Machine Translation for Highly Low-Resource Language: A Case Study of Ainu, a Critically Endangered Indigenous Language in Northern Japan

So Miyagawa
National Institute for Japanese Language and Linguistics
Midoricho 10-2, Tachikawa, Tokyo
miyagawa.so.36u@kyoto-u.jp

Abstract

This paper explores the potential of Machine Translation (MT) in preserving and revitalizing Ainu, an indigenous language of Japan classified as critically endangered by UNESCO. Through leveraging Marian MT, an open-source Neural Machine Translation framework, this study addresses the challenging linguistic features of Ainu and the limitations of available resources. The research implemented a meticulous methodology involving rigorous preprocessing of data, prudent training of the model, and robust evaluation using the SacreBLEU metric. The findings underscore the system's efficacy, achieving a SacreBLEU score of 32.90 for Japanese to Ainu translation. This promising result highlights the capacity of MT systems to support language preservation and aligns with recent research emphasizing the potential of computational techniques for low-resource languages. The paper concludes by affirming the significant role of MT in the broader context of language preservation, serving as a crucial tool in the fight against language extinction. The study paves the way for future research to harness advanced MT techniques and develop more sophisticated models for endangered languages.

1 Introduction

The Ainu language, a polysynthetic and culturally rich language, has been traditionally spoken by the Ainu people in the northern regions of Japan, such as Hokkaido, Southern Sakhalin, and the Kuril Islands. Despite its intricate structure, the Ainu language faces significant endangerment. In 2009, UNESCO classified Ainu as a "critically endangered" language (Moseley, 2010), underscoring the critical need for efforts towards its preservation. The language's vulnerability is further highlighted by the dwindling number of native Ainu speakers and the loss of many Ainu dialects, including Sakhalin Ainu and Kuril Ainu.

The language itself has linguistic uniqueness such as polysynthesis and noun incorporation, which are characteristics of many indigenous languages in North America. Example 1 exemplifies its polysynthesis and noun incorporation.

(1) Hokkaido Ainu (Shibatani, 1990, 72)

Usa-opuspe
various-rumors
a-e-yay-ko-tuyma-si-ram-suy-pa
1SG-APL-REFL-APL-far-REFL-heart-sway-ITR

"I wonder about various rumors."[1]

Against this backdrop, this study aims to employ the advancements in Natural Language Processing (NLP) and Machine Translation (MT) to further our understanding and translation of the Ainu language. This research endeavors to leverage these technologies to contribute to the survival and revival of the Ainu language, especially given the urgency emphasized by its UNESCO status.

The ultimate objective of this study is to develop an AI-assisted educational program and a teaching robot to facilitate the learning and preservation of the Ainu language. The proposed program intends to incorporate several components like speech recognition, speech generation, part-of-speech tagging, and Universal Dependencies tagging, among other linguistic technologies, based on recent studies on the Ainu language.

Previous studies, such as the work of Nowakowski et al. (2019) on the Mingmatch—an n-gram model for Ainu word segmentation, and the creation of an Ainu folklore speech corpus by Matsuura et al. (2020b), the works mentioned in

[1] The list of abbreviations in the gloss: 1SG = first-person singular, APL = applicative, REFL = reflective, ITR = iterative.

Proceedings of the Joint 3rd NLP4DH and 8th IWCLUL, pages 120–124
December 1-3, 2023 ©2023 Association for Computational Linguistics

the next section, have laid the groundwork for this research. Building on these pivotal studies, this research aims to develop a robust NLP model that leverages the Marian MT as the primary translating model for Ainu to Japanese and Japanese to Ainu. The insights gained from this endeavor will inform the design of AI-assisted educational tools, thereby fostering the preservation and understanding of the Ainu language and culture.

2 Previous Literature

The potential for leveraging advanced computational techniques such as NLP and MT for language revitalization is gradually being explored. Previous work by Nowakowski et al. (2019) showcased a fast n-gram model for word segmentation of the Ainu language. This work signaled the potential of computational approaches for improving the accessibility and study of Ainu. Further efforts in this direction were made by Nowakowski (2020), who developed a digital corpus and core language technologies for Ainu. In another study, Nowakowski et al. (2017) proposed better text-processing tools for the Ainu language. These seminal works laid the groundwork for applying NLP techniques to Ainu, facilitating its digitalization.

Nowakowski's later work (Nowakowski et al., 2023) adapted a multilingual speech representation model for under-resourced languages through multilingual fine-tuning and continued pretraining. This showcased how techniques in NLP could be adjusted for low-resource languages like Ainu.

Efforts have also been made to apply speech recognition technology to the Ainu language. Matsuura et al. (2020b) developed a speech corpus of Ainu folklore and end-to-end speech recognition for the Ainu language. These authors also successfully utilized generative adversarial training data adaptation for very low-resource automatic speech recognition (Matsuura et al., 2020a). These studies significantly contribute to the field and provide a solid foundation for further exploration of NLP applications in low-resource languages. In terms of the linguistic study of Ainu, the work of Senuma and Aizawa (2017) in developing universal dependencies for Ainu and Ptaszynski et al. (2016) in improving part-of-speech tagging of the Ainu language have contributed significantly to the understanding of Ainu syntax and morphology, which is essential in developing accurate NLP tools.

The broader challenges of MT are aptly highlighted in the works of Koehn and Knowles (2017), which underscored the need for advanced techniques to address these challenges effectively. In language education, an innovative application of NLP tools was demonstrated by Nowakowski et al. (2020) through developing an Ainu language-speaking Pepper robot, indicating the potential of such technologies in promoting and preserving endangered languages. The insights and methodologies proposed in these studies pave the way for further exploration into using MT and other NLP technologies for language preservation, particularly for endangered languages such as Ainu.

3 Methodology

In this study, we utilized Marian MT. This efficient and adaptable open-source MT framework has demonstrated excellent performance in numerous research projects, particularly in scenarios involving low-resource languages (Ponti et al., 2021). Our choice for Marian MT was also informed by its inherent capacity to handle different language structures, an essential feature for polysynthetic languages such as Ainu (Ortega et al., 2020).

Our methodology commenced with data preprocessing obtained from multiple digital Ainu text sources. These included the Ainugo Archive from the National Ainu Museum[2], the Glossed Audio Corpus of Ainu Folklore from the National Institute for Japanese Language and Linguistics [3], and the ILCAA Ainu Language Resource from the Tokyo University of Foreign Studies[4]. We converted all the Katakana transcription into the Roman alphabet. The collected corpus was subsequently cleansed to eliminate redundancies and inconsistencies. Following this, we tokenized the data and segmented it into sentence pairs. Given the polysynthetic structure of Ainu (see Example 1), we took extra caution during the tokenization process to correctly separate individual morphemes. The number of sentence pairs of the Ainu original text and the Japanese translation is around 100,000.

We trained the Marian MT model with our prepared corpus, translating Ainu to Japanese and Japanese to Ainu directions. The model parameters were optimized through a learning rate schedule

[2] `https://ainugo.nam.go.jp/` (accessed June 24, 2023)

[3] `https://ainu.ninjal.ac.jp/folklore/` (accessed June 24, 2023)

[4] `https://ainugo.aa-ken.jp/` (accessed June 24, 2023)

combined with early stopping. The learning rate schedule gradually reduced the learning rate during the training process, thereby preventing the overfitting of the model to the training data. The early stopping technique mitigated overfitting by terminating the training when the model's performance on a validation set stopped improving (Almansor and Al-Ani, 2018).

We utilized the SacreBLEU metric to evaluate the performance of our MT system (Kim and Kim, 2022b). SacreBLEU provides a reliable and uniform method for comparing different MT systems or versions, implementing identical tokenization and detokenization procedures across all systems evaluated. It also accounts for multiple reference translations, thereby offering a more comprehensive evaluation of the translation quality (Kim and Kim, 2022a). This feature is especially beneficial for languages like Ainu, where the availability of parallel corpora is limited, and a given sentence could have multiple valid translations.

Our methodology thus encapsulated a combination of the Marian MT framework, rigorous preprocessing of the Ainu corpus, meticulous model training in Ainu to Japanese and Japanese to Ainu directions, and robust evaluation using the SacreBLEU metric. With this method, we developed a robust MT system capable of translating between Ainu and Japanese with significant accuracy.

4 Results

This chapter elucidates the outcomes of our MT experiments and provides an extensive discussion of their implications. The core of our study centered around two translation tasks: from Japanese to Ainu, from Ainu to Japanese, and both directions[5] (Table 1).

The model was trained on an extensive dataset gathered from various Ainu digital text sources for the Japanese-to-Ainu translation task. This resulted in a SacreBLEU score of 32.90, which implies a significant level of translation quality. This achievement showcases the model's ability to translate between these two disparate languages precisely. Notably, these results were obtained despite the inherent challenges posed by developing an MT system for a low-resource language like Ainu.

	Jpn.-Ain.	**Ain.-Jpn.**	**Bi-dir.**
Num. pairs	97,161	95,232	220,023
SacreBLEU	32.90	10.45	29.91

Table 1: Number of sentence pairs in used corpora and SacreBLEU scores of the best MT models in each case

The Ainu-to-Japanese translation task brought additional challenges, mainly due to the limited resources available for the Ainu language. Regardless, the MT system achieved a SacreBLEU score of 10.45. We also trained Marian MT for bidirectional translations, namely Japanese-Ainu and Ainu-Japanese, with a doubled corpus but reversed in the order of two languages in the latter half. The SacreBLEU score of this bi-directional experiment was 29.91, and the input can be both Japanese and Ainu, but the output is in the other language, which was not typed in the input.

Our research's relatively high SacreBLEU scores underline the feasibility of utilizing MT to aid language preservation and revitalization efforts. The results demonstrate that, even with limited resources, MT models can achieve a level of proficiency that renders them practical tools for Ainu learners and researchers (Kim and Kim, 2022b).

Additionally, our study supports the successes of previous attempts to apply computational techniques to the Ainu language. A notable instance is the Ainu speech recognition project by the Kawahara Lab at Kyoto University, whose results were documented in Matsuura et al. (2020b). Together, these studies underscore the potential contributions of NLP and MT technologies to preserve and revitalize endangered languages.

The outcomes of our study should inspire further exploration of MT applications in low-resource language contexts. Future endeavors could focus on refining the model's performance, expanding the dataset, and investigating how this technology can be integrated into interactive language learning platforms. Such efforts would further contribute to the revitalization of the Ainu language and culture.

5 Conclusions

This research project was undertaken to unlock the potential of MT in the preservation and revitalization of Ainu, a critically endangered and low-resource indigenous language of Japan. Grounded

[5] The models made in this study were published on HuggingFace. Ainu-to-Japanese: `https://huggingface.co/SoMiyagawa/ainu-2-japanese`, Japanese-to-Ainu: `https://huggingface.co/SoMiyagawa/japanese2ainu`, and bi-directional: `https://huggingface.co/SoMiyagawa/AinuTrans-2.0` (all accessed on June 24, 2023).

in the neural MT framework, Marian MT, and powered by a comprehensive dataset sourced from various Ainu digital text corpora, our study has made significant strides in demonstrating the feasibility and efficacy of MT in language preservation efforts.

The outcomes of our study are promising. With a SacreBLEU score of 32.90 for the Japanese to Ainu translation task, the quality of translation produced is commendable, particularly considering the challenging polysynthetic nature of the Ainu language (Ortega et al., 2020). Even more impressively, the model achieved a respectable SacreBLEU score of 29.91 for the Japanese-Ainu and Ainu-Japanese bi-directional translation task, underlining the robustness of the neural MT framework when dealing with complex, low-resource languages (Kim and Kim, 2022b).

These findings contribute to the expanding body of research that explores the potential of MT in bridging linguistic gaps and aiding in the preservation of endangered languages. Our results align with studies such as those conducted by Ranathunga et al. (2023), which emphasized the potential of neural MT for low-resource languages, and Kumar et al. (2021), which explored MT in low-resource language varieties.

Our research underscores the necessity for further work to leverage advanced MT techniques for low-resource languages, particularly where traditional linguistic databases may be limited or non-existent. By contributing to the intersection of Natural Language Processing (NLP), MT, and language preservation, our study offers a replicable methodology and highlights the importance of continuous innovation in these areas (Pilch et al., 2022).

Revitalizing endangered languages is complex and multifaceted, necessitating collaborative efforts across linguists, educators, technologists, and communities. Our research reiterates that MT and other language technologies are crucial in this process. While further refinement of models and expansion of datasets can enhance translation quality, our current findings underscore the significance of MT in the broader context of language preservation.

In conclusion, our study suggests that MT can make even the most resource-limited languages, like Ainu, more accessible. By facilitating communication, preserving cultural heritage, and fostering a deeper understanding of diverse human experiences, our research reaffirms the profound value of language preservation and the transformative power of technology in these endeavors. In light of UNESCO's classification of Ainu as "critically endangered," we believe our research can add a crucial layer of defense in the fight against language extinction and contribute to celebrating our shared linguistic heritage (Moseley, 2010).

References

Ebtesam H Almansor and Ahmed Al-Ani. 2018. A hybrid neural machine translation technique for translating low resource languages. In *Machine Learning and Data Mining in Pattern Recognition: 14th International Conference, MLDM 2018, New York, NY, USA, July 15-19, 2018, Proceedings, Part II 14*, pages 347–356. Springer.

Ahrii Kim and Jinhyeon Kim. 2022a. Vacillating human correlation of sacrebleu in unprotected languages. In *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 1–15.

Ahrii Kim and Jinhyun Kim. 2022b. Guidance to Pre-tokeniztion for SacreBLEU: Meta-Evaluation in Korean.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*.

Sachin Kumar, Antonios Anastasopoulos, Shuly Wintner, and Yulia Tsvetkov. 2021. Machine translation into low-resource language varieties. *arXiv preprint arXiv:2106.06797*.

Kohei Matsuura, Masato Mimura, Shinsuke Sakai, and Tatsuya Kawahara. 2020a. Generative adversarial training data adaptation for very low-resource automatic speech recognition. *arXiv preprint arXiv:2005.09256*.

Kohei Matsuura, Sei Ueno, Masato Mimura, Shinsuke Sakai, and Tatsuya Kawahara. 2020b. Speech corpus of Ainu folklore and end-to-end speech recognition for Ainu language. *arXiv preprint arXiv:2002.06675*.

Christopher Moseley. 2010. *Atlas of the World's Languages in Danger*. Unesco.

Karol Nowakowski, Michal Ptaszynski, and Fumito Masui. 2017. Towards better text processing tools for the Ainu language. In *Language and Technology Conference*, pages 131–145. Springer.

Karol Nowakowski, Michal Ptaszynski, and Fumito Masui. 2019. Mingmatch—a fast n-gram model for word segmentation of the Ainu language. *Information*, 10(10):317.

Karol Nowakowski, Michal Ptaszynski, and Fumito Masui. 2020. Spicing up the game for underresourced language learning: Preliminary experiments with Ainu language-speaking Pepper robot. In *The 6st workshop on linguistic and cognitive approaches to dialog agents*.

Karol Nowakowski, Michal Ptaszynski, Kyoko Murasaki, and Jagna Nieuważny. 2023. Adapting multilingual speech representation model for a new, underresourced language through multilingual fine-tuning and continued pretraining. *Information Processing & Management*, 60(2):103148.

Karol Piotr Nowakowski. 2020. Development of a digital corpus and core language technologies for the Ainu language.

John E Ortega, Richard Castro Mamani, and Kyunghyun Cho. 2020. Neural machine translation with a polysynthetic low resource language. *Machine Translation*, 34(4):325–346.

Agnieszka Pilch, Ryszard Zygała, and Wiesława Gryncewicz. 2022. Quality assessment of translators using deep neural networks for polish-english and english-polish translation. In *2022 12th International Conference on Advanced Computer Information Technologies (ACIT)*, pages 227–230. IEEE.

Edoardo Maria Ponti, Julia Kreutzer, Ivan Vulić, and Siva Reddy. 2021. Modelling latent translations for cross-lingual transfer. *arXiv preprint arXiv:2107.11353*.

Michal Ptaszynski, Karol Nowakowski, Yoshio Momouchi, and Fumito Masui. 2016. Comparing multiple dictionaries to improve part-of-speech tagging of Ainu language. In *Proceedings of the 22nd Annual Meeting of The Association for Natural Language Processing, Sendai, Japan*, pages 7–11.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*, 55(11):1–37.

Hajime Senuma and Akiko Aizawa. 2017. Toward universal dependencies for Ainu. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 133–139.

Masayoshi Shibatani. 1990. *The languages of Japan*. Cambridge University Press.