Mind the Gap: Diverse NMT Models for Resource-Constrained Environments

Ona de Gibert[1]* Dayyán O'Brien[2] Dušan Variš[3] Jörg Tiedemann[1]
[1]University of Helsinki [2]University of Edinburgh [3]Charles University
*Corresponding author: ona.degibert@helsinki.fi

Abstract

We present fast Neural Machine Translation models for 17 diverse languages, developed using Sequence-level Knowledge Distillation. Our selected languages span multiple language families and scripts, including low-resource languages. The distilled models achieve comparable performance while being 10x times faster than transformer-base and 35x times faster than transformer-big architectures. Our experiments reveal that teacher model quality and capacity strongly influence the distillation success, as well as the language script. We also explore the effectiveness of multilingual students. We release publicly our code and models in our Github repository: https://github.com/hplt-project/bitextor-mt-models.

1 Introduction

Neural Machine Translation (NMT) has seen significant advancements with the advent of Large Language Models (LLMs; Zhu et al., 2024). Although LLMs often perform exceptionally well on high-resource languages, their performance on low-resource languages lags behind (Stap and Araabi, 2023; Kocmi et al., 2023; Robinson et al., 2023). Nevertheless, recent advancements suggest that this gap may be narrowing (Enis and Hopkins, 2024).

Despite their high quality performance, LLMs come with substantial computational costs, requiring significant amount of traning data, high-end hardware and extensive energy consumption (Rae et al., 2021). These limitations make LLMs unsuitable for many real-world scenarios where resources are constrained, such as on-device translation, low-latency requirements, or environments with privacy concerns.

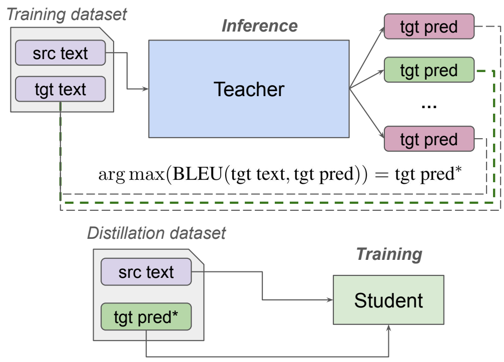

Figure 1: Conceptual overview of interpolated Sequence-Level Knowledge Distillation.

The traditional sequence-to-sequence (seq2seq) Transformer architecture (Vaswani et al., 2017), though not as versatile as LLMs, offers considerable advantages in terms of computational efficiency. These models can be optimized to run faster, consume less memory, and require fewer resources, making them a practical solution for many NMT applications (Kim et al., 2019; Aji and Heafield, 2020).

In this work, we leverage Knowledge Distillation (KD) (Hinton et al., 2015; Kim and Rush, 2016) to train compact seq2seq NMT models. KD allows the transfer of knowledge from a large, high-performing *teacher* model to a smaller, more efficient *student* model.

We present fast NMT models for 17 diverse languages with English as the target language. The selected languages vary widely in terms of script, language family, and resource availability, including low-resource languages like North Azerbaijani and high-resource languages like Hindi.

In our experiments, we address the following Research Questions (RQ): *RQ1: How does the capacity gap affect the distillation quality?*, *RQ2: To what extent does script influence the transfer of knowledge?* and *RQ3: Can we train multilingual students effectively?*.

Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025), pages 209–216
March 3-4, 2025

2 Related Work

We use Sequence-level KD (Seq-KD, Kim and Rush, 2016), which has which has proven to be effective to do KD for NMT (Gumma et al., 2023; Team et al., 2024). In Seq-KD, the teacher model is used to forward-translate all the sentences in the training data to create a distilled dataset. In the interpolated Seq-KD variant, the teacher generates K-candidate translations, selecting the one with the highest smoothed sentence BLEU (Chen and Cherry, 2014) with the reference. Then, the student model is trained on the synthetically generated data. Figure 1 illustrates this procedure. In this way, the lightweight student retains much of the teacher's performance while being optimized for speed and efficiency.

Several studies explore how to build compact NMT models. With the motivation of testing the time-efficiency of NMT systems, a shared task on NMT efficiency was organized for several years within the Workshop on Neural Generation and Translation (Hayashi et al., 2019; Heafield et al., 2020, 2021). Research has focused on various aspects, including compressing multilingual systems (Tan et al., 2018), investigating different architectures for student models (Bogoychev et al., 2020), and understanding the effectiveness of KD (Zhou et al., 2020). One widely adopted approach is the thin and deep architecture (Gala et al., 2023; Gumma et al., 2023), characterized by a deep encoder and a shallow decoder (Mohammadshahi et al., 2022; Kasai et al., 2020), which has become a standard for compressing NMT models. We follow that approach in this work.

3 Methodology

Next, we describe the selected languages, datasets, tools, and teacher and student architectures used for our experiments.

Languages The 17 selected languages are listed in Table 1. To highlight their diversity, we provide the language family (spanning 13 distinct families) and the script, representing seven different scripts: Arabic (Arab), Latin (Latn), Hebrew (Hebr), Devangari (Deva), Japanese (Jpan), Cyrillic (Cyrl), Hangul (Hang). We also include the taxonomy class proposed by Joshi et al. (2020) to classify languages according to their available resources. It ranges from 1 (resources for that language are limited) to 5 (rich-resource languages).

Language	Family	Class	Data (M)
Arabic (arb_Arab)	Semitic	5	10.44
Basque (eus_Latn)	Isolate	4	6.40
Catalan (cat_Latn)	Romance	4	29.23
Galician (glg_Latn)	Romance	3	7.78
Hebrew (heb_Hebr)	Semitic	3	28.90
Hindi (hin_Deva)	Indo-Iranian	4	13.62
Japanese (jpn_Jpan)	Japonic	5	15.81
Kazakh (kaz_Cyrl)	Turkic	3	21.28
Korean (kor_Hang)	Koreanic	4	7.56
Latvian (lvs_Latn)	Baltic	3	24.73
Lithuanian (lit_Latn)	Baltic	3	34.70
Slovak (slk_Latn)	Slavic	3	53.66
Swahili (swh_Latn)	Bantu	2	6.27
Malay (zsm_Latn)	Austronesian	3	42.65
N. Azerbaijani (azj_Latn)	Turkic	1	44.46
N. Uzbek (uzn_Latn)	Turkic	3	17.55
Vietnamese (vie_Latn)	Austro-Asiatic	4	2.83

Table 1: Overview of the selected languages, including their script, language family, class as defined by Joshi et al. (2020) and training data (in millions of sentences).

Datasets We use the Tatoeba Challenge dataset, a compilation of all datasets available in OPUS (Tiedemann et al., 2024), de-duplicated and shuffled. Other datasets include: MaCoCu (Bañón et al., 2022, 2023) for Catalan; CLUVI (Universidade de Vigo, 2012) for Galician; SAWA (De Pauw et al., 2009) and Gourmet (Sánchez-Martínez et al., 2020) for Swahili. We use a combination of OpusCleaner (Bogoychev et al., 2023) and OpusFilter (Aulamo et al., 2020) for cleaning the corpora. We list the clean training data sizes for each language pair in Table 1. For development and evaluation, we use Flores-200 (Goyal et al., 2022).

Tools We train our models with interpolated Seq-KD with three different tools: we follow recipes from the Bergamot project[1], the Firefox Translations training pipeline[2] and its extended multilingual version, OpusDistillery (de Gibert et al., 2025). All tools perform a forward translation of the training data to create the distilled dataset, generating an 8-best list of candidate translations, as illustrated in Figure 1. Using the distilled dataset, we train a new, shared 32k subword vocabulary with SentencePiece (Kudo and Richardson, 2018), alignments with fast_align (Dyer et al., 2013) and lexical shortlists for faster

[1] https://github.com/browsermt/students/tree/master/train-student

[2] https://github.com/mozilla/firefox-translations-training

decoding with extract_lex[3]. Then, we train the student with guided alignment using the Marian NMT toolkit (Junczys-Dowmunt et al., 2018). Finally, we quantize the student models using an 8-bit integer representation, which significantly reduces memory usage while maintaining translation quality.

OPUS-MT teacher models All teachers are OPUS-MT transformers (tf). We use one single teacher for each student model. Five teachers are tf-base (∼70M parameters) while the remaining are tf-big (∼209M params). We show the size of each teacher in Table 3. We train our own tf-big teachers for Galician and Swahili. For the other languages, we use the OPUS-MT dashboard (Tiedemann and De Gibert, 2023) to choose the best available teacher.

Tiny student models Our student models adopt the tiny architecture proposed by Bogoychev et al. (2020), consisting of a transformer encoder with 6 layers and a lightweight RNN-based decoder with the Simpler Simple Recurrent Unit (SSRU, Kim et al., 2019) with 2 layers. In a pilot study, we initially trained both small and tiny student models, with a detailed comparison of their architectures provided in Table 2. Results from this study showed that the translation quality loss in tiny models was minimal compared to the small models. Consequently, we opted to focus exclusively on the tiny models, which offer substantial inference speedups. After training, we quantize the model. **On average, the tiny architecture is 10x times faster than tf-base and 35x times faster than tf-big architectures.**

We train bilingual student models for all language pairs except for the Baltic and Turkic families, for which we train multilingual many-to-one students.

Evaluation We use COMET[4] (Rei et al., 2020) and spBLEU (Goyal et al., 2022) for evaluation. COMET is a neural metric that demonstrates the highest correlation with human judgments in translation quality assessment. It covers all tested languages. Additionally, we use SacreBleu (Post, 2018) to compute spBLEU, which refers to the BLEU (Papineni et al., 2002) metric on the tokenized text with SentencePiece.

	Teachers		Students	
	big	base	small	tiny
N_{enc}	6	6	6	6
N_{dec}	6	6	2	2
d_{emb}	1024	512	512	256
d_{ff}	4096	2048	2048	1536
h	16	8	8	8
Params (M)	213	65	39	17
Size (MB)	798	277	42	17
Speed (tok/s)	814.8	2758.5	18649.5	28854.7

Table 2: Comparison of tf architectures used for teachers (big, base) and students (small, tiny). The table lists the number of encoder and decoder layers (N_{enc} and N_{dec}), embedding dimensions (d_{emb}), feed-forward dimensions (d_{ff}), number of attention heads (h), parameters in millions, model size in MB, and decoding speed in tokens per second. Speed values are averaged across all models on 32 CPU cores.

4 Results

Tables 3 and 4 summarize the results of our distillation experiments in COMET scores for bilingual and multilingual settings, respectively. We report spBLEU scores in Tables 5 and 6 in the Appendix.

On average, the students exhibit a drop of 2.9 COMET points compared to their teachers. In general, we observe that our students maintain competitive performance, with high scores for several languages, including Catalan, Galician, Hebrew, Slovak, and Malay. These results indicate that, despite the reduction in model size and complexity, these students still capture a significant portion of the teacher's knowledge. However, for languages like Arabic, Korean and Japanese, the scores drop significantly. For Japanese, Table 5 reveals that the teacher model performs the worst among all selected languages, with a spBLEU score of 19.2. **This suggests that a low-performing teacher is not capable of knowledge transfer.** Therefore, we exclude Japanese from our analysis in the next section.

We expect that our students do not outperform their teachers, due to the capacity limitations of the students when compared to their larger teachers, known as the capacity gap problem (Jafari et al., 2021). However, our Catalan student achieves a COMET score 1.1 point higher than its teacher, correlating with a 90% human agreement that it outputs better translations (Kocmi et al., 2024).

[3]https://github.com/marian-nmt/extract-lex

[4]We use the model Unbabel/wmt22-comet-da.

Language		ara	cat	eus	glg	heb	hin	jpn	kor	slk	swh	vie	zsm
Teacher	Params (M)	76.4	69.4	235.4	209.1	238.1	75.9	77.5	209.2	235.5	209.1	63.9	237.1
	Performance	83.7	84.3	83.8	87.6	86.2	81.9	80.3	85.3	85.1	82.9	79.2	85.6
Student	Compression	4.5	4.1	13.9	12.4	14.1	4.5	4.6	12.4	13.9	12.4	3.8	14.0
	Performance	76.7	85.4	80.6	84.4	85.2	81.8	62.8	78.9	85.2	79.3	79.8	85.6
	Δ	-7.0	+1.1	-3.3	-3.2	-1.0	-0.1	-17.6	-6.4	+0.1	-3.6	+0.6	+0.1

Table 3: COMET score results of our bilingual distillation experiments. For the teacher models, we report parameters in millions and performance. We provide results for the students, as well as their compression ratio. Δ shows the difference in COMET scores with the teacher.

Family		Baltic		Turkic		
Language		lit	lvs	azn	kaz	uzj
Teacher	Params (M)	236.9	236.9	238.8	238.8	238.8
	Performance	83.5	84.0	82.0	81.7	81.7
Student	Compression	14.02	14.02	14.13	14.13	14.13
	Performance	82.7	83.7	80.2	78.5	78.9
	Δ	-0.8	-0.3	-1.8	-3.2	-2.7

Table 4: COMET score results of our multilingual distillation experiments.

We also find an improved score for Vietnamese, Slovak and Malay, though these improvements were less significant.

5 Discussion

In this section, we address the research questions (RQs) posed in the introduction based on the results of our distillation experiments.

RQ1: How does the capacity gap between the teacher and student models affect the distillation quality? The capacity gap between the teacher and student models is a critical factor in distillation quality. We find that larger teachers (tf-big) lead to a more significant performance drop, with an average COMET reduction of 2.2 compared to tf-base teachers, which exhibit an average of 1.1 COMET. **This directly correlates with the capacity gap problem: the smaller the gap in model size, the better the distillation**. The compression ratios for tf-big teachers are 3.2 times larger, underscoring the complexity of transferring knowledge from a high-capacity teacher to a smaller student.

RQ2: To what extent does script influence the transfer of knowledge? We compare Latin vs. non-Latin scripts because English (the target language in all models) is in the Latin script. Students trained for Latin script languages have an average of 1.2 COMET, while non-Latin script languages have a similar average of 3.5 COMET. **This difference indicates that script plays a role in the transfer of knowledge during distillation.** With a fixed vocabulary size, a shared script between source and target lets SentencePiece build longer, more semantically rich subwords. In contrast, non-Latin script languages yield shorter subwords, making knowledge transfer more difficult and reducing translation quality.

RQ3: Can we train multilingual students effectively? The student models for the language families in Table 4 maintain relatively high scores. For example, Lithuanian and Latvian demonstrate that multilingual training can compensate for some of the limitations of model compression, particularly for closely related languages. The Turkic family has a combination of scripts that may hinder knowledge transfer. **Even with the reduced size of the tiny model, we are able to fit multiple languages into a single student.**

6 Conclusions and Future Work

In this paper, we introduced fast MT models for 17 diverse languages, leveraging interpolated SeqKD to compress large teacher models into more efficient students. Our experiments reveal that low-performing teachers struggle to transfer knowledge effectively. We also demonstrate that the capacity gap between teacher and student models, as well as language script, significantly affect distillation performance. Additionally, our results highlight the effectiveness of multilingual distillation for related languages.

For future work, we plan to develop student models for additional languages. We also aim to expand our approach by distilling from a broader range of teacher models available on the HuggingFace Hub[5] and to further investigate cross-script knowledge transfer.

[5] https://huggingface.co/

Acknowledgements

This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant agreement No 101070350 and from UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant number 10052546]. The contents of this publication are the sole responsibility of its authors and do not necessarily reflect the opinion of the European Union.

References

Alham Fikri Aji and Kenneth Heafield. 2020. Compressing neural machine translation models with 4-bit precision. In *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 35–42, Online. Association for Computational Linguistics.

Mikko Aulamo, Sami Virpioja, and Jörg Tiedemann. 2020. Opusfilter: A configurable parallel corpus filtering toolbox. In *2020 Annual Conference of the Association for Computational Linguistics*, pages 150–156. The Association for Computational Linguistics.

Marta Bañón, Malina Chichirau, Miquel Esplà-Gomis, Mikel L. Forcada, Aarón Galiano-Jiménez, Cristian García-Romero, Taja Kuzman, Nikola Ljubešić, Rik van Noord, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Peter Rupnik, Vít Suchomel, Antonio Toral, and Jaume Zaragoza-Bernabeu. 2023. Catalan-english parallel corpus MaCoCu-ca-en 1.0. Slovenian language resource repository CLARIN.SI.

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 *EAMT 2022 - Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 303–304. European Association for Machine Translation.

Nikolay Bogoychev, Roman Grundkiewicz, Alham Fikri Aji, Maximiliana Behnke, Kenneth Heafield, Sidharth Kashyap, Emmanouil-Ioannis Farsarakis, and Mateusz Chudyk. 2020. Edinburgh's submissions to the 2020 machine translation efficiency task. In *The 4th Workshop on Neural Generation and Translation*, pages 218–224. Association for Computational Linguistics (ACL).

Nikolay Bogoychev, Jelmer van der Linde, Graeme Nail, Barry Haddow, Jaume Zaragoza-Bernabeu, Gema Ramírez-Sánchez, Lukas Weymann, Tudor Nicolae Mateiu, Jindřich Helcl, and Mikko Aulamo. 2023. Opuscleaner and opustrainer, open source toolkits for training machine translation and large language models. *arXiv preprint arXiv:2311.14838*.

Boxing Chen and Colin Cherry. 2014. A systematic comparison of smoothing techniques for sentence-level bleu. In *Proceedings of the ninth workshop on statistical machine translation*, pages 362–367.

Guy De Pauw, Peter Waiganjo Wagacha, and Gilles-Maurice de Schryver. 2009. The sawa corpus: a parallel corpus english-swahili. In *Proceedings of the First Workshop on Language Technologies for African Languages*, pages 9–16.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 644–648.

Maxim Enis and Mark Hopkins. 2024. From llm to nmt: Advancing low-resource machine translation with claude. *arXiv preprint arXiv:2404.13813*.

Jay Gala, Pranjal A Chitale, AK Raghavan, Varun Gumma, Sumanth Doddapaneni, Janki Atul Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, et al. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Transactions on Machine Learning Research*.

Ona de Gibert, Tommi Nieminen, Yves Scherrer, and Jörg Tiedemann. 2025. OpusDistillery: A Configurable End-to-End Pipeline for Systematic Multilingual Distillation of Open NMT Models. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, Tallinn, Estonia.

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.

Varun Gumma, Raj Dabre, and Pratyush Kumar. 2023. An empirical study of leveraging knowledge distillation for compressing multilingual neural machine translation models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 103–114.

Hiroaki Hayashi, Yusuke Oda, Alexandra Birch, Ioannis Konstas, Andrew Finch, Minh-Thang Luong, Graham Neubig, and Katsuhito Sudoh. 2019. Findings of the third workshop on neural generation and translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 1–14, Hong Kong. Association for Computational Linguistics.

Kenneth Heafield, Hiroaki Hayashi, Yusuke Oda, Andrew Finch, Graham Neubig, Xian Li, and Alexandra Birch. 2020. Findings of the fourth workshop on neural generation and translation. In *The 4th Workshop on Neural Generation and Translation*, pages 1–9. Association for Computational Linguistics (ACL).

Kenneth Heafield, Qianqian Zhu, and Roman Grundkiewicz. 2021. Findings of the wmt 2021 shared task on efficient translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 639–651.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network.

Aref Jafari, Mehdi Rezagholizadeh, Pranav Sharma, and Ali Ghodsi. 2021. Annealing knowledge distillation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2493–2504, Online. Association for Computational Linguistics.

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*.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in c++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121.

Jungo Kasai, Nikolaos Pappas, Hao Peng, James Cross, and Noah Smith. 2020. Deep encoder, shallow decoder: Reevaluating non-autoregressive machine translation. In *International Conference on Learning Representations*.

Yoon Kim and Alexander M Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327.

Young Jin Kim, Marcin Junczys-Dowmunt, Hany Hassan, Alham Fikri Aji, Kenneth Heafield, Roman Grundkiewicz, and Nikolay Bogoychev. 2019. From research to production and back: Ludicrously fast neural machine translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 280–288, Hong Kong. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, OndÅ™ej 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 PopoviÄ‡, and Mariya Shmatova. 2023. Findings of the 2023 conference on machine translation (wmt23): Llms are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Tom Kocmi, Vilém Zouhar, Christian Federmann, and Matt Post. 2024. Navigating the metrics maze: Reconciling score magnitudes and accuracies. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1999–2014, Bangkok, Thailand. 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.

Alireza Mohammadshahi, Vassilina Nikoulina, Alexandre Bérard, Caroline Brun, James Henderson, and Laurent Besacier. 2022. Small-100: Introducing shallow multilingual machine translation model for low-resource languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8348–8359.

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 on Association for Computational Linguistics*, ACL '02, page 311–318, USA. 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, Belgium, Brussels. Association for Computational Linguistics.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, H. Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant M. Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew J. Johnson, Blake A. Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Edward Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2021. Scaling language models: Methods, analysis & insights from training gopher. *CoRR*, abs/2112.11446.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for high- (but not low-) resource languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. Association for Computational Linguistics.

Felipe Sánchez-Martínez, Víctor M Sánchez-Cartagena, Juan Antonio Pérez-Ortiz, Mikel L Forcada, Miquel Espla-Gomis, Andrew Secker, Susie Coleman, and Julie Wall. 2020. An english-swahili parallel corpus and its use for neural machine translation in the news domain. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 299–308.

David Stap and Ali Araabi. 2023. ChatGPT is not a good indigenous translator. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 163–167, Toronto, Canada. Association for Computational Linguistics.

Xu Tan, Yi Ren, Di He, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2018. Multilingual neural machine translation with knowledge distillation. In *International Conference on Learning Representations*.

NLLB Team et al. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841.

Jörg Tiedemann, Mikko Aulamo, Daria Bakshandaeva, Michele Boggia, Stig-Arne Grönroos, Tommi Nieminen, Alessandro Raganato, Yves Scherrer, Raúl Vázquez, and Sami Virpioja. 2024. Democratizing neural machine translation with opus-mt. *Language Resources and Evaluation*, 58(2):713–755.

Jörg Tiedemann and Ona De Gibert. 2023. The opus-mt dashboard–a toolkit for a systematic evaluation of open machine translation models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 315–327.

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*, volume 30.

Grupo de investigación TALG Universidade de Vigo. 2012. Cluvi parallel corpus.

Chunting Zhou, Jiatao Gu, and Graham Neubig. 2020. Understanding knowledge distillation in non-autoregressive machine translation. In *International Conference on Learning Representations*.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781.

A spBLEU results

Language		ara	cat	eus	glg	heb	hin	jpn	kor	slk	swh	vie	zsm
Teacher	Params (M)	76.4	69.4	235.4	209.1	238.1	75.9	77.5	209.2	235.5	209.1	63.9	237.1
	Performance	37.6	45.1	33.6	44.5	46.5	32.1	19.2	30.1	43.2	41.0	28.7	44.4
Student	Compression	4.5	4.1	13.9	12.4	14.1	4.5	4.6	12.4	13.9	12.4	3.8	14.0
	Performance	29.7	43.3	26.7	39.5	41.2	29.8	7.4	22.2	38.4	35.4	29.9	41.4
	Δ	-7.9	-1.8	-6.9	-5.0	-5.3	-2.3	-11.8	-7.9	-4.8	-5.6	+1.2	-3.0

Table 5: spBLEU score results of our bilingual distillation experiments. For the teacher models, we report parameters in millions and performance. We provide results for the students, as well as their compression ratio. Δ shows the difference in spBLEU scores with the teacher.

Family		Baltic		Turkic		
Language		lit	lvs	azn	kaz	uzj
Teacher	Params (M)	236.9	236.9	238.8	238.8	238.8
	Performance	34.0	36.2	24.2	30.0	32.0
Student	Compression	14.02	14.02	14.13	14.13	14.13
	Performance	31.3	32.9	20.2	24.3	26.1
	Δ	-2.9	-3.3	-4.0	-5.7	-5.9

Table 6: spBLEU score results of our multilingual distillation experiments.