Open and Competitive Multilingual Neural Machine Translation in Production

Andre Tättar[1], Taido Purason[1], Hele-Andra Kuulmets[1], Agnes Luhtaru[1], Liisa Rätsep[1], Maali Tars[1], Mārcis Pinnis[2], Toms Bergmanis[2], Mark Fishel[1]

[1]University of Tartu

[2]Tilde

Introduction

MTee Project

Estonian governmental project (April 2021 to January 2022) carried out by **University of Tartu** and **Tilde**.

Organised by Estonian Ministry of Education and Research as a public procurement via the Language Technology Competence Center (Institute of the Estonian Language)

Enable **faster distribution of information** in times of crisis with open and competitive multilingual neural machine translation.

Introduction

Translation directions:

ESTONIAN ⟷ ENGLISH / GERMAN / RUSSIAN

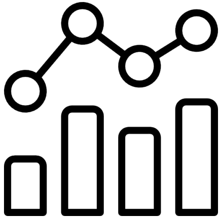
Domains:

General, Legal, Crisis, Military, Spoken

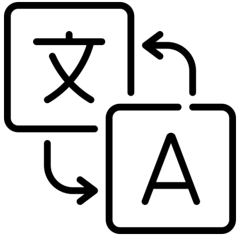
Introduction

Outcomes

Parallel and monolingual corpora



Public benchmarks



Open-source NMT systems

Data Sources

1) **Open Sources:**
   - OPUS
   - ELRC-SHARE
   - EU Open Data Portal
   - Meta-Share
   - CLARIN
   - ELRA

2) **Web Scraping**
   - E.g. state news and other governmental sites

3) **Data Donors and Industry Partners**

Pre-processing

Filtering using OpusFilter

Parallel data:

- Duplicates
- Sentence length ratio
- Maximum sentence length
- Maximum word length
- Maximum word count
- Foreign word
- Digit mismatch
- Statistical word alignment
- Test data overlap

Monolingual data:

- Maximum sentence length
- Maximum word length
- Parallel filters after back-translation

Pre-processing

Normalization

Normalize punctuation and whitespace.

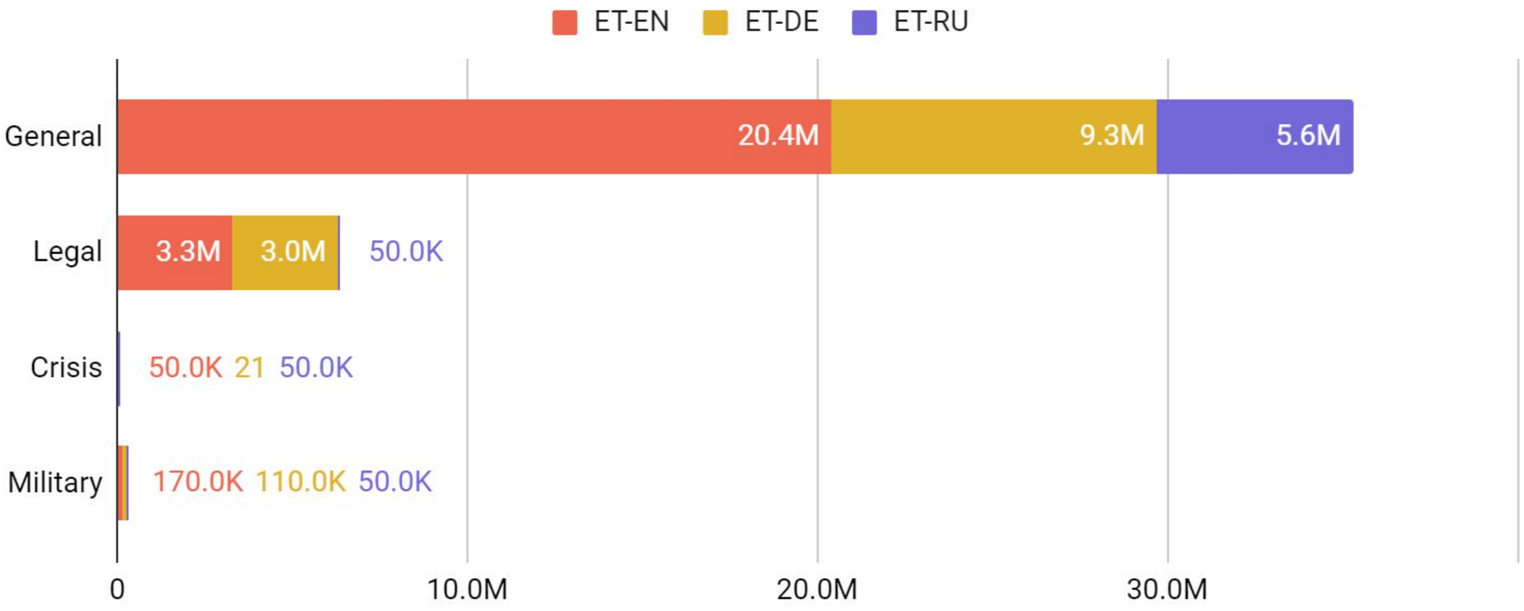
Customized **Moses Statistical MT** normalization script.

«Text» → "Text"

(Text) . → (Text).

Text’s. → Text's.

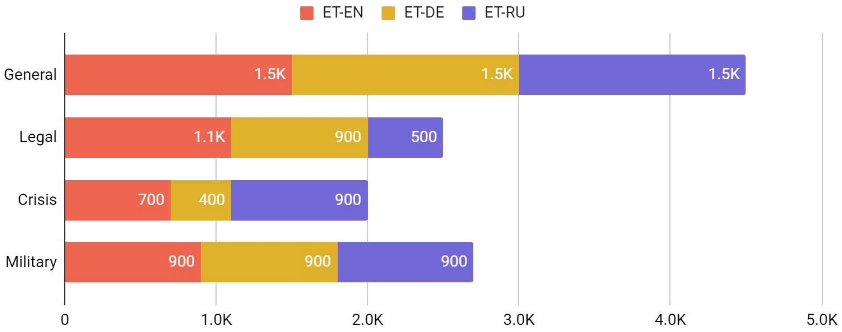
Training Data



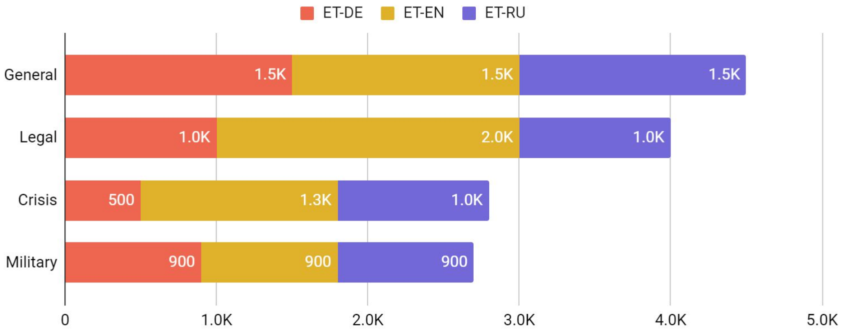
Test Data

Manually filtered/corrected the data with annotators

Validation dataset



Test dataset



Monolingual data

	General	Military	Legal	Crisis
ET	50M	0.9M	0.5M	0.6M
EN	48.9M	1.5M	0.3M	10M
DE	49.3M	130K	0.6M	3.4M
RU	49.6M	8K	5.4M	142K

Segmentation model

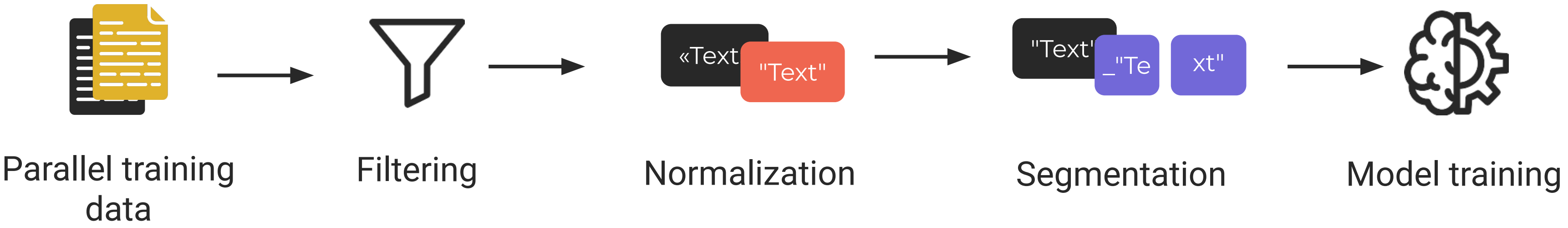
SentencePiece BPE

Separate **model for each language**.

- Trained on 10,000,000 sentences sampled from the dataset
- Vocabulary size of 24,000
- Character coverage of 0.9999
- Finally, add top-500 characters (across whole dataset) to each model

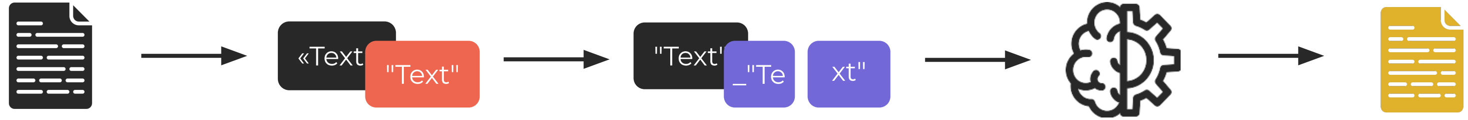
Data Processing Overview

Training



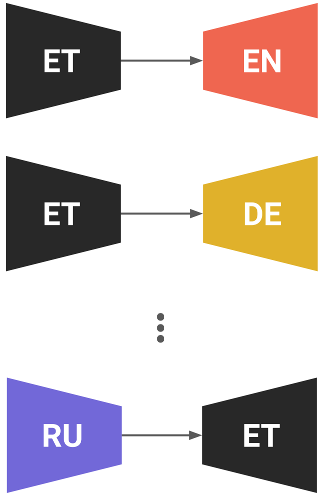
Data Processing Overview

Translation

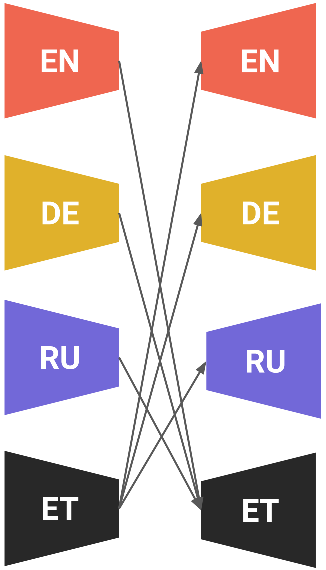


Source language sentences → Normalization → Segmentation → Translation model → Target language translations

Model Architectures



Unidirectional models



Language-specific encoders/decoders (our approach)

Model Training

Jointly trained language-specific encoders-decoders
(modular)

Custom **Fairseq** implementation (open-sourced)
Transformer base encoders-decoders (6-6)

Steps:
1) Train general model (whole dataset inc. domain)
2) Fine-tune domain models
3) Back-translate and repeat

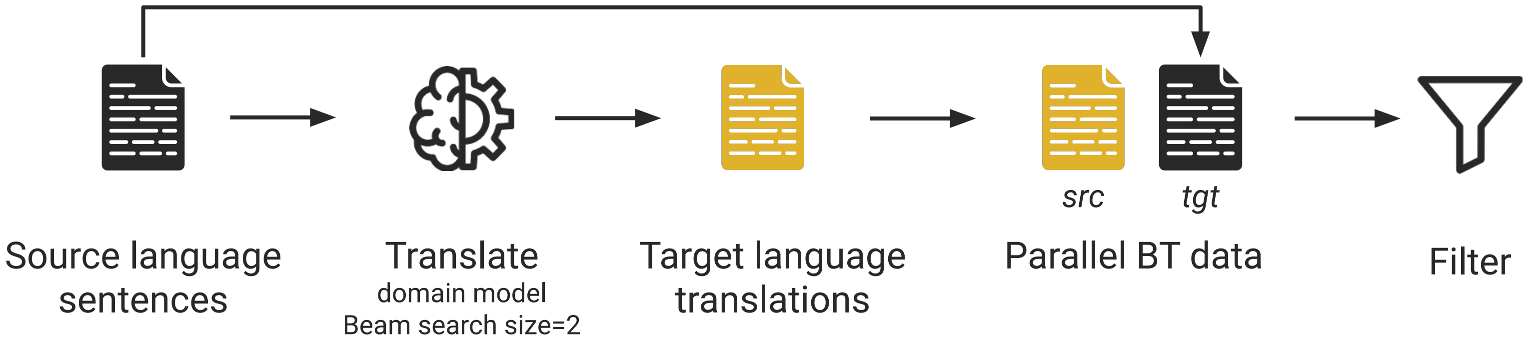
Data augmentation

Back-translations

Estonian Proper Nouns

Spoken language

Back-translation



Resulting in ~54M new parallel sentences per direction (325M in total)

Estonian Proper Nouns

Data for some languages contains no diacritics common in Estonian (**õ, ä, ö, ü, š, ž**). Thus the model does not know how to translate them when they occur.

Augment the dataset using Tatoeba (Tom & Mary) and collected Estonian proper nouns containing the diacritics.

- 1650 sentence pairs for DE-ET
- 20241 sentence pairs for EN-ET

You know who **Tom** is, don't you? - Sa ju tead, kes on **Tom**?

↓

You know who **Tõnis** is, don't you? - Sa ju tead, kes on **Tõnis**?

Spoken Language

Sub-word level **insertion**, **substitution**, and **deletion operations** with fixed probabilities derived from speech recognition output.

Validation	baseline	ft 95-5	ft 90-10	ft 75-25	ft 50-50
MT	39.9	39.7	39.7	39.6	39.4
ASR translation	32.4	32.8	32.7	32.4	32.2

BLEU scores

Speech translation fine-tuning this way is not beneficial, use general model.

Training Summary

1) **Training** on whole **parallel dataset** and **augmented NE data**
2) **Fine-tune general model** with parallel **domain data**
3) **Second training iteration** with whole **data from (1)**, and the whole **back-translated dataset** (yielding final general model)
4) **Fine-tune final general model** on **domain data**, sample back-translated domain data if there are fewer than 50,000 sentences

Domain Detection

Fine-tuned XLM-Roberta

Metric	General	Legal	Crisis	Military
Precision	0.61	0.77	0.88	0.85
Recall	0.84	0.80	0.57	0.49
Recall*	0.84	0.97	0.94	0.87

Recall* - True positive is either correct domain or general domain

Evaluation

Benchmarks

Selected monolingual data and translated by translators.

Domain	ET-EN	ET-DE	ET-RU
General	1152	1166	1126
Legal	500	500	500
Crisis	500	500	500
Crisis-doc	177	177	177
Military	500	500	500
Military-doc	194	194	194
Spoken	1602	1602	1602

Translation Evaluation

Automatic metrics

BLEU

chrF

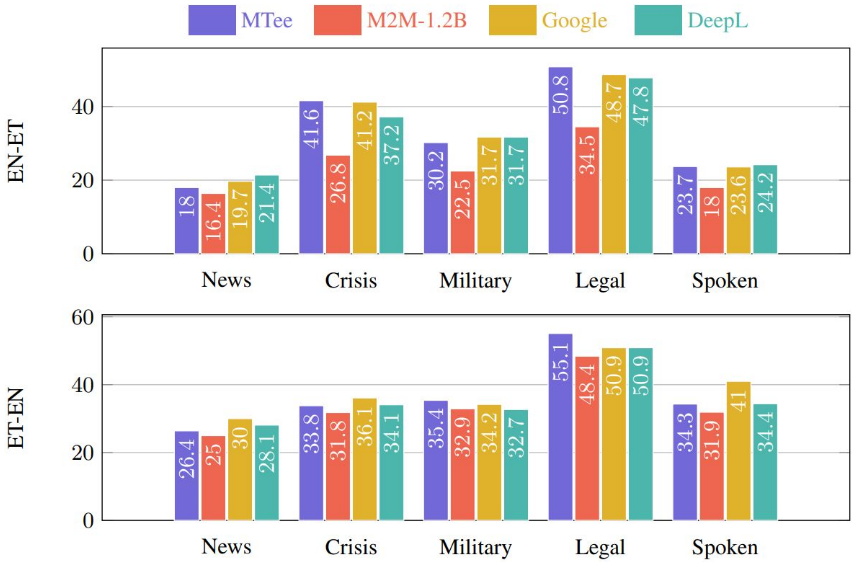
COMET

TARTUNLP

Results

EN ↔ ET

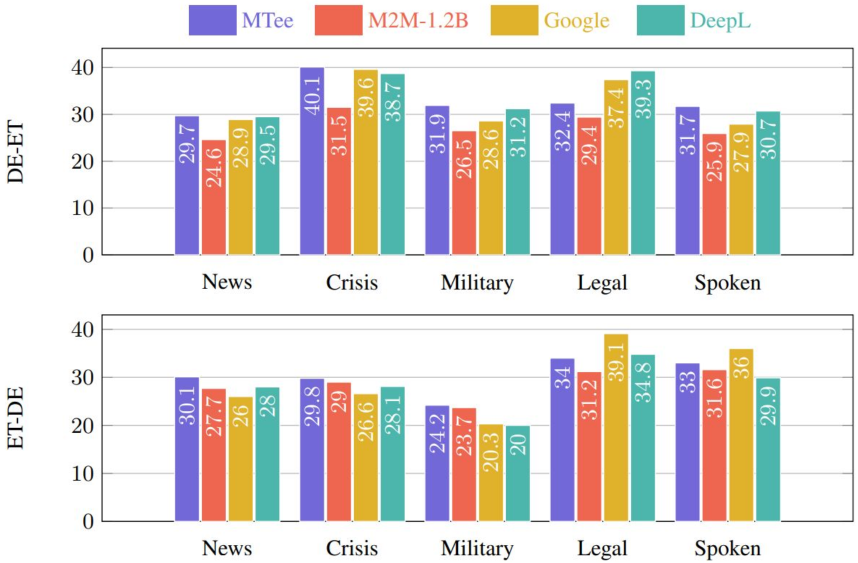
DeepL and Google outperform MTee except for legal and EN-ET crisis.



Results

DE ↔ ET

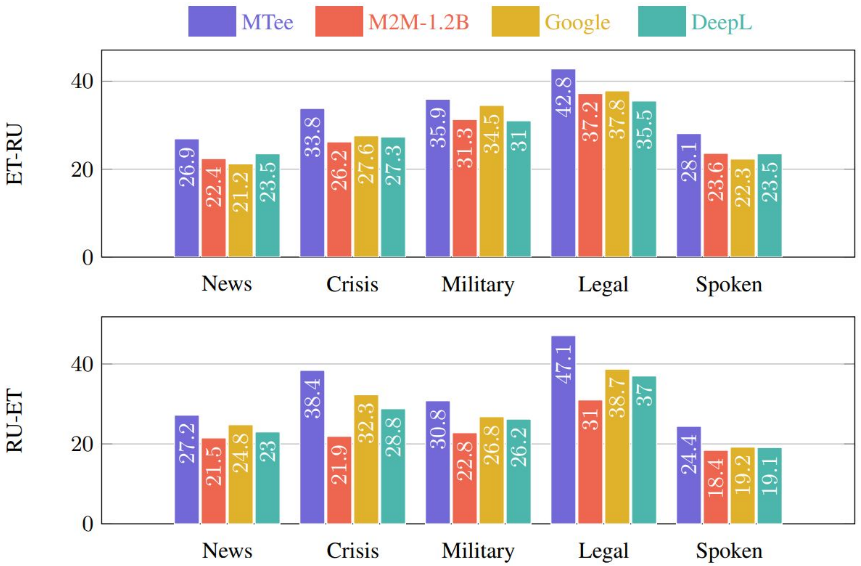
MTee achieves the best results in every domain except legal.



Results

RU ↔ ET

MTee outperforms the other systems in all domains.



Results

With domain detection (crisis)

Apply domain detection (*dd*) before inference

base - general model
ft - fine-tuned with domain data
ft+gen - fine-tuned with domain data and general data

	BLEU				
	base	ft	ft+gen	dd+ft	dd+ft+gen
ET-EN	34.3	**36.1**	35.9	35.6	35.9
ET-DE	29.8	**31.3**	29.8	30.7	29.7
ET-RU	34.7	**35.7**	33.7	35.4	33.7
EN-ET	41.9	**42.5**	35.5	41.8	36.5
DE-ET	46.6	**49.1**	43.8	40.2	39.7
RU-ET	39.0	**39.2**	33.7	38.1	33.6
avg	37.7	**39.0**	35.4	37.0	34.9

Live Demo

https://mt.cs.ut.ee/

Conclusion

As a result of this project we have made available (**Open-source**):

- Monolingual and parallel data
- Benchmarks
- Translation models
- Demo

TARTUNLP

Thank you!

www.tartunlp.ai