Bootstrapping Moksha-Erzya Neural Machine Translation from Rule-Based Apertium

Khalid Alnajjar
Rootroo Ltd
khalid@rootroo.com

Mika Hämäläinen
Metropolia University of Applied Sciences
mika.hamalainen@metropolia.fi

Jack Rueter
University of Helsinki
jack.rueter@helsinki.fi

Abstract

Neural Machine Translation (NMT) has made significant strides in breaking down language barriers around the globe. For lesser-resourced languages like Moksha and Erzya, however, the development of robust NMT systems remains a challenge due to the scarcity of parallel corpora. This paper presents a novel approach to address this challenge by leveraging the existing rule-based machine translation system Apertium as a tool for synthetic data generation. We fine-tune NLLB-200 for Moksha-Erzya translation and obtain a BLEU of 0.73 on the Apertium generated data. On real world data, we got an improvement of 0.058 BLEU score over Apertium.

1 Introduction

A significant number of the world's languages are currently at risk of becoming endangered to varying degrees (Moseley, 2010). This endangered status presents particular challenges when it comes to conducting modern NLP research with these languages. The primary issue stems from the fact that many endangered languages lack extensive textual resources that are readily accessible online. Moreover, even when some resources are available, there are concerns regarding the quality of the data, which can be influenced by factors like the author's fluency level, spelling accuracy, and basic character encoding inconsistencies, as discussed in Hämäläinen 2021.

For over two centuries, scholars have been investigating the cohesion and variety within the contemporary Mordvin literary languages, namely Erzya (myv) and Moksha (mdf). The first comprehensive grammatical works on these languages were published in the 1830s, with Moksha in 1838 (Ornatov, 1838) and Erzya in 1838-1839 (Gabelentz, 1839). In the subsequent 180 years, researchers have engaged in extensive fieldwork, compiled grammars, created dictionaries, and worked towards popularizing these languages. Notably, in 2002, the inaugural monolingual Erzya dictionary was published, authored by Abramov (Abramov, 2002), with plans for future expansion. Recent years have also witnessed continued academic interest in the Mordvin languages (Luutonen, 2014; Hamari and Aasmäe, 2015; Kashkin and Nikiforova, 2015; Grünthal, 2016), highlighting their enduring significance in linguistic research.

It is crucial to provide newcomers to language documentation with chances to enhance their comprehension of languages by involving them in projects. A noteworthy period to highlight in this regard is the years spanning from 1988 to 1997 when many of today's researchers were engaged in word processing for the extensive 'Dictionary of Mordvin Dialects', compiled on the basis of language materials whose collection was originated and orchestrated by Prof. Heikki Paasonen at the turn of the twentieth century, and which, when completed, comprised a substantial 2073 pages.

The research conducted in this paper is based on data generated using the rule-based machine translation system Apertium. While rule-based tradition has influenced the current NLP for endangered Uralic languages (cf. Pirinen et al. 2023), our aim is to study the degree to which more modern neural models can be incorporated into the existing paradigm. The largest obstacle in using machine learning models is the scarcity of data available in these languages. Using Apertium to generate training data is our attempt at overcoming this problem.

2 Related work

Many contemporary machine translation models heavily rely on the presence of parallel texts. However, finding parallel texts is a challenging endeavor, particularly when dealing with less-

Proceedings of the Joint 3rd NLP4DH and 8th IWCLUL, pages 213–218

Erzya input	Moksha output
Лей чиресэ пандыне, */* Пандонть прясо кудыне...	Ляй ширеса пандоня, */* Пандть пряса куданя...
Леесь чуди чипельде пелеве ёнов, лемезэ Ока.	Ляйсь шуди чипельде веньгучка шири, лемезэ Ока.
Сынь кадноз ваномс ды налсо леднемс.	Сань кадондозь ваномс налса лядендемс.

Table 1: Example of Apertium generated training data

resourced languages across various domains. This challenge becomes even more pronounced when attempting to train data-intensive models for these languages.

A method proposed by Munteanu and Marcu (2005) addresses this issue by utilizing a large, non-parallel but comparable corpus, such as news articles, in conjunction with relatively small parallel corpora from a different domain, like the United Nations corpus. By matching sentences from comparable articles that share the same topic, this method attempts to determine if two sentences are translations of each other. Although this approach enhances translations within the news domain, it's not always viable, especially for extremely low-resource languages like Erzya and Moksha, which lack the necessary comparable corpora.

To develop machine translation models for languages with limited resources, another approach involves leveraging a resource-rich language closely related to the low-resource one as a parent language. This entails acquiring some of the resource-rich language's characteristics, such as syntax and morphology, and transferring them to the low-resource machine translation model (Zoph et al., 2016; Nguyen and Chiang, 2017; Passban et al., 2017; Karakanta et al., 2018). These techniques don't necessarily require parallel texts in the low-resource language but rely on the resource-rich language, which may result in limited coverage of the low-resource language's morphology.

Researchers have also explored methods for constructing parallel texts through crowdsourcing, wherein online workers are tasked with translating expressions into another language (Ambati and Vogel, 2010; Ambati, 2012; Zaidan and Callison-Burch, 2011). Crowdsourcing, however, proves to be a challenging endeavor when dealing with low-resource languages due to the limited number of native speakers. Additionally, the absence of a standardized language form or even linguistic variation further complicates the quality control of crowdsourced translation tasks, even if a significant number of workers are involved.

A different approach proposed by Chahuneau et al. (2013) involves translating English into morphologically complex languages by creating a model that predicts word inflections in the target language. This model is then used to generate synthetic phrases, potentially with new inflections, which are incorporated into the training data alongside a parallel corpus to train a machine translation model.

Hämäläinen and Alnajjar (2019) introduced a method for creating parallel data for low-resource endangered languages with complex morphology. They demonstrated their approach using Finnish as a pilot language, matching the resource limitations to those of Erzya. Additionally, the authors described a technique for automatically aligning the abstract morphosyntactic structures of two languages to generate a set of parallel templates. However, the system could only translate phrases as opposed to complete sentences.

3 Apertium in data generation

Apertium (Forcada et al., 2011) is rule-based machine translation system that is available for several language pairs. There is a special version of the system for endangered languages hosted by GiellaLT (Trosterud, 2017) that has support for Erzya and Moksha. The Erzya-Moksha translation has been developed through a shallow transfer approach (Rueter and Hämäläinen, 2020) and it utilizes FST transducers developed for these languages (Rueter et al., 2020).

We use a monolingual Erzya corpus of around 220 000 sentences (the Erzya-language novel Purgaz (Abramov, 1988)). We feed this corpus to Apertium translator and translate the sentences into Moksha. This way, we will have a parallel corpus of Erzya-Moksha sentences where the target side, namely Erzya is of a high quality and the source side, namely Moksha is synthetically generated using the rule-based system. Apertium is not able to inflect all words, so it tags such words with different tags such as # and @. We remove these extra characters from the data.

Table 1 shows an example of the data we used for training our model. The key notion is that we train

Moksha input	Apertium	Our Model	Gold standard
Минь карматама природать тонафнемонза.	Минь карматано *природать тонавтнемензэ.	Минек карматано природанть тонавтнемензэ.	Минь карматано природанть тонавтнеме.
Природать колга наукати мярьгихть естествознания.	*Природать содалмонтень мерить *естествознания.	Природантьдонть наукантень мерить естествознания.	Природадо наукантень мерить естествознания.
Естествознаниять тейнек пяк оцю значенияц.	*Естествознаниять тенек пек покш *значенияц.	Естествознаниять тенек пек покш значенияц.	Естествознаниянть значениязо миненек пек покш.
Сон лезды лац шарьхкодемс природать.	Сон лезды парсте чарькодемс *природать.	Сон лезды лазтнэнь чарькодиця природанть.	Сон лезды тенек природанть видестэ-парсте чарькодеме.

Table 2: Results on the real world data

our NMT model in an inverse direction. We treat the Moksha output from Apertium as input and the Erzya sentences as output when training the model. This ensures a good and grammatical target representation. This is similar to the back translation methodology described by Sennrich et al. (2016).

The data is split randomly into 80 % training, 10 % validation and 10 % testing. The model is trained only on this Apertium generated dataset.

4 Neural machine translation

We use the NLLB-200 model by Meta (Costa-jussà et al., 2022) to conduct our experiments. The model has been trained to support translation for over 200 languages. Erzya and Moksha are not supported by the model by default. In fact, the only Uralic languages that are supported are Finnish, Estonian and Hungarian - none of which are endangered.

NLLB-200 is a 54.5B parameter Mixture of Experts (MoE) model that has been trained on a dataset containing more than 18 billion sentence pairs. On benchmark evaluations, NLLB-200 outperforms other state-of-the-art models by up to 44%. The model's performance has been validated through extensive evaluations for each of the 200 languages it supports.

We use Transformers Python library (Wolf et al., 2020) to fine-tune the 600M distilled NLLB-200 model[1] for Moksha to Erzya translation. We add two new languages to the tokenizer: mdf_Cyrl and myv_Cyrl for Moksha and Erzya.

We trained the model for 5 epochs. We used weight decay of 0.1 and an initial learning rate of 2e-5. The model was validated after each epoch using BLEU as the validation metric. The final validation BLEU score was 0.85.

5 Results

We have withheld 10% of the data for testing proposes. **Our model achieves a BLEU score of 0.73**. When looking at the results, we can perceive errors coming from the fact that we used Apertium translator such as missing vocabulary resulting in wrong word choice in the translation output as well as minimal transfer rules to address diversity in verbal government and idiomatic expressions. This calls for further comparison of the model with Apertium translations.

We take a relatively small parallel corpus of a natural science book that has been translated into Erzya and Moksha from Russian[2]. The corpus consists of a little over 2000 sentences of human-authored translations. We test both our model and Apertium on this dataset. Neither of the models reaches very good performance when measured by BLEU scores. The Apertium translator gets a BLEU score of 0.037 whereas our model gets a BLEU of 0.095. It is important to note that our model got an improvement of 0.058 BLEU score. The results can be seen in Table 2.

The biggest problem Apertium generated data has is a lack of vocabulary coverage. Our model gets the grammar and morphology correct more frequently than the rule-based Apertium translator because of the good target representation from the high-quality Erzya sentences used in training, although there are instances of multiexponence in morphology as in Природанть+донть 'the nature+about the' which might be explained by incomplete transfer rules in the Apertium translator. However, both Apertium and our model struggle with vocabulary and many of the words are either not translated at all or are translated to wrong words.

However, our result suggest that using Apertium to generate data for an NMT model is a viable way

[1] https://huggingface.co/facebook/nllb-200-distilled-600M

[2] The test example is taken from Russian to Erzya and Moksha translations, see https://urn.fi/urn:nbn:fi:lb-2023042421

of combining the rule-based tradition with latest neural models. In particular, the fact that the NMT model was able to produce better results makes this a worthwhile approach for any endangered language machine translation project. This is even more so in cases where the rule-based translation system has reached a higher level of maturity.

6 Conclusions

In conclusion, our study has revealed significant insights into the performance of our neural machine translation (NMT) model when compared to the rule-based Apertium translator in the context of translation between two endangered languages, Moksha and Erzya.

The primary challenge observed in the Apertium-generated translations was the limited vocabulary coverage. Our NMT model, on the other hand, demonstrated better accuracy in terms of grammar and morphology due to the high-quality Erzya sentences used during training. Nevertheless, both Apertium and our model struggled with vocabulary, resulting in untranslated or incorrectly translated words.

In light of our findings, our results suggest that employing Apertium to generate data for an NMT model represents a viable approach that combines the rule-based translation tradition with state-of-the-art neural models. Notably, our NMT model outperformed Apertium, making this approach valuable for endangered language machine translation projects, particularly when the rule-based translation system has reached a higher level of maturity. This study highlights the potential for leveraging advanced technology to revitalize and preserve endangered languages through more accurate and efficient translation methods.

The findings of this study open up several promising avenues for future research and development in the field of machine translation, particularly for endangered languages. Addressing the challenge of limited vocabulary coverage observed in both the rule-based Apertium translator and the NMT model is critical. Future research could focus on methods to expand and enrich the vocabulary used in training data, possibly through the inclusion of additional linguistic resources or domain-specific terminology.

Augmenting the parallel corpus with more diverse and representative data, including various text genres and dialects, can contribute to better training NMT models. Collecting and curating additional language resources could be a valuable step. Furthermore, we could engage with the community of speakers and experts in Erzya and Moksha languages to gather feedback, improve resources, and establish collaborative efforts in language preservation and machine translation.

These future work directions reflect the ongoing efforts needed to advance the field of machine translation, particularly in the context of preserving and revitalizing endangered languages, and hold the potential to significantly improve the quality and accessibility of translation services for these linguistic communities.

7 Limitations

Our approach requires that there is an existing method of producing translated text in the source language. Furthermore, our approach requires a considerably sized monolingual corpus in the target language. The limitations of the overall system come from limitations in the existing translation system and the monolingual corpus.

The model does not require large computational resources. We trained the model on a desktop PC on an Nvidia RTX3090 GPU. The training was completed in less than a day.

Acknowledgments

This work has been supported by a grant from Oskar Öflunds Stiftelse.

References

Kuzma Abramov. 2002. Валонь ёвтнема валкс. Mordovskoj knizhnoj izdateljstvasj. The manuscript of this dictionary was compiled by the Erzya national writer Kuz'ma Grigorievich Abramov, 1914-2008, whose activities as an Erzya writer spanned nearly 70 years.

Kuzma Abramov. 1988. *Purgaz*. Mordovskoj knižnoj izdatelstvas, Saransk. Online version: https://urn.fi/urn:nbn:fi:lb-2023021601.

Vamshi Ambati. 2012. *Active Learning and Crowdsourcing for Machine Translation in Low Resource Scenarios*. Ph.D. thesis, Pittsburgh, PA, USA. AAI3528171.

Vamshi Ambati and Stephan Vogel. 2010. Can crowds build parallel corpora for machine translation systems? In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10,

pages 62–65, Stroudsburg, PA, USA. Association for Computational Linguistics.

Victor Chahuneau, Eva Schlinger, Noah A. Smith, and Chris Dyer. 2013. Translating into morphologically rich languages with synthetic phrases. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1677–1687. Association for Computational Linguistics.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Mikel L Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, 25(2):127–144.

Herr Conon von der Gabelentz. 1839. Versuch einer mordwinischen grammatik. In *Zeitschrift für die Kunde des Morgenlandes.*, II. 2–3., pages 235–284, 383–419. Druck und Verlag der Dieterlichschen Buchhandlung., Göttingen.

Riho Grünthal. 2016. *Transitivity in Erzya: Second language speakers in a grammatical focus*, Uralica Helsingiensia, page 291–318. Finno-Ugrian Society, Finland.

Mika Hämäläinen. 2021. Endangered languages are not low-resourced! *Multilingual Facilitation*.

Mika Hämäläinen and Khalid Alnajjar. 2019. A template based approach for training nmt for low-resource uralic languages-a pilot with finnish. In *Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence*, pages 520–525.

Arja Hamari and Niina Aasmäe. 2015. Negation in erzya. *Negation in Uralic languages*, 108:293.

Alina Karakanta, Jon Dehdari, and Josef van Genabith. 2018. Neural machine translation for low-resource languages without parallel corpora. *Machine Translation*, 32(1):167–189.

Egor Kashkin and Sofya Nikiforova. 2015. Verbs of sound in the moksha language: a typological account. *Nyelvtudományi Közlemények*, 111:341–362.

Jorma Luutonen. 2014. Kahden sukupolven ersää – kielenhuoltoa ja muutoksen merkkejä. *Memoires de la Societe Finno-Ougrienne*, 270:187—201.

Christopher Moseley. 2010. *Atlas of the World's Languages in Danger*, 3rd edition. UNESCO Publishing. Online version: http://www.unesco.org/languages-atlas/.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Comput. Linguist.*, 31(4):477–504.

Toan Q. Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. *CoRR*, abs/1708.09803.

Pavel Ornatov. 1838. *Mordovskaja grammatika / sostavlennaja na narechij mordvy mokshi Pavlom Ornatovym.* V Sinodalnoj tip., Moskva.

Peyman Passban, Qun Liu, and Andy Way. 2017. Translating low-resource languages by vocabulary adaptation from close counterparts. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 16(4):29:1–29:14.

Flammie Pirinen, Sjur Moshagen, and Katri Hiovain-Asikainen. 2023. GiellaLT — a stable infrastructure for Nordic minority languages and beyond. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 643–649, Tórshavn, Faroe Islands. University of Tartu Library.

Jack Rueter and Mika Hämäläinen. 2020. Prerequisites for shallow-transfer machine translation of mordvin languages: Language documentation with a purpose. In Материалы Международного образовательного салона, pages 18–29. Ижевск: Институт компьютерных исследований.

Jack Rueter, Mika Hämäläinen, and Niko Partanen. 2020. Open-source morphology for endangered mordvinic languages. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 94–100.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96.

T Trosterud. 2017. Language technology in russia. In ЭЛЕКТРОННАЯ ПИСЬМЕННОСТЬ НАРОДОВ РОССИЙСКОЙ ФЕДЕРАЦИИ: ОПЫТ, ПРОБЛЕМЫ И ПЕРСПЕКТИВЫ, pages 294–298.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Omar F. Zaidan and Chris Callison-Burch. 2011. Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies -*

Volume 1, HLT '11, pages 1220–1229, Stroudsburg, PA, USA. Association for Computational Linguistics.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. *CoRR*, abs/1604.02201.