Multi-dialect Neural Machine Translation and Dialectometry

Kaori Abe[1], Yuichiroh Matsubayashi[2], Naoaki Okazaki[3], and Kentaro Inui[1,2]
[1]Tohoku University, [2]RIKEN, [3]Tokyo Institute of Technology
abe-k@ecei.tohoku.ac.jp, yuichiro.matsubayashi@riken.jp
okazaki@c.titech.ac.jp, inui@ecei.tohoku.ac.jp

Abstract

We present a multi-dialect neural machine translation (NMT) model tailored to Japanese. While the surface forms of Japanese dialects differ from those of standard Japanese, most of the dialects share fundamental properties such as word order, and some also use many of the 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 geographical and typological analyses of dialects.

1 Introduction

Since 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) has become more widespread, demand has also grown to bridge the gap between dialects and the standard form of a given language. The importance of handling dialects is especially evident in a rapidly aging society like 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 this can provide correct translations in the dialect-to-standard direction, then, other natural language processing systems (e.g., information retrieval or semantic analysis) that take standard Japanese as input could be applied to dialects as well. In addition, if a standard-to-dialect translation system were also available, smart speakers could respond to the native speakers of a dialect using that dialect. We believe that friendly interactions of this sort might lead to such systems gaining more widespread acceptance of in the Japanese society.

In this paper, we present a multi-dialect neural MT (NMT) system tailored to Japanese. Specifically, we employ *kana*, a Japanese phonetic lettering system, as basic units in the encoder–decoder framework to avoid the following: 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 handling large numbers of words originating from different dialects. Since dialects almost always use the same word order as standard Japanese, we employ *bunsetsu* (Japanese phrase units) as a unit of sequences rather than sentence which is more commonly used in NMT.

One issue for Japanese dialects is a lack of training data. To deal with this, we build a unified NMT model covering multiple dialects, inspired by work on multilingual NMT (Johnson et al., 2016). This approach utilizes *dialect embeddings*, namely 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 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 the multilingual architecture for multiple related languages is it can enable the system to acquire knowledge of their lexical and syntactic similarities. For example, Lakew et al. (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 (namely Japanese dialects) in multilingual NMT.

2 Related Work

Little dialectal text is available since dialects are generally spoken rather than written. For this reason, many dialect MT researchers work in low-resource settings (Zbib et al., 2012; Scherrer and Ljubešić, 2016; Hassan et al., 2017).

However, 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 for closely-related dialects (Meftouh et al., 2015; Honnet et al., 2018). For example, Honnet et al. (2018) reported that a character-level NMT model trained on one Swiss-German dialect performed moderately well at translating sentences in closely-related dialects.

Therefore, given this, we use multilingual NMT (Johnson et al., 2016) to learn parameters that encode knowledge of dialects' shared lexical and syntactic structure. Gu et al. (2018) demonstrated that multilingual NMT can be useful for low-resource language pairs, while Lakew et al. (2018) found that a multilingual NMT system trained on multiple related languages showed improved zero-shot translation performance. We believe that multilingual NMT can be effective for closely-related dialects, and can compensate for a lack of translation data for the different dialects.

Multilingual NMT can also help us to analyze the characteristics of each language. Östling and Tiedemann (2016) 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 analyze our dialect embeddings to investigate whether our multi-dialect model can capture similarities between dialects (Section 5).

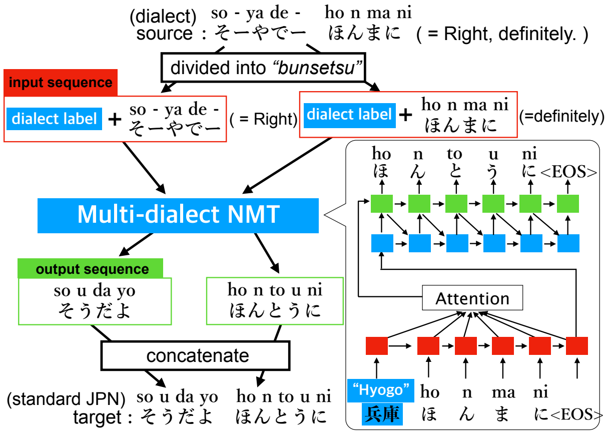

(dialect)
source : so - ya de - そーやでー ho n ma ni ほんまに (= Right, definitely.)
divided into "bunsetsu"
input sequence
dialect label + so - ya de - そーやでー (= Right)
dialect label + ho n ma ni ほんまに (=definitely)
Multi-dialect NMT
output sequence
so u da yo そうだよ
ho n to u ni ほんとうに
concatenate
(standard JPN)
target : so u da yo そうだよ ho n to u ni ほんとうに
ho ほ n ん to と u う ni に <EOS>
Attention
"Hyogo" 兵庫 ho ほ n ん ma ま ni に <EOS>

Figure 1: Proposed multi-dialect NMT model.

Previous work reported that character-level statistical machine translation (SMT) using words as translation units was effective for translating between closely-related languages (Nakov and Tiedemann, 2012; Scherrer and Ljubešić, 2016). There are two reasons for this: the character-level information enables the system to exploit lexical overlaps, while using words as translation units takes advantage of related languages' syntactic overlaps.

In this study, we present a method of translating between Japanese dialects that combines three ideas: multilingual NMT, character-level NMT, and the use of 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 in low-resource settings.

3 Data: Japanese Dialect Corpus

Japanese is a dialect-rich language, with dozens of dialects that are used for everyday conversations in most Japanese regions. They can be characterized in terms of differences in their content words (vocabulary) and regular phonetic shifts, mostly in their postpositions and suffixes. That said, they nonetheless share most words with standard Japanese, as well as mostly using common grammatical rules, such as for word ordering, syntactic marker categories, and connecting syntactic markers. Some dialects also share the dialect-specific vocabulary.

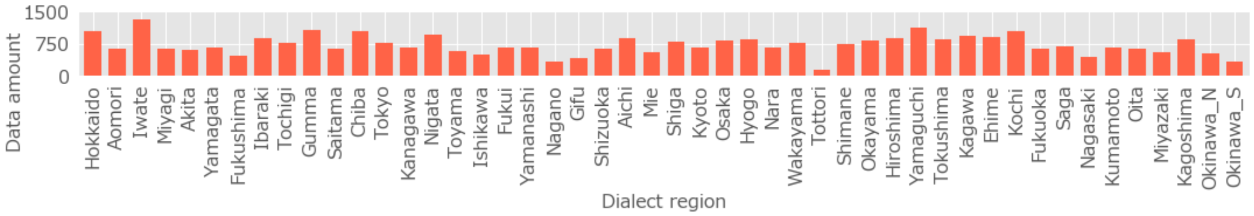


Figure 2: Number of sentences in each dialect in *The National Dialect Discourse Database Collection*.

In this study, we used a collection of parallel textual data for dialects and standard Japanese, *the National Dialect Discourse Database Collection* (NINJAL, 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, and Figure 2 shows the number of sentences in each dialect.

Japanese texts are generally written in a mix of *kanji* and *kana*; therefore, we converted the *kanji* in the sentences into *kana* and, then, segmented them into *bunsetsu*.[1] After preprocessing, the average sentence lengths were 14.62 and 15.57 characters for the dialects and standard Japanese, respectively, and the average number of *bunsetsu*s per sentence was 3.42.

4 NMT Model

Figure 1 gives an overview of our multi-dialect NMT system's network structure. Since our focus is on examining the effectiveness of multi-dialect NMT and its detailed behavior, rather than on creating a novel translation model, we used OpenNMT (Klein et al., 2017), a stacking LSTM encoder–decoder model with a multilingual extension similar to that of Johnson et al. (2016). However, to improve its direct translation accuracy, we make the following three modifications.

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

ID	Encoder input order
a	source label, sentence
b	target label, sentence
c	source label, target label, sentence
d	source label, sentence, target label

Table 1: Dialect label input order variants.

Syllable-to-syllable translation As mentioned in Section 3, one of the keys to translating between two closely-related languages is to model the phonetic correspondences between them. Thus, to consider syllable-level translation rules that may be shared by similar dialects, we define our translation task as syllable-to-syllable translation.

We realize syllable-to-syllable translation by representing the inputs and outputs as *kana* sequences and performing character-based MT. A similar approach has been used to normalize Japanese text from Twitter, where the main issue was phonological transliteration (Saito et al., 2017). In our dataset, the dialect utterances have all been transcribed using *kana*; however, the standard Japanese translations use a mix of *kanji* and *kana* characters. Therefore, we converted these into *kana* sequences as well by automatically analyzing the pronunciation of each *kanji* character and replacing it with the corresponding *kana* sequence.

Translation without distortion Standard MT methods take a single sentence as input and yield

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

a translated sentence in an appropriate word order for the target language. However, in 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, finding no changes in ordering (distortion). This fact could enable the model to be more efficiently trained on the translation pairs, and hence require less supervision data, because it does not need to learn a distortion model. Based on this intuition, we split each input sentence into base-phrase chunks, namely a *bunsetsu* sequence, translate each chunk from the source to the target language and, then, 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 by BLEU scores. 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 proportions of 8:1:1. We over-sampled 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.

Since Japanese dialects mostly share the same vocabulary and grammatical rules and there are few distortions (word order changes), translation between a Japanese dialect and standard Japanese is relatively easy compared with translating between different languages. Given that, the main focus of these experiments was to evaluate how well the model captured the phonological shifts between dialects and standard Japanese. For this reason, we employed syllable-level BLEU scores as an evaluation measure and, then, macro-averaged the scores over all dialects. Note that, this evaluation measure generally gives higher scores than if we had calculated it at the word-level. In fact, the macro-averaged BLEU score reached 35.10 even when we simply output the dialect sentences as they were (without translation).

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)	64.04

Table 2: Syllable-level BLEU scores for all models

We used OpenNMT-py[2] with its default hyperparameter settings, except for the number of the training epochs (which we set to 20) and selected the model that performed best on the development set. In addition, we employed Moses[3] (Koehn et al., 2007) as the baseline SMT model and set the distortion limit to 0. The standard Japanese language model used in Moses was trained with KenLM (Heafield, 2011).

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

Finally, we used MeCab 0.996[4] to analyze the *kanji* characters' pronunciations.

5.2 Multi-Dialect NMT Model Performance

Table 2 shows the dialect translation performance of all the models considered, with the first row group comparing their scores for dialect-to-standard translation with different input settings.

Mono-lingual vs multi-lingual For comparison, we first consider a model that was trained using only a single set of dialect-standard parallel data (Mono NMT). This performed quite poorly compared with the other models that used data for all the dialects (Multi NMT), and was even worse than simply outputting the dialect sentences unchanged

[2] https://github.com/OpenNMT/OpenNMT-py

[3] http://www.statmt.org/moses

[4] http://taku910.github.io/mecab/

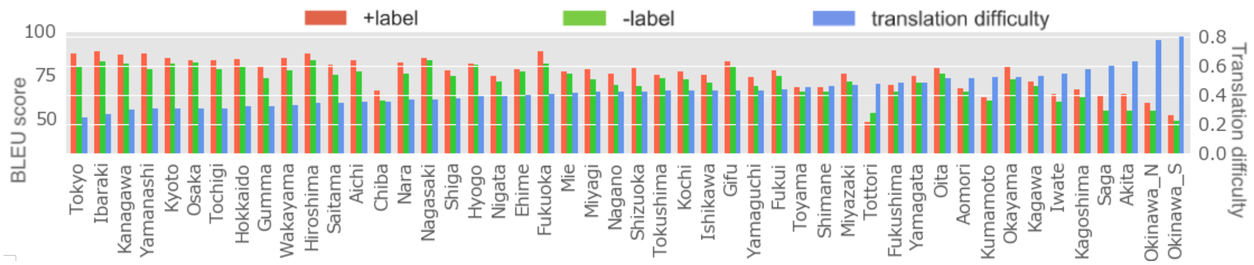


Figure 3: Multi NMT models' BLEU scores and translation difficulty for all dialect.

(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 demonstrated drastically improved the translation performance.

The Multi NMT model's performance (BLEU score) for standard-to-dialect translation was slightly lower than that for dialect-to-standard translation (last row of Table 2). This is consistent with a previously-reported multilingual NMT result (Johnson et al., 2016).

Dialect labels Including dialect labels improved the Multi NMT model's macro-averaged BLEU score by 4.37 points (forth row of Table 2) compared with the same model without dialect labels (second row). Figure 3 shows these two models' BLEU scores for all dialects in ascending order of translation difficulty. Here, the translation difficulty is defined as the average normalized Levenshtein distance over all sentence pairs (dialect and standard Japanese) for a given dialect. As expected, the BLEU scores for each dialect show a strong negative correlation ($\rho = -0.82$) with its translation difficulty. In addition, we can observe that the model with language labels consistently outperforms the model without labels, except for the Tottori dialect. This indicates that explicitly providing the source and target dialects 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 chunk-wise translation is it cannot capture the context beyond each chunk's boundary. However, despite this disadvantage, our Multi NMT model was still able to outperform a model that had access to a broader context (Multi NMT-sentence), indicating that our fixed-order translation approach is suitable for translating Japanese dialects despite its limited context sensitivity.

NMT vs SMT Zoph et al. (2016) found that SMT models largely outperformed state-of-the-art NMT models for low-resource languages. Therefore, for comparison, the second row group in Table 2 shows results for a fixed-order character-based SMT baseline. In these experiments, even though the NMT model trained using a single dialect (Mono NMT) gave the poorest performance, the one with dialect labels outperformed the baseline Multi SMT model, achieving the best performance overall.

5.3 Example Translation Results

To demonstrate how each of the proposed components contributed to generating accurate translations, we now present some concrete examples of our models' translation results for the Hyogo, Kagoshima, and Nigata dialects (Table 3).

Comparing the Multi NMT models with and without dialect labels, we can see that adding labels enabled the model to better translate the chunks that required dialect-specific knowledge. In Example 1, the source sentence includes a local name (*O* -) for a certain area (*Aioi*) in Hyogo, which only the model with dialect labels could successfully translate. In addition, in Example 2, including labels enabled 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 "ta ra").

Similarly, since the Multi SMT model could not

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)

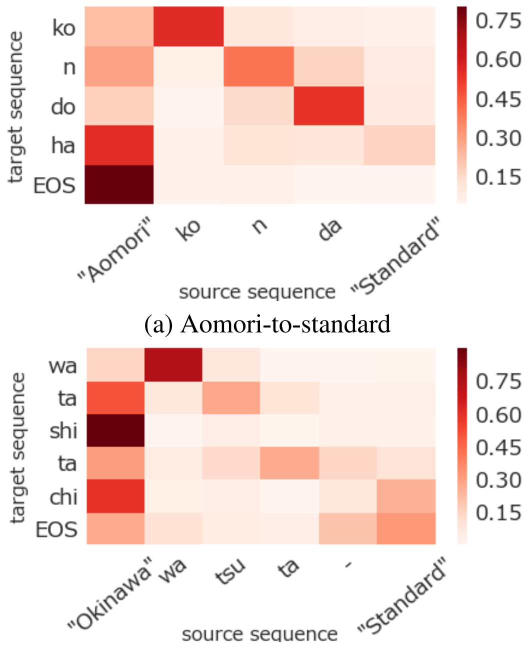
Example 2	Kagoshima region
Meaning	After a few days, then it was...
Source	so i ga / mo / na n ni k ka / shi ta ya (そいが / も / なんにっか / したや)
Reference	so re ga / mo u / na n ni chi ka / shi ta ra (それが / もう / なんにちか / したら)
Multi NMT (w/o label)	so re ga / mo u / na n ni chi ka / shi ta da (それが / もう / なんにちか / 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 / なんにっか / 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)

Table 3: Example dialect-to-standard translations for the Hyogo, Kagoshima, and Nigata dialects.

take advantage of the dialect labels, it failed to capture dialect-specific translation rules.

In the previous section, we saw that our fixed-order translation approach is suitable for translating Japanese dialects, despite its limited context sensitivity. However, this became a problem in Example 3, where the proposed chunk-wise translation models could not correctly translate a phrase due to the lack of context. Here, none of the models, except Multi NMT-sentence, could translate the phrase "mi zu n" in the Nigata dialect to the correct standard Japanese phrase "mi zu no", since the translation of



(a) Aomori-to-standard

(b) Okinawa-to-standard

Figure 4: Attention weight examples. (a) *Aomori-to-standard* translation of "next time". The Aomori word *konda* is made by linking the syllables from two standard Japanese words, *kondo* (next time) and *ha* (topic marker). (b) *Okinawa-to-standard* translation of "we". The 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.

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 a potential future directions: extending fixed-order translation to incorporate contextual information.

5.4 Visualizing Attention Weights

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

Figure 4(a) shows the model's attention history for an example where the target language syllables were derived from those in the source language according to a regular rule. In such cases, the model tended to weight the dialect label heavily when applying the rule. Conversely, Figure 4(b) shows the attention history for an example where almost all of the syllables were transliterated. In these cases, the model needed to disambiguate the morpheme-level meanings to create a correct translation and thus tended to focus on the entire sequence of semantically- or grammatically-related morphemes.

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

Table 4: Impact of excluding the nearest or farthest five dialect regions from the training data when calculating the BLEU score for each diaect region. Here, "Avg. Δ" denotes the average BLUE score difference compared with using all the data.

5.5 Visualizing Language Labels

Östling and Tiedemann (2016) reported that clustering the language embeddings used to train a multilingual language model produced a language cluster structure similar to the established relationships among language families. Inspired by their work, we decided to examine the relationships between the dialect embeddings and the dialects' typology.

Figure 5 shows a t-SNE projection of the dialect embeddings. This 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 a dialectological typology theory known as *center versus periphery* (Yanagida, 1980), where new language use trends gradually propagate from the cultural center (the old capital, Kyoto) to less culturally-influential areas. This potentially explains why the dialects in the Tohoku region (E) are similar to those of 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 of the different dialect regions. As we saw 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

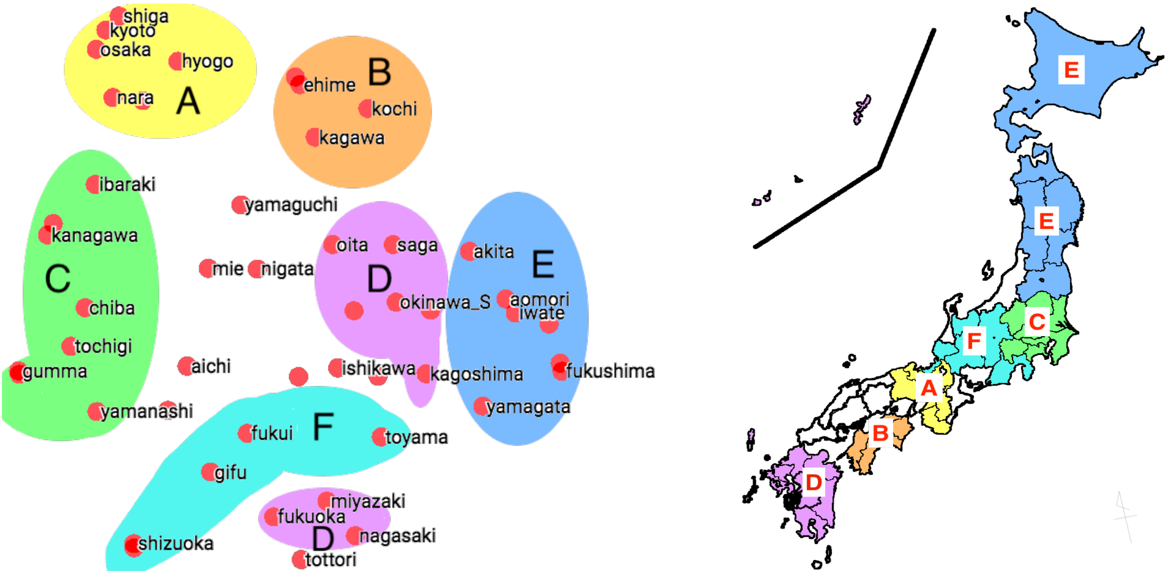


Figure 5: t-SNE projection of the dialect label vectors. Dialects belonging to the same region are shaded using the same background color.

from other dialects will differ depending on their geographical distances from the target dialect.

To investigate this assumption, we prepared two Multi NMT models per dialect, trained on data that excluded the five geographically nearest or farthest dialects[5] for the given dialect region, and calculated the differences in BLEU score between these models and the original model that used all the dialect data. For example, one of the Tokyo dialect models excluded training data from the nearby Chiba, Kanagawa, Saitama, Gumma, and Ibaraki dialects but was otherwise trained the same way as the full Multi NMT model. Then, we compared this model's BLEU score for Tokyo dialect instances in the test set to that of the full model.

Table 4 shows the average results over all 48 models for both 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 with the full model, indicating that even very distant dialects still helped with training other dialects. In addition, we can see that removing the nearest five dialects had a more significant impact than removing the farthest five, implying that similar dialects contribute more to helping a multi-dialect NMT learn effectively.

6 Conclusion

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

One limitation of the proposed model is it cannot consider longer-range dependencies beyond the chunk level. Therefore, our future research plans include incorporating contextual information into fixed-order translation models and investigating the dialect embeddings' characteristics further.

Acknowledgments

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

[5] The distances between dialect pairs were calculated using the Euclidean distances between the points where the dialogs were recorded.

References

Jiatao Gu, Hany Hassan, Jacob Devlin, Victor O K Li, and Google Research. 2018. Universal Neural Machine Translation for Extremely Low Resource Languages. In *Proceedings of The 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 344–354.

Chinnappa Guggilla. 2016. Discrimination between similar languages, varieties and dialects using cnn- and lstm-based deep neural networks. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 186–194.

Hany Hassan, Mostafa Elaraby, and Ahmed Y Tawfik. 2017. Synthetic Data for Neural Machine Translation of Spoken-Dialects. *arXiv preprint arXiv:1707.00079.*

Kenneth Heafield. 2011. KenLM : Faster and Smaller Language Model Queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, 2009, pages 187–197.

Pierre-Edouard Honnet, Andrei Popescu-Belis, Claudiu Musat, and Michael Baeriswyl. 2018. Machine Translation of Low-Resource Spoken Dialects: Strategies for Normalizing Swiss German. In *Proceedings of 11th edition of the Language Resources and Evaluation Conference*, pages 3781–3788.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. In *Proceedings of Transactions of the Association for Computational Linguistics*, volume 5, pages 339–351.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. *arXiv preprint arXiv:1701.02810.*

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran Mit, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of 45th Annual Meeting of the Association for Computational Linguistics Demo and Poster Sessions*, pages 177–180.

Yasuo Kumagai. 2016. Developing linguistic atlas of japan database and advancing analysis of geographical distributions of dialects. *The future of dialects: Selected papers from Methods in Dialectology XV.*, pages 333–362.

Surafel M Lakew, Cettolo Mauro, and Marcello Federico. 2018. A Comparison of Transformer and Recurrent Neural Networks on Multilingual Neural Machine Translation. *arXiv preprint arXiv:1806.06957.*

Karima Meftouh, Salima Harrat, Salma Jamouss, Mourad Abbas, and Kamel Smaili. 2015. Machine Translation Experiments on PADIC: A Parallel Arabic DIalect Corpus. In *Proceedings of 29th Pacific Asia Conference on Language, Information and Computation*, pages 26–34.

Preslav Nakov and Jörg Tiedemann. 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)*, volume 10, pages 301–305.

John Nerbonne and William A. Kretzschmar. 2011. Dialectometry++. *Literary and Linguistic Computing*, 28(1):2–12.

NINJAL. 1980. *"Zenkoku Hougen Danwa Database Nihon no Furusato Kotoba Syusei" (The National Dialect Discourse Database Collection).* Kokushokankokai Inc.

Robert Östling and Jörg Tiedemann. 2016. Continuous multilinguality with language vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 2, pages 644–649.

Taraka Rama and Çağrı Çöltekin. 2016. LSTM autoencoders for dialect analysis. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 25–32.

Itsumi Saito, Jun Suzuki, Kyosuke Nishida, and Kugatsu Sadamitsu. 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*, pages 257–262.

Yves Scherrer and Nikola Ljubešić. 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)*, pages 248–255.

Kunio Yanagida. 1980. *"Kagyuko"*. Iwanami Shoten, Publishers.

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar F Zaidan, and Chris Callison-Burch. 2012. Machine Translation of Arabic Dialects. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*, pages 49–59.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer Learning for Low-Resource Neural Machine Translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575.