TARTUNLP

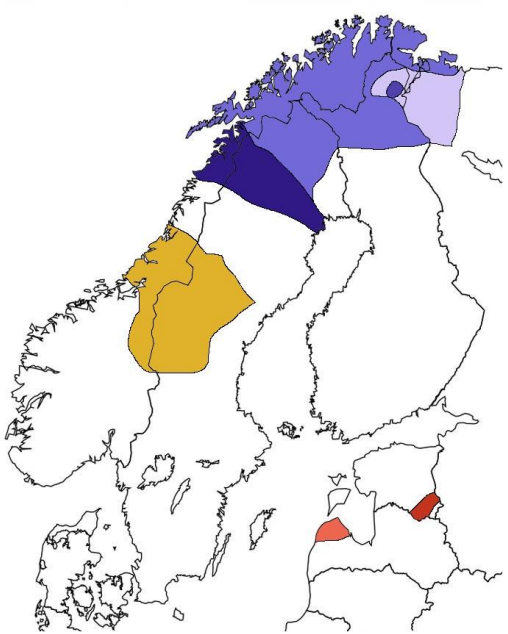
Cross-lingual Transfer to Unseen Languages

Maali Tars, Andre Tättar, Mark Fišel

Background

Low-resource Finno-Ugric neural machine translation

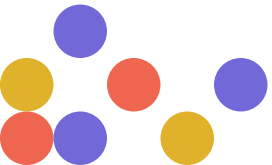
Further developments of previous efforts



Pre-trained models

M2M-100, MT5, NLLB

M2M-100 as an example for adapting pre-trained model to cross-lingual transfer

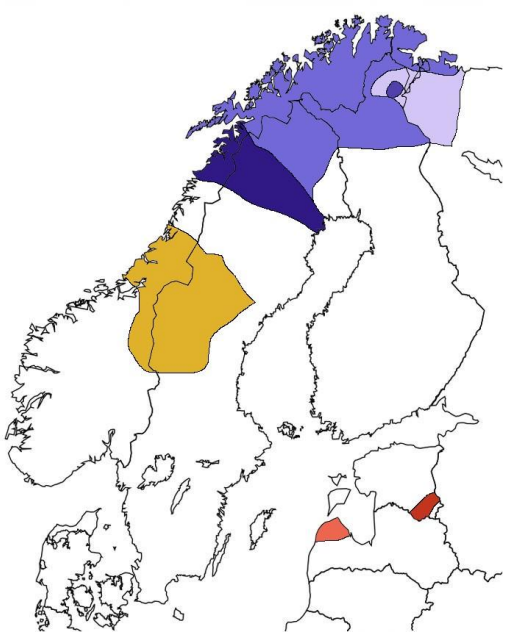


Data

Livonian, Võro, South Sami, North Sami, Lule Sami, Skolt Sami, Inari Sami

Most parallel data: ~ 200 000
Least parallel data: ~ 300

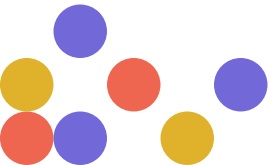
No monolingual data used



Data

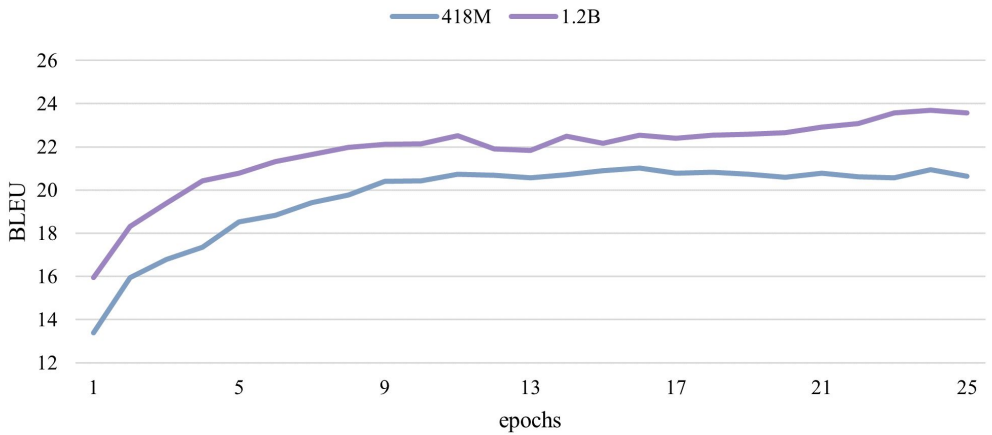
Created a new benchmark (except for Livonian)

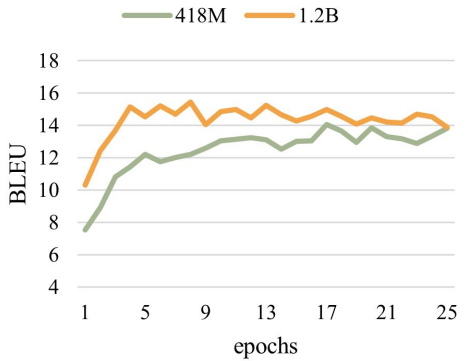
Detokenization, normalization, filtering

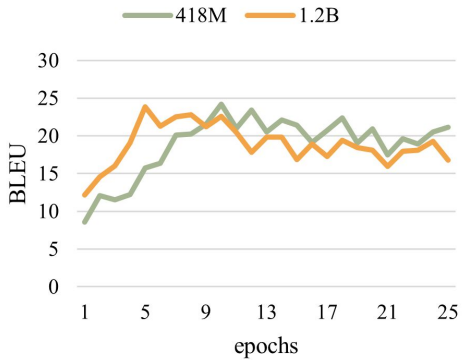


Experiments

M2M-100 418M vs 1.2B





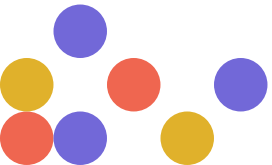


Experiments

First NMT results for **Skolt Sami, Lule Sami, Inari Sami**: in range of **33-75 BLEU**

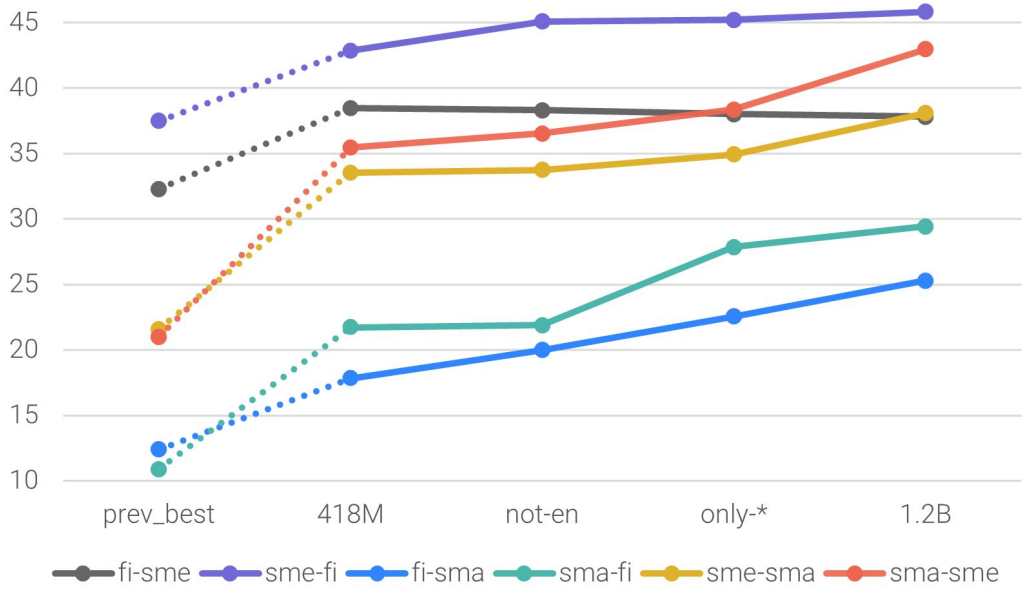
1.2B parameter model performs best

Comparing different 418M models: **dividing data into smaller groups helps to improve** (+4 BLEU for Võro)



Experiments

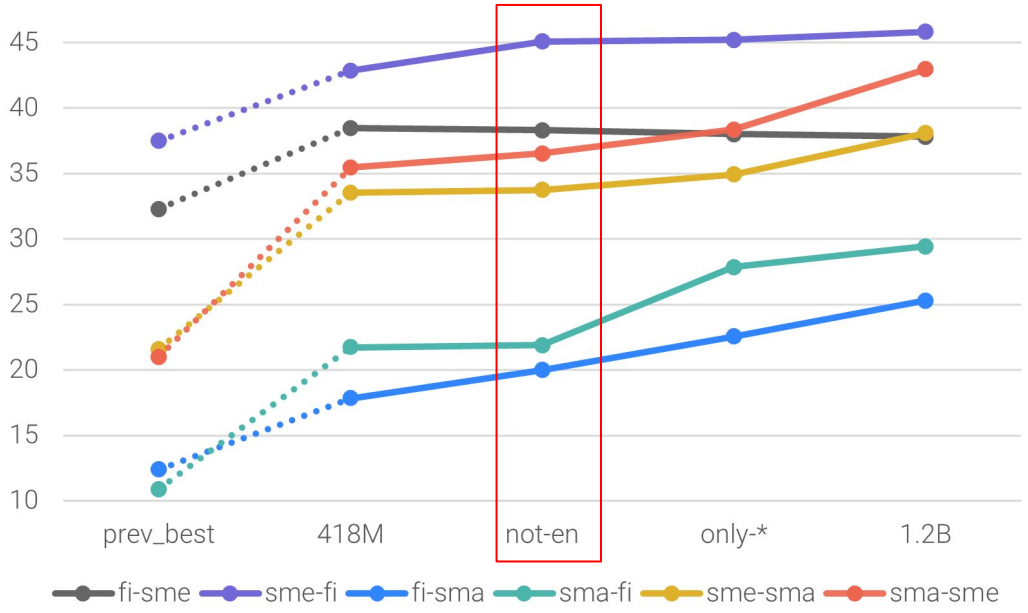
Results compared to previous work



Experiments

Results compared to previous work

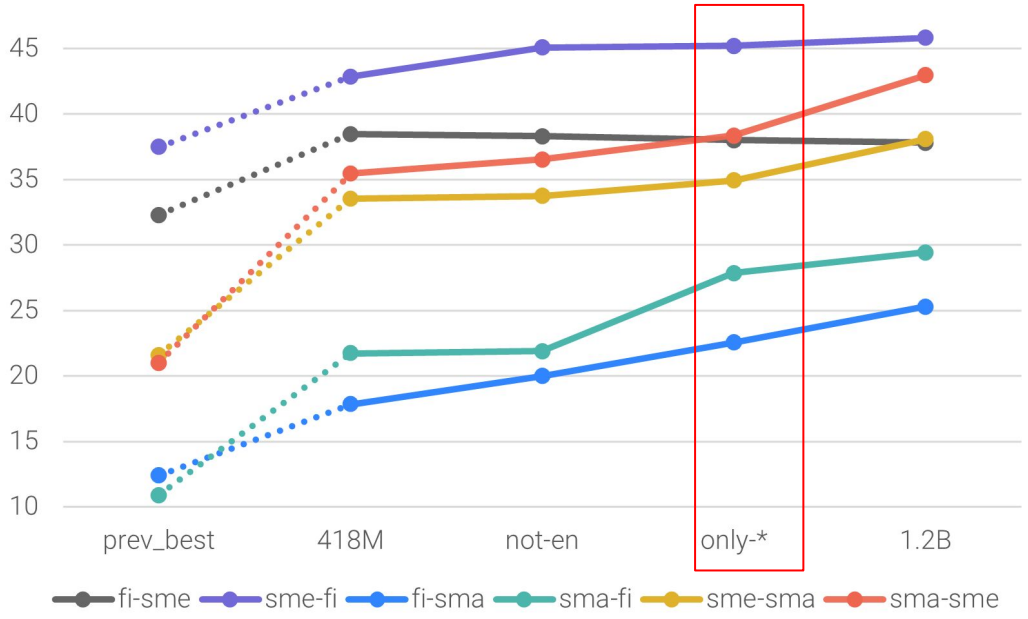
Removing English



Experiments

Results compared to previous work

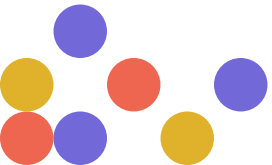
Smaller language groups



Conclusions

Large pre-trained multilingual NMT models are very helpful

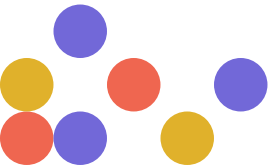
Comparable results achieved with **little parallel data** and **no monolingual data** (e.g. for Livonian)



Conclusions

Smaller language groups within Finno-Ugric family **improve** on **quality**

Created a **new benchmark** and **first NMT results** for Skolt Sami, Lule Sami, Inari Sami

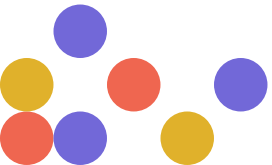


Future

Human evaluation, additional automatic evaluation

Use of **monolingual data**

Currently collecting data for **other Finno-Ugric languages**



TARTUNLP

Thank you!

www.tartunlp.ai