftouh, Harrat, Jamouss, Abbas, and Smaili 2015; Honnet, Popescu-Belis, Musat, and Baeriswyl 2018). For example, Honnet et al. (2018) reported that a character-level NMT model trained on one Swiss-German dialect performed moderately well for

translating sentences in closely related dialects.

Therefore, in view of the above, we use multilingual NMT (Johnson et al. 2016) to learn the parameters that encode the knowledge of the shared lexical and syntactic structures of dialects. Some researchers (Gu, Hassan, Devlin, and Li 2018; Arivazhagan, Bapna, Firat, Lepikhin, Johnson, Krikun, Chen, Cao, Foster, Cherry, Macherey, Chen, and Wu 2019) demonstrated that multilingual NMT could be useful for low-resource language pairs, additionally Lakew et al. (2018) found that a multilingual NMT system trained on multiple related languages showed an improved zero-shot translation performance. We believe that multilingual NMT can be effective for closely related dialects, and can compensate for the lack of translation data for the different associated dialects.

Multilingual NMT can also assist in analyzing the characteristics of each considered language. Östling and Tiedemann (2017) found that clustering the language embeddings learned by a character-level multilingual system provided an illustration of the language families involved. In the light of this, we also examine our dialect embeddings to investigate whether our multi-dialect model can capture the similarities between the dialects (Section 5).

Previous studies reported that character-level statistical machine translation (SMT) using words as translation units is effective for translating between closely related languages (Nakov and Tiedemann 2012; Scherrer and Ljubešić 2016). There are two reasons for this: character-level information enables the system to exploit lexical overlaps, whereas using words as translation units takes the advantage of the syntactic overlaps of the related languages. To utilize these overlaps, Pointer Networks (Vinyals, Fortunato, and Jaitly 2015; Gülçehre, Ahn, Nallapati, Zhou, and Bengio 2016), which can copy some parts of input sequences to output sequences, also seem to be effective for dialect translation. In this study, we conduct an experiment with a simple long short-term memory (LSTM) architecture to train a multilingual NMT model for multiple dialects. Adopting a copy architecture similar to Pointer Networks will be conducted in a future study.

In this study, we present a method of translating between Japanese dialects by combining three ideas: multilingual NMT, character-level NMT, and using base phrases (i.e., *bunsetsu*) as translation units. We believe this enables our approach to fully exploit the similarities among dialects and standard Japanese, even under low-resource settings.

Fig. 1 Number of sentences in each dialect in *National Dialect Discourse Database Collection*

3 Data: Japanese Dialect Corpus

Japanese is a dialect-rich language, with dozens of dialects used in everyday conversations in most Japanese regions. They can be characterized in terms of the differences in their content words (vocabulary) and regular phonetic shifts, mostly in their postpositions and suffixes. Specifically, they share most words with standard Japanese, and mostly use common grammatical rules, such as for the word order, syntactic marker categories, and connecting syntactic markers.¹ Some dialects also share the dialect-specific vocabulary. For example, the word, *しゃっこい* (*shakko*, meaning “cold”), is shared among some dialects in the Tohoku region, such as Aomori and Akita.

In this study, we used a collection of parallel textual data for the dialects and standard Japanese, called as the *National Dialect Discourse Database Collection* (National Institute for Japanese Language and Linguistics 1980). This corpus includes 48 dialects, one from each of the 47 prefectures and an additional dialect from the Okinawa Prefecture. For each dialect, the texts consist of transcribed 30-minute conversations between two native speakers of that dialect. The total number of dialect sentences (each paired with a translation into standard Japanese) is 34,117. Figure 1 shows the number of sentences in each dialect. The amount of each dialect data is less than 1,500 sentences and vary.

Japanese texts are generally written in a mix of *kanji* and *kana*; therefore, we converted the *kanji* in the sentences into *kana*, and subsequently, segmented them into *bunsetsu*.² In this study, we used the *bunsetsu* segmentation annotated in the original corpus. After preprocessing, the average sentence lengths were of 14.62 and 15.57 characters for the dialects and the standard Japanese, respectively. The average number of *bunsetsu* per sentence was 3.42.³

¹ For details, *Linguistic Atlas of Japan Database* (<https://www.lajdb.org/TOP.html>) published by NINJAL provides an overview of the Japanese dialect distribution.

² This is the smallest Japanese phrase unit, containing a single content word and attached postpositions.

³ The total number of *bunsetsu* is 116,928.

4 NMT Model

Figure 2 presents an overview of our network structure of the multi-dialect NMT system. Because our focus is on examining the effectiveness of the multi-dialect NMT and its detailed behavior, rather than on creating a novel translation model, we used OpenNMT (Klein, Kim, Deng, Senellart, and Rush 2017). OpenNMT is a stacking LSTM encoder-decoder model with a multilingual extension similar to that of Johnson’s method (Johnson et al. 2016). However, to improve its direct translation accuracy, we introduce the following three modifications.

Dialect labels: Following a previous multilingual NMT study (Johnson et al. 2016), we train the unified model that deals with all the 48 dialects simultaneously using the dialect embeddings including auxiliary dialect labels. Johnson et al. (2016) added a label to the beginning of each sequence to specify the output language. We modify this approach to specify both the input and output dialects of the model, and examine the four different placements for these labels, as listed in Table 1.

Syllable-to-syllable translation: As mentioned in Section 3, a key to translate between two closely related languages, especially in our case, dialects, is modeling the phonetic correspondences between them. Thus, to consider syllable-level translation rules that may be shared by similar dialects, we defined our translation task as a syllable-to-syllable translation.

Fig. 2 Proposed multi-dialect NMT model

Table 1 Dialect label input order variants

ID	Encoder input order
a	<SRC>, sequence
b	<TGT>, sequence
c	<SRC>, <TGT>, sequence
d	<SRC>, sequence, <TGT>

For example, when we translate Aomori dialect into standard language, we test using input sentence that replaced <SRC> with 青森 (*Aomori*) and <TGT> with 標準語 (*Hyoujungo*).

We realize syllable-to-syllable translation by representing the inputs and the outputs as *kana* sequences and performing character-based MT. A similar approach was used to normalize Japanese text from Twitter, where the main issue was phonological transliteration (Saito, Suzuki, Nishida, and Sadamitsu 2017). In our dataset, all the dialect expressions are transcribed using *kana*; however, the standard Japanese translations use a mix of *kanji* and *kana* characters. Therefore, in order to conform to the syllable-to-syllable task, we also convert them into *kana* sequences by automatically analyzing the pronunciation of each *kanji* character and replacing it with the corresponding *kana* sequence.

Translation without distortion: Finally, we attempt to remove the word-order distortion modeling from NMT. In a standard MT, systems adopt a single sentence as the input and yield a translated sentence in an appropriate word order for the target language. However, in the dialect translation, the input and output word orders are mostly the same. To test this, we manually checked 100 randomly-selected sentence pairs from the training set, and found no differences in the ordering (distortion). This fact suggests that we do not require sentence-by-sentence supervision data, because it does not need to learn a distortion model. Based on this intuition, we split each input sentence into base-phrase parts, i.e., *bunsetsu* sequences, translate each chunk from the source to the target language and, subsequently, output the translated chunks in the same order.

5 Experiments

Using parallel text data (standard Japanese and 48 regional dialects), we trained both a single dialect-to-standard translation model and a reverse (standard-to-dialect) model, measuring the translation quality using BLEU scores (Papineni, Roukos, Ward, and Zhu 2002). In addition, we analyzed the trained dialect embeddings in detail and conducted data ablation tests.

5.1 Experimental Setup

For these experiments, we split the corpus into training, development, and the test sets in an 8:1:1 ratio. We oversampled the translation pairs to ensure that every dialect had the same amount of training data, because there were different numbers of training and test instances for each dialect (in Figure 1). For the oversampling, we randomly sampled the existing sentence in each dialect dataset. Finally, all the training sets for each dialect consisted of 1,042 sentences, the same size as the largest original training set of the Iwate dialect.

Because Japanese dialects mostly share the same vocabulary and there are few distortions (word order changes), we expect that the translation between a Japanese dialect and standard Japanese is relatively easy compared to that between other languages. Thus, the main focus of the following experiments was to evaluate how well the model captured the phonological shifts between the dialects and the standard Japanese. Therefore, we employed syllable-level (i.e., character-level) BLEU scores as the evaluation measures. We calculated the syllable-level BLEU for each sentence by concatenating the chunk-wised translations. Note that this evaluation measure generally yields higher scores than those calculated at the word level. Finally, we macro-averaged the scores over all the dialects. For Multi NMT or SMT models, we generated the translation output for the entire test set, which contained all the dialects, and we divided it into the 48 dialect test sets. Subsequently, we calculated the macro-averaged BLEU scores using the 48 local BLEU scores obtained on the test set for all the dialects. For Mono NMT or SMT models, we trained 48 local NMT/SMT models using a local training/valid set (also a subset of the entire training/valid set) and subsequently evaluated it with a local test set. For all the settings, for the evaluation, we used the test set written as one sentence per line. The difference between the dialect-to-standard and standard-to-dialect translations is simply the exchange of the source and target languages. In fact, the macro-averaged BLEU score reached 35.10 even when we simply output the dialect sentences without translation.

We used OpenNMT-py⁴ with its default hyper-parameter settings, except for the number of training epochs (which we set to 20), and selected the model that performed best on the development set. For the details, we list the hyperparameter settings in Table 2. In addition, we employed Moses⁵ (Koehn, Hoang, Birch, Callison-Burch, Federico, Bertoldi, Cowan, Shen, Moran Mit, Zens, Dyer, Bojar, Constantin, and Herbst 2007) as the baseline SMT model and set the distortion limit to 0. The standard Japanese language model used in Moses was trained

⁴ <https://github.com/OpenNMT/OpenNMT-py>

⁵ <http://www.statmt.org/moses>

Table 2 Experimental settings of OpenNMT such as hyper parameters

Train	
epoch	20
layer (encoder / decoder)	2
batch size	64
valid batch size	32
word embedding dim.	500
hidden dim.	500
dropout rate	0.3
optimizer	SGD
learning rate	1.0
Decode	
Beam-size	5

Table 3 Descriptions of each model

System	Translation unit	Model	Distinction between dialects
None (w/o translation)	—	—	—
Mono NMT	<i>bunsetsu</i>	local NMT $\times$ 48	True
Multi NMT (w/o labels)	<i>bunsetsu</i>	multilingual NMT	False
Multi NMT-sentence (w/ labels)	sentence	multilingual NMT	True
Multi NMT (w/ labels)	<i>bunsetsu</i>	multilingual NMT	True
Mono SMT	<i>bunsetsu</i>	local SMT $\times$ 48	True
Multi SMT (w/o labels)	<i>bunsetsu</i>	multilingual SMT	False

with KenLM (Heafield 2011). For the syllable-to-syllable translation, we used MeCab 0.996⁶ to analyze the pronunciations of the *kanji* characters.

Regarding the dialect label order used for the input, our preliminary experiments on the validation set indicated that the best models were obtained using input sequence (d) (Table 1) for the dialect-to-standard translation and input sequence (b) for the standard-to-dialect translation.⁷

A brief description of each model we used in our experiments is provided in Table 3. Except for the multi-sentence NMT models, we used each *bunsetsu* as a translation unit. Note that, as we mentioned in Sec. 4, it is practically unnecessary to model the word order distortion in Japanese dialect translation. The individual dialects are distinguished by two methods: addition of dialect

⁶ <http://taku910.github.io/mecab/>

⁷ See Appendix B for more details.

Table 4 Syllable-level BLEU scores of all models

System	BLEU
dialect-to-standard	
None (w/o translation)	35.10
Mono NMT	22.45
Multi NMT (w/o labels)	71.29
Multi NMT-sentence (w/ labels)	69.74
Multi NMT (w/ labels)	75.66
Mono SMT	52.98
Multi SMT (w/o labels)	73.54
standard-to-dialect	
Multi NMT (w/ labels)	65.30

labels to the multilingual NMT, and training the local models for each dialect separately. In comparison, the multilingual NMT and SMT systems without the labels (Multi NMT w/o labels, Multi SMT w/o labels) do not distinguish dialects, and therefore, are disadvantageous.

5.2 Multi-Dialect NMT Model Performance

Table 4 summarizes the results of the dialect translation performance of all the considered models, with the first row group comprising their scores for dialect-to-standard translation under different input settings.

Monolingual vs. multilingual: For comparison, we first considered a model that was trained using only a single set of dialect-standard parallel data (Mono NMT). It performed quite poorly compared to the other models that used data for all the dialects (Multi NMT) and was even worse than simply outputting the dialect sentences unchanged (35.10). This indicates that training independent NMT models for each language pair with a limited amount of training data is extremely inefficient. In contrast, the multi-dialect model presented a drastically improved the translation performance.

Dialect labels: Including dialect labels improved the Multi-NMT BLEU score by 4.37 points (fifth row of Table 4) compared to that of the same model without the dialect labels (third row). Figure 3 shows for these two models the BLEU scores for all the dialects in ascending order of translation difficulty. Here, the translation difficulty is defined as the average normalized Levenshtein distance over all the sentence pairs (dialect and standard Japanese) for a given dialect. As expected, the BLEU scores for all the dialects present a strong negative correlation

Fig. 3 BLEU scores of Multi NMT models and translation difficulty for all dialects

($\rho = -0.82$) with the translation difficulty. In addition, we can observe that the model with language labels consistently outperforms that without the labels, except for the Tottori dialect, for which there is an extremely small amount of text data (Figure 1). This result indicates that explicit information of the source and target dialects with dialect labels can improve the encoding and decoding accuracy.

Fixed-order translation: Comparing the proposed model (Multi NMT) with the same model trained via the standard approach of using entire sentences as input/output sequences (Multi NMT-sentence) shows that Multi NMT outperforms Multi NMT-sentence by 5.92 points. One disadvantage of the chunk-wise translation is that it cannot capture the context beyond the boundary of each chunk; however, despite this disadvantage, our Multi NMT model can still outperform the model with an access to a broader context (Multi NMT-sentence). This indicates that our fixed-order translation approach is suitable for translating Japanese dialects, despite its limited context sensitivity.

NMT vs. SMT: Zoph, Yuret, May, and Knight (2016) found that SMT models largely outperformed state-of-the-art NMT models for low-resource languages. The second-row group in Table 4 summarizes the results for a fixed-order character-based SMT baseline. In these experiments, the NMT model trained using a single dialect (Mono NMT) yielded the poorest performance; however, the one with dialect labels outperformed the baseline Multi SMT model, achieving the best performance overall.

5.3 Example of Translation Results

To demonstrate how each of the proposed component contributes to generating accurate translations, we now present some concrete examples of the translation results of our models for the Hyogo, Kagoshima, and Nigata dialects (Table 5).

Comparing the Multi NMT models with and without dialect labels, we noted that adding labels enables the models to better translate the chunks that required dialect-specific knowledge.

Table 5 Example dialect-to-standard translations for Hyogo, Kagoshima, and Nigata dialects

Example 1	Hyogo region
Meaning	Yes, until then, in Aioi ...
Source	n - / so re ma de / o - ni wa (んー / それまで / おーにわ)
Reference	u n / so re ma de / a i o i ni ha (うん / それまで / あいおいには)
Multi NMT (w/o label)	u n / so re ma de / o o ni ha (うん / それまで / おおには)
Multi NMT-sentence (w/ label)	n - / so re ma de / a t ta n da (んー / それまで / あったんだ)
Multi NMT (w/ label)	n - / so re ma de / a i o i ni ha (んー / それまで / あいおいには)
Multi SMT (w/o label)	u n / so re ma de / o o ni ha (うん / それまで / おおには)
Example 2	Kagoshima region
Meaning	After a few days, then it was...
Source	so i ga / mo / na n ni k ka / shi ta ya (そいが / も / なんにか / したや)
Reference	so re ga / mo u / na n ni chi ka / shi ta ra (それが / もう / なんにちか / したら)
Multi NMT (w/o label)	so re ga / mo u / na n ni chi ka / shi ta da (それが / もう / なんにちか / しただ)
Multi NMT-sentence (w/ label)	so re ga / mo u / na n ni tsu ka / shi ta yo (それが / もう / なんにか / したよ)
Multi NMT (w/ label)	so re ga / mo u / na n ni chi ka / shi ta ra (それが / もう / なんにちか / したら)
Multi SMT (w/o label)	so re ga / mo u / na ni ka / shi ta de (それが / もう / なにか / したで)
Example 3	Nigata region
Meaning	I want to go to the water park as soon as possible, but...
Source	ha yo - / mi zu n / do ko e / i ko - to / o mo u ke do (はよー / みずん / どこえ / いこーと / おもうけど)
Reference	ha ya ku / mi zu no / to ko ro he / i ko u to / o mo u ke re do (はやく / みずの / ところへ / いこうと / おもうけれど)
Multi NMT (w/o label)	ha ya ku / mi zu ga / do ko he / i ko u to / o mo u ke do (はやく / みずが / どこへ / いこうと / おもうけど)
Multi NMT-sentence (w/ label)	ha ya ku / mi zu no / to ko ro he / i ko u to / o mo u ke do (はやく / みずの / ところへ / いこうと / おもうけど)
Multi NMT (w/ label)	ha ya ku / mi zu ga / do ko he / i ko u to / o mo u ke re do (はやく / みずが / どこへ / いこうと / おもうけれど)
Multi SMT (w/o label)	ha ya ku / mi zu ga / do ko he / i ko u to / o mo u ke do (はやく / みずが / どこへ / いこうと / おもうけど)

In Example 1, the source sentence includes a local name, おー (*O -*), for a certain area, あいおい (*Aioi*), in Hyogo, which only the model with dialect labels can successfully translate. In addition, in Example 2, the dialect labels enable the model to capture a dialect-specific transliteration rule for the functional suffix (“ta ra”), a conditional-mood marker in the last chunk of the reference sentence (i.e., “ta ya” to “a ra”).

Similarly, because the Multi SMT model could not take advantage of the dialect labels, it failed to capture dialect-specific translation rules.

In the previous section, we noted that our fixed-order translation approach is suitable for translating Japanese dialects, despite its limited context sensitivity. However, this becomes a problem in Example 3, where the chunk-wise translation models cannot correctly translate a *bunsetsu* owing to the lack of context. Here, none of the models, except the Multi NMT-sentence, can translate the *bunsetsu*, “mi zu n”, in the Nigata dialect to the correct standard Japanese *bunsetsu* “mi zu no.” Because the translation of the functional word, “n,” in the Nigata dialect is ambiguous, it can be translated as either “ga” (nominative marker) or “no” (of) depending on the following context. This example exposes the limitations of our chunk-wise translation models and suggests potential future directions: extending the fixed-order translation to incorporate contextual information.

5.4 Visualizing Attention Weights

Here, to investigate how the proposed model translated the *kana* sequences in various dialects, we visualized the attention weights of the best-performing model for some correctly-translated examples.

Figure 4(a) shows the attention history of the model for an example where a part of the target language (standard Japanese) *bunsetsu* changes from the source language (Aomori dialect) according to a simple regular rule. In such cases, the model tends to weight the dialect label heavily when applying the rule (“da” → “do ha”). Conversely, Figure 4(b) shows the attention history for an example where almost all the syllables are transcribed. In these cases, the model needs to disambiguate the morpheme-level definitions to create a correct translation, and thus, tended to focus on the entire sequence of semantically- or grammatically-related morphemes.

5.5 Visualizing Dialect Embeddings

Östling and Tiedemann (2017) reported that clustering the language embeddings used to train a multilingual language model produced a language cluster structure similar to those of established relationships among the language families. Motivated by their work, we decided to

Fig. 4 Attention weight examples. (a) *Aomori-to-standard* translation of “next time”. Aomori word *konda* is formed by linking syllables from two standard Japanese words, *kondo* (next time) and *ha* (topic marker). (b) *Okinawa-to-standard* translation of “we”. Okinawa word *watta* combines two standard Japanese words, *watashi* (I) and *tachi* (plural marker), with *watt* and *ta*- roughly corresponding to *watashi* and *tachi*, respectively.

Fig. 5 T-SNE projection of dialect label vectors. Dialects belonging to same region are shaded using same background color.

examine the relationships between the dialect embeddings and the typology of the dialects.

Figure 5 shows the t-Distributed Stochastic Neighbor Embedding (t-SNE) projection of the dialect embeddings. It indicates that dialects from neighboring regions tend to form a single cluster. Furthermore, we can observe an interesting agreement between the cluster distances and the predictions of the dialectological typology theory, known as *center versus periphery* (Yanagida 1980), wherein new language use trends gradually propagate from the cultural center (the old

Table 6 Impact of excluding nearest or farthest five dialect regions from training data when calculating BLEU score for each dialect region

Dataset	Avg. Δ	#Regions BLEU decreased
–nearest 5	–0.94	34 / 48 (71%)
–farthest 5	–0.22	31 / 48 (65%)

“Avg. Δ ” denotes the average BLEU score difference compared to that using all the data.

capital, Kyoto) to less culturally influential areas. This potentially explains why the dialects in the Tohoku region (E) are similar to those in the Kyushu region (D), despite their large geographical separation.

5.6 Effect of the Nearby Dialects

To investigate in more detail how jointly learning multiple dialects contributed to the dialect-to-standard translations for each dialect, we performed an ablation study on all the dialect regions. As presented in the previous section, the dialects in geographically close regions are generally more similar to each other than those in other regions. Therefore, we assumed that the impact of sharing data from other dialects would differ depending on their geographical distances from the target dialect.

To investigate this assumption, we prepared two Multi NMT models per dialect, trained on the data that excluded the five geographically nearest or farthest dialects⁸ for the given dialect region, and calculated the differences in the BLEU scores of these models and the original model, for all the dialect regions. For example, the *–nearest5* model for the Tokyo dialect is the Multi NMT model learned with the training data excluding the Chiba, Kanagawa, Saitama, Gumma, and Ibaraki dialects. Subsequently, we compared this BLEU score of the model for the Tokyo dialect in the test set to that of the original model trained with full data.

Table 6 lists the average results over all the 48 models for both the cases. Both the models trained without the nearest five dialects, and those without the farthest five dialects yielded lower average BLEU scores for their target dialects compared to the full models. This suggests that even very distant dialects still assist in training other dialects. In addition, we note that removing the nearest five dialects had a more significant impact than removing the farthest five dialects, indicating that similar dialects contribute more to assisting a multi-dialect NMT to learn effectively.

⁸ The distances between the dialect pairs were calculated using the Euclidean distances between the points where the dialogs were recorded.

6 Conclusion

We have examined the effectiveness of a multilingual, syllable-based, fixed-phrase-order NMT model for translating Japanese dialects into standard Japanese. The results showed that each component of our multi-dialect NMT model successfully improved the translation accuracy when using a limited amount of supervised training data. In addition, we demonstrated the potential benefit of analyzing dialect embeddings for dialectological analysis applications, and have also analyzed how the multi-dialect NMT leverages the training data involving similar dialects to translate a given dialect.

One limitation of the proposed model is that it cannot consider longer-range dependencies beyond the chunk level. Therefore, our future research plans include incorporating contextual information, e.g., n-to-1 translation (Tiedemann and Scherrer 2017), into fixed-order translation models and investigating the characteristics of the dialect embeddings further.

Acknowledgement

We would like to thank the Center for the Study of Dialectology, Tohoku University for sharing the dataset with us. This work was supported by JSPS KAKENHI Grant Numbers JP15K16045, JP15H05318, and JP15H01702.

References

- Arivazhagan, N., Bapna, A., Firat, O., Lepikhin, D., Johnson, M. G., 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 preprint arXiv:1907.05019*.
- Gu, J., Hassan, H., Devlin, J., and Li, V. O. (2018). “Universal Neural Machine Translation for Extremely Low Resource Languages.” In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 185–194. Association for Computational Linguistics.
- Guggilla, C. (2016). “Discrimination Between Similar Languages, Varieties and Dialects Using CNN- and LSTM-based Deep Neural Networks.” In *Proceedings of the 3rd Workshop on NLP for Similar Languages, Varieties and Dialects*, pp. 186–194. The COLING 2016 Organizing

Committee.

- Gülçehre, Ç., Ahn, S., Nallapati, R., Zhou, B., and Bengio, Y. (2016). “Pointing the Unknown Words.” *CoRR*, **abs/1603.08148**.
- Hassan, H., Elaraby, M., and Tawfik, A. Y. (2017). “Synthetic Data for Neural Machine Translation of Spoken-Dialects.” *arXiv preprint arXiv:1707.00079*.
- Heafield, K. (2011). “KenLM : Faster and Smaller Language Model Queries.” In *Proceedings of the 6th Workshop on Statistical Machine Translation*, pp. 187–197. Association for Computational Linguistics.
- Honnet, P.-E., Popescu-Belis, A., Musat, C., and Baeriswyl, M. (2018). “Machine Translation of Low-Resource Spoken Dialects: Strategies for Normalizing Swiss German.” In *Proceedings of the 11th edition of the Language Resources and Evaluation Conference*, pp. 3781–3788. European Language Resources Association.
- Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., and Dean, J. (2016). “Google’s Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation.” In *Proceedings of Transactions of the Association for Computational Linguistics*, Vol. 5, pp. 339–351.
- Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. M. (2017). “OpenNMT: Open-Source Toolkit for Neural Machine Translation.” *arXiv preprint arXiv:1701.02810*.
- Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran Mit, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). “Moses: Open Source Toolkit for Statistical Machine Translation.” In *Proceedings of 45th Annual Meeting of the Association for Computational Linguistics*, pp. 177–180. Association for Computational Linguistics.
- Kumagai, Y. (2016). “Developing the Linguistic Atlas of Japan Database and Advancing Analysis of Geographical Distributions of Dialects.” In Côté, M. H., Knooihuizen, R., and Nerbonne, J. (Eds.), *The future of dialects: Selected papers from Methods in Dialectology XV*, pp. 333–362. Berlin: Language Science Press.
- Lakew, S. M., Mauro, C., and Federico, M. (2018). “A Comparison of Transformer and Recurrent Neural Networks on Multilingual Neural Machine Translation.” *arXiv preprint arXiv:1806.06957*.
- Meftouh, K., Harrat, S., Jamouss, S., Abbas, M., and Smaili, K. (2015). “Machine Translation Experiments on PADIC: A Parallel Arabic Dialect Corpus.” In *Proceedings of 29th Pacific Asia Conference on Language, Information and Computation*, pp. 26–34. Association for Computational Linguistics.

- Nakov, P. and Tiedemann, J. (2012). “Combining Word-Level and Character-Level Models for Machine Translation Between Closely-Related Languages.” In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 301–305. Association for Computational Linguistics.
- National Institute for Japanese Language and Linguistics (1980). “National Dialect Discourse Database Collection (全国方言談話データベース 日本のふるさと言葉集成)”. Kokushokankokai Inc. (国書刊行会) .
- Nerbonne, J. and Kretzschmar, W. A. (2011). “Dialectometry++.” *Literary and Linguistic Computing*, **28** (1), pp. 2–12.
- Östling, R. and Tiedemann, J. (2017). “Continuous Multilinguality with Language Vectors.” In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 644–649. Association for Computational Linguistics.
- 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 on Association for Computational Linguistics*, ACL ’02, pp. 311–318, USA. Association for Computational Linguistics.
- Rama, T. and Çöltekin, Ç. (2016). “LSTM Autoencoders for Dialect Analysis.” In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects*, pp. 25–32. The COLING 2016 Organizing Committee.
- Saito, I., Suzuki, J., Nishida, K., and Sadamitsu, K. (2017). “Improving Neural Text Normalization with Data Augmentation at Character- and Morphological Levels.” In *Proceedings of the 8th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 257–262. Asian Federation of Natural Language Processing.
- Scherrer, Y. and Ljubešić, N. (2016). “Automatic Normalisation of the Swiss German ArchiMob Corpus Using Character-level Machine Translation.” In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, pp. 248–255. Bochumer Linguistische Arbeitsberichte.
- Tiedemann, J. and Scherrer, Y. (2017). “Neural Machine Translation with Extended Context.” In *Proceedings of the 3rd Workshop on Discourse in Machine Translation*, pp. 82–92, Copenhagen, Denmark. Association for Computational Linguistics.
- Vinyals, O., Fortunato, M., and Jaitly, N. (2015). “Pointer Networks.” In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R. (Eds.), *Advances in Neural Information Processing Systems 28*, pp. 2692–2700. Curran Associates, Inc.

Yanagida, K. (1980). “Kagyuko (蝸牛考)”. Iwanami Shoten, Publishers (岩波書店).

Zbib, R., Malchiodi, E., Devlin, J., Stallard, D., Matsoukas, S., Schwartz, R., Makhoul, J., Zaidan, O. F., and Callison-Burch, C. (2012). “Machine Translation of Arabic Dialects.” In *Proceedings of the North American Chapter of the Association for Computational Linguistics*, pp. 49–59. Association for Computational Linguistics.

Zoph, B., Yuret, D., May, J., and Knight, K. (2016). “Transfer Learning for Low-Resource Neural Machine Translation.” In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1568–1575. Association for Computational Linguistics.

Appendix

A Data size for each dialect

We show the details of the dataset size in Figure 6. In our experiments, we used the largest available Japanese dialect dataset; however, the number of sentences in some dialects (e.g., Tottori) was small in our dataset. Owing to the lack of the training or test dataset, the BLEU score of the Tottori dialect is actually inconsistent with those of the other dialects, as shown in Figure 3. However, this result suggests that the proposed method is effective for almost all the dialects, except in a very low-resource scenario.

Fig. 6 Size of each dataset (training/valid/test) in our experimental setting per dialect

B Experiments of dialect label order variants

We summarize the results of the preliminary experiments to examine the best setting of the dialect labels in Table 7 (validation set) and Table 8 (test set). As can be seen from both the tables, the best models are obtained using input sequence (d) for the dialect-to-standard translation and input sequence (b) (in Table 1) for the standard-to-dialect translation. Label set (b) in the dialect-to-standard translation and set (a) in the standard-to-dialect translation do not contain the dialect information, in contrast with the other settings. They present a quite lower performances.

Table 7 Syllable-level BLEU scores for each dialect label in Multi NMT in both translation directions on the validation sets. We showed the details of dialect labels (a)–(d) in Table 1.

Seed	(a)	(b)	(c)	(d)
dialect-to-standard				
0	90.16	89.30	90.28	91.03
1	90.92	89.45	90.93	91.16
standard-to-dialect				
0	79.32	85.37	84.35	84.68
1	79.36	85.34	84.39	84.83

Table 8 Syllable-level BLEU scores for each dialect label in Multi NMT in both translation directions on the test sets.

Seed	(a)	(b)	(c)	(d)
dialect-to-standard				
0	74.35	71.59	74.25	75.57
1	75.31	71.44	75.43	75.50
standard-to-dialect				
0	51.30	65.30	64.18	63.17
1	51.61	65.13	64.06	63.19

Kaori Abe: is a first-year doctoral student at the Graduate School of Information Sciences, Tohoku University. She received the M.Sc degree in information science from Tohoku University, Japan, in 2020. She is supported by Research Fellowships for Young Scientists (DC1), Japan Society for the Promotion of Science. Her research interests include natural language processing, especially dialogue systems and machine translation.

Yuichiro Matsubayashi: is an associate professor at the Graduate School of Education, Tohoku University. He received a Ph.D. degree in Information Science and Technology from the University of Tokyo in 2010. His work focuses specifically on modeling a communication process of humans, building a semantics of the natural language, and semantic processing of natural language text.

Naoaki Okazaki: is a professor of the School of Computing, Tokyo Institute of Technology. He received a Ph.D. degree in the Graduate School of Information Science and Technology, the University of Tokyo in 2007. He is also a visiting research scholar of the Artificial Intelligence Research Center (AIRC), AIST. His research interests include natural language processing, text mining, and machine learning.

Kentaro Inui: is a professor of the Graduate School of Information Sciences at Tohoku University, where he is head of the Natural Language Processing Lab. He also leads the Natural Language Understanding Team at the RIKEN Center for the Advanced Intelligence Project. His research areas include natural language processing and artificial intelligence. He currently serves as Vice-chairperson of the Association for Natural Language Processing, Member of the Science Council of Japan, and Director of NPO FactCheck Initiative Japan.

(Received May 1, 2020)

(Revised August 1, 2020)

(Accepted August 27, 2020)