Development of Text Translation System from Tsugaru Dialect into Common Japanese

Taiki Niida
Hirosaki University
h21ms420@hirosaki-u.ac.jp

Masasi Imai
Hirosaki University
miyabi@hirosaki-u.ac.jp

Abstract – Tsugaru dialect can be an obstacle to communication between Aomori residents and residents who have transferred there for work and tourists from outside the prefecture. We are developing a bidirectional voice and text translation system between Tsugaru-ben and common Japanese utilizing artificial intelligence. In this paper, our research project is firstly introduced and the developed text translation system from Tsugaru dialect into common Japanese is explained. Some evaluation results of the morphological analysis and translation tools are also shown.

I. Introduction

Tsugaru-ben, which is a dialect particular to the Tsugaru region in Aomori prefecture, can be an obstacle to communication between Tsugaru residents and residents who have transferred there for work and tourists from outside the prefecture and abroad. For example, interactions between patients from the region and a doctor from outside the region in the medical field and communications between elder farmers in the region and student part-time workers from outside the prefecture in the agricultural field can be considered. In particular, in medical and disaster situations, even if a patient complains of symptoms, if the doctor cannot understand the symptoms, the appropriate medical treatments cannot be given, resulting in life-threatening. Therefore, it is desirable to develop a system that performs voice and text translation between Tsugaru-ben and the common Japanese.

The Tsugaru region includes four areas: Chunan-Tsugaru, Kitago-Tsugaru, Nishi-Tsugaru, and Tosei-Tsugaru, and the Tsugaru-ben used in each area is slightly different. It is often seen that even young people in the Tsugaru region either cannot understand the Tsugaru-ben spoken by older people, or the young people can understand it but not use it themselves, resulting in the disappearance of the old Tsugaru culture.

With the advance of VLSI manufacturing technology, the computing performance and data storage spaces have been significantly increased. Various services utilizing Artificial Intelligence (AI) have also been provided. Among them, there are several services that translate between multiple languages, such as Japanese to English, French, and Chinese. However, very few attempts have been made at such dialect translation and no system has been found to fully support the Tsugaru-ben.

We are conducting the “(omitted) x AI x Tsugaru-ben Project” with the aim of developing a bidirectional voice and text translation system between Tsugaru-ben and the common Japanese utilizing AI. We will extend the existing text translation method to Tsugaru dialect and try to evaluate it. Tsugaru-ben in various academic and cultural fields is widely collected by a cross-faculty team and used for AI learning. The collected Tsugaru-ben is also systematically organized to establish a data infrastructure that can be used by the next generation. In this paper, our current work is explained and some evaluation results of morphological analysis and translation tools are shown.

II. Text translation from Tsugaru-ben into common Japanese

Out text translation system from Tsugaru-ben into common Japanese consists of two sub-systems as shown in Fig. 1. The former performs morphological analysis using a Tsugaru Dialect Dictionary which generates word-separated sentences from Tsugaru-ben sentences in which there is no break point except punctuation characters in Japanese. The latter performs translation from word-separated Tsugaru-ben into the corresponding word-separated common Japanese using AI which is implemented by neural networks with deep learning.

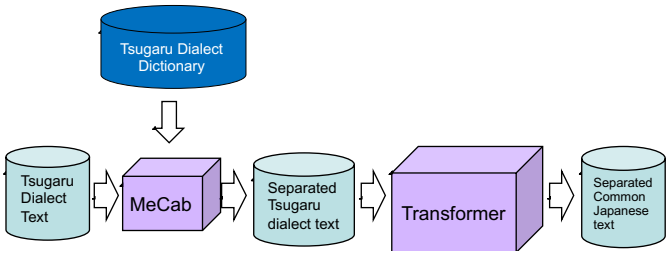

Fig. 1. Text translation system from Tsugaru-ben into common Japanese.

A. Morphological analysis

In natural languages such as Japanese and English, the smallest components that make sense in a sentence are called morphemes. Morphological analysis is a process of breaking down a sentence into a list of morphemes and clarifying each grammatical attributes, which represent the type of parts of speech, conjugation of verbs, and etc., in order to find out what kind of elements the sentence is composed of. There is a morphological analysis tool for Japanese called MeCab[1]. MeCab is an open-source morphological analysis engine developed through a joint research unit project between the

Graduate School of Informatics, Kyoto University and the Nippon Telegraph and Telephone Corporation Communication Science Laboratories.

Figure 2 and Fig.3 show morphological analysis results of a common Japanese sentence "今日は良い天気だ。" which means that "It's fine weather today." and a Tsugaru-ben sentence "へごまなひとだ。" which means that "He is a diligent person." using MeCab without any expansion for Tsugaru-ben.

```
今日 名詞,副詞可能,*,*,*,今日,キョウ,キョー
は　助詞,係助詞,*,*,*,は,ハ,ワ
良い 形容詞,自立,*,*,形容詞・アウオ段,基本形,良い,ヨイ,ヨイ
天気 名詞,一般,*,*,*,天気,テンキ,テンキ
だ　助動詞,*,*,*,特殊・ダ,基本形,だ,ダ,ダ
。　記号,句点,*,*,*,*,。,。,。
```

Fig. 2. A morphological analysis example of a common Japanese sentence.

```
へ　助詞,格助詞,一般,*,*,*,へ,へ,エ
ごま　名詞,一般,*,*,*,*,ごま,ゴマ,ゴマ
な　助動詞,*,*,*,特殊・ダ,体言接続,だ,ナ,ナ
人　名詞,一般,*,*,*,*,人,ヒト,ヒト
だ　助動詞,*,*,*,特殊・ダ,基本形,だ,ダ,ダ
。　記号,句点,*,*,*,*,。,。,。
```

Fig. 3. A morphological analysis example of a Tsugaru-ben sentence.

As shown in Fig. 2, the common Japanese sentence is correctly morphologically analyzed and separated into a list of the corresponding morphemes. On the other hand, it can be seen that the Tsugaru-ben sentence is separated into the incorrect list of morphemes. Correctly, "へごま" must not be separated since it means "diligent" in Tsugaru dialect. This is because the general MeCab's library does not have words of Tsugaru dialect.

MeCab supports extra libraries generated by users. Thus, we build a library for Tsugaru dialect analysis. The format of MeCab library is as follows;

- Surface form, left context ID, right context ID, cost, part of speech, part-of-speech classification 1, part-of-speech classification 2, part-of-speech classification 3, conjugation, conjugation type, prototype, reading, pronunciation

In addition, if a verb has a conjugation, it is necessary to register all the conjugated forms in the library.

In our research project, a Tsugaru-ben database has been constructed which contains about 10,000 Tsugaru words and their related information such as the corresponding common Japanese word, example sentences in which the word is used, and user information like area, age, and gender. The construction of the MeCab library for Tsugaru dialect analysis is based on the following steps.

- Extract Tsugaru-ben word and its corresponding common Japanese word from the database.
- Determine whether the part of speech of the word is a verb or a non-verb.
  - ➢ If it is a verb, expand all the conjugated forms.
- If morphemes of the corresponding common Japanese word exist in the general MeCab library, the stored parameters are obtained and applied to the original Tsugaru-ben morphemes since it can be considered that the same parameters can be used.
- If the morphemes of the corresponding common Japanese word do not exist in the general MeCab library, the appropriate parameters for the specific Tsugaru-ben morphemes are set.
- Save them into a CSV file and compile it.

An in-house tool which supports the above steps is implemented in Python. As a result, the Tsugaru Dialect Dictionary for MeCab which contains 7,720 non-verb morphemes and 20,980 verb morphemes can be developed. Some evaluation and comparison results are shown in the next section.

B. Text translation using AI

The second sub-system that translates word-separated Tsugaru-ben sentences into the corresponding word-separated common Japanese sentences is implemented in Python using the machine learning software framework called TensorFlow[2] and Keras[4]. Figure 4 shows the translation flow. In this study, the Transformer model[3][5] is used to evaluate translation accuracy.

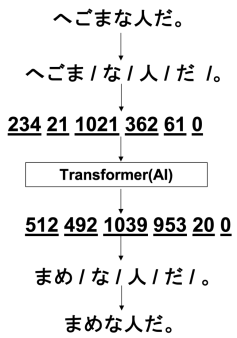

Fig. 4. Translation flow.

Transformer is a deep learning model that Google published in 2017. The model handles variable-length input using a self-attention layer instead of an RNN (Recurrent Neural Network) and CNN (Convolutional Neural Network). It has the following advantages[4].

- No required assumption of temporal / spatial relationships in the data
- Layer outputs which are not in series like RNNs can be computed in parallel.
- Distant elements can influence each other's outputs without a number of RNN steps and convolutional layers.
- Long range dependencies can be learned.

Figure 5 shows the architecture of Transformer. It mainly consists of an Encoder and a Decoder. The Encoder is

responsible for extracting features of source sentences. The Decoder is responsible for generating the translated sentences from features of the source sentences obtained from the Encoder. In the learning phase a translation source sentence is input to the Encoder and the corresponding translated sentence is input to the Decoder as shown in Fig. 5. However, both the sentences should be preprocessed before input.

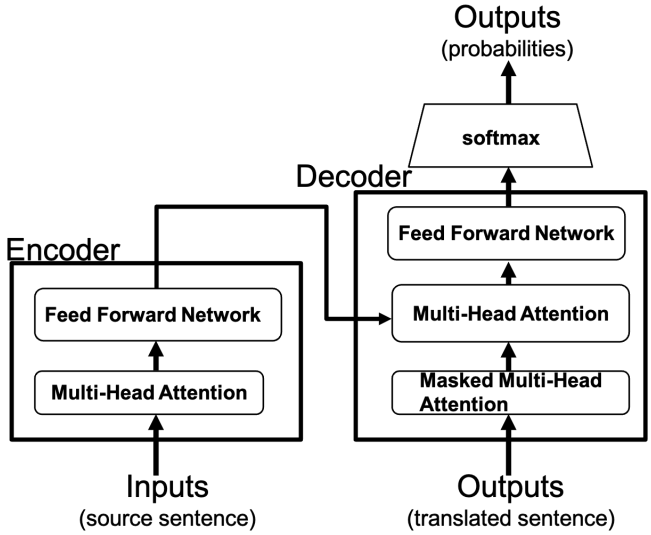

Fig. 5. The architecture of Transformer.

The preprocess steps are as follows;

1. Obtain a list of word-separated sentences using MeCab. When Tsugaru-ben sentences are applied, the developed Tsugaru Dialect Dictionary is used.
2. Assign start and end tokens to indicate the beginning and ending of the sentence, respectively.
3. Replace words with IDs.
4. Perform zero-filling in order to uniform the length of the sentences since the Transformer model requires almost the same number of words in a sentence to efficiently process them.

In addition, the maximum number of words in a sentence should be determined in advance.

In Fig.5, the Attention plays a role to connect between the Encoder and the Decoder. It determines the importance of words in the source sentence and send it to the Decoder. The Decoder can correctly select the appropriate words according to the received information. In the Attention model, input data are calculated using a vector which contains Query, Key, and Value. The Attention score of each word is calculated Query and Key and its weighted sum is calculated using Value and the calculated Attention score. As a result, the latent expression of the word can be obtained. In the Multi-Head Attention model, the specified multiple words are assumed as a Head and its vector which represents the latent expression of the Head is calculated. Finally, the latent expression of the target word is represented as the following equations using the obtained multiple Attentions of Heads. The latent expression space is different each other. Thus, the effective information can be collected in this scheme.

$$MultiHead(Q,K,V) = concat(head_1,\cdots,head_h)W^o$$
$$head_i = Attntion\left(QW_i^Q, KW_i^K, VW_i^V\right)$$

An in-house tool which supports the above steps is also implemented in Python.

III. Evaluation

In this section, some evaluation and comparison results are shown. TABLE I shows the numbers of Tsugaru dialect words and sentences used for evaluation. They have been collected through our research project.

TABLE I
The numbers of Tsugaru dialect words and sentences.

# words	10,150
# sentences	6,066

First, the effectiveness of the developed Tsugaru Dialect Dictionary for MeCab is evaluated. 3 sets of 50 sentences are randomly selected and the accuracy of the morphological analysis using MeCab with / without the Tsugaru Dialect Dictionary is manually countered. TABLE II shows evaluation results of the accuracy of the morphological analysis.

TABLE II
The accuracy of morphological analysis using MeCab.

	Without Tsugaru Lib.	With Tsugaru Lib.
1st set	28%	62%
2nd set	36%	68%
3rd set	20%	62%
Average	28%	64%

As shown in TABLE II, the accuracy of the morphological analysis is improved from 28% to 64% on average with the developed Tsugaru Dialect Dictionary. However, the absolute values are still small in order to practically use in field. It is necessary to further improve the percentage of the correct answers since the accuracy of translation with the developed system is greatly affected by the morphological results. For this purpose, we will expand the Tsugaru Dialect Dictionary for MeCab by adding unregistered words and modifying the parameters.

Second, the accuracy of translation is evaluated. BLEU (BiLingual Evaluation Understudy) score is a metric used to evaluate machine translation results. Values are represented as

real numbers between 0 to 1. TABLE III shows the approximate values[6].

TABLE III
Approximate values.

0.1 or less	Little or no help
0.1 ~ 0.2	Difficult to understand the main idea
0.2 ~ 0.3	The main idea is clear, but there are serious grammatical errors
0.3 ~ 0.4	Understandable, moderate quality translation
0.4 ~ 0.5	High quality translation
0.5 ~ 0.6	Very high quality translation that is adequate and fluent
0.6 or higher	Higher quality than human translation

BLEU score calculates the similarity between machine translation result and the reference translation by the following formula based on the number of *n*-gram which matches between the two.

$$BLEU = BP_{BLEU} \cdot exp\left(\sum_{n=1}^{N} \omega_n \log p_n\right)$$

where P_n and ω_n denote the following terms;

$$P_n = \frac{\sum_i \text{The number of } n\text{-gram which matches between machine translation}_i \text{ and reference translation}_i}{\sum_i \text{The number of total } n\text{-gram in machine translation}_i}$$

$$\omega_n = \frac{1}{N}$$

P_n is the n-gram (2-gram) agreement rate calculated by comparing the translation and the reference translation for the entire evaluation corpus. The score is calculated by finding the geometric mean for *n*-gram from 1-gram. *N* is usually taken as 4. 1-gram is a metric for word translation correctness, and the higher dimension n-gram is a metric for translation fluency, so the BLEU score is a combination of the two. BP_{BLEU} represents the given penalty when the translation is shorter than the reference translation. On the other hand, it is 1 when the translation is longer than the reference translation[7].

TABLE IV shows the model definition and learning parameters. The number of nodes of the input layer, 4 middle layers, and the output layer are 64. The word number of Multi-Head Attention is assumed as 4. In this evaluation, 6066 sentences shown in TABLE I are used for evaluation. 95 percent of the data is used for learning and the remaining 5 percent is used for evaluation, *i.e.*, 5763 sentences and 303 sentences are randomly selected, respectively. In this evaluation, 10 pairs of the learning and evaluation data are prepared. The epoch number and the batch size to learn the data are taken as shown in TABLE IV.

TABLE IV
Model and learning parameters.

# middle layers	4
# nodes of layers	64
Multi-Head Attention	4
# epoch	100
Batch size	16, 32, 64

TABLE V shows evaluation results of the average BLEU scores in each trial. The values in the parentheses represent the minimum and maximum values, respectively. It is observed that the BLEU score is considerably varied. For example, when the second pair is used and the batch size is assumed as 32, the maximum BLEU score is 0.960 and the minimum BLEU score is 0, resulting in 0.183 on average. The following sentences represent the input-and-output pair whose BLEU score is 0.960. In this case, it is confirmed that the translated output sentence is understandable.

Original input:
「せっかく日本さきたはんで、相撲とば見に行くべし。」
Translated output:
「せっかく日本にきたのだから、相撲を見に行きましょう。」

TABLE V
Average BLEU score.

	Batch 16	Batch 32	Batch 64
Pair 1	0.084(0~0.8)	0.112(0~0.77)	0.124(0~0.75)
Pair 2	0.115(0~0.90)	0.104(0~0.87)	0.104(0~0.90)
Pair 3	0.187(0~0.83)	0.210(0~0.96)	0.225(0~0.96)
Pair 4	0.170(0~0.84)	0.208(0~0.85)	0.214(0~0.84)
Pair 5	0.189(0~0.90)	0.197(0~0.75)	0.228(0~0.93)
Pair 6	0.166(0~0.82)	0.193(0~0.82)	0.218(0~0.82)
Pair 7	0.173(0~0.86)	0.213(0~0.85)	0.221(0~0.85)
Pair 8	0.168(0~0.85)	0.204(0~0.84)	0.230(0~0.84)
Pair 9	0.163(0~0.83)	0.202(0~0.81)	0.227(0~0.84)
Pair 10	0.139(0~0.66)	0.186(0~0.96)	0.206(0~0.97)
Average	0.155	0.183	0.200
Min	0	0	0
Max	0.90	0.96	0.97

However, the average BLEU score evaluated is up to 0.230 which is classified as “the main idea is clear, but there are serious grammatical errors.” The main reason for the low score can be attributed to the small volume of datasets. Thus, it is necessary to build other models with different number of layers and nodes. In addition, while the Transformer model was used in this evaluation, it is in the scope of our future work to construct and evaluate models using the seq2seq model[8] and other models.

The Tsugaru-ben corpus which includes Tsugaru dialect words, sentences, and voice data is widely collected from the public through the website (http://tgrb.jp/) shown in Fig. 6. It is continued to collect the Tsugaru-ben corpus in order to enhance the learning data and improve the translation accuracy.

Fig. 6. The project website.

IV. Conclusion

We have developed a prototype translation system from Tsugaru-ben text to the corresponding common Japanese text in order to support communications between the Tsugaru residents and residents who have transferred there for work and tourists from outside the prefecture and abroad. The translation system uses the morphological analysis tool MeCab and the deep learning framework TensorFlow and Keras. For morphological analysis, the Tsugaru Dialect Dictionary which contains 28,000 morphemes was constructed. In this paper, the evaluation result of the first trial of translation using the Transformer model is shown. Its accuracy is still low and many issues remain to be solved. We try to improve the accuracy by modifying the developed prototype tools and enhancing learning data.

Acknowledgments

This work is partially supported by Hirosaki University's next-generation institutional research and the Mutsu Ogawara Regional and Industrial Promotion Foundation.

References

[1] Kudo, T. “MeCab: Yet another part-of-speech and morphological analyzer,” https://mecab.sourceforge.net/, 2010.

[2] Dillon J.V., Langmore I., Tran D., Brevdo E., Vasudevan S., Moore D., Saurous R. A. “TensorFlow distributions,” arXiv preprint arXiv:1711.10604.

[3] Vswani Ashish, et al., “Attention is all you need,” Advances in neural information processing systems 30, 2017.

[4] https://keras.io/

[5] https://www.tensorflow.org/tutorials/text/transformer

[6] https://cloud.google.com/translate/automl/docs/evaluate#bleu

[7] Kenji Imamura, et al., “The comparison of automatic evaluation criteria of machine translation,” 10th Annual Convention of the Association for Natural Language Processing, pp.452-455, 2004 (in Japanese).

[8] Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks." *Advances in neural information processing systems* 27, 2014.