Hybrid Distillation from RBMT and NMT: Helsinki-NLP's Submission to the Shared Task on Translation into Low-Resource Languages of Spain

Ona de Gibert[1] and **Mikko Aulamo**[1] and **Yves Scherrer**[1,2] and **Jörg Tiedemann**[1]
[1]University of Helsinki, Dept. of Digital Humanities
[2]University of Oslo, Dept. of Informatics
[1]firstname.lastname@helsinki.fi
[2]firstname.lastname@ifi.uio.no

Abstract

The Helsinki-NLP team participated in the 2024 Shared Task on Translation into Low-Resource languages of Spain with four multilingual systems covering all language pairs in the open submission track. The task consists in developing Machine Translation (MT) models to translate from Spanish into Aragonese, Aranese and Asturian. Our models leverage known approaches for multilingual MT; namely, data filtering, fine-tuning, data tagging, and distillation. We use distillation to merge the knowledge from neural and rule-based systems and explore the trade-offs between translation quality and computational efficiency. We demonstrate that our distilled models can achieve competitive results while significantly reducing computational costs. Our best models ranked 4th, 5th, and 2nd in the open submission track for Spanish–Aragonese, Spanish–Aranese, and Spanish–Asturian, respectively. We release our code and data publicly at https://github.com/Helsinki-NLP/lowres-spain-st.

1 Introduction

In this work, we describe the participation of our team to the Shared Task on Translation into Low-Resource Languages of Spain 2024 (Sánchez-Martínez et al., 2024), the first edition of its kind. The task involves developing Machine Translation (MT) systems for translating from Spanish (*spa*) into three closely related Romance target languages: Aranese (*arn*), Aragonese (*arg*) and Asturian (*ast*). Aranese is a variety of Occitan spoken in the northwestern part of Catalonia; Aragonese is spoken in Aragon, in northwest Spain; and Asturian is spoken in Asturias, in northeast Spain.

Although these minority languages have some form of official status in their respective regions, they are all considered endangered. According to the linguistic taxonomy proposed by Joshi et al. (2020), these languages fall into the category of the "Scraping Bys". This means that, while there is some available unlabeled data, substantial and coordinated efforts are necessary to raise awareness and gather labeled datasets to improve the prospects of these languages in the future. This task is designed precisely to address these challenges by fostering the development of resources and tools for these under-resourced languages.

In terms of current technological support, some linguistic resources are available for these languages, including online dictionaries and established orthographic standards. Apertium (Forcada et al., 2011) is an open-source Rule-Based MT (RBMT) toolkit initially developed for related languages, that offers substantial coverage for the three target languages. Nevertheless, resources remain notably sparse for data-driven approaches like Neural Machine Translation (NMT). By contributing to this task, we aim to change this picture.

We focus our participation efforts on data collection – by gathering additional data from Wikipedia and online dictionaries –, data augmentation – by producing back-translations (Sennrich et al., 2016) of monolingual data –, and data preparation – by carrying out corpus-targeted cleaning. We also experiment with different data tagging strategies. We submit four multilingual models, which arise from fine-tuning and applying Knowledge Distillation (KD), by leveraging both neural and RBMT outputs, similarly to Aulamo et al. (2021). We evaluate our models both for translation quality and efficiency, resulting in a diverse set of submissions that balance accuracy and speed. Our contributions, including our code and data, are publicly available for further research in our Github repository.

The rest of the paper is organised as follows. Section 2 describes the benchmarking of existing models. Section 3 provides a detailed description of our data collection and preparation efforts. Section 4 describes the submitted models in detail. Section 5 outlines the results and, finally, section 6 concludes our work.

Proceedings of the Ninth Conference on Machine Translation, pages 908–917
November 15-16, 2024 ©2024 Association for Computational Linguistics

			spa–xxx			xxx–spa		
Model	Params (M)	spa–arg	spa–arn	spa–ast	arg–spa	arn–spa	ast–spa	Speed (s)
Apertium	–	53.8	32.5	14.4	56.4	31.3	17	99.00
opus-mt/itc-itc	212	42.0	9.4	15.3	70.1	32.9	22	284.24
opus-mt/deu+eng+fra+por+spa-itc*	222	**42.6**	**9.6**	**16.9**	–	–	–	307.06
opus-mt/roa-deu+eng+fra+por+spa**	222	–	–	–	**71.1**	37.5	22.4	289.09
opus-mt/itc-deu+eng+fra+por+spa	222	–	–	–	70.4	**37.7**	22.3	245.22
nllb-200-distilled-600M	600	–	8.55	13.38	–	30.38	21.66	442.94
nllb-200-distilled-1.3B	1,300	–	8.66	13.95	–	32.59	22.48	800.38
nllb-200-1.3B	1,300	–	8.62	12.46	–	34.35	22.85	822.73
nllb-200-3.3B	3,300	–	8.75	13.38	–	35.16	**23.65**	914.43

Table 1: BLEU scores on the development set for all language pairs in both directions of existing MT models. We also report the average decoding speed in seconds on a single Nvidia V100 GPU. The asterisks (* and **) indicate that we use those models for our work.

2 Benchmarking of Existing Models

The first step we took when approaching this task was to benchmark existing models for the target languages. This process enabled us to assess the current landscape of available models, identify those suitable for fine-tuning, and determine which models could be utilized for back-translation.

We evaluate three types of models: OPUS-MT models[1], the smaller NLLB variants (Costa-jussà et al., 2022), and the rule-based Apertium systems. OPUS-MT models are trained with the Tatoeba Translation Challenge dataset[2] (Tiedemann, 2020) and data from the massively parallel Bible corpus (Mayer and Cysouw, 2014) as part of the JHUBC corpus (McCarthy et al., 2020). These are all *transformer-big* (Vaswani et al., 2017) systems. The NLLB models are trained on a diverse collection of multilingual text and come in different sizes. Apertium and OPUS-MT cover all three target languages, whereas NLLB does not support Aragonese.

We evaluate the systems with the provided development set by the organizers and BLEU (Papineni et al., 2002), as implemented in sacreBLEU (Post, 2018). The development set consists on a manually-crafted revision of the 997 sentences from Flores+ (Goyal et al., 2022). Results are shown in Table 1.

We can see how OPUS-MT models, although much smaller in size, outperform all NLLB variants when the target language is the Romance minority language. All NLLB variants perform similarly, independently of their size. We attribute the lower score on the Spanish–Aranese language pair to the models being trained on Occitan data rather than Aranese. Additionally, the remarkable performance of the Apertium models in translating from Spanish stands out, as they surpass the neural systems, except in the case of Asturian.[3] This demonstrates the effectiveness of rule-based systems in handling closely related languages.

With Spanish as the target language, the NLLB models follow the scaling laws and their score increases along their size, as would be expected. The OPUS-MT models exhibit comparable performance. As expected, the compact OPUS-MT models are much faster when considering the decoding speed of the different models. OPUS-MT models have been trained using Marian (Junczys-Dowmunt et al., 2018), while the NLLB family was trained using Fairseq (Ott et al., 2019).

Taking this into account, we decide to select two OPUS-MT models from Table 1 for our work: the model marked with * for fine-tuning; and model ** for producing back-translations. Given that the NMT models significantly outperform the Apertium systems for translation into Spanish, the rule-based back-translation strategy employed by Aulamo et al. (2021) did not suit our context.

3 Data

The data used to train our NMT systems consists of parallel and monolingual datasets provided by the organizers, as well as additional Wikipedia and dictionary data. We utilize the monolingual datasets by back-translating them to create synthetic parallel

[1] https://github.com/Helsinki-NLP/OPUS-MT-train

[2] Version v2023-09-26.

[3] Asturian sentences were professionally translated from English, while Aragonese and Aranese sentences were machine translated from Spanish using Apertium and later post-edited. Hence, the higher score for Apertium and the language pairs involving Aragonese and Aranese.

	Monolingual						**Parallel**	
	PILAR crawled	PILAR literary	PILAR cat–arn	Dictionary	Wikipedia	Wikipedia Discussions	Tatoeba Challenge	TOTAL
arg								
raw	60,028	24,675	–	–	255,149	–	41,623	381,475
langid	60,028	20,241	–	–	241,415	–	22,354	344,038
filtered	56,103	19,328	–	–	237,793	–	19,479	**332,703**
arn/oci								
raw	7,358	229,886	85,491	14,874	616,530	14,591	744,731	1,713,461
arn								
langid	7,358	228,512	64,141	14,874	29,627	2,429	106,248	453,189
filtered	7,243	213,960	64,141	14,874	27,160	2,249	87,189	**337,801**
oci								
langid	0	474	0	0	511,713	11,415	354,202	877,804
filtered	0	357	0	0	493,216	10,810	299,440	**803,823**
ast								
raw	14,776	24,093	–	82,009	2,230,855	–	5,511,336	7,863,069
langid	10,538	17,112	–	82,009	1,920,758	–	3,705,483	5,735,900
filtered	9,975	16,072	–	82,009	1,862,821	–	991,617	**2,880,485**

Table 2: Number of sentence pairs in training datasets. The "raw" line shows the sizes before any filtering, the "langid" line shows the number of sentence pairs in the correct language according to Idiomata Cognitor, and the "filtered" line shows the final sizes of the clean datasets. The Aranese and Occitan data are separated into two sets after language identification. The final cleaned data size for each language is shown in bold.

training data. We remove noise from the training data using the Idiomata Cognitor (Galiano-Jiménez et al., 2024a) language identification tool, OpusCleaner (Bogoychev et al., 2023) and its visual user interface, and the configurable filtering toolbox OpusFilter (Aulamo et al., 2020).

3.1 Data Collection

Table 2 shows the sizes of the datasets used for training. As original parallel data, we use only the Tatoeba Challenge data[4] (Tiedemann, 2020), which contains all data in OPUS (Tiedemann, 2012), deduplicated and shuffled. We also use the crawled and literary PILAR corpora (Galiano-Jiménez et al., 2024b) as monolingual data for all three language pairs. Additionally, we use the Aranese side of the Catalan-Aranese PILAR corpus also as monolingual data.

We also leverage monolingual data that is not provided by the organizers, which puts our models in the open track: Wikipedia and online dictionaries for Aranese[5] and Asturian[6]. From Wikipedia, we obtain the latest dump per language. Moreover, for Occitan, we also make use of OcWikiDisc (Miletic and Scherrer, 2022), a corpus extracted from the talk pages associated with the Occitan Wikipedia. We assume that Occitan datasets include Aranese data, since it is a variety of Gascon, one of the main dialects of Occitan. For the online dictionaries, we develop our own scraping scripts to gather definitions. The scripts can be found in our Github repository. For the monolingual data, we produce back-translations into Spanish with the openly available OPUS-MT model (marked with ** in Table 1) to produce synthetic parallel data.

3.2 Data Cleaning

The first step of our data cleaning pipeline is language identification. We use the Idiomata Cognitor tool (Galiano-Jiménez et al., 2024a) to identify the correct target languages in all data sets. Idiomata Cognitor also allows us to distinguish Aranese from other Occitan varieties. Hence, from this point onwards, we treat Aranese and (non-Aranese) Occitan data separately in order to experiment with different model training strategies as described in Section 4.

Next, we create customized filtering configurations for each corpus (and for each subcorpus in Tatoeba) to apply optimal data cleaning based on the style and domain of the texts. To this end, we use OpusCleaner (Bogoychev et al., 2023), which is a parallel data cleaning tool that allows the user to add and adjust filters and see their effects on a sample of the corpus in real time in a graphical interface. The filters most commonly applied

[4] We use the same version as the original OPUS-MT model.

[5] https://www.diccionari.cat/cerca/diccionari-der-aranes

[6] https://diccionariu.alladixital.org/

Model	Submission	oci tag	arn tag	spa–arg	spa–arn	spa–ast
OPUS-MT	-	-	-	42.6	9.6	16.9
A.1	1	`»oci«`	`»oci«`	**54.8**	12.3	18.5
A.2	1	–	`»oci«`	51.5	**28.2**	18.5
B.1	–	`»oci«`	`»xxx«`	51.6	26.1	18.5
B.2	–	`»xxx«`	`»oci«`	51.5	**26.8**	18.5
B.3	–	`»oci«»oci«`	`»oci«»xxx«`	**55.2**	26.0	18.5
B.4	–	`»oci«»xxx«`	`»oci«»oci«`	52.8	25.7	18.5

Table 3: BLEU scores on the development set of fine-tuning the OPUS-MT model with different tagging strategies. We provide the scores of the OPUS-MT model for reference. We report the best checkpoint score per language pair. Model A.2 does not use Occitan data.

to our training sets are: (1) `src_trg_ratio`: The ratio between the number of source and target tokens. (2) `num_mismatch`: The ratio between the number of overlapping and differing numerals. (3) `alpha_ratio`: The ratio between the number of words and non-words, and the ratio between the number of language and non-language characters.

Additionally, some corpora contain unwanted structures, such as HTML tags or transcription content between double square brackets in Wikipedia data, which we remove from the sentences. Finally, we apply OpusFilter to concatenate the different corpora, normalize whitespace characters, remove all sentences shorter than 3 or longer than 150 words and remove all duplicate sentence pairs. Table 2 shows the data sizes for each language pair after applying language identification and corpus cleaning. The final size of our corpus is 4,35M sentences, with 66.14% spa–ast, 7.63% spa–arg, 7.75% spa–arn, and 18.45% spa–oci. All of our data cleaning configuration files can be found in our Github repository.

4 Models

In this section, we detail our modeling choices for the four submissions, all of which employ one-to-many multilingual models. Our models leverage fine-tuning and data tagging, and the integration of RBMT with neural models via Sequence-Level KD (Seq-KD) (Kim and Rush, 2016). All models are based on the Transformer architecture (Vaswani et al., 2017) and use the OPUS-MT model, as described earlier, as the initial checkpoint in some form. For tokenization, we use the OPUS-MT model's SentencePiece vocabularies (Kudo and Richardson, 2018), two distinct 32k piece vocabularies: one shared among all source languages (in our case, only Spanish) and another shared among all targets. All models are trained on 4 Nvidia V100 GPUs, except models C.2 and D.2, which are trained on 8 AMD MI250x GPUs. Further configuration details are provided in Appendix A.

4.1 Models A: Fine-tuning

As an initial step, we use the openly available OPUS-MT model described in Section 2 and fine-tune it using different data sampling schemes. We train one model with all available training data (model A.1) and another excluding the Occitan data (A.2). The decision to exclude Occitan data was made because both languages share the same language tag, which could potentially confuse the model, since there is much more training data on Occitan than on Aranese. The development set scores are presented in Table 3.

Compared to the original OPUS-MT model, we observe a significant increase in BLEU scores for the spa–arg language pair (+12.2) and for spa–arn (+18.6). However, for Asturian, the increase is more modest (+1.6). Removing the Occitan data results in an increased score of almost +16 BLEU points for the spa–arn language pair. Interestingly, despite having the largest amount of new data for Asturian, the model quickly reaches a performance plateau during training, as shown in Figure 2 in Appendix B. This trend persists throughout our experiments, leading us to conclude that the spa–ast language pair is the most challenging task.

For our **Submission #1**, we ensemble the best n checkpoints per language pair across both A.1 and A.2 models.[7]

[7] We perform ensembling using the top 10 best checkpoints for each language pair and submit the ensemble with the highest score on the development set.

Model	Teacher(s)	Size	Submission	spa–arg	spa–arn	spa–ast
C.1	A	base	2	53.6	**28.1**	18.5
C.2	A	tiny	–	51.3	25.5	18.2
C.3	B.3	base	–	55.4	26.6	18.5
D.1	A + RBMT	base	2	**54.2**	27.3	18.5
D.2	A + RBMT	tiny	3	**52.8**	**27.1**	18.2
D.3	B.3 + RBMT	base	4	56.9	**30.2**	18.5
D.3_fixed	B.3 + RBMT	base	–	**57.0**	26.9	18.5
D.4	RBMT	base	–	62.4	36.8	16.9

Table 4: Comparison of BLEU scores for our distillation experiments between NMT-only models and NMT+RBMT systems across different language pairs on the development set. We report the best checkpoint score per language pair, except for models D, where we use the same single checkpoint.

4.2 Models B: Data tagging

In multilingual systems, it is a common practice to prepend a language tag to the source sentence to indicate the target language. For consistency, we applied uniform tagging across all models. Nevertheless, for Aranese, we experimented with different tagging schemes, given that it is a variety of Gascon, a dialect of Occitan.

Exploring the OPUS-MT vocabulary, we identified an unused tag, »xxx«, which prompted us to experiment with various combinations of the »oci« and »xxx« tags, including the use of double tags. We fine-tune the original OPUS-MT model with all available training data and different tagging schemes. Results of are provided in Table 3.

While the performance of Asturian remained unaffected, using different tags for Aranese and Occitan led to a much higher BLEU on the spa–ara language pair compared to model A.1; due to the effectiveness of the data tagging schemes. Notably, Aragonese appeared to be the most impacted by data tagging, although we are unsure why. On average, the best performing model leverages double tags (B.3). We do not submit any of these, but use B.3 as a teacher in our distillation experiments.

4.3 Sequence-Level Distillation

Seq-KD (Kim and Rush, 2016) is a technique where a student model is trained using translations generated by one or more teacher model(s), with the goal of transferring knowledge from a large, powerful teacher model to a smaller, more efficient student model. We experiment with Seq-KD to train fast students.

Models C: NMT-distilled

First, we distill student models using the previously fine-tuned *transformer-big* NMT systems as teachers. We leverage Sequence-Level Interpolation (Kim and Rush, 2016), generating 8-best candidate translations for all the training data using the best checkpoint for each language pair. From these, we select the translation with the highest ChrF (Popović, 2015) with the reference to create a distilled dataset, which is then used to train the student model. We use ChrF instead of BLEU as a more fine-grained metric at character level.

To explore the tradeoff between translation quality and speed, we use models A[8] as the teachers and train two students of different sizes. Model C.1 is a *transformer-base* model (67.5M parameters), while model C.2 follows the *tiny* architecture described in Bogoychev et al. (2020), its size is 20.4M parameters (3.3 times smaller). We train model C.2 using the OpusDistillery[9], a pipeline for multilingual Seq-KD of open NMT models. In addition, to investigate the effect of multi-teacher distillation, we distill another *transformer-base* model (C.3) using a single NMT teacher, in this case, model B.3. The development set scores for these student systems are shown in Table 4.

When comparing models C.1 and C.2, it becomes evident that the capacity gap between the teacher and student models significantly impacts student performance. In KD, student models are typically smaller than their teacher counterparts, which can hinder their ability to effectively learn and fit noisy data. This is reflected in the lower scores of the smaller C.2 model across all language pairs compared to C.1. On the other hand, models C.1 and C.3 share the same size, but their training

[8] For Occitan, we use the original OPUS-MT model, as our fine-tuned model has a lower score on Occitan, due to the catastrophic forgetting phenomenon (Goodfellow et al., 2013).

[9] https://github.com/Helsinki-NLP/OpusDistillery

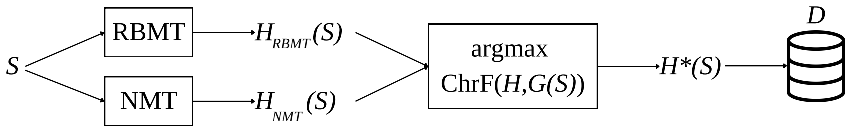

Figure 1: Overview of our Seq-KD distillation process to merge NMT and RBMT data. Given a source sentence *(S)*, we produce a hypothesis translation *(H)* with both our RBMT and NMT models. Then, we choose the translation *(H*)* that has the maximum ChrF with the ground truth *(G)* to create the distilled dataset *(D)*.

strategies differ. Model C.1 is distilled from multiple teachers, while C.3 is distilled from a single teacher. Notably, distilling from a single model in C.3 appears to offer greater stability.

Models D: Hybrid-distilled

Since rule-based translation models are remarkably good for the given language pairs, as shown in Table 1, we further experiment with Seq-KD to train student models that benefit from both RBMT and NMT outputs.

In this case, we use two types of teachers: (1) the best checkpoint per language pair of the NMT model(s) (as in the previous section) and (2) the Apertium RBMT models. We forward translate the training data with both teachers. For each source sentence, we select the translation that has the highest ChrF score with the ground truth to create the distilled dataset. Finally, we train a new student model on the distilled dataset. An overview of this process is depicted in Figure 1.

For each of the former models C, we train a comparable hybrid-distilled student using a combination of the NMT and RBMT data. The development set scores for these models (D.1–D.3) are shown in Table 4. The proportions of RBMT data selected for the final distilled dataset are provided in Appendix C.

The inclusion of RBMT data in the distillation process leads to better performance across all language pairs overall. For model D.1, the addition of rule-based distilled data results in a slight decrease for spa–arn, in comparison to C.1. For spa–ast, the performance is identical across all models. It is remarkable to note that model D.3 surpasses the performance of its own teacher with +1.7 BLEU for spa–arg and +4.2 BLEU for spa–arn.

After the submission deadline, we discovered that the NMT distilled dataset for model D.3 had been generated using incorrect language tags for Aranese and Occitan (»oci« instead of »oci«»oci« and »oci«»xxx«, respectively). We provide the corrected results for the D.3 model (model D.3_fixed in Table 4). Interestingly, the initial D.3 model performed better for Aranese due to the higher proportion of RBMT data for that language (as shown in Appendix C, 15% vs. 2.3%), which favored the RBMT-heavy development set. Motivated by this finding, we trained a student using RBMT-only distilled data after the submission deadline (model D.4 in Table 4), which outperforms all other models except for Asturian. This opens up a new avenue of research, leveraging linguistically informed methods for distillation.

Table 4 demonstrates the effectiveness of using distillation to train a single model that performs well across all three language pairs. Among the language pairs, Aragonese shows the most significant improvement when RBMT data is incorporated, highlighting the particular benefit of combining rule-based and neural translation methods for this language. This aligns with our expectations since as can be seen from Table 1, the spa–arg Apertium model achieves the highest BLEU score.

Out of our distillation experiments, we make three submissions. For **submission #2**, we ensemble the best n checkpoints per language pair across models C.1 and D.2. For **submission #3**, we submit model D.2. In this case, we do not use ensembling because we want to test it for speed. Finally, for **submission #4**, we ensemble the best n checkpoints per language pair from model D.3.

5 Results

We make four submissions in the open submission track. The test set corresponds to the 1,012 lines of Flores+ evaluation set. We summarize our submissions' test results in Table 5, as provided by the organizers of the Shared Task. For comparison, we also include the scores of the top-performing competitor, overall and in the open submission track. The official evaluation metrics of the task are BLEU and ChrF. Additionally, we report the average decoding speed and the model sizes.

Our best models ranked 4th, 5th, and 2nd in the open submission track for spa–arg, spa–ara, and

#	Method	BLEU / ChrF			Params (M)	Speed (s)
		arg	arn	ast		
1	Fine-tuning Data Sampling Ensembling	51.5 / 75.6	22.1 / 45.1	18.2 / 51.6	222.9	852.22
2	Distillation RBMT+NMT Ensembling	50.6 / 75.4	22.4 / 45.7	18.0 / 51.6	65.7	361.33
3	Distillation RBMT+NMT	49.1 / 75.4	21.6 / 45.0	17.9 / 51.4	20.4	4.06
4	Distillation Data Tagging RBMT+NMT Ensembling	52.7 / 75.9	24.3 / 46.6	18.0 / 51.5	67.5	891.76
Best (overall)	–	63.0 / 80.3	30.4 / 50.1	23.2 / 55.2	–	–
Best (open)	–	62.7 / 80.0	28.8 / 49.4	23.2 / 55.2	–	–

Table 5: Summary of our submissions. BLEU refers to the score obtained by the best ensemble on the development set; Speed refers to the averaged decoding speed for submission across language pairs on one single AMD MI250x GPU. In addition, we provide the best competitor scores for each target language.

spa–ast, respectively. On average, our best submission for each language pair falls short of the top competitor by 4 BLEU points and 3.8 ChrF points. This narrow margin reflects the competitive nature of this year's task, which saw over 178 submissions.

Our best model is submission #4, followed closely by submissions #1, #2 and, finally #3, in that order. It is noteworthy that our distilled models perform really well compared to their teachers. Submission #2, a distilled model from Submission #1, demonstrates an increase of +0.3 BLEU for spa–arn over its teacher, highlighting the potential of distillation to not only preserve but even enhance translation quality. Moreover, our smallest model, Submission #3, although showing a slight average decrease of –1.1 BLEU compared to its teacher, offers a significant advantage in terms of speed—it is 210 times faster.

6 Conclusions

In this work, we have presented our participation in the Shared Task of Translation into Low-Resource Languages of Spain 2024. We have described our data collection and preparation efforts, as well as our four submissions based on multilingual models. We explore fine-tuning of an existing open model with different data tagging schemes and use Seq-KD to train small efficient student models. Furthermore, to our knowledge, we are the first to leverage RBMT to improve distillation for similarly related languages and prove its effectiveness.

This study opens up new research directions for advancing in low-resource MT by demonstrating the potential of data tagging strategies and hybrid distillation methods, ensuring these languages are both preserved and accessible in the digital age.

7 Ethical Considerations

In addition to evaluating the performance of our models in terms of translation quality, it is equally important to consider the computational resources required for their training and deployment. By analyzing the GPU consumption of our experiments, including the time spent and energy consumed for each task, we aim to provide a comprehensive assessment of the efficiency and sustainability of our approaches. This will allow the community to take informed decisions about model selection and optimization in real-world applications, where computational efficiency is often as critical as accuracy. We report the energy consumption of the totality of our experiments in Table 6, which amounts to 508 kWh.

Task	Model	Time (h)	Energy (kWh)
Back-translation		18.9	19.5
Fine-tune	A.1	35.3	37.0
	A.2	22.4	23.0
	B.1	50.5	52.2
	B.2	27.4	28.9
	B.3	28.9	29.6
	B.4	28.0	28.8
Forward translation		7.9	6.6
Train via Seq-KD	C.1	64.6	66.9
	C.2	11.5	18.1
	C.3	57.1	54.2
	D.1	53.4	55.3
	D.2	11.2	17.8
	D.3	66.6	68.6
	D.3_fixed	55.9	57.1
	D.4	56.0	57.8
Ensembling		30.2	3.52
Submission		1.7	0.19
	Total	627.3	**625.1**

Table 6: Energy consumption of our work. We report the time (hours) and energy consumption across the different tasks of our experiments, run on 4 Nvidia V100 GPUs. The training of models D has been run on 8 AMD MI250x GPUs. Ensembling and translations for submission have been run on 1 Nvidia V100 GPUs.

Acknowledgments

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. The authors wish to thank CSC – IT Center for Science, Finland for computational resources and support.

References

Mikko Aulamo, Sami Virpioja, Yves Scherrer, and Jörg Tiedemann. 2021. Boosting neural machine translation from Finnish to Northern Sámi with rule-based backtranslation. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 351–356, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

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.

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 *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 218–224, Online. Association for Computational Linguistics.

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. *Preprint*, arXiv:2311.14838.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

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

Aarón Galiano-Jiménez, Felipe Sánchez-Martínez, and Juan Antonio Pérez-Ortiz. 2024a. Idiomata cognitor.

Aarón Galiano-Jiménez, Felipe Sánchez-Martínez, and Juan Antonio Pérez-Ortiz. 2024b. Pan-iberian language archival resource.

Ian J. Goodfellow, Mehdi Mirza, Xia Da, Aaron C. Courville, and Yoshua Bengio. 2013. An empirical investigation of catastrophic forgeting in gradient-based neural networks. *CoRR*, abs/1312.6211.

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.

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. 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, et al. 2018. Marian: Fast neural machine translation in c++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121.

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.

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.

Thomas Mayer and Michael Cysouw. 2014. Creating a massively parallel Bible corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3158–3163, Reykjavik, Iceland. European Language Resources Association (ELRA).

Arya D. McCarthy, Rachel Wicks, Dylan Lewis, Aaron Mueller, Winston Wu, Oliver Adams, Garrett Nicolai, Matt Post, and David Yarowsky. 2020. The Johns Hopkins University Bible corpus: 1600+ tongues for typological exploration. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2884–2892, Marseille, France. European Language Resources Association.

Aleksandra Miletic and Yves Scherrer. 2022. OcWikiDisc: a corpus of Wikipedia talk pages in Occitan. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 70–79, Gyeongju, Republic of Korea. Association for Computational Linguistics.

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.

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.

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.

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.

Felipe Sánchez-Martínez, Juan Antonio Pérez-Ortiz, Aarón Galiano-Jiménez, and Antoni Oliver. 2024. Findings of the WMT 2024 shared task on translating into low-resource languages of spain: Blending rule-based and neural systems. In *Proceedings of the Ninth Conference on Machine Translation*, Miami, USA. Association for Computational Linguistics.

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.

Jörg Tiedemann. 2020. The Tatoeba Translation Challenge – realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

A Hyperparameters

All models are based on the transformer architecture. They all share the following: the Adam optimizer is used with β_1=0.9 and β_2=0.998. The models are trained until convergence with early-stopping on development data after BLEU has stalled 10 times. Next, we specify each model's unique configuration details.

Models A and B are *transformer-big* models. They use a 6-layered transformer with 16 heads, 1024 dimensions in the embeddings and 4,096 dimensions in the feed-forward layers.

Models C.1, C.3, D.1 and D.3 use a 6-layered transformer with 8 heads, 512 dimensions in the embeddings and 2,048 dimensions in the feed-forward layers.

Models C.2 and D.2 are trained using tiny architecture proposed in Bogoychev et al. (2020). The student model has a transformer encoder with 6 layers and a light-weight RNN based decoder with Simpler Simple Recurrent Unit (SSRU) (Kim et al., 2019) with 2 layers; 8 heads, 256 dimensions in the embeddings and 1,536 dimenstions in the feed-forward layers.

B Learning Curves

Figure 2 shows the BLEU score progression over training updates per language pair for model A.2. It shows how the performance for spa–ast quickly reaches a plateau.

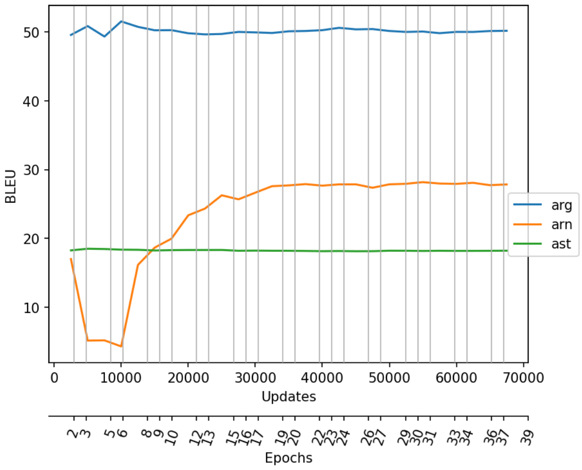

Figure 2: BLEU score progression over training updates and epochs for model A.2.

C Rule-based MT Data

For the distilled models D, we use a combination of NMT and RBMT teachers to build a distilled dataset. The RBMT teachers are the Apertium models. For each source sentence, we generate a hypothesis translation using both teachers and then compute the ChrF score against the ground truth. We retain the hypothesis with the highest ChrF score for each sentence. Table 7 shows the proportion of sentences originating from RBMT across our experiments.

Teacher(s)	A.1, A.3	B.3	B.3
Model(s)	C.1, C.2	D.3	D.3_fixed
Submission	#2 #3	#4	-
Pair	%	%	%
spa–arg	4.39	7.37	7.37
spa–arn	1.32	15.32	2.33
spa–ast	3.85	3.75	3.75
spa–oci	8.95	1.69	1.64

Table 7: Distribution of distilled data coming from RBMT in sentence count and percentage (%).