SMUGRI-MT - Machine Translation System for Low-Resource Finno-Ugric Languages

Taido Purason* **Aleksei Ivanov*** **Lisa Yankovskaya** **Mark Fishel**
Institute of Computer Science
University of Tartu, Estonia
{taido.purason,aleksei.ivanov,lisa.yankovskaya,mark.fisel}@ut.ee

Abstract

SMUGRI is a research project supported by an Estonian Research Council grant, aiming to develop natural language processing tools for Finno-Ugric languages and varieties. In this paper, we describe SMUGRI-MT, the part of the project that focuses on developing neural machine translation for this language family. Currently 20 low-resource Finno-Ugric languages are covered, along with seven high-resource languages.

1 Introduction

This project focuses on neural machine translation (MT) for the Finno-Ugric languages. Besides three mid-resource languages (Estonian, Finnish and Hungarian), this family includes dozens more that range from low-resource (e.g. Komi, Veps) to extremely endangered and under-supported languages (e.g. Livonian, Votic). Our goal is to include as many of these languages and varieties as possible and provide them with reliable MT models and methodology.

Our work on developing MT systems for low-resource Finno-Ugric languages started in 2021, initially focusing on Võro as well as Southern and Northern Sami (Tars et al., 2021). One year later, we added Inari, Skolt and Lule Sami languages (Tars et al., 2022) and developed an MT system for Livonian (Rikters et al., 2022). Last year, we significantly expanded the scope of our MT system to include a total of 20 low-resource languages and dialects (Yankovskaya et al., 2023). Since the last version, we have collected more data, transitioned to a different multilingual pretrained model (NLLB Team et al., 2022, 1.2B parameters, distilled) and implemented language identification and hallucination detection tests to ensure a cleaner dataset.

Below we describe the current state of the developed MT system and outline the future challenges.

2 The current stage

The version currently available online[1] is tailored for 20 low-resource languages: Mansi, Khanty, Komi, Komi-Permyak, Udmurt, Meadow and Hill Mari, North Sami, South Sami, Inari Sami, Lule Sami, Skolt Sami, Erzya, Moksha, Ludian, Proper Karelian, Võro, Veps, Livvi Karelian, Livonian along with seven high-resource languages: English, Estonian, Finnish, Hungarian, Latvian, Norwegian (Bokmål), and Russian.

Benchmark test sets are currently available for only nine low-resource languages: Komi, Udmurt, Hill and Meadow Mari, Erzya, Moksha, Livonian, Mansi, and Livvi Karelian (Yankovskaya et al., 2023). Table 1 presents the average chrF++ (Popović, 2017) and BLEU (Papineni et al., 2002) scores for these languages along with scores for seven high-resource languages presented in our MT system. In comparison with our previous MT system, the current system shows better performance, with improvements of 5-7 points in terms of chrF++ score. Translations from low-resource to high-resource languages achieved the highest performance. In contrast, translations into low-resource languages, whether from other low-resource languages or high-resource languages,

*Equal contribution

[1] https://translate.ut.ee/

demonstrated notably lower performance.

	chrF++	BLEU
low-low	32.4 (+5.2)	6.1
low-high	43.8 (+7.2)	16.8
high-low	33.9 (+7.1)	6.7

Table 1: Average chrF++ (Popović, 2017) and BLEU (Papineni et al., 2002) scores across different language pair clusters; numbers in brackets indicate the improvement of the current system over the preivous one. Low-low - translations from low-resource to low-resource, low-high - from low-resource to high-resource, high-low - from high-resource to low-resource languages.

3 Future challenges

Although this work started in 2021, the project SMUGRI-MT has received funding only recently, thus there is still a lot of work to be done more systematically. Our efforts will focus on three directions of future research:

Data: We will continue to collect parallel and monolingual data for currently supported languages, as well as expand our dataset to include new languages and varieties. In addition to data collection, we will prioritize preprocessing steps to produce cleaner corpora.

Several Finno-Ugric languages (already included and these to be added) do not have a normalized orthography and additionally represent a mixture of dialects and varieties. Therefore an important future direction is separating the varieties from each other and deciding what to do about orthographic variation.

Analysis: We plan to conduct a detailed qualitative analysis of the translations to identify errors overlooked by automatic metrics. Additionally, we aim to develop test sets across various domains and language varieties, enabling a more comprehensive qualitative and quantitative analysis.

Architecture: The current system translates one sentence at a time. We have plans to transition to paragraph-level and document-level systems to reduce at least gender-related issues. This is mainly due to the Finno-Ugric languages being gender-neutral, making it impossible to determine the correct gender without context and making it important to translate genders consistently throughout the document when generating a language with grammatical gender (like English).

Acknowledgments

This work was partially supported by the Estonian Research Council grant PRG2006 as well as the National Programme of Estonian Language Technology grant EKTB67. All computations were performed on the LUMI Supercomputer through the University of Tartu's HPC center.

References

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Popović, Maja. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation (WMT'17)*, pages 612–618.

Rikters, Matīss, Marili Tomingas, Tuuli Tuisk, Valts Ernštreits, and Mark Fishel. 2022. Machine translation for Livonian: Catering to 20 speakers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 508–514.

Tars, Maali, Andre Tättar, and Mark Fišel. 2021. Extremely low-resource machine translation for closely related languages. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 41–52.

Tars, Maali, Andre Tättar, and Mark Fišel. 2022. Cross-lingual transfer from large multilingual translation models to unseen under-resourced languages. *Baltic Journal of Modern Computing*, 10.3:435–446.

Yankovskaya, Lisa, Maali Tars, Andre Tättar, and Mark Fishel. 2023. Machine translation for low-resource Finno-Ugric languages. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 762–771.