Estonian-Centric Machine Translation: Data, Models, and Challenges

Elizaveta Korotkova and **Mark Fishel**
Institute of Computer Science
University of Tartu, Estonia
{elizaveta.korotkova, mark.fisel}@ut.ee

Abstract

Machine translation (MT) research is most typically English-centric. In recent years, massively multilingual translation systems have also been increasingly popular. However, efforts purposefully focused on less-resourced languages are less widespread. In this paper, we focus on MT from and into the Estonian language. First, emphasizing the importance of data availability, we generate and publicly release a back-translation corpus of over 2 billion sentence pairs. Second, using these novel data, we create MT models covering 18 translation directions, all either from or into Estonian. We re-use the encoder of the NLLB multilingual model and train modular decoders separately for each language, surpassing the original NLLB quality. Our resulting MT models largely outperform other open-source MT systems, including previous Estonian-focused efforts, and are released as part of this submission.

1 Introduction

The majority of work on neural machine translation (NMT) is nowadays primarily English-centric, with some notable work on (massively) multilingual MT (Fan et al., 2020; NLLB Team et al., 2022; Kudugunta et al., 2023). In recent years, some attention has been directed at translation directions out of English (e.g. this is the primary focus of the WMT'2024 evaluation campaign[1]) or at pairs that do not include English: for instance, recent WMT and IWSLT shared tasks included one or two such pairs (Kocmi et al., 2023; Kocmi et al., 2022; Agarwal et al., 2023).

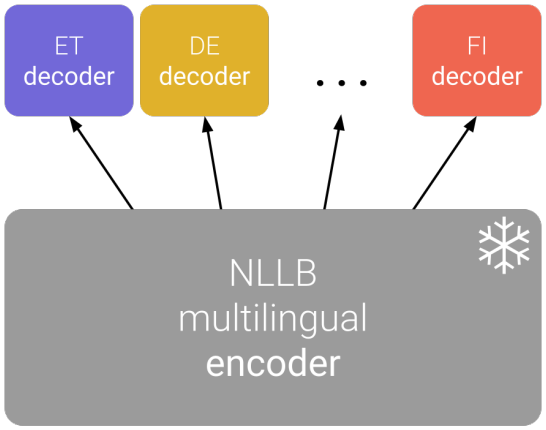

Figure 1: Model architecture. We reuse the multilingual Transformer encoder of NLLB-1.3B and train a new Transformer decoder for each target language.

In this work, we present our recent efforts on advancing Estonian-centric machine translation. In a broader scope the work is part of the Neurotõlge project, which develops open machine translation for Estonian.[2] The name Neurotõlge means *Neural translation* in Estonian and the work on its development has started in 2017 and is ongoing.

The present contribution covers 18 new translation directions for Neurotõlge from and into Estonian. We openly release a massive back-translation corpus for these language pairs, extending the Synthetic Corpus of Parallel Estonian (SynEst) (Korotkova et al., in press), and release translation models trained using these data.

We employ a partially modular approach (Escolano et al., 2021; Lyu et al., 2020) in creating

[1] http://www2.statmt.org/wmt24/translation-task.html

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

Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1), pages 647–660
June 24-27, 2024 ©2024 European Association for Machine Translation

translation models. Specifically, we use the encoder of an existing massively multilingual translation system NLLB (NLLB Team et al., 2022) and create the decoders for each target language as separate modules (the architecture is shown in Figure 1). This setup makes it possible to train the decoders independently, and any subset of the decoders can be deployed afterwards. The achieved translation quality is better than the original NLLB system and also surpasses other open systems on the included translation directions.

The main contributions of this paper are thus:

- we extend the SynEst corpus to cover 12 new translation directions and 4 new data sources, adding over 2 billion filtered sentence pairs to the corpus, and make the full corpus publicly available;[3]
- we create new MT systems for Estonian translation, covering 6 translation directions from Estonian and 12 translation directions into Estonian. Our systems demonstrate stronger translation performance than previous open-source efforts, including Estonian-centric ones, on most language pairs when translating from Estonian into other languages, and show especially noticeable and consistent improvements for translation into Estonian (up to 13 BLEU (Papineni et al., 2002) depending on translation direction and text domain). The models are released for open use.[4]

2 Related Work

In our work, we focus on strengthening the capabilities of open-source MT systems focused on the Estonian language. This builds upon previous efforts centered on Estonian public translation, most recently, the MTEE governmental project (Tättar et al., 2022), and, more generally, the Neurotõlge project and online translation engine.[2] MTEE covered translation between Estonian and three other languages: English, German, and Russian, and achieved state-of-the-art translation quality at the time (Tättar et al., 2022). In this work, we train Estonian-centric models for more language pairs, outperforming the MTEE models in most cases.

Instead of training models from scratch, we use the NLLB multilingual translation model (NLLB Team et al., 2022) as a starting point for our systems. NLLB is a massive effort utilizing the multilingual MT approach (Dong et al., 2015; Johnson et al., 2017), and covering 200 languages, which makes it a convenient base on which to build systems tailored to a smaller number of languages.

In this work, we mostly rely on creating large amounts of new training data to improve Estonian translation. Specifically, we use the back-translation technique (Sennrich et al., 2016). Existing MT systems are used to generate translations of monolingual corpora into desired languages. The obtained parallel data is then reversed and used to augment the training corpus. Thus, the noisy, automatically translated text is on the source side, and the target side contains the cleaner original data, which allows the model to learn text generation based on genuine data. Specifically, we use and extend the SynEst corpus (Korotkova et al., in press), an Estonian-focused back-translation dataset, to cover new translation directions and source corpora.

In terms of model architecture, our systems are inspired by modular approaches (Lyu et al., 2020; Escolano et al., 2021), where multilingual MT models share encoder and decoder modules for each input and output language instead of having one encoder and one decoder covering all languages. More specifically, we use an existing multilingual encoder module from NLLB and train a new decoder for each target language from scratch, somewhat similarly to concurrent work on "mix-and-match translation" by Purason et al. (2024), where encoders and decoders from different models are unified to form a new model.

3 Extending the SynEst Corpus

Synthetic Corpus of Parallel Estonian, or SynEst (Korotkova et al., in press), includes data from the NewsCrawl monolingual corpus (Kocmi et al., 2023) automatically translated into Estonian from 11 languages (Arabic, Chinese, English, Finnish, French, German, Latvian, Lithuanian, Russian, Spanish, and Ukrainian). The dataset can be used as a back-translated corpus to facilitate training MT models which include Estonian.

In this work, we significantly extend SynEst to

[3] https://metashare.ut.ee/repository/search/?q=SynEst, for direct DOI links to each language pair, see Appendix B.

[4] https://huggingface.co/tartuNLP/synest-models

code	target language	parallel	back-translated corpus				total
			NewsCrawl	ParaCrawl	UNPC	OpenSubtitles	
DE	German	9.3	332.6	159.3	–	–	501.2
EN	English	19.6	254.7	433.2	19.4	61.0	787.9
FI	Finnish	15.0	23.5	19.1	–	–	57.6
RU	Russian	5.1	86.8	2.2	13.1	–	107.2
UK	Ukrainian	2.6	1.8	6.7	–	–	11.1
ZH	Chinese	5.8	10.4	4.7	–	–	20.9

Table 1: Sizes of training corpora for models translating from Estonian into other languages (filtered, in millions of sentence pairs). Parallel shows the total size of all parallel corpora used for each language pair. For back-translated corpora, the source side (Estonian) is the automatically translated data, while the target side is the original data. UNPC denotes the United Nations Parallel Corpus.

code	source language	parallel	ENC	total
AR	Arabic	6.3	94.3	100.6
DE	German	9.3	143.8	153.1
EN	English	19.6	144.7	164.3
ES	Spanish	19.5	126.8	146.3
FI	Finnish	15.0	136.8	151.8
FR	French	18.8	132.1	150.9
LT	Lithuanian	10.5	132.7	143.2
LV	Latvian	7.1	132.2	139.3
RU	Russian	5.1	112.1	117.2
SV	Swedish	13.4	127.8	141.2
UK	Ukrainian	2.6	115.7	118.3
ZH	Chinese	5.8	113.6	119.4
total				1,645.6

Table 2: Sizes of training corpora for models translating into Estonian from other languages (filtered, in millions of sentence pairs). Parallel shows the total size of all parallel corpora used for each language pair. ENC denotes the Estonian Parallel Corpus. The Estonian Parallel Corpus was back-translated: the source side is the data automatically translated from Estonian into other languages, while the target side is the original Estonian data.

include more source corpora and translation directions, most importantly, introducing translation directions *from* Estonian. We make the updated dataset publicly available for unrestricted use.[3]

3.1 Translation Directions into Estonian

For translation directions into Estonian, we extend the corpus with three new data sources: ParaCrawl (Bañón et al., 2020), the United Nations Parallel Corpus (Ziemski et al., 2016), and OpenSubtitles (Lison and Tiedemann, 2016).

In case of ParaCrawl, we use 10 language pairs present in this parallel corpus: one side is always English, and the other one of German, Spanish, Finnish, French, Lithuanian, Latvian, Russian, Swedish, Ukrainian, and Chinese. We automatically translate both sides of the corpora into Estonian. The sizes of the resulting corpora range from 5.4 million sentence pairs for Russian–Estonian to a total of 878.4 million pairs for English–Estonian. As both sides of the parallel corpus are translated into a third language (Estonian), this setup opens the possibility of exploring triangular MT approaches; however at present we treat the corpora we translate as monolingual and leave investigation of this direction for future work.

For the United Nations Parallel Corpus, we translate its English and Russian monolingual subsets into Estonian, obtaining 33.4 million and 28.5 million sentence pairs before filtering, respectively. Finally, we translate the English OpenSubtitles corpus into Estonian as well, resulting in 441.4 million sentence pairs before filtering.

The total sizes of the generated dataset for each source corpus and translation direction are given in Table 8 in Appendix A.

3.2 Translation Directions from Estonian

Most importantly, we focus on extending the SynEst synthetic corpus to include translation directions from Estonian. This will allow to use the corpus to train models for translation into Estonian. We translate the Estonian National Corpus (Koppel and Kallas, 2022) into 12 languages: Arabic, Chinese, English, Finnish, French, German, Latvian, Lithuanian, Russian, Spanish, Swedish, and Ukrainian. The resulting back-translation corpus contains between 171.4 million and 196.6 million sentence pairs per translation direction (see Table 7 in Appendix A for approximate numbers

	target language					
	DE	EN	FI	RU	UK	ZH
NLLB-1.3B	24.4	36.7	15.5	22.4	18.7	25.0
MTEE	25.8	37.0	–	22.4	–	–
MADLAD-3B	26.0	37.8	20.1	20.0	15.5	33.5
Ours	27.5	38.1	21.9	23.5	21.3	31.6
DeepL	**30.9**	39.9	**24.4**	**26.7**	**25.6**	40.5
Google	30.8	**41.7**	22.9	26.6	24.4	**42.2**

Table 3: BLEU scores on the FLORES-devtest benchmark for models translating from Estonian into other languages. The best scores overall are shown in **bold**, and the best scores among open-source models are underlined. For our models, we report the score of the checkpoint with the best validation BLEU. With MTEE, we use the general-domain model to translate the FLORES benchmark.

for each translation direction).

3.3 Translation Models

For generating the synthetic side of the SynEst corpus we translate from and into English, German, and Russian with the MTee models (Tättar et al., 2022), using the domain-specific engines MTee-legal for the United Nations Parallel Corpus and MTee-general for all other corpora. For translation directions not involving these languages we use the M2M-100 1.2B-parameter model (Fan et al., 2020). In all cases, we use beam search with beam size 5.

4 Experiments

4.1 Models

We replicate the model setup used in previous exploratory experiments (Korotkova et al., in press). We base our systems on the multilingual NLLB-1.3B dense model (NLLB Team et al., 2022). We freeze the NLLB encoder and train a new, randomly initialized Transformer decoder (Vaswani et al., 2017) for each target language. We keep the dimensions of the decoder layers the same as in the encoder, but use 6 decoder layers instead of the encoder's 24. Keeping the encoder parameters fixed allows to reduce the training-time costs, while reducing the size of the decoder lowers both training- and inference-time costs compared to full fine-tuning of the base model. Freezing the encoder parameters also maintains the multilingual properties of the encoder, meaning that after fine-tuning the model on a certain translation direction it can still translate from any of the 200 languages of NLLB. As all models share the same encoder parameters, final models can be built in a modular fashion, with a single decoder for all translation directions, and one encoder per target language.

We focus on creating Estonian-centric MT models: all translation directions in our experiments include Estonian as either the source or the target language. Specifically, for translation from Estonian into other languages, we train models that translate into German, English, Finnish, Russian, Ukrainian, and Chinese. For translation into Estonian, as the encoder is shared between all models and Estonian is the common target language, we train a single model on the concatenation of data representing 12 language pairs (see Table 2).

We use FairSeq (Ott et al., 2019) to train our models; details on model and training hyperparameters can be found in Appendix D.

4.2 Training Data

To train our models, we use two types of data: parallel corpora and the extended SynEst back-translated corpus.

We use the concatenation of 10 parallel corpora: CCMatrix (Schwenk et al., 2021b), WikiMatrix (Schwenk et al., 2021a), MultiParaCrawl (Bañón et al., 2020), Europarl (Koehn, 2005), OpenSubtitles (Lison and Tiedemann, 2016), JRC-Acquis (Steinberger et al., 2006), TED2020 (Reimers and Gurevych, 2020), EMEA, infopankki, and DGT (Tiedemann, 2012). For the Estonian–English language pair, MultiParaCrawl is replaced with ParaCrawl (Bañón et al., 2020). Not all of these corpora exist for each language pair in our experiments; we use each of the corpora whenever it is available for a language pair.

For SynEst, we use all source corpora available for a given translation direction. As the dataset is used as additional back-translation data, the auto-

	ET–DE	ET–EN	ET–RU
News			
NLLB-1.3B	25.8	25.6	22.8
MTEE	30.1	26.4	**26.9**
MADLAD-3B	26.3	28.7	19.7
Ours	**30.5**	25.9	26.5
DeepL	28.0	28.1	23.5
Google	26.0	**30.0**	21.2
Crisis			
NLLB-1.3B	26.3	21.4	26.2
MTEE	29.8	33.8	33.8
MADLAD-3B	22.1	35.0	25.0
Ours	**30.3**	33.2	**34.7**
DeepL	28.1	34.1	27.3
Google	26.6	**36.1**	27.6
Military			
NLLB-1.3B	21.0	31.1	30.1
MTEE	24.2	**35.4**	**35.9**
MADLAD-3B	19.6	33.2	28.8
Ours	**25.4**	32.9	35.7
DeepL	20.0	32.7	31.0
Google	20.3	34.2	34.5
Legal			
NLLB-1.3B	27.1	48.9	35.5
MTEE	34.0	**55.1**	42.8
MADLAD-3B	32.1	47.8	39.9
Ours	34.7	53.7	**43.0**
DeepL	34.8	50.9	35.5
Google	**39.1**	50.9	37.8
Spoken			
NLLB-1.3B	29.3	30.5	23.3
MTEE	33.0	34.3	**28.1**
MADLAD-3B	33.1	35.2	22.8
Ours	33.2	32.2	28.0
DeepL	29.9	34.4	23.5
Google	**36.0**	**41.0**	22.3

Table 4: BLEU scores on the MTEE domain benchmark sets for models translating from Estonian into other languages. The best scores overall are shown in **bold**, and the best scores among open-source models are underlined. For our models, we report the score of the checkpoint with the best validation BLEU. With MTEE, we show the scores reported by Tättar et al. (2022).

matically generated side of the corpus is always used as the source and the cleaner original data as the target during training.

We concatenate all corpora to create our full training dataset. Approximate sizes of the full training corpora and their components are shown in Tables 1 and 2 for model translation directions from Estonian and into Estonian, respectively. (The sizes are shown after filtering; details on data filtering can be found in Appendix C).

The dev split of the FLORES dataset (Goyal et al., 2022) is used as the validation set.

4.3 Evaluation

We compare the performance of our Estonian-centric models to that of three other open-source MT systems:

- the NLLB-1.3B (NLLB Team et al., 2022) multilingual translation model, which also serves as the starting model in our experiments;
- the models trained within the MTEE project (Tättar et al., 2022), which was the previous effort of public Estonian-centric MT. These models cover the Estonian↔German, Estonian↔English, and Estonian↔Russian translation directions, and employ a fully modular approach;
- the more recent MADLAD-400 3B (Kudugunta et al., 2023).

For additional comparison, we also show the results of DeepL[5] and Google Translate,[6] two widely used proprietary online translation engines.

The test sets we employ for evaluation are the FLORES evaluation benchmark (Goyal et al., 2022) (the devtest split), and the MTEE domain-specific benchmark sets (Tättar et al., 2022). FLORES is useful in providing a benchmark for multilingual translation between many languages, which is based on Wikipedia. MTEE, while covering fewer language pairs (Estonian–English, Estonian–German, and Estonian–Russian), is centered on language pairs which include Estonian, and allows to estimate model performance on text belonging to 5 distinct domains.

We use the sacreBLEU implementation (Post, 2018) of the BLEU score (Papineni et al., 2002) to

[5] https://www.deepl.com/translator
[6] https://translate.google.com

	source language											
	AR	DE	EN	ES	FI	FR	LT	LV	RU	SV	UK	ZH
NLLB-1.3B	15.7	17.8	22.7	13.8	16.1	17.3	15.1	16.1	15.8	18.4	16.9	11.6
MTEE	–	21.7	27.6	–	–	–	–	–	20.2	–	–	–
MADLAD-3B	20.3	21.7	26.2	16.3	19.2	19.9	19.3	22.8	17.7	21.3	16.2	15.4
Ours	20.0	23.0	29.4	16.7	20.9	23.3	19.3	21.0	20.1	24.0	21.4	14.6
DeepL	**23.4**	24.4	30.2	**19.0**	**22.5**	23.7	**22.1**	**23.6**	**22.6**	**26.3**	**24.1**	18.0
Google	23.2	**25.3**	**30.7**	18.5	22.4	**24.5**	21.5	**23.6**	**22.6**	25.7	23.3	**18.8**

Table 5: BLEU scores on the FLORES-devtest benchmark for models translating from other languages into Estonian. The best scores overall are shown in **bold**, and the best scores among open-source models are underlined. For our model, we report the score of the checkpoint with the best validation loss (the same checkpoint is used for all source languages). With MTEE, we use the general-domain model to translate the FLORES benchmark.

measure the models' performance.[7] Additionally, we report COMET scores (Rei et al., 2020) in Appendix E. For models translating from Estonian, we choose the checkpoint which shows the best BLEU score on FLORES-dev for the language pair in question. For the models translating into Estonian, we use the checkpoint showing the best loss on the combined validation set; we do not choose a best checkpoint for each source language separately.

5 Results

BLEU scores of NLLB-1.3B, MTEE, MADLAD-3B, our model, DeepL, and Google Translate on FLORES-devtest for translation directions from Estonian into other languages (our experiments cover German, English, Finnish, Russian, Ukrainian, and Chinese as target languages) are shown in Table 3. In this setting, our model shows the strongest results among the open-source systems for five out of six language pairs, outperforming the next best open-source models by 0.3 to 2.6 BLEU points. On the MTEE domain benchmarks (Table 4), our model consistently outperforms other open-source ones on the Estonian–German language pair, while for Estonian–English it shows lower scores than the MTEE and, for most domains, MADLAD models. For Estonian–Russian, results are more mixed, with our models being the best among all models on the crisis and legal domains (with a small margin of 0.2 BLEU over MTEE for legal and a more noticeable one of 0.9 BLEU for crisis) and falling slightly behind MTEE on the news, military, and spoken domains (by up to 0.4 BLEU).

Table 5 shows results on FLORES-devtest for translation into Estonian. Our model noticeably improves upon the NLLB model for all translation directions, while also outperforming all compared open-source models on 7 out of 12 translation directions. On three more directions, the difference between our model and the best performing one among open systems does not exceed 0.3 BLEU points.

From Table 6 we see that our into-Estonian model performs consistently well on different domains. It outperforms all models, including proprietary ones and the MTEE models fine-tuned to these domains, on all language pairs and domains, with the exception of EN–ET news, with margins to the next best models ranging from 0.2 to 13 BLEU for different language pairs and domains. This consistently strong performance can be attributed to the fact that this single model has encountered a vast amount of training data, with 12 input languages and Estonian as the output language, leading it to learn generating Estonian output very well.

6 Deployment and Known Issues

The models are made publicly available on the HuggingFace model hub[4] and can be run using the TartuNLP translation worker.[8] The models are set up in a modular fashion, with one encoder covering all input languages and a separate decoder for each output language.

We have found that the models are not robust to some inputs, such as single words; while full

[7] sacreBLEU signature for all target languages except Chinese: `nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1`. For Chinese: the same with `tok:zh`.

[8] `https://github.com/TartuNLP/translation-worker/tree/nllb-based-est`

	DE–ET	EN–ET	RU–ET
News			
NLLB-1.3B	22.0	15.6	19.5
MTEE	29.7	18.0	27.2
MADLAD-3B	24.9	19.0	22.5
Ours	**33.2**	19.7	**30.0**
DeepL	29.5	**21.4**	23.0
Google	28.9	19.7	24.8
Crisis			
NLLB-1.3B	27.4	24.3	20.1
MTEE	40.1	41.6	38.4
MADLAD-3B	36.2	31.1	27.2
Ours	**53.1**	**45.8**	**40.8**
DeepL	38.7	37.2	28.8
Google	39.6	41.2	32.3
Military			
NLLB-1.3B	22.6	21.6	20.1
MTEE	31.9	30.2	30.8
MADLAD-3B	28.0	24.6	24.1
Ours	**37.1**	**31.9**	**32.7**
DeepL	31.2	31.7	26.2
Google	28.6	31.7	26.8
Legal			
NLLB-1.3B	25.0	31.1	26.9
MTEE	32.4	50.8	47.1
MADLAD-3B	31.1	31.7	37.9
Ours	**48.0**	**52.1**	**50.3**
DeepL	39.2	47.8	37.0
Google	37.4	48.7	38.7
Spoken			
NLLB-1.3B	23.0	18.0	16.9
MTEE	31.7	23.7	24.4
MADLAD-3B	27.5	22.2	19.5
Ours	**37.5**	**26.1**	**27.3**
DeepL	30.7	24.2	19.1
Google	27.9	23.6	19.2

Table 6: BLEU scores on the MTEE domain benchmark sets for models translating from other languages into Estonian. The best scores overall are shown in **bold**, and the best scores among open-source models are underlined. For our model, we report the score of the checkpoint with the best validation loss (the same checkpoint is used for all source languages). With MTEE, we show the scores reported by Tättar et al. (2022).

sentence translation works reasonably well, with single-word or isolated phrase input the models may start severely overgenerating.

7 Future Work

So far the efforts of the project have focused on sentence-level NMT. The next iterations of development and model training will likely focus on document-level MT, either with sequence-to-sequence or decoder-only models. Moreover, we are looking into instruction-tuned sequence-to-sequence models: this approach should yield translation-specific emergent abilities and would thus enable the integration of terminologies, on-the-fly domain adaptation, and other types of translation output control. We also plan to dedicate more attention to the robustness of the developed translation engines, for instance, by including upper-cased data in the training dataset for smoother handling of headlines and other all-caps segments, as well as including phrase and word pairs to enhance translation performance when the input is not a complete sentence.

8 Conclusion

In this work, we have made a contribution towards open-source machine translation centered on the Estonian language.

First, we presented an extended version of the SynEst synthetic corpus. The new version introduces 12 translation directions from Estonian, in addition to previously present directions into Estonian. In total, we have generated over 2 billion filtered sentence pairs. We release the full corpus for public use and hope that the availability of this resource will facilitate further work on Estonian translation.

Second, we created new MT models for translation from Estonian into 6 languages and from 12 languages into Estonian and made them publicly available. Evaluation on two benchmarks covering 6 domains has shown that our models are comparable to or outperform previous open efforts on translation from Estonian, depending on the language pair and domain, and perform especially well on translation into Estonian, outperforming not only previous open-source but also proprietary systems by up to 13 BLEU on some domains. These consistent improvements are likely due to the use of massive amounts of synthetic data we created.

References

Agarwal, Milind, Sweta Agrawal, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Mingda Chen, William Chen, Khalid Choukri, Alexandra Chronopoulou, Anna Currey, Thierry Declerck, Qianqian Dong, Kevin Duh, Yannick Estève, Marcello Federico, Souhir Gahbiche, Barry Haddow, Benjamin Hsu, Phu Mon Htut, Hirofumi Inaguma, Dávid Javorský, John Judge, Yasumasa Kano, Tom Ko, Rishu Kumar, Pengwei Li, Xutai Ma, Prashant Mathur, Evgeny Matusov, Paul McNamee, John P. McCrae, Kenton Murray, Maria Nadejde, Satoshi Nakamura, Matteo Negri, Ha Nguyen, Jan Niehues, Xing Niu, Atul Kr. Ojha, John E. Ortega, Proyag Pal, Juan Pino, Lonneke van der Plas, Peter Polák, Elijah Rippeth, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Yun Tang, Brian Thompson, Kevin Tran, Marco Turchi, Alex Waibel, Mingxuan Wang, Shinji Watanabe, and Rodolfo Zevallos. 2023. Findings of the IWSLT 2023 evaluation campaign. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 1–61, Toronto, Canada (in-person and online).

Aulamo, Mikko, Sami Virpioja, and Jörg Tiedemann. 2020. OpusFilter: A configurable parallel corpus filtering toolbox. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 150–156, Online, July. Association for Computational Linguistics.

Bañón, Marta, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online, July. Association for Computational Linguistics.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Dong, Daxiang, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1723–1732, Beijing, China, July. Association for Computational Linguistics.

Escolano, Carlos, Marta R. Costa-jussà, José A. R. Fonollosa, and Mikel Artetxe. 2021. Multilingual machine translation: Closing the gap between shared and language-specific encoder-decoders. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 944–948, Online, April. Association for Computational Linguistics.

Fan, Angela, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Çelebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22:107:1–107:48.

Goyal, Naman, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Johnson, Melvin, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Kingma, Diederik and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, San Diego, CA, USA.

Kocmi, Tom, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid).

Kocmi, Tom, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In Koehn, Philipp, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore, December. Association for Computational Linguistics.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of*

Machine Translation Summit X: Papers, pages 79–86, Phuket, Thailand, September 13-15.

Koppel, Kristina and Jelena Kallas. 2022. Eesti keele ühendkorpuste sari 2013–2021: mahukaim eesti-keelsete digitekstide kogu. *Eesti Rakenduslingvistika Ühingu aastaraamat = Estonian papers in applied linguistics*, 18:207–228.

Korotkova, Elizaveta, Taido Purason, Agnes Luhtaru, and Mark Fishel. in press. Multilinguality or back-translation? A case study with Estonian. In *Accepted for publication at the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*. European Language Resources Association (ELRA).

Kudo, Taku and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium, November. Association for Computational Linguistics.

Kudugunta, Sneha, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: A multilingual and document-level large audited dataset. In Oh, A., T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 67284–67296. Curran Associates, Inc.

Lison, Pierre and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Lyu, Sungwon, Bokyung Son, Kichang Yang, and Jaekyoung Bae. 2020. Revisiting Modularized Multilingual NMT to Meet Industrial Demands. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5905–5918, Online, November. Association for Computational Linguistics.

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.

Ott, Myle, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota, June. Association for Computational Linguistics.

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, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Post, Matt. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium, October. Association for Computational Linguistics.

Purason, Taido, Andre Tättar, and Mark Fishel. 2024. Mixing and matching: Combining independently trained translation model components. In *Proceedings of the 1st Workshop on Modular and Open Multilingual NLP (MOOMIN 2024)*, pages 44–56, St Julians, Malta.

Rei, Ricardo, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In Webber, Bonnie, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online, November. Association for Computational Linguistics.

Rei, Ricardo, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In Koehn, Philipp, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid), December. Association for Computational Linguistics.

Reimers, Nils and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the*

2020 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 11.

Schwenk, Holger, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021a. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online, April. Association for Computational Linguistics.

Schwenk, Holger, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021b. CCMatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6490–6500, Online, August. Association for Computational Linguistics.

Sennrich, Rico, 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, Berlin, Germany, August. Association for Computational Linguistics.

Srivastava, Nitish, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.

Steinberger, Ralf, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, and Dániel Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May. European Language Resources Association (ELRA).

Tättar, Andre, Taido Purason, Hele-Andra Kuulmets, Agnes Luhtaru, Liisa Rätsep, Maali Tars, Mārcis Pinnis, Toms Bergmanis, and Mark Fishel. 2022. Open and competitive multilingual neural machine translation in production. In *Baltic Journal of Modern Computing*, volume 10, pages 422–434.

Tiedemann, Jörg. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Ziemski, Michał, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations parallel corpus v1.0. In Calzolari, Nicoletta, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534, Portorož, Slovenia, May. European Language Resources Association (ELRA).

A Back-translated Dataset Sizes

The approximate number of sentence pairs in each of our back-translated corpora before filtering are shown in Table 7 (translated from Estonian into other languages) and Table 8 (translated from other languages into Estonian).

target language	ENC
Arabic	183.7
German	196.6
English	196.4
Spanish	172.7
Finnish	177.7
French	173.7
Lithuanian	174.0
Latvian	174.3
Russian	196.3
Swedish	171.4
Ukrainian	175.5
Chinese	189.0

Table 7: Sizes of the back-translation corpora translated from Estonian (unfiltered, in millions of sentence pairs). ENC stands for the Estonian National Corpus.

	corpus			
source language	NC	PC	UNPC	OS
Arabic	42.3	–	–	–
German	427.1	278.3	–	–
English	314.3	878.4	33.4	441.4
Spanish	72.1	208.4	–	–
Finnish	28.8	31.3	–	–
French	104.8	217.6	–	–
Lithuanian	7.6	13.2	–	–
Latvian	14.9	13.1	–	–
Russian	126.6	5.4	28.5	–
Swedish	–	49.1	–	–
Ukrainian	2.3	13.2	–	–
Chinese	13.9	14.2	–	–

Table 8: Sizes of the back-translation corpora translated into Estonian (unfiltered, in millions of sentence pairs). NC, PC, UNPC, and OS denote the NewsCrawl, ParaCrawl, United Nations Parallel Corpus, and OpenSubtitles corpora, respectively.

B Digital Object Identifiers for the Extended SynEst Corpus

The DOIs for each language pair of the extended SynEst corpus are shown in Table 9.

C Data Filtering

The back-translation datasets are filtered based on log probability of the generated translations. We only keep the examples that where log probability is higher than $\mu - 1.5\sigma$ where μ is the mean and σ is the standard deviation over all translation log probabilities for a given translation direction and corpus.

All data, both synthetic and parallel, are normalized with the MTee normalization script (Tättar et al., 2022) and filtered with OpusFilter (Aulamo et al., 2020). The following filters are used:

1. `LongWordFilter`: filter examples with words longer than 40 characters (default).
2. `LengthFilter`: filter examples longer than 1000 characters or shorter than 10 characters.
3. `LengthFilter`: filter examples longer than 100 words.
4. `LengthRatioFilter`: filter examples where the source and target sentence lengths differ more than 3 times in terms of number of words.
5. `CharacterScoreFilter` with threshold 1 (default) for the respective scripts.
6. `LanguageIDFilter` with fastText (Bojanowski et al., 2017) language identification model.
7. `LanguageIDFilter` with CLD2 language identification.
8. `TerminalPunctuationFilter` with the default parameters.
9. `NonZeroNumeralsFilter` with the default parameters.

This configuration is applied to all language pairs with the following exceptions:

- Arabic–Estonian, which uses filters 1 – 6 and uses minimal sentence length of 3 characters in filter 2;

language pair	DOI
Arabic–Estonian	doi.org/10.15155/y746-qa68
German–Estonian	doi.org/10.15155/2fy2-2k14
English–Estonian	doi.org/10.15155/5r1e-6r35
Spanish–Estonian	doi.org/10.15155/sqk9-ze70
Finnish–Estonian	doi.org/10.15155/hjw7-m565
French–Estonian	doi.org/10.15155/4vb6-ab11
Lithuanian–Estonian	doi.org/10.15155/7at2-jv07
Latvian–Estonian	doi.org/10.15155/erkh-k466
Russian–Estonian	doi.org/10.15155/4e20-vs27
Swedish–Estonian	doi.org/10.15155/jfws-ed89
Ukrainian–Estonian	doi.org/10.15155/xmpv-ft58
Chinese–Estonian	doi.org/10.15155/m6ww-j693
Estonian–all	doi.org/10.15155/ctz5-1d43

Table 9: DOIs for the extended SynEst corpus

- Chinese–Estonian, which only uses `LengthFilter` with maximal sentence length of 750 characters (no minimal length), `CharacterScoreFilter`, and `LanguageIDFilter` with fastText as language identification model.

Duplicates and test set overlaps are removed from the training dataset.

D Training Details

The models are trained with FairSeq (Ott et al., 2019). The NLLB-1.3B encoder consists of 24 transformer layers with embedding dimension 1024, feed-forward dimension 8192, and 16 attention heads. The decoders are randomly initialized and have 6 transformer layers; the dimensions of the decoders are the same as those of the encoder. The input and output embeddings of the decoder are shared. The vocabulary size is 256,000 for the encoder and 32,000 for the decoder (we train a separate SentencePiece (Kudo and Richardson, 2018) model for each output language). Models are trained on 8 GPUs (4 AMD MI250x 128GB GPU modules, each acting as 2 GPUs) with batch size 4,096 tokens per GPU. Models are trained for 2,000,000 updates, with checkpoints saved after every 2,000 updates. We use the inverse square root learning rate scheduler with 4,000 warm-up updates from initial learning rate 1×10^{-7} to maximum learning rate 5×10^{-4}. We use Adam optimizer (Kingma and Ba, 2015) with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Dropout probability (Srivastava et al., 2014) is 0.1, attention dropout 0.1, and activation dropout is not used. The loss function is cross-entropy.

E COMET Scores

Tables 10, 11, and 12 show COMET scores (Rei et al., 2020) for translation from and into Estonian on the FLORES benchmark. Tables 13 and 14 contain results of translating the MTEE test sets from and into Estonian, respectively.

COMET scores were calculated with the default `wmt22-comet-da` model (Rei et al., 2022).

	target language					
	DE	EN	FI	RU	UK	ZH
NLLB-1.3B	84.19	88.33	86.65	86.71	85.90	80.01
MTEE	84.88	88.49	–	87.33	–	–
MADLAD-3B	84.64	89.19	89.03	85.55	82.23	85.57
Ours	85.95	88.92	90.25	88.26	87.79	84.51
DeepL	87.08	89.54	**91.44**	89.67	**90.07**	87.69
Google	**87.21**	**89.75**	90.70	**89.74**	89.77	**87.78**

Table 10: COMET scores on the FLORES-devtest benchmark for models translating from Estonian into other languages. The best scores overall are shown in **bold**, and the best scores among open-source models are underlined. For our models, we report the score of the checkpoint with the best validation BLEU. With MTEE, we use the general-domain model to translate the FLORES benchmark.

	source language					
	AR	DE	EN	ES	FI	FR
NLLB-1.3B	84.08	87.37	89.36	86.13	87.23	87.00
MTEE	–	88.82	89.34	–	–	–
MADLAD-3B	87.65	88.86	90.65	87.78	88.84	88.01
Ours	87.34	90.42	91.60	88.67	90.58	89.76
DeepL	**89.02**	**91.25**	**92.54**	**89.78**	**91.13**	**90.67**
Google	88.35	90.34	91.77	89.29	90.72	90.17

Table 11: COMET scores on the FLORES-devtest benchmark for models translating from Arabic, German, English, Spanish, Finnish, and French into Estonian. The best scores overall are shown in **bold**, and the best scores among open-source models are underlined. For our model, we report the score of the checkpoint with the best validation loss (the same checkpoint is used for all source languages). With MTEE, we use the general-domain model to translate the FLORES benchmark.

	source language					
	LT	LV	RU	SV	UK	ZH
NLLB-1.3B	85.36	85.78	86.27	87.50	85.69	84.03
MTEE	–	–	88.28	–	–	–
MADLAD-3B	87.82	90.27	86.07	88.54	83.44	88.48
Ours	88.72	89.92	89.37	90.57	89.08	88.18
DeepL	**90.23**	**91.05**	**89.92**	**91.55**	**90.15**	**89.91**
Google	89.68	90.46	89.42	90.77	89.24	89.55

Table 12: COMET scores on the FLORES-devtest benchmark for models translating from Lithuanian, Latvian, Russian, Swedish, Ukrainian, and Chinese into Estonian. The best scores overall are shown in **bold**, and the best scores among open-source models are underlined. For our model, we report the score of the checkpoint with the best validation loss (the same checkpoint is used for all source languages). With MTEE, we use the general-domain model to translate the FLORES benchmark.

	ET–DE	ET–EN	ET–RU
News			
NLLB-1.3B	83.35	83.64	85.15
MTEE	85.12	84.03	86.70
MADLAD-3B	83.74	85.07	82.41
Ours	85.41	84.19	87.36
DeepL	86.32	**85.51**	**88.71**
Google	**86.64**	85.25	88.70
Crisis			
NLLB-1.3B	83.79	85.08	87.60
MTEE	85.62	86.76	90.18
MADLAD-3B	81.26	87.00	85.75
Ours	85.50	86.65	**90.77**
DeepL	86.18	87.88	90.62
Google	**86.26**	**88.39**	90.24
Military			
NLLB-1.3B	83.05	86.35	88.72
MTEE	84.26	87.14	89.88
MADLAD-3B	80.62	87.34	85.85
Ours	**85.54**	87.04	90.53
DeepL	84.68	87.51	90.51
Google	85.12	**88.12**	**90.60**
Legal			
NLLB-1.3B	84.51	87.12	90.84
MTEE	86.72	**88.17**	92.33
MADLAD-3B	85.01	88.01	90.85
Ours	87.04	88.14	**92.39**
DeepL	**87.09**	87.91	91.07
Google	86.68	87.62	91.32
Spoken			
NLLB-1.3B	80.55	81.65	83.30
MTEE	82.22	82.19	84.04
MADLAD-3B	81.85	83.75	81.44
Ours	82.92	81.96	84.37
DeepL	83.21	83.94	**86.08**
Google	**83.98**	**84.26**	85.67

Table 13: COMET scores on the MTEE domain benchmark sets for models translating from Estonian into other languages. The best scores overall are shown in **bold**, and the best scores among open-source models are underlined. For our models, we report the score of the checkpoint with the best validation BLEU. With MTEE, we calculate the scores on the same model outputs as used by Tättar et al. (2022).

	DE–ET	EN–ET	RU–ET
News			
NLLB-1.3B	85.80	86.61	87.41
MTEE	87.83	85.85	89.34
MADLAD-3B	87.32	87.96	87.00
Ours	89.88	**88.93**	**91.07**
DeepL	**90.45**	**89.93**	90.53
Google	90.00	88.47	90.36
Crisis			
NLLB-1.3B	89.55	91.08	87.75
MTEE	91.00	93.96	91.91
MADLAD-3B	91.48	93.02	88.86
Ours	**93.83**	**94.51**	**92.44**
DeepL	92.52	94.36	91.81
Google	92.25	94.07	91.49
Military			
NLLB-1.3B	88.73	92.26	89.24
MTEE	90.81	93.40	92.00
MADLAD-3B	90.33	92.92	89.16
Ours	**92.56**	93.55	**92.57**
DeepL	92.19	**94.28**	91.81
Google	91.43	93.92	91.54
Legal			
NLLB-1.3B	90.07	92.88	91.13
MTEE	91.96	95.50	94.23
MADLAD-3B	92.84	93.50	93.54
Ours	**94.51**	**95.62**	**94.72**
DeepL	93.49	95.45	93.49
Google	92.35	94.54	92.22
Spoken			
NLLB-1.3B	86.59	88.31	84.26
MTEE	89.56	90.15	87.51
MADLAD-3B	89.11	90.13	84.33
Ours	**90.88**	90.75	**88.47**
DeepL	90.06	**90.98**	87.23
Google	89.58	90.72	87.30

Table 14: COMET scores on the MTEE domain benchmark sets for models translating from other languages into Estonian. The best scores overall are shown in **bold**, and the best scores among open-source models are underlined. For our model, we report the score of the checkpoint with the best validation loss (the same checkpoint is used for all source languages). With MTEE, we calculate the scores on the same model outputs as used by Tättar et al. (2022).