Cogs in a Machine, Doing What They're Meant to Do – The AMI Submission to the WMT24 General Translation Task

Atli Jasonarson, Hinrik Hafsteinsson, Bjarki Ármannsson, Steinþór Steingrímsson
The Árni Magnússon Institute for Icelandic Studies
Reykjavík, Iceland
atli.jasonarson,hinrik.hafsteinsson,bjarki.armannsson,
steinthor.steingrimsson@arnastofnun.is

Abstract

This paper presents the submission of the Árni Magnusson Institute's team to the WMT24 General translation task. We work on the English→Icelandic translation direction. Our system comprises four translation models and a grammar correction model. For training our models we carefully curate our datasets, aggressively filtering out sentence pairs that may detrimentally affect the quality of our system's output. Some of our data are collected from human translations and some are synthetically generated. A part of the synthetic data is generated using an LLM, and we find that it increases the translation capability of our system significantly.

1 Introduction

We describe our submission to the 2024 WMT general translation task. Large Language Models (LLMs) have become near-ubiquitous in the field of Natural Language Processing (NLP) in the last couple of years. They have shown remarkable translation capabilities (see e.g. Xu et al., 2024a), but require significantly larger computational resources than previous neural MT (NMT) models, both for training and inference. Most openly available LLMs are primarily trained on English texts and may therefore need further training in order to be able to translate from or into less-resourced languages, such as Icelandic.

The ALMA models (Xu et al., 2024a) are LLM-based translation models, built on LLaMA-2. They have been trained to translate ten directions, including English↔Icelandic. We explore the capabilities of some of these models, the 7B and 13B parameter versions of ALMA-R (Xu et al., 2024b), and find that they generate very competitive translations as measured against the English–Icelandic WMT21 test sets (Akhbardeh et al., 2021), especially from Icelandic into English. Unfortunately, using our settings the translation speed was quite slow (approximately one sentence per second) on an NVIDIA A100 GPU card.

We are interested in building faster models so we use the more traditional encoder-decoder Transformer architecture described in Vaswani et al. (2017). We collect all parallel data available to us for our language pair, generate additional synthetic pairs using the ALMA-R 13B parameter model and apply iterative back-translation using our own models. We apply filters to remove sentence pairs that may have detrimental effects on the models output.

We train four Transformer models[1] of varying sizes and let each model generate five translation candidates. A spelling and grammar checking model is then applied to the translations to generate "corrected" versions of the sentences. Finally the best candidate is selected from the pool of translations, corrected or not, using a reranking model.

We evaluate our models and approaches on the WMT21 test set for English→Icelandic.

2 Related Work

We only submit a system for the English→Icelandic translation direction. This language pair was previously one of the pairs for the WMT General Translation shared task in 2021 but prior to that, limited work had been published on MT for Icelandic. Brandt et al. (2011) describe a rule-based system for translating Icelandic→English, based on Apertium (Forcada et al., 2011). Jónsson et al. (2020) was the first published work describing SMT and NMT for Icelandic. Since 2021 the WMT21 evaluation data, as well as various parallel corpora projects, have made it more accessible to train and evaluate MT systems translating to or from Icelandic, and with that the language has been included in various research projects. We believe this is an indicator of the importance of evaluation campaigns, such

[1]Models available at https://huggingface.co/arnastofnun.

Proceedings of the Ninth Conference on Machine Translation, pages 253–262
November 15-16, 2024

as the ones run in association with the WMT conferences, for less prominent languages.

Our approach uses an ensemble of four different translation models and a reranking model to select the best candidate. This is a common approach, motivated by the intuition that different systems may have different strengths. In recent work, Toral et al. (2023) use this approach in their experiments with literary translations. In their work on bidirectional reranking, Imamura and Sumita (2017) discuss reranking and ensembling for MT in some detail. Examples from the period of statistical MT include the work of Olteanu et al. (2006) and Wang et al. (2007), describing language model-based reranking on hypotheses generated by phrase-based SMT systems.

3 Data Selection and Filtering

Various parallel data are available for the English–Icelandic language pair. ParIce (Barkarson and Steingrímsson, 2019) is partly a collection of parallel corpora available elsewhere, which has been realigned and refiltered, and partly data compiled for that project, the largest source being regulatory texts published in relation with the European Economic Area (EEA) agreement. Data for the English–Icelandic language pair were collected within the Paracrawl project (Bañón et al., 2020), CCMatrix (Schwenk et al., 2021), MaCoCu (Bañón et al., 2022) and HPLT (Aulamo et al., 2023). Data for the language pair are also available from multiple smaller datasets distributed on OPUS (Tiedemann and Thottingal, 2020). We utilize all these datasets in training our models.

We also use synthetic data: Backtranslations made available by Jónsson et al. (2022), translations generated using the ALMA-R 13B parameter model and backtranslations generated by our trained models. We describe these in more detail in Section 3.3.

Khayrallah and Koehn (2018) show that incorrect translations, untranslated target text, misalignments, and other noisy segments in training data can have a detrimental effect on the quality of translations generated by NMT systems trained on that data. By filtering our training data rather aggressively, we try to minimize such noise.

3.1 ParIce

Even though care has been taken to realign and refilter data for the ParIce corpus, Steingrímsson et al. (2023) show that it still contains noise, such as misalignments and mistranslations, that may be detrimental when training NMT systems. They refilter the data using a combination of approaches: Shallow filters based on simple heuristics, by using Bicleaner (Sánchez-Cartagena et al., 2018; Ramírez-Sánchez et al., 2020) and by employing classifiers (support vector machine-based ones (Cortes and Vapnik, 1995) had the best outcome) with a combination of scoring mechanisms, including LASER (Artetxe and Schwenk, 2019), LaBSE (Feng et al., 2022), NMTScore (Vamvas and Sennrich, 2022) using the M2M100 multilingual translation model (Fan et al., 2021), and WAScore, a word alignment-based score devised to measure word-level parallelism, introduced in Steingrímsson et al. (2021). In Steingrímsson (2023) these data are processed further by realigning the EEA texts in the ParIce corpus using SentAlign (Steingrímsson et al., 2023).

As the basis for our training we use the ParIce dataset, processed as described above, as well as parallel data extracted from Wikipedia using the comparable corpora mining approach described in (Steingrímsson et al., 2021) and sentence pairs extracted from version 9 of Paracrawl using the filtering approaches described above and in Steingrímsson et al. (2023).

3.2 Filtering the OPUS Datasets

An overview of the data for Icelandic-English parallel texts sourced from the OPUS catalog is provided in Appendix A. This data, accounting for redundant sentence pairs, amounts to 21.167.708[2] sentence pairs. At face value, this is a substantial amount of available data. However, the quality of these parallel texts is not reliable, with noisy and incorrect pairs being prevalent throughout most individual datasets in the catalog. To remedy this, and thus ensure that the data sourced via OPUS can be used effectively in our project, we applied an aggressive, sequential filtering process, with the goal of whittling away the majority of the low-quality sentence pairs.

Our sequential filtering process consists of ten individual steps, most of which only remove sentences from the data without modifying the content of other sentences. The process is *sequential*, in that the input of a filtering step is the output of the previous filtering step. Furthermore, the order of

[2]This applies to the state of the OPUS catalog at the time of development, i.e., April 2024.

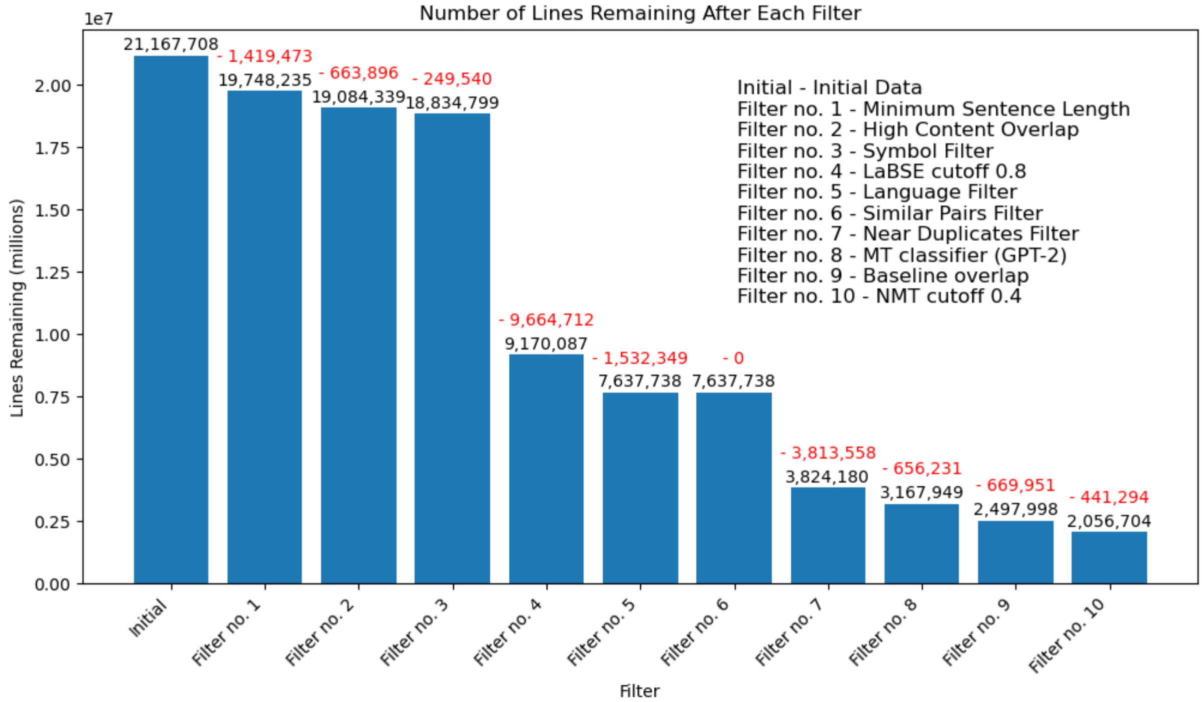

Figure 1: Each filtering step's effect on OPUS dataset size

these steps is decided to ensure optimal processing time of the filters so that computationally heavy filtering steps process the least amount of data, which minimizes run time. For a detailed overview of each filtering step, see Appendix B.

The effects of each filtering step on the data amount is shown in Fig. 1. To ensure that our filtering methods affected our implementation positively, we intermittently added the output of the filtering process to our training pipeline and evaluated the performance. In particular, we used this approach to dial in the optimal LaBSE and NMT score cutoffs in our filters.

The final output of our filtering process produces a relatively high-quality data set of 2.056.704 English-Icelandic sentence pairs (roughly 9.71% of the original 21.167.708 raw sentence pairs sourced from the OPUS catalog), which we then add to our training data.

3.3 Synthetic Data

The dataset made available by Jónsson et al. (2022) contains translations from Europarl, Newscrawl, Wikipedia and the IGC. We perform a filtering step similar to the one used applied on the OPUS data, consisting of a length filter, removing all sentences that have fewer than four word tokens and more than 150, an overlap filter, removing all sentence pairs that share 40% or more of word tokens, and a symbol filter removing all sentence pairs where more than 20% of characters in one of the sentences is non-alphabetical. Furthermore we use two scoring mechanisms for filtering, LaBSE, using a score threshold of 0.8, and NMTScore with a threshold of 0.4. These scores are selected based on the evaluation in (Steingrímsson et al., 2023). After filtering, we are left with 4.4M sentence pairs from this dataset.

We use the 13B parameter ALMA-R model to translate English sentences from Newscrawl to Icelandic and Icelandic texts from the Icelandic Gigaword Corpus (IGC) (Steingrímsson et al., 2018) to English. The Icelandic texts are sampled from three different subcorpora of the IGC, comprising news, scholarly journals, and literary texts. For each source sentence we generate five translations and use LaBSE to select the two best ones, granted that they exceed a threshold of a LaBSE score of 0.8 and pass through the three shallow filters described above: length, overlap and symbol filters. Our final set contains 8.9M sentence pairs translated from Icelandic to English and 700K sentence pairs translated from English to Icelandic.

Finally, we do iterative back-translation. We use the same training data as described above to train models to translate texts from the IGC to English. For the back-translations we use Transformer$_{\text{BIG}}$

model	d_{model}	d_{ff}	h	N_{enc}	N_{dec}
$Base$	512	2048	8	6	6
$Base_{deep}$	512	2048	8	36	12
Big	1024	4096	16	6	6
Big_{deep}	1024	4096	16	36	12

Table 1: Model dimensions, heads and number of layers.

models (Vaswani et al., 2017), as described in Table 1. We use the same approach as before, generate five translations for each sentence and use LaBSE to select the two best ones, as long as they exceed the threshold of 0.8 and are not filtered out by the other filters. We do two iterations of translating and training models in both translation directions using backtranslated data. This results in a total of approximately 60M sentence pairs.

3.4 Other Data

To decide which datasets to use, we trained Transformer$_{\mathrm{BASE}}$ models as described in Vaswani et al. (2017) and evaluated the models using the test set from WMT21. We started by training a baseline system using the dataset described in Section 3.1. We then added different datasets to the baseline data, trained new systems and evaluated them. If the new dataset seemed to improve the output we used that for our final system. In addition to previously described datasets we tried generating backtranslations using SMT and to add data from a bilingual lexicon using token-pair training as described by Jones et al. (2023). Table 2 shows chrF scores (Popović, 2015) for our different experiments.

Dataset	chrF
Baseline	50.4
Baseline+lexicon	50.4
Baseline+OPUS	53.7
Baseline+Jónsson	53.5
Baseline+Jónsson+SMT	53.2
Baseline+Jónsson+ALMA	54.7
Baseline+Jónsson+ALMA+OPUS	55.1
Baseline+Jónsson+ALMA+OPUS+BT1	56.4
Baseline+Jónsson+ALMA+OPUS+BT2	56.8

Table 2: The table shows that when most of the datasets in our experiments are added to the training data the quality, as measured by chrF, increases. Exceptions to that are the experiments with adding token-pairs from an English-Icelandic lexicon and with using backtranslations generated by an SMT system. These two datasets are therefore not used in our final systems.

Dataset	Sentence Pairs
Base	2,277,023
OPUS-filtered	2,056,704
Miðeind-BT	2,559,806
Miðeind-FT	1,837,945
ALMA-BT	8,927,720
ALMA-FT	700,253
IGC-BT-1	27,794,398
IGC-BT-2	33,465,175

Table 3: Datasets used for training and number of sentence pairs in each dataset.

The total number of sentence pairs used for training is shown in Table 3

4 System Description

Our motivation for using multiple models is twofold: First, we want to use models that are computationally inexpensive to run and so we train models that can run on one consumer grade GPU. Second, systems of different sizes may have complementary strengths and so training multiple systems and reranking the results may give us better results than any one model.

We train four encoder-decoder Transformer models, all of which play a part in the translation pipeline. Two of the models follow the exact architecture described in Vaswani et al. (2017), i.e. the ‘base’ and ‘big’ versions of the original Transformer model, while the other two are deeper, using 36 encoder layers and 12 decoder layers instead of six. The difference between the four models is shown in Table 1.

The outputs from the translation models undergo two post-processing steps. First, they are run through a grammatical error correction model, a version of the byte-level sequence-to-sequence model ByT5 (Xue et al., 2022) that has been finetuned by Ingólfsdóttir et al. (2023) to correct spelling errors in Icelandic as well as handling more complex grammatical, semantic and stylistic issues. Second, we fix punctuation errors which translation models are prone to making when translating into Icelandic (mostly to do with quotation marks, which are different in Icelandic and English) as well as some that might be unique to our system, such as their incapability to translate emojis. As the grammatical error correction model proved too aggressive for our purposes, merging and splitting

model	chrF
Base	56.8
$Base_{deep}$	57.1
Big	57.7
Big_{deep}	57.7
Ensemble+COMETKIWI	58.3
Ensemble+error correction +COMETKIWI	**58.4**
ALMA-R 7B	52.2
ALMA-R 13B	53.4

Table 4: chrF scores for each of our models, compared with scores for the model ensembles and for the ALMA-R models. The scores are calculated on the WMT21 evaluation set.

some sentences, normalizing informal language usage and hashtags, etc., we also revert some of the changes it introduced.

Using the WMT21 test set we experiment with an ensemble approach, using COMETKIWI-DA-22 (Rei et al., 2022) to select the best sentence out of 20 hypotheses made by the four models (each model generates five hypotheses using beam search with beam size 12). This raises the chrF score to 58.3 for our evaluation set. On top of this we add the spelling and grammar error correction, which gives us a very modest increase in quality as measured by chrF, shown in Table 4.

We investigate whether the COMETKIWI-DA-22 model prefers the output from some of the translation models over the others. Table 5 shows which translation models generated the translations ultimately chosen by the scoring model when experimenting on the WMT21 evaluation set of 1000 sentences. While translations by the deeper model are more likely to be selected, it is evident that all models are contributing, with the final selection containing 753 translation generated by only one model, and of these all models contribute over 150 translations each. 247 of the selected translations were generated by more than one model (non-unique translations). An ensemble approach thus seems to be likely to improve overall translation quality.

4.1 The pipeline

Basing our system on the most succesful approach in our experiments, our translation pipeline consists of three steps: First, using each of our four models, we generate five translation hypotheses using beam search for all source paragraphs, resulting in a total of 20 candidates.

model	Selected	Unique
Base	293	158
$Base_{deep}$	347	186
Big	287	163
Big_{deep}	419	246

Table 5: The number of sentences generated by each model selected for the final output when translating the WMT21 test set.

Furthermore, each paragraph is segmented into sentences, $s_1, \ldots, s_n$. For each sentence, every model produces five hypotheses. These hypotheses are evaluated using COMETKIWI-DA-22, and the highest-scoring hypothesis is selected for each sentence. The selected hypotheses are concatenated to form a new paragraph. Finally, a single paragraph is created by combining the best translation of each sentence, leaving us with 25 translation candidates.

Each of these candidates is then corrected with regard to grammar, spelling and style using the ByT5 model described above.

These two steps, translating the source text and correcting the translations, result in a total of 50 translation candidates. In order to find the best candidate we use COMETKIWI-DA-22 to score all candidates. The highest scoring one is the selected translation of our system.

5 Results

We evaluate our system on the test data from WMT21. As expected, the bigger models perform better, but the best results are achieved by selecting translations from an ensemble of differently trained Transformer models. We use COMETKIWI-DA-22 to select the best translation out of 20 hypotheses made by the four models, five hypotheses by each using beam search with beam size 12. This raises the chrF score to 58.3 and when we add error correction on top, the score is slightly higher, 58.4, as shown in Table 4.

In the WMT24 general translation task, systems were evaluated using two automatic metrics, MetricX-23-XL (Juraska et al., 2023) and COMETKIWI-DA-XL (Rei et al., 2023), as well as by human evaluation. According to the automatic metrics, reported in Kocmi et al. (2024), our model is competitive among the open systems, although four closed systems achieve better scores. Results

System Name	Type	AutoRank ↓	MetricX ↓	CometKiwi ↑
Unbabel-Tower70B	Closed	1.0	2.5	0.740
Claude-3.5	Closed	2.3	3.6	0.697
Dubformer	Closed	2.5	3.4	0.685
IKUN	Open	3.2	4.3	0.666
GPT-4	Closed	3.4	4.7	0.673
AMI	**Open**	**3.7**	**4.9**	**0.663**
IKUN-C	Constrained	3.7	4.9	0.657
TranssionMT	Closed	4.2	5.5	0.653
ONLINE-B	Closed	4.2	5.5	0.652
IOL-Research	Open	4.3	5.7	0.655
ONLINE-A	Closed	5.5	6.4	0.603
Llama3-70B	Open	6.7	8.0	0.586
ONLINE-G	Closed	6.9	7.9	0.573
CommandR-plus	Closed	9.8	10.6	0.487
Mistral-Large	Closed	10.4	10.9	0.465
Aya23	Open	15.2	14.9	0.311
Phi-3-Medium	Closed	16.2	15.7	0.278
ONLINE-W	Closed	18.1	19.5	0.296
TSU-HITs	Constrained	19.2	18.4	0.192
CycleL	Constrained	21.0	20.2	0.148

Table 6: Preliminary WMT24 General MT automatic ranking for English-Icelandic. Our system is in bold.

for the automatic metrics are shown in Table 6.

6 Conclusions and Future Work

We show that while Large Language Models have become nearly ubiquitous in Natural Language Processing, traditional encoder-decoder Transformer models remain a viable approach to machine translation, particularly when computational efficiency is a priority.

Nevertheless, our findings also reveal that integrating LLMs can be advantageous during the training process. Specifically, ALMA-R 13B proved to be an important part of our training pipeline, as the synthetic data it generated increased the quality of our translation systems.

Furthermore, our results indicate that while more training data usually result in a better translation system, low-quality data, such as the backtranslations generated with an SMT system, can have a detrimental impact on performance. Similarly, our experiments with a bilingual lexicon using token-pair training negatively affected the system's output. This may be due to a variety of reasons. Our SMT system could probably be improved as well as our approach to include data from a bilingual lexicon in the training data. This warrants further investigation.

Our filtering method, as described in Sections 3.2, 3.3 and Appendix B, has proven effective, even though it may be argued that it is still somewhat crude and more work into minimizing the loss of useful sentence pairs and more effectively remove detrimental sentence pairs would very likely improve the training data and in turn the translation models. For example, while we use LaBSE, LASER and NMT to evaluate sentence pairs, we apply individual cutoff values for each score. A better approach could entail using a classifier to combine all metrics for an optimal result.

Although currently impractical at production-scale, genetic algorithms, as shown by Jon and Bojar (2023) and Jon et al. (2023), show promising results in generating translation candidates. Given larger computational resources, similar approaches might prove useful and await future study.

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.

Mikel Artetxe and Holger Schwenk. 2019. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Mikko Aulamo, Nikolay Bogoychev, Shaoxiong Ji, Graeme Nail, Gema Ramírez-Sánchez, Jörg Tiedemann, Jelmer van der Linde, and Jaume Zaragoza. 2023. HPLT: High performance language technologies. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 517–518, Tampere, Finland. European Association for Machine Translation.

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.

Marta Bañón, Miquel Esplà-Gomis, Mikel L. Forcada, Cristian García-Romero, Taja Kuzman, Nikola Ljubešić, Rik van Noord, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Peter Rupnik, Vít Suchomel, Antonio Toral, Tobias van der Werff, and Jaume Zaragoza. 2022. MaCoCu: Massive collection and curation of monolingual and bilingual data: focus on under-resourced languages. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 303–304, Ghent, Belgium. European Association for Machine Translation.

Starkaður Barkarson and Steinþór Steingrímsson. 2019. Compiling and Filtering ParIce: An English-Icelandic Parallel Corpus. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 140–145, Turku, Finland. Linköping University Electronic Press.

Martha Dís Brandt, Hrafn Loftsson, Hlynur Sigurþórsson, and Francis M. Tyers. 2011. Apertium-IceNLP: A rule-based Icelandic to English machine translation system. In *Proceedings of the 15th Annual conference of the European Association for Machine Translation*, pages 217–224, Leuven, Belgium. European Association for Machine Translation.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond English-Centric Multilingual Machine Translation. *J. Mach. Learn. Res.*, 22(1).

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT Sentence Embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Mikel Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis Tyers. 2011. Apertium: A free/open-source platform for rule-based machine translation. *Machine Translation*, 25:127–144.

Kenji Imamura and Eiichiro Sumita. 2017. Ensemble and reranking: Using multiple models in the NICT-2 neural machine translation system at WAT2017. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, pages 127–134, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Svanhvít Lilja Ingólfsdóttir, Petur Ragnarsson, Haukur Jónsson, Haukur Simonarson, Vilhjalmur Thorsteinsson, and Vésteinn Snæbjarnarson. 2023. Byte-level grammatical error correction using synthetic and curated corpora. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7299–7316, Toronto, Canada. Association for Computational Linguistics.

Josef Jon and Ondřej Bojar. 2023. Breeding machine translations: Evolutionary approach to survive and thrive in the world of automated evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2191–2212, Toronto, Canada. Association for Computational Linguistics.

Josef Jon, Martin Popel, and Ondřej Bojar. 2023. CUNI at WMT23 general translation task: MT and a genetic algorithm. In *Proceedings of the Eighth Conference on Machine Translation*, pages 119–127, Singapore. Association for Computational Linguistics.

Alexander Jones, Isaac Caswell, Orhan Firat, and Ishank Saxena. 2023. GATITOS: Using a new multilingual lexicon for low-resource machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 371–405, Singapore. Association for Computational Linguistics.

Haukur Páll Jónsson, Haukur Barri Símonarson, Pétur Orri Ragnarsson, Svanhvít Lilja Ingólfsdóttir, Vilhjálmur Þorsteinsson, and Vésteinn Snæbjarnarson. 2022. Long context synthetic translation pairs for english and icelandic (22.09). CLARIN-IS.

Haukur Páll Jónsson, Haukur Barri Símonarson, Vésteinn Snæbjarnarson, Steinþór Steingrímsson, and Hrafn Loftsson. 2020. Experimenting with Different Machine Translation Models in Medium-Resource Settings. In *Text, Speech, and Dialogue - 23rd International Conference, TSD 2020, Brno, Czech Republic, September 8-11, 2020, Proceedings*, volume 12284 of *Lecture Notes in Computer Science*, pages 95–103. Springer.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767, Singapore. Association for Computational Linguistics.

Huda Khayrallah and Philipp Koehn. 2018. On the Impact of Various Types of Noise on Neural Machine Translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 74–83, Melbourne, Australia. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondrej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popovic, Mariya Shmatova, Steinþór Steingrímsson, and Vilém Zouhar. 2024. Preliminary WMT24 Ranking of General MT Systems and LLMs. *ArXiv*, abs/2407.19884.

Shuyo Nakatani. 2010. Language detection library for java.

Marian Olteanu, Pasin Suriyentrakorn, and Dan Moldovan. 2006. Language models and reranking for machine translation. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 150–153, New York City. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Dario Amodei, Daniela Amodei, Jack Clark, Miles Brundage, and Ilya Sutskever. 2019. Better language models and their implications. *OpenAI blog*, 1(2).

Gema Ramírez-Sánchez, Jaume Zaragoza-Bernabeu, Marta Bañón, and Sergio Ortiz Rojas. 2020. Bifixer and Bicleaner: two open-source tools to clean your parallel data. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 291–298, Lisboa, Portugal. European Association for Machine Translation.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André Martins. 2023. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848, Singapore. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Víctor M. Sánchez-Cartagena, Marta Bañón, Sergio Ortiz-Rojas, and Gema Ramírez. 2018. Prompsit's submission to WMT 2018 Parallel Corpus Filtering shared task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 955–962, Belgium, Brussels. Association for Computational Linguistics.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021. 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.

Peter M. Stahl. 2024. Lingua - an accurate natural language detection library for short and mixed-language text. `https://github.com/pemistahl/lingua-py`. Accessed: 2024-08-21.

Steinþór Steingrímsson, Sigrún Helgadóttir, Eiríkur Rögnvaldsson, Starkaður Barkarson, and Jón Guðnason. 2018. Risamálheild: A Very Large Icelandic Text Corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, LREC 2018, pages 4361–4366, Miyazaki, Japan.

Steinþór Steingrímsson, Hrafn Loftsson, and Andy Way. 2023. Filtering matters: Experiments in filtering training sets for machine translation. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 588–600, Tórshavn, Faroe Islands. University of Tartu Library.

Steinþór Steingrímsson, Pintu Lohar, Hrafn Loftsson, and Andy Way. 2021. Effective Bitext Extraction From Comparable Corpora Using a Combination of Three Different Approaches. In *Proceedings of the 14th Workshop on Building and Using Comparable Corpora (BUCC 2021)*, pages 8–17, Online (Virtual Mode). INCOMA Ltd.

Steinþór Steingrímsson. 2023. *Effectively compiling parallel corpora for machine translation in resource-scarce conditions*. Ph.D. thesis, Reykjavik University.

Steinþór Steingrímsson, Hrafn Loftsson, and Andy Way. 2023. SentAlign: Accurate and scalable sentence alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 256–263, Singapore. Association for Computational Linguistics.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Antonio Toral, Andreas Cranenburgh, and Tia Nutters. 2023. Literary-adapted machine translation in a well-resourced language pair. In Andrew Rothwell, Andy Way, and Roy Youdale, editors, *Computer-Assisted Literary Translation*, pages 27–52. Routledge.

Jannis Vamvas and Rico Sennrich. 2022. NMTScore: A multilingual analysis of translation-based text similarity measures. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 198–213, Abu Dhabi, United Arab Emirates.

Ashish Vaswani, 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 30 (NIPS 2017)*, pages 5999–6009, Long Beach, California.

Wen Wang, Andreas Stolcke, and Jing Zheng. 2007. Reranking machine translation hypotheses with structured and web-based language models. In *2007 IEEE Workshop on Automatic Speech Recognition Understanding (ASRU)*, pages 159–164.

Titus Wormer. 2024. Franc - a natural language detection library. `https://github.com/wooorm/franc`. Accessed: 2024-08-21.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024a. A Paradigm Shift in Machine Translation: Boosting Translation Performance of Large Language Models. In *The Twelfth International Conference on Learning Representations*.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024b. Contrastive Preference Optimization: Pushing the Boundaries of LLM Performance in Machine Translation. *ArXiv*, abs/2401.08417.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

A OPUS Texts

The parallel texts we sourced from the OPUS catalog are listed in this section. The format of the list is as follows:

Index. **Name**; *version*; sentence pairs

For brevity, the *ELRC* parallel text names are abbreviated after the first entry in the list, with the *ditto* symbol ('"') replacing the 'ELRC' part of the name.

1. **CCAligned**; *v1*; 1,192,542
2. **CCMatrix**; *v1*; 8,723,145
3. **ECDC**; *v2016-03-16*; 2,512
4. **ELRC-2718-EMEA**; *v1*; 542,624
5. **"-3206-antibiotic**; *v1*; 816
6. **"-4295-www.malfong.is**; *v1*; 12,634
7. **"-4324-Government_Offices_I**; *v1*; 18,185
8. **"-4327-Government_Offices_I**; *v1*; 36,290
9. **"-4334-Rkiskaup_2020**; *v1*; 10,236
10. **"-4338-University_Iceland**; *v1*; 10,164
11. **"-502-Icelandic_Financial_**; *v1*; 1,525
12. **"-504-www.iceida.is**; *v1*; 1,055
13. **"-505-www.pfs.is**; *v1*; 2,866
14. **"-506-www.lanamal.is**; *v1*; 1,140
15. **"-5067-SciPar**; *v1*; 110,831
16. **"-508-Tilde_Statistics_Ice**; *v1*; 2,427
17. **"-509-Gallery_Iceland**; *v1*; 577
18. **"-510-Harpa_Reykjavik_Conc**; *v1*; 1,197
19. **"-511-bokmenntaborgin_is**; *v1*; 330
20. **"-516-Icelandic_Medicines**; *v1*; 711
21. **"-517-Icelandic_Directorat**; *v1*; 1,536
22. **"-597-www.nordisketax.net**; *v1*; 1,065
23. **"-718-Statistics_Iceland**; *v1*; 2,361
24. **"-728-www.norden.org**; *v1*; 41,073
25. **"-EMEA**; *v1*; 542,624
26. **"-antibiotic**; *v1*; 816
27. **"-www.norden.org**; *v1*; 41,073
28. **"-www.nordisketax.net**; *v1*; 1,065
29. **EUbookshop**; *v2*; 9,783
30. **GNOME**; *v1*; 28,776
31. **HPLT**; *v1*; 2,148,876
32. **KDE4**; *v2*; 98,989
33. **MaCoCu**; *v2*; 267,366

34. **MultiCCAligned**; *v1*;	1,192,537
35. **MultiHPLT**; *v1*;	2,148,855
36. **MultiMaCoCu**; *v2*;	267,366
37. **MultiParaCrawl**; *v7.1*;	2,392,423
38. **NLLB**; *v1*;	8,723,145
39. **OpenSubtitles**; *v1*;	7,138
40. **OpenSubtitles**; *v2016*;	1,359,224
41. **OpenSubtitles**; *v2018*;	1,569,189
42. **ParIce**; *v1*;	2,097,022
43. **ParaCrawl**; *v7.1*;	2,392,422
44. **ParaCrawl**; *v8*;	5,724,373
45. **ParaCrawl**; *v9*;	2,967,579
46. **QED**; *v2.0a*;	27,611
47. **TED2020**; *v1*;	2,430
48. **Tatoeba**; *v2*;	8,139
49. **Tatoeba**; *v20190709*;	9,436
50. **Tatoeba**; *v2020-05-31*;	9,438
51. **Tatoeba**; *v2020-11-09*;	9,440
52. **Tatoeba**; *v2021-03-10*;	9,443
53. **Tatoeba**; *v2021-07-22*;	9,443
54. **Tatoeba**; *v2022-03-03*;	9,522
55. **Tatoeba**; *v2023-04-12*;	9,600
56. **TildeMODEL**; *v2018*;	420,712
57. **Ubuntu**; *v14.10*;	2,155
58. **WikiMatrix**; *v1*;	85,992
59. **WikiTitles**; *v3*;	50,176
60. **XLEnt**; *v1*;	962,661
61. **XLEnt**; *v1.1*;	962,661
62. **XLEnt**; *v1.2*;	962,661
63. **bible-uedin**; *v1*;	62,163
64. **wikimedia**; *v20190628*;	581
65. **wikimedia**; *v20210402*;	2,625
66. **wikimedia**; *v20230407*;	4,471

B Filtering steps

Filter 1. Sentence length
Sentences should contain at minimum four characters and at maximum 150 characters.

Filter 2. High inter-pair content overlap
Sentence pairs where the content of the source and target sentences are highly similar should be removed from the dataset.

Filter 3. Character symbol filtering
All characters in the English and Icelandic alphabets (along with punctuation and numbers) designated as a set of allowed characters. Sentences containing less than 60% of these characters removed from the data and all characters outside the allowed set removed from the remaining sentences.[3]

Filter 4. LaBSE scoring
We use score each sentence pair using LaBSE (Feng et al., 2022) and remove all sentences with a score lower than 0.8[4].

Filter 5. Language detection
We use various language detection software to gauge whether both the source and target sentences are in the correct language. The software we used was *fasttext* (Joulin et al., 2016), *franc* (Wormer, 2024), *lingua* (Stahl, 2024) and *langdetect* (Nakatani, 2010).

Filter 6. Similar dataset pairs
As a safeguard, we remove any duplicate entries of our dataset if, for any reason, there remain duplicate instances after the previous filters. In our final experiment, this was rendered redundant, but was required in previous iterations and may prove useful in future iterations.

Filter 7. Near-duplicate dataset pairs
Sentences are compared by removing content-specific words that are likely proper names and dates, etc., and comparing the remainder.

Filter 8. Likely machine-translated target sentences
A GPT-2 (Radford et al., 2019) classifier is used to evaluate whether a given target sentence is machine-translated, based on a 10.000 sentence hand-evaluated reference set. If this is true for the target sentence, that pair is removed from the dataset.

Filter 9. Existing datasets
As a final safeguard check, we remove any sentence pair that we already have on file in other datasets, as touched on in section 3.2.

Filter 10. NMTScore cross-likelyhood 0.4
Finally, we use a translation cross-likelyhood NMTScore (Vamvas and Sennrich, 2022) to determine the translation quality of a given sentence pair. This step is computationally heavy and was therefore saved for last. Our experiments suggest that 0.4 is a suitable cutoff for our dataset.

[3]This is the last filtering step that inherently modifies the content inside individual sentences.

[4]This is a higher cutoff than the original LaBSE authors suggest to use, but our experiments suggets it better suits our data.