Multilinguality or Back-translation? A Case Study with Estonian

Elizaveta Korotkova, Taido Purason, Agnes Luhtaru, Mark Fishel
Institute of Computer Science
University of Tartu, Estonia
{elizaveta.korotkova,taido.purason,agnes.luhtaru,mark.fisel}@ut.ee

Abstract

Machine translation quality is highly reliant on large amounts of training data, and, when a limited amount of parallel data is available, synthetic back-translated or multilingual data can be used in addition. In this work, we introduce SynEst, a synthetic corpus of translations from 11 languages into Estonian which totals over 1 billion sentence pairs. Using this corpus, we investigate whether adding synthetic or English-centric additional data yields better translation quality for translation directions that do not include English. Our results show that while both strategies are effective, synthetic data gives better results. Our final models improve the performance of the baseline No Language Left Behind model while retaining its source-side multilinguality.

Keywords: machine translation, synthetic corpus, less-resourced languages

1. Introduction

The quality of neural machine translation systems heavily depends on the availability and quality of training data. While for some languages (first and foremost English) vast amounts of suitable resources are often readily available, for less-resourced languages that is not often the case (Joshi et al., 2020). In such cases, one can resort to generating synthetic data and/or leveraging multilingual resources for transfer learning in order to mitigate the lack of parallel data.

In this work, we directly compare these two data augmentation approaches for machine translation (MT). We focus on Estonian, a mid-resourced European language of the Finno-Ugric language group, with no genealogically or geographically close languages that are particularly resource-rich. We introduce a novel large-scale synthetic parallel corpus, SynEst, consisting of translations from 11 other languages into Estonian. The choice of source languages is motivated both globally, with languages such as English or Chinese, and regionally, for e.g. Finnish, Latvian, and Lithuanian. The resulting corpus contains over 1 billion parallel sentences and is 6 times larger than the monolingual national corpus of Estonian (Koppel and Kallas, 2022) and more than twice the size of the Estonian part of the CulturaX corpus (Nguyen et al., 2023).

With the help of this new resource, we carry out a pilot experiment focused on machine translation from Estonian into other languages, intentionally exploring non-English-centric translation directions. We aim to determine whether a more substantial gain in translation quality can be achieved by using synthetic Estonian–other (ET–X) data or multilingual data, specifically, English–other (EN–X), and show that, while both approaches are successful, augmenting with synthetic data leads to better performance. The final result is an MT system which uses our new synthetic dataset for augmentation and surpasses the baseline No Language Left Behind (NLLB Team et al., 2022) model in quality while increasing its inference speed and retaining its support of multilingual input.

We first briefly outline related work in Section 2, then describe our novel synthetic corpus in Section 3, present the pilot experiments in Section 4, describe their empirical results in Section 5, and discuss them in Section 6. Section 7 concludes the paper.

This work's main contributions are:

- we release a synthetic parallel corpus of over 1 billion sentence pairs with translations from 11 languages into Estonian (Section 3);[1]
- we directly compare two data augmentation methods, namely, leveraging synthetic back-translated data and English-centric data, by performing experiments focused on training machine translation systems for translation from Estonian with limited parallel resources (Section 4);
- we empirically show the usefulness and satisfactory quality of our synthetic dataset for out-of-Estonian machine translation (Section 5).

2. Related Work

Non-English-centric MT has been underexplored in machine translation research compared to language pairs involving English. Recently, however, there has been some shift towards including more pairs without English. For instance, the general MT task at WMT 2020 included 2 translation directions without English out of 22 in total (Barrault et al., 2020), while in 2021 6 out of 20, and in 2022 6 out

[1] https://doi.org/10.15155/a4q3-ma56

LREC-COLING 2024, pages 11838–11848

of 21 directions did not include English (Akhbardeh et al., 2021; Kocmi et al., 2022). In the space of multilingual MT, works such as Fan et al. (2020) have stressed the utility of many-to-many training data as opposed to purely English-centric. In our work, we intentionally focus on experiments with non-English-centric translation directions.

Methods for low-resource MT can be used when a limited amount of parallel data for a translation direction is available. Haddow et al. (2022) outline using back-translated data (Sennrich et al., 2016) and multilingual models (Dong et al., 2015; Johnson et al., 2017) as two such methods. While the language pairs in our experiments are not low-resource but rather mid-resource, in the absence of abundant parallel data, we draw inspiration from low-resource MT, and use both back-translated data and a multilingual model to improve MT performance in our experiments.

Synthetic parallel corpora have proven effective, but can be costly to produce, especially on a massive scale. Thus, efforts similar to ours have published back-translated corpora to be reused. CzEng 2.0 (Kocmi et al., 2020) is a Czech-English parallel corpus that includes automatic translations of 127M total sentences crawled from news servers. Símonarson et al. (2021) create an English-Icelandic parallel corpus of 76M sentences. In this work, we also focus on one relatively under-resourced language and produce a massive synthetic corpus, including 11 translation directions.

Modular architectures for MT were introduced by Escolano et al. (2021) and Lyu et al. (2020). Our choice of a modular architecture with a fixed encoder and language-specific decoders draws inspiration from these works.

3. SynEst: Synthetic Corpus of Parallel Estonian

To create synthetic back-translated data for augmenting our parallel corpus, we translate the whole NewsCrawl monolingual corpus[2] (Kocmi et al., 2022; Haddow et al., 2022) up to year 2021[3] into Estonian. The NewsCrawl corpus contains monolingual text extracted from online newspapers and released for the WMT series of shared tasks. We select 11 languages to translate from: 6 globally wide-spread languages (English, German, Spanish, French, Chinese, Arabic), Finnish as a language closely related to Estonian, and 4 regionally important languages of neighbors and Estonian minorities (Latvian, Lithuanian, Ukrainian, Russian).

[2] https://data.statmt.org/news-crawl/

[3] At the time of translation, NewsCrawl data for 2022 was not available yet and its translation is left for future work.

code	source language	snt count (millions)	word count (billions)
AR	Arabic	42.3	1.0
DE	German	427.1	6.0
EN	English	314.3	5.3
ES	Spanish	72.1	1.3
FI	Finnish	28.8	0.3
FR	French	104.8	1.5
LT	Lithuanian	7.6	0.1
LV	Latvian	14.9	0.2
RU	Russian	126.6	1.6
UK	Ukrainian	2.3	0.03
ZH	Chinese	13.9	0.3

Table 1: Sizes of the synthetic back-translation corpora (unfiltered): snt count gives the number of sentences, and word count gives the number of words in the Estonian output.

The number of the resulting translated sentences and words is shown in Table 1.

We translate from English, German, and Russian with the MTee general-domain model as the most high-performing MT model for these language pairs (Tättar et al., 2022). For all other source 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.

The dataset is available to download via MetaShare[1] under the CC BY license. For each language pair, we provide an unfiltered and filtered corpus (data filtering details are described in Appendix A). Each unfiltered corpus is a file in tab-separated format with three columns: the original sentence, the translation, and the translation's log-probability score. The filtered corpora are provided as `.tar` archives which contain parallel text files.

4. Experiments

In this section, we present a pilot study that uses 3 translation directions from the SynEst corpus. The aim is to see if an existing massively multilingual MT system (NLLB: NLLB Team et al., 2022) can be efficiently improved for the chosen translation directions without losing its multilinguality using modular MT (Escolano et al., 2021; Lyu et al., 2020). Below we describe the parallel data used in addition to SynEst, the modular approach, and evaluation details. Results can be found in the next section.

4.1. Training Data

We focus on translation from Estonian into three target languages: Finnish (closely related to Estonian), German (resource-rich, unrelated language but has some similarities with Estonian on lexical

	original	augmentation	
language pair	ET–X	synth ET–X	EN–X
ET–FI	15.0	23.5	80.3
ET–DE	9.3	332.6	398.6
ET–ZH	5.8	10.4	63.8

Table 2: Sizes (in millions of sentence pairs) of original parallel ET–X corpora used for training and total sizes of available augmentation corpora (after filtering). In our experiments, we always use all original ET–X data and mix it 1:7 with augmentation data, under/oversampling additional data as needed.

and grammatical level), and Chinese (entirely unrelated, language pair is data-scarce for contrast).

We use the concatenation of 10 parallel corpora in our experiments: 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). Each corpus is used whenever it is available for a particular translation direction.

For all three translation directions, there is not an overwhelming amount of parallel data available. To mitigate this, we explore augmenting our parallel data with two types of additional data:

- the synthetic back-translation data of SynEst, making use of available monolingual data in the target languages,
- parallel data between English and the target languages (EN–X), leveraging the relative abundance of English-centric data and the base model's ability to translate from multiple source languages.

As sources of EN–X data, we use the same 10 corpora for EN–X as for the original parallel ET–X data, except that MultiParaCrawl is replaced by ParaCrawl (Bañón et al., 2020) in this case. The total sizes of our parallel and additional training corpora are shown in Table 2.

Using the original ET–X parallel data and the two varieties of additional data, we obtain 4 different training datasets:

1. only original parallel ET–X data,
2. parallel ET–X data mixed with SynEst back-translation data,
3. parallel ET–X data mixed with EN–X English-centric data,
4. parallel ET–X data mixed with both SynEst and EN–X data.

In augmentation scenarios 2 and 3, we mix the original parallel and additional data 1:7, always using all available original data. When using all types of data (scenario 4), we mix the original parallel, SynEst, and EN–X data 1:7:7. For information on preliminary experiments with other original to additional data proportions, see Appendix E.

4.2. Models

As the base model, we use the multilingual NLLB-1.3B dense model (NLLB Team et al., 2022). We freeze the parameters of the original model's encoder, retaining the multilinguality of the model on the source side. While focusing primarily on one input language, this allows us to not lose, and, in some cases, improve translation quality from other source languages, while also reducing the training-time costs. This also contributes to improved inference-time efficiency, as the same encoder is reused for multiple language pairs, and the models can be built in a modular fashion (Lyu et al., 2020; Escolano et al., 2021).

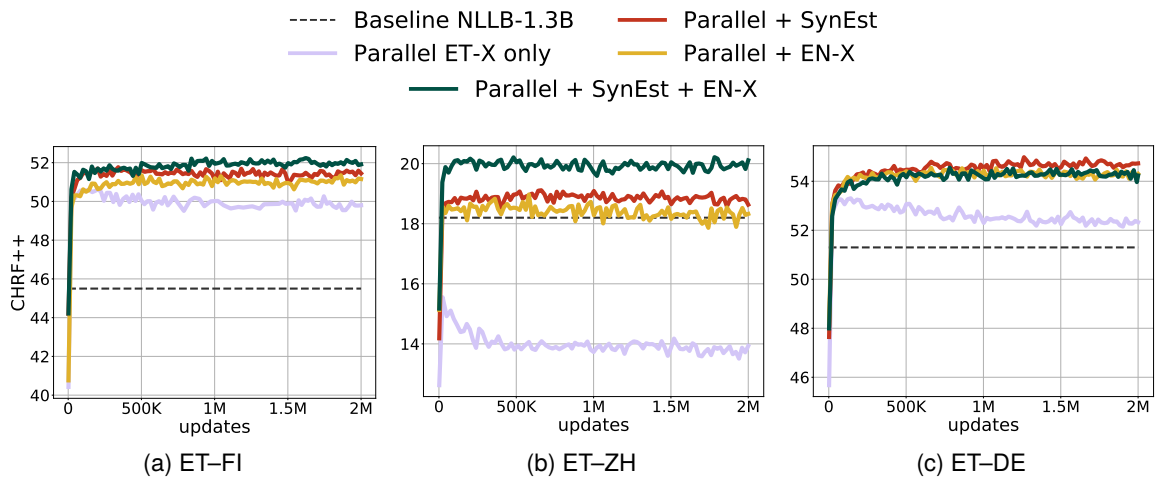
For each target language, we train a new randomly initialized decoder with 6 transformer layers of the same dimensions as in the original NLLB-1.3B. (For details on the decoder size choice, see Appendix D.) This allows us to train specialized decoders for each target language while making them more lightweight and reducing the training and inference costs. We also reduce the target vocabulary size to 32k (from 256k in NLLB). We use FairSeq (Ott et al., 2019) to train the models. For further model training details, see Appendix B.

4.3. Hyperparameter Search

We perform grid search for data mixing proportions and decoder size.

We experiment with 2:1, 1:1, 1:3, and 1:7 parallel to augmentation data proportions for SynEst and EN–X data, and find 1:7 to be the best performing on average. Details of this experiment are shown in Appendix E.

To choose the number of decoder layers, we train models on parallel ET–FI data for 200k updates. The model fails to train with 18- and 24-layer decoders due to the amount of training data being insufficient to match the large number of parameters; 3, 6, and 12 layers show results comparable to each other, with 6 slightly outperforming the others; having 1 layer leads to noticeably worse performance. See Figure 3 in Appendix D for more detailed results of this experiment.



(a) ET–FI (b) ET–ZH (c) ET–DE

Figure 1: chrF++ curves of models trained only using ET-X parallel data (purple), with added EN–X data (yellow), with added synthetic data (red), and with both types of data added (dark green) on FLORES-dev. Dashed lines (black) show original NLLB-1.3B scores. Parallel ET–X to additional data proportions are 1:7 (1:7:7 when both synthetic and EN–X data are added).

4.4. Evaluation

We evaluate the performance of our models using the FLORES benchmark dataset (Goyal et al., 2022). The dev split of FLORES is also used as the development set during training.

Following NLLB Team et al. (2022) and the recommendations of Kocmi et al. (2021), we report both BLEU (Papineni et al., 2002) and chrF++ (Popović, 2017) scores for evaluation. Specifically, we use the sacreBLEU implementation (Post, 2018) for both metrics.[4,5]

5. Results

For each translation direction, we compare four models: one trained using only ET–X parallel data, the second augmented with SynEst as back-translation data, the third augmented with EN–X parallel data and, finally, a model trained using parallel and both types of augmentation data. The original NLLB-1.3B model serves as the baseline. Table 3 shows each scenario's BLEU and chrF++ scores on the devtest split of the FLORES benchmark dataset for the three translation directions. To calculate the scores of our models, we use the checkpoint with the best score on FLORES-dev for each model.

For all three translation directions, both augmentation strategies prove useful. The performance when adding synthetic and English-centric data is similar, with models trained with added synthetic data showing slightly better scores. This is also evident in Figure 1, which shows the chrF++ scores of our models on FLORES-dev as the training progresses. Models trained with added synthetic data (shown in red) consistently show better results and less stagnation in the later stages of training than those trained with added EN–X data (yellow). This confirms that both multilingual and synthetic data can be used to support translation in directions with limited parallel resources, and shows the practical usefulness of the introduced synthetic corpus. Augmenting with both kinds of data at the same time improves the results slightly for ET–FI and more noticeably for ET–ZH, while for ET–DE using both

	ET–FI	ET–ZH	ET–DE
NLLB-1.3B	15.5/45.8	25.0/18.7	24.4/51.0
Parallel	19.9/50.5	24.8/16.4	25.7/52.6
+ SynEst	20.8/51.3	30.2/19.6	26.6/53.7
+ EN–X	20.8/51.0	29.9/19.3	26.4/53.6
+ Both	20.9/51.6	31.8/20.7	26.4/53.6
GPT-4	20.8/51.8	35.9/24.1	28.2/54.8

Table 3: BLEU/chrF++ scores on FLORES-devtest. "Parallel" indicates models trained using only ET–X parallel data, "+ SynEst" indicates models trained on parallel and SynEst back-translation data, "+ EN–X" on parallel and English-centric data, and "+ Both" on parallel and both back-translation and English-centric data. We choose the checkpoint with the highest dev score and report its score on devtest.

[4]sacreBLEU signature for chrF++: `nrefs:1|case:mixed|eff:yes|nc:6|nw:2|space:no|version:2.3.1`

[5]sacreBLEU signutare for BLEU: `nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1` (FI & DE), `nrefs:1|case:mixed|eff:no|tok:zh|smooth:exp|version:2.3.1` (ZH)

	FR–FI (ET–FI model)	RU–DE (ET–DE model)
NLLB-1.3B	16.9/46.5	24.8/51.5
Parallel ET–X	18.9/49.2	24.1/51.8

Table 4: BLEU/chrF++ scores on FLORES-dev (for our models, we report the score of the best checkpoint). Note that the FR–FI dataset was translated with the model trained on ET–FI data, and RU-DE with the model trained on ET–DE data, which makes the shown translation directions zero-shot.

types of augmentation data does not yield a better result than adding only SynEst data.

While the scores of the models augmented with SynEst synthetic data and with English-centric data are mostly very close, for synthetic data the result is reached with fewer unique sentence pairs in the training set. The original and augmentation data are always mixed 1:7, using all of the original data once and under- or oversampling augmentation data as needed. While the augmentation data for ET–DE is always undersampled, for ET–FI and ET–ZH the SynEst data is oversampled more times than the EN–X data, since EN–X is more abundant. A SynEst sentence pair occurs, on average, around 4.5 times in the ET–FI training corpus, while each EN–FI sentence pair occurs only 1.3 times. For ET–ZH, the figure is 3.9 for SynEst, while the EN–ZH corpus is undersampled and thus its sentence pairs are not repeated at all. This shows that unique SynEst data is likely more valuable for translation performance than English-centric data.

Although the models are trained with only one source language (or two, in the case when EN–X data is added), they also maintain or even improve NLLB's translation quality when translating from other languages, due to the encoder being frozen and the decoder being focused on a specific target language. Table 4 shows two examples of this, namely, the scores of the model trained on Estonian–Finnish parallel data translating from French into Finnish, and the Estonian–German model translating from Russian into German. While translation from Russian, which uses a different script than Estonian, has a lower BLEU but slightly higher chrF++ score than NLLB, translation from French into Finnish is improved compared to the baseline NLLB performance.

6. Discussion

While for ET–FI and ET–DE the non-augmented models trained only on ET–X parallel data easily outperform the multilingual NLLB-1.3B baseline, for ET–ZH that is not the case, with the new model trailing behind the baseline. The augmented models manage to beat the baseline, which suggests the need for further exploration in this direction.

For further context, we also apply GPT-4 (OpenAI, 2023) to translating FLORES-devtest. The evaluation setup is described in Appendix F. While GPT-4 outperforms our models on translation into German and Chinese, it shows a similar BLEU score for Finnish. This suggests that it may not be optimal for less represented languages, given that, to the best of our understanding, GPT-4 uses several orders of magnitude more parameters and training data than our models. The main issue, however, is the instability of its content moderation system: GPT-4 refused to translate around 1.7% of FLORES-devtest sentences, including ones of innocent nature, such as "Today, the only insects that cannot fold back their wings are dragon flies and mayflies" (reference English translation). This presents a significant challenge for using closed models.

7. Conclusion

In this work, we present two main contributions. First, we release SynEst, a large synthetic corpus comprised of translations from 11 languages into Estonian and totaling over 1 billion parallel sentences. Second, we perform experiments with this corpus, training translation systems based on the NLLB multilingual model for three language pairs, with a focus on translation directions which do not involve English and for which limited resources are available. Our models retain NLLB's multilinguality on the source side while improving translation quality for the translation directions of interest. We compare three data augmentation methods, namely, using our novel synthetic data, using English-centric parallel data, and a combination of the two. We demonstrate the usefulness of our corpus for training out-of-Estonian MT systems.

8. Bibliographical References

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos

Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Mikko Aulamo, 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. Association for Computational Linguistics.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

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

Daxiang Dong, 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. Association for Computational Linguistics.

Carlos Escolano, 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. Association for Computational Linguistics.

Angela Fan, 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.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732.

Melvin Johnson, 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.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online.

Diederik Kingma 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.

Tom Kocmi, 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). Association for Computational Linguistics.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Taku Kudo 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. Association for Computational Linguistics.

Sungwon Lyu, 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. 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.

OpenAI. 2023. Gpt-4 technical report.

Myle Ott, 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. Association for Computational Linguistics.

Kishore Papineni, 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. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post. 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. Association for Computational Linguistics.

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, Berlin, Germany. Association for Computational Linguistics.

Andre Tättar, 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.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study.

9. Language Resource References

Marta Bañón, 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. Association for Computational Linguistics.

Naman Goyal, 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.

Haddow, Barry and others. 2022. *NewsCrawl*. PID https://data.statmt.org/news-crawl/.

Tom Kocmi, Martin Popel, and Ondrej Bojar. 2020. Announcing czeng 2.0 parallel corpus with over 2 gigawords.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

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

Pierre Lison 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. European Language Resources Association (ELRA).

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2023. Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages.

Nils Reimers 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.

Holger Schwenk, 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. Association for Computational Linguistics.

Holger Schwenk, 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. Association for Computational Linguistics.

Haukur Barri Símonarson, Vésteinn Snæbjarnarson, Pétur Orri Ragnarson, Haukur Jónsson, and Vilhjalmur Thorsteinsson. 2021. Miðeind's WMT 2021 submission. In *Proceedings of the Sixth Conference on Machine Translation*, pages 136–139, Online. Association for Computational Linguistics.

Ralf Steinberger, 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. European Language Resources Association (ELRA).

Jörg Tiedemann. 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. European Language Resources Association (ELRA).

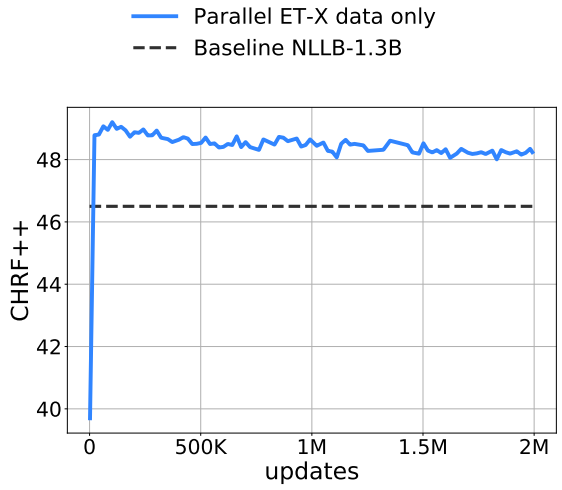
A. Pre-processing

The synthetic dataset was first 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 language and corpus.

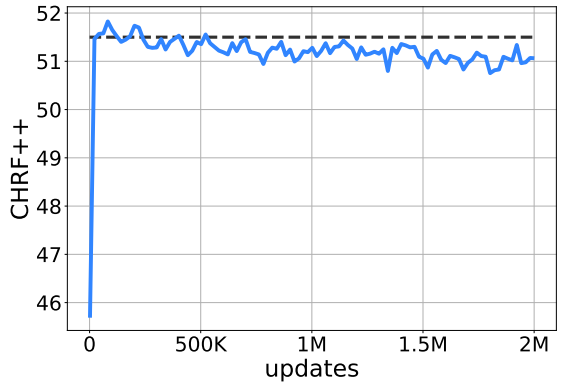
Both synthetic and parallel data are normalized with MTee normalization script (Tättar et al., 2022) and filtered with OpusFilter (Aulamo et al., 2020). The OpusFilter configuration is a modified version of filters used in MTee. The following filters are used:

1. `LongWordFilter`: filter examples with words longer than 40 characters (default).

2. `LengthFilter`: filter examples longer than 1000 characters or shorter than 10 characters.



(a) FR–FI translated with ET–FI model



(b) RU–DE translated with ET–DE model

Figure 2: Flores-dev chrF++ curves; models translate from source languages unseen during decoder training. Freezing the encoder allows us to retain its multilingual properties and possibly improve translation quality from other source languages as well.

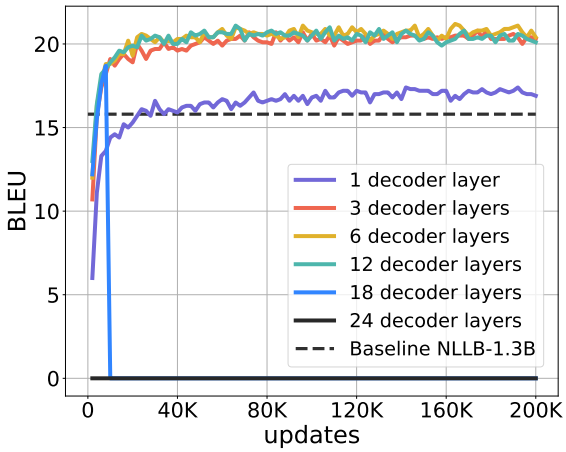


Figure 3: Flores-dev BLEU curves of models with different number of decoder layers trained on ET–FI parallel data for 200k updates. Best viewed in color.

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 the language-pairs with the following exceptions:

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

- 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.

Furthermore, duplicates and test set overlaps are removed from the training dataset.

B. Training

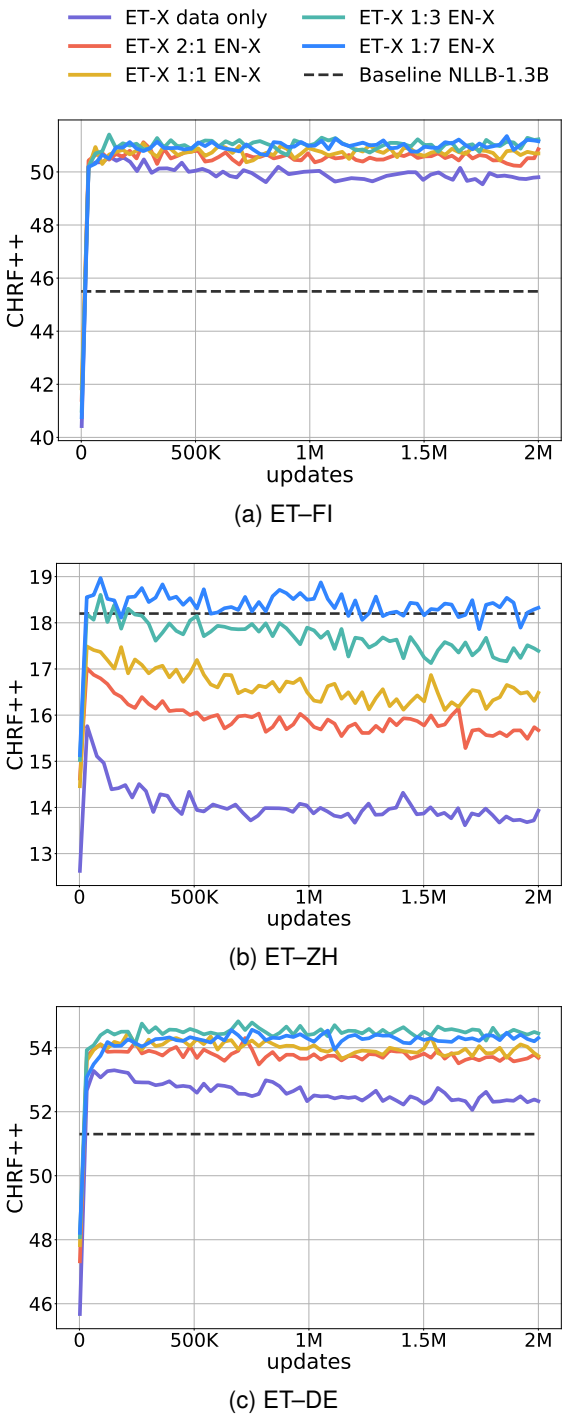
The models in the main experiments use the NLLB-1.3B encoder, which has 24 transformer layers with embedding dimension 1024, feed-forward dimension 8192, and 16 attention heads. The decoders are randomly initialized and have 6 transformer layers with the same dimensions as the encoder. The input and output embeddings of the decoder are shared. The vocabulary size is 256,000 for the encoder and 32,000 for the decoder. Model size is approximately 950M parameters; only 184M of these parameters are trained, the rest are not updated and are re-used in all our models. The original NLLB SentencePiece (Kudo and Richardson, 2018) model and vocabulary are used for the encoder, and a new model is trained for the decoder for each target language. The new subword models are trained using the non-augmented parallel data for each translation direction, and are re-used for all models for that translation direction.

We use FairSeq to train the models (Ott et al., 2019). All models are trained on 8 GPUs (4 AMD MI250x 128GB GPU modules, each acting as 2 GPUs). The batch size is 4,096 tokens per GPU. FP16 floating-point format is used. All models are trained for 2,000,000 updates.

The initial learning rate is 1×10^{-7}, with inverse square root learning rate scheduler with 4,000 warm-up updates to a maximum learning rate of 5×10^{-4}. We use Adam optimizer (Kingma and Ba, 2015) with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Dropout probability is set to 0.1, attention dropout to 0.1, and activation dropout is not used. The loss function is cross-entropy.

C. Retaining Source-Side Multilinguality

While we focus on Estonian as the main source language, freezing the parameters of the NLLB-1.3B encoder allows us to retain the model's multilinguality on the source side. Figure 2 shows how the chrF++ score on the dev split of the FLORES dataset progresses during training. In Figure 2a, we show scores on the French-Finnish translation direction when the model is trained using only Estonian-Finnish parallel data; similarly, Figure 2b demonstrates results of the model trained on Estonian-German parallel data when applied to Russian-German translation. While for RU–DE the ET–DE model starts lagging behind the baseline NLLB-1.3B after around 250,000 updates, the ET–FI model consistently outperforms the baseline on FR–FI translation.



(a) ET–FI

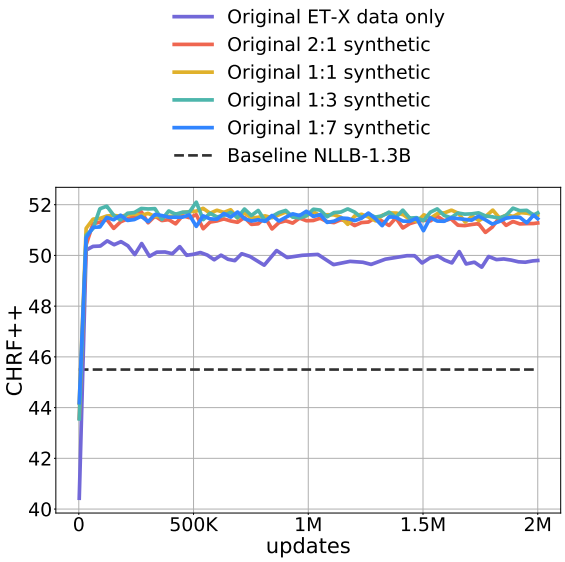
(b) ET–ZH

(c) ET–DE

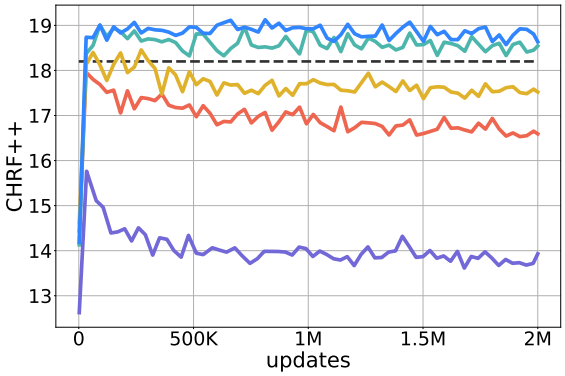
Figure 4: chrF++ curves (on FLORES dev) of models trained only using ET–X parallel data (purple), and models trained with EN–X parallel added in different proportions. Dashed horizontal lines show original NLLB-1.3B scores. Best viewed in color.

D. Decoder Size

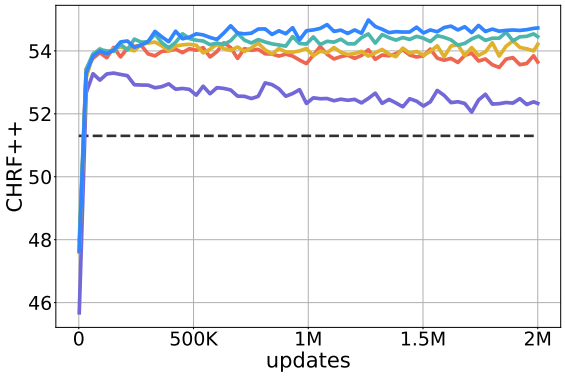
To explore possible choices of decoder size, we first use the parallel ET–FI data to train models with different number of transformer layers in the decoder (1, 3, 6, 12, 18, and 24 layers). The decoder



(a) ET–FI



(b) ET–ZH



(c) ET–DE

Figure 5: chrF++ curves (on FLORES dev) of models trained only using ET–X parallel data (purple), and models trained with synthetic data added in different proportions. Dashed horizontal lines show original NLLB-1.3B scores. Best viewed in color.

layers have the same dimensions as the encoder layers (embedding dimension 1024, feed-forward dimension 8192, 16 attention heads). The models are trained for 200,000 updates. Figure 3 shows the progress of BLEU score (Papineni et al., 2002) on the FLORES dev set during training. 18- and 24-layer decoders fail to train with the amount of parallel data available; the model with 1 decoder layer slightly outperforms the original NLLB-1.3B model, while 3-, 6-, and 12-layer decoders show comparable results. In subsequent experiments, we train models with 6 layers in decoder.

E. Data Proportions

Figures 4 and 5 show FLORES-dev chrF++ scores during training of models with different proportions of original and augmentation data (EN–X data in Figure 4, and synthetic data in Figure 5). We mix original and additional data 2:1, 1:1, 1:3, and 1:7, the latter being the main experiments described in Section 4.

F. GPT-4 Evaluation

To evaluate translation performance of GPT-4 (OpenAI, 2023), we follow Zhang et al. (2023) and choose a simple prompt template:

```
[src]: [input]
[tgt]:
```

`[src]` and `[tgt]` denote source and target language names, respectively, and `[input]` denotes the input test sentence. The translations were retrieved on 16 October 2023.

The main issue with using the GPT-4 API for translation is that some prompts trigger the content management policy, and no translation is provided at all. This moderation system seems to be unstable; the reference English translation of one of the FLORES-devtest sentences which it refused to translate is "Today, the only insects that cannot fold back their wings are dragon flies and mayflies." This presents a significant challenge for using closed models. Where a translation could not be generated, it was replaced with an empty line.

For translation into Chinese, GPT-4 noticeably outperforms all our models and NLLB-1.3B. At the same time, ET-ZH is also the weakest translation for our models in comparison with the original NLLB model. There is also some difference in favor of GPT-4 in translation into German, while for Finnish our augmented models and GPT-4 show the same BLEU score, with our models, to the best of our understanding, having orders of magnitude fewer parameters and training data examples. This suggests that, while GPT-4 shows high quality of translation when generating widespread languages, for less represented languages it might not be optimal. It also not possible to establish whether the model has encountered the FLORES test set before.