Findings of the AmericasNLP 2024 Shared Task on Machine Translation into Indigenous Languages

Abteen Ebrahimi[◇*] **Ona de Gibert**[♠*] **Raúl Vázquez**[♠*] **Rolando Coto-Solano**[Ω*]
Pavel Denisov[∇] **Robert Pugh**[ψ] **Manuel Mager**[♣] **Arturo Oncevay**[♡]
Luis Chiruzzo[Φ] **Katharina von der Wense**[◇♯] **Shruti Rijhwani**[≁]
[◇]University of Colorado Boulder [♠]University of Helsinki [Ω]Dartmouth College
[∇]University of Stuttgart [ψ]Indiana University [♣]Amazon AWS AI
[♡]Pontificia Universidad Católica del Perú [Φ]Universidad de la República, Uruguay
[♯]Johannes Gutenberg University Mainz [≁]Google Deepmind
abteen.ebrahimi@colorado.edu

Abstract

This paper presents the findings of the third iteration of the AmericasNLP Shared Task on Machine Translation. This year's competition features eleven Indigenous languages found across North, Central, and South America. A total of six teams participate with a total of 157 submissions across all languages and models. Two baselines – the Sheffield and Helsinki systems from 2023 – are provided and represent hard-to-beat starting points for the competition. In addition to the baselines, teams are given access to a new repository of training data which consists of data collected by teams in prior shared tasks. Using ChrF++ as the main competition metric, we see improvements over the baseline for 4 languages: Chatino, Guarani, Quechua, and Rarámuri, with performance increases over the best baseline of 4.2 ChrF++. In this work, we present a summary of the submitted systems, results, and a human evaluation of system outputs for Bribri, which consists of both (1) a rating of meaning and fluency and (2) a qualitative error analysis of outputs from the best submitted system.

1 Introduction

Though the field of natural language processing (NLP) has seen a steep increase in interest and impressive performance improvements over the past decade, a large performance gap still remains between a handful of so-called "high-resource," mostly colonial, languages and the remaining majority of the world's languages (Blasi et al., 2022). The Indigenous languages of the Americas exemplify this reality, representing nearly 15% of the world's linguistic diversity (Eberhard et al., 2024) and yet, until recently, receiving little attention in NLP research.

Language	Family	Train	Extra	Syn.	Dev.
Asháninka (cni)	Arawak	3,883	-	13,195	883
Aymara (aym)	Aymaran	6,531	24,331	16,750	996
Bribri (bzd)	Chibchan	7,508	-	-	996
Chatino (ctp)	Oto-Manguean	357	2,246	-	499
Guarani (gn)	Tupi-Guarani	26,032	42,186	40,516	995
Nahuatl (nah)	Uto-Aztecan	16,145	2,493	9,222	672
Otomí (oto)	Oto-Manguean	4,889	9,012	-	599
Quechua (quy)	Quechuan	125,008	6,469	60,399	996
Rarámuri (tar)	Uto-Aztecan	14,720	2,254	-	995
Shipibo-Konibo (shp)	Panoan	14,592	16,721	23,595	996
Wixarika (hch)	Uto-Aztecan	8,966	2,653	511	994

Table 1: Languages of the shared task, their ISO codes, language families, and dataset statistics.

The AmericasNLP Shared Task on Machine Translation (MT), now in its third iteration (2021, 2023, and 2024), is focused on pushing the performance of MT on this group of languages through two main avenues: by applying modeling and architectural advancements, and through the creation of new linguistic resources which support the training and evaluation of these systems.

This year's shared task continues to focus on the eleven Indigenous languages from the last competition. While this year's competition does not feature new data for evaluation, competitors are given access to a new repository of training data which extends the original set of parallel examples with additional data collected by teams in prior years. This repository represents the first step in creating a new living source of data which can grow through contributions from teams participating in future iterations of the shared task. This year's competition also features two baselines: the University of Sheffield (Gow-Smith and Villegas, 2023) and University of Helsinki (De Gibert et al., 2023) systems which each achieved the best performance for a subset of languages in 2023 (Ebrahimi et al., 2023). These baselines are strong and hard-to-beat; across 157 submissions from 6 different teams, we see improvements for only 4 of the 11

[*] Equal contribution.

Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 236–246
June 21, 2024 ©2024 Association for Computational Linguistics

languages: Chatino, Guarani, Quechua, and Rarámuri. As two of these four languages are the relatively highest-resourced, this finding may indicate that we are approaching a plateau in performance gains achievable purely through modeling and architectural approaches; therefore, a focus on collecting additional training data may yield the most future improvements.

The paper is structured as follows. In Section 2, we provide a brief overview of the data and languages provided by the organizers at the beginning of the competition. Section 3 contains summary descriptions of the approaches used by each team. Section 4 discusses the results of the competition. In Section 5, we conduct a human evaluation of system outputs for Bribri. In the first part of this evaluation, we follow the prior shared tasks in quantitatively rating a sample of outputs on two axes: meaning and fluency. For the second part, we conduct a qualitative error analysis, comparing baseline systems to the best submitted system. In Section 6, we conclude with a brief discussion of future directions in improving MT quality for Indigenous languages of the Americas.

2 Data and Languages

The shared task features 11 Indigenous languages of the Americas. The language direction we are interested in is from Spanish into the low-resource language.

We use the AmericasNLP 2021 data for development and evaluation. It consists of a multi-way parallel dataset of the Spanish XNLI test set into 10 languages of the Americas (Asháninka, Aymara, Bribri, Guarani, Nahuatl, Otomí, Quechua, Rarámuri, Shipibo-Konibo, and Wixarika). The task also includes Chatino, for which the data comes from Mexican court proceedings. Chatino was introduced as a surprise language in last year's edition (Ebrahimi et al., 2023). For an in-depth review of development and evaluation data, please refer to Ebrahimi et al. (2022) and Mager et al. (2021).

For training data, besides the data used in previous editions, this year we include the data collected by De Gibert et al. (2023) as part of their Helsinki-NLP submission. This consists of `extra` data, made up of different sources listed in their system description paper, as well as `syn`, which refers to synthetic data obtained through backtranslation. Table 1 provides an overview of the languages, their linguistic families, and the total number of parallel sentences with Spanish. While there is no new data for Bribri, this year's data sizes increased considerably for Shipibo-Konibo, Aymara, Quechua and Guarani, with over 40k added sentences (although the majority comes from backtranslations). The test data for all languages consists of 1,003 sentences, except for Chatino, which contains 1,000 sentences.

We publicly release the training and development data in our Github repostitory.[1]

3 Metrics

For evaluation, we use the automatic metric ChrF++ (Popović, 2017) as implemented in SACREBLEU (Post, 2018). It is an overlap-based metric at the character-level, which is adequate for our task since most languages are morphologically rich.

While teams are not required to submit a system for all languages, the final score for each submission (ChrF++ column in Table 3) is calculated by taking an average over all eleven languages; if there is no model output for a given language, the score is taken as 0.

4 Baselines and Submitted Systems

In this section, we describe the 2024 baseline systems and each team's approach. We present a summary of all approaches in Table 2.

4.1 Baselines

This year, we consider two different baselines, based on the strongest submissions of the previous edition of our shared task, shown to be competitive among each other. The overall winning team in the previous edition was Sheffield (Gow-Smith and Villegas, 2023). They exploited the knowledge from different distilled versions of NLLB (Costa-jussà et al., 2022), a large pretrained model. We use their Submission 3, which chooses a single checkpoint with best average ChrF across all languages.

We also include Helsinki-NLP's Submission 6 (De Gibert et al., 2023), given that it outperforms the previous system on several language pairs. Their winning model is a multilingual one-to-many system, pretrained on Spanish–English data.

4.2 Submitted Systems

BSC The BSC team submitted systems for two languages: Quechua and Guarani, and followed the

[1] `https://github.com/AmericasNLP/americasnlp2024/tree/master/ST1_MachineTranslation`

Team	Models	Data	Overview
BSC Gilabert et al. (2024)	• NLLB-200 (1.3B)	• Length-based data filtering • Train set deduplication • Embedding-based sentence similarity	• Multilingual and bilingual fine-tuning of NLLB-200 • Low-Rank Adaptation (LoRA; 15% trainable params.) and full finetuning achieves
NordicAlps Attieh et al. (2024)	• From-scratch transformer encoder–decoder models	Various tokenizations: • Byte-level BPE • SentencePiece • Redundancy-driven tokenization	2 stage training: • First focus on Spanish-English data • Second, reduce Spanish-English to 50% with the other 50% sampled to equal amounts from the 11 TGT languages
DC_DMV DeGenaro and Lupicki (2024)	• NLLB-200 (600M, distil.) • State-space model (Mamba) from scratch	• Partition data into three stages, with deduplication	• Fully fine-tune a distilled NLLB-200 model using two data stages • Train a 3-layered Mamba network from scratch followed by a language model head using three data stages
Edinburgh Iyer et al. (2024)	• Llama-2 (7B) • Mistral (7B) • MaLA-500	• Collect additional data through OCR • Grammar and Education books, Scientific Papers, Dictionaries, and Books as sources	• Fine-tune LLama-2, Mistral and MaLA-500 models using a 2-stage training • LoRA fine-tuning with monolingual data first, then continue with instruction tuning • Regularize outputs using model averaging of the 4 last checkpoints

Table 2: Summary overview of each team's approach.

prior year's baseline approach of finetuning NLLB-200. In addition to the data provided by the organizers, the team collected new data from multiple sources, including the Monolingual-Quechua-IIC dataset (Zevallos et al., 2022), Flores-200 (Team et al., 2022), and other online datasets.[2] After collection, the data is cleaned in a multi-step process to remove duplicates and filter sentences. In the first step, sentences with more than 150 tokens and sentence pairs with a length ratio greater than 3 are removed. Next, various libraries are used to further clean the data, including Bifixer (Ramırez-Sanchez and Zaragoza-Bernabeu, 2020) and NLPDedup.[3] Finally, an embedding-based approach is used to calculate similarities between the source and targets side of a sentence pair; similarity scores are used with various thresholds to determine the final training examples.

NLLB is finetuned separately for each target language, and parallel sentences between each target and English, Portuguese, and Spanish are used. Two model sizes are considered: the 3.3B and 1.3B parameter version. Interestingly, the larger model only shows improvements for Quechua while performance decreases for Guarani; this relationship depends on the finetuning method used. Increasing the similarity score threshold offers better performance up to a point, after which performance begins to decrease, likely due to the greatly reduced amount of available data for finetuning. Overall, the best performance is found by using NLLB 1.3B with full finetuning for Guarani, improving over the prior best model by 1.91 ChrF++. For Quechua, NLLB 1.3B + LoRA (Hu et al., 2021) finetuning improves over the prior best score by 4.2 ChrF++. For these two languages, both systems achieved the highest performance across all submitted systems in this year's shared task.

NordicAlps The NordicAlps team submitted systems for all eleven languages of the shared task, building on the Helsinki system (De Gibert et al., 2023) from the 2023 shared task. The final models are one-to-many, trained to output translations in any of the competition languages as well as English. Target language tags are used to specify the output language. Data used is similar to the prior year's system, but this year's submission does not include Bible data. Preprocessing steps include whitespace normalization, Unicode normalization, and punctuation tokenization; these steps were implemented using the Moses tokenizer as well as through handwritten rules. The models do not make use of additional meta-data tags describing the language variant or quality on the input side. Of the three submitted systems, the main difference lies in the tokenization: a traditional byte-level BPE tokenization, SentencePiece tokenization, and `BPE-MR` tokenization, which consists of a BPE subword tokenizer trained using only 300 merges. `BPE-MR` tokenization is motivated by prior work on text compression through tokenization, and the finding that monolingual text can be compressed optimally using a small number of merge operations. Model training is carried out in stages, with the first stage covering a high-resource language pair (Spanish–English), and the second stage introducing more Indigenous language pairs (up to 50% of the examples used for training). Of the three sub-

[2] https://huggingface.co/datasets/somosnlp-hackathon-2022/spanish-to-quechua

[3] https://github.com/saattrupdan/NLPDedup

Rank	Team	Ver.	Count	Tot. BLEU	Tot. ChrF	Tot. ChrF++	Avg. BLEU	Avg. ChrF	Avg. ChrF++	BLEU	ChrF	**ChrF++**
1	NordicsAlps	1	11	55.48	321.69	287.60	5.04	29.24	26.15	5.04	29.24	26.15
2	DC_DMV	4	11	40.14	288.67	256.51	3.65	26.24	23.32	3.65	26.24	23.32
3	DC_DMV	3	11	39.63	287.64	255.49	3.60	26.15	23.23	3.60	26.15	23.23
4	DC_DMV	1	11	38.97	284.66	252.38	3.54	25.88	22.94	3.54	25.88	22.94
5	DC_DMV	5	11	37.84	284.26	252.21	3.44	25.84	22.93	3.44	25.84	22.93
6	DC_DMV	6	11	37.95	284.04	251.77	3.45	25.82	22.89	3.45	25.82	22.89
7	DC_DMV	2	11	34.15	272.59	243.83	3.10	24.78	22.17	3.10	24.78	22.17
8	NordicsAlps	2	11	27.28	265.46	232.41	2.48	24.13	21.13	2.48	24.13	21.13
9	UEdin	3	11	23.41	236.36	208.56	2.13	21.49	18.96	2.13	21.49	18.96
10	UEdin	1	11	24.04	235.42	208.34	2.19	21.40	18.94	2.19	21.40	18.94
11	UEdin	2	11	19.62	224.50	198.44	1.78	20.41	18.04	1.78	20.41	18.04
12	NordicsAlps	3	11	18.03	195.03	171.81	1.64	17.73	15.62	1.64	17.73	15.62
13	Z-AGI_Labs	1	4	8.35	103.03	87.32	2.09	25.76	21.83	0.76	9.37	7.94
14	DC_DMV	9	11	2.08	96.67	83.69	0.19	8.79	7.61	0.19	8.79	7.61
15	BSC	3	2	16.48	85.68	76.95	8.24	42.84	38.47	1.50	7.79	7.00
16	BSC	4	2	16.10	84.56	75.83	8.05	42.28	37.91	1.46	7.69	6.89
17	BSC	2	2	16.09	84.56	75.73	8.04	42.28	37.86	1.46	7.69	6.88
18	BSC	1	2	15.89	84.42	75.63	7.95	42.21	37.82	1.44	7.67	6.88
19	BSC	5	1	11.53	38.37	35.73	11.53	38.37	35.73	1.05	3.49	3.25
20	ND-NAIST	1	1	2.60	38.51	32.88	2.60	38.51	32.88	0.24	3.50	2.99

Table 3: Main ranking of all submitted systems. COUNT denotes the number of languages a particular system was submitted for, with the AVG.* columns showing the average metric score across submitted systems. The final three columns represent the average over all 11 languages of the shared task, with CHRF++ being used to calculate the overall ranking.

missions, the model using `BPE-MR` tokenization offered the best performance and achieved the best result for 5 of the shared task languages, and 2nd for 2 other languages.

DC_DMV The DC_DMV team submitted a system for each of the eleven languages, and followed two main approaches: finetuning a single version of the distilled 600m version of NLLB-200 for all the languages, as well as using a state-space model based on the Mamba architecture (Gu and Dao, 2023). Similar to the BSC team, duplicate examples are filtered, and the data is split into mutually exclusive stages. Stage 1 contains the largest set of data with over 700k examples, while Stages 2 and 3 have 100k and 200k examples, respectively. For the NLLB approach, the model is fully finetuned using data from the latter two stages, and the various submitted systems following this approach differ in the amount of training done using data from each stage. For the Mamba approach, a model is trained from scratch using all available data. While this approach did not yield strong results, likely due to the lack of pretraining, an NLLB-based submission achieved the best result across all submitted systems for Aymara, Shipibo-Konibo, and Rarámuri, while a different NLLB model achieved the best results for Bribri.

University of Edinburgh The University of Edinburgh participated with three system submissions for each of the eleven languages. These are the best performing systems in a series of experiments where the authors explore finetuning three well-known open-source LLMs: Llama-2 7B (Touvron et al., 2023), Mistral 7B (Jiang et al., 2023) and MaLA-500 (Lin et al., 2024). The finetuning consists of a two-stage training process employing Low-Rank Adaptation (LoRA) (Hu et al., 2021) and instruction tuning. In a nutshell, the first stage consists of finetuning LoRA adapters by continued pretraining on the LLM monolingual data, to adapt the models to specific linguistic features of each of the target languages. This setup includes using diverse data sets such as MADLAD-400 (Kudugunta et al., 2023) and Glot500 (ImaniGooghari et al., 2023). The second stage focuses on instruction tuning where models are finetuned using a combination of human-annotated and synthetic cross-lingual data, which helps improve the models' efficiency in real translation tasks. Furthermore, the authors explore *n*-last checkpoint averaging, with different beam search, and sampling setups to boost model performance at inference time.

5 Results

The overall ranking for the shared task can be found in Table 3, and the best per-language performance for each team can be found in Table 4. The full results for all submissions and teams can be found in Table 6.

Lang. Team	AYM	BZD	CNI	CTP	GN	HCH	NAH	OTO	QUY	SHP	TAR
Baseline - Helsinki	29.36	23.47	**24.92**	29.84	37.02	**28.67**	22.78	**13.32**	28.81	**30.21**	16.98
Baseline - Sheffield	**31.84**	**25.58**	24.76	37.05	35.76	28.28	**23.28**	12.87	34.01	30.06	16.25
BSC	-	-	-	-	**38.93**	-	-	-	**38.21**	-	-
DC_DMV	30.97	23.47	22.98	16.52	33.31	26.46	21.63	12.63	36.02	29.37	**17.03**
ND-NAIST	-	-	-	-	-	-	-	-	32.88	-	-
NordicsAlps	29.39	23.32	23.20	**37.38**	36.23	27.64	22.87	12.98	32.98	27.04	14.57
UEdin	21.89	16.54	14.82	20.70	29.20	24.41	18.98	9.19	25.23	22.86	9.65
Z-AGI_Labs	11.89	-	22.65	-	-	-	21.71	-	31.07	-	-

Table 4: The best ChrF++ scores for each team (across all submitted systems) across all languages. Bold values represent the best performing system overall, while underlined values are the best performing submission to this year's shared task.

The first place in the shared task, across all eleven language pairs, is awarded to the NordicAlps team (Submission 1). Their overall score significantly surpasses those of the second and third place teams, DC_DMV and UEdin, respectively. Notably, only three of the six teams submit entries for all eleven languages.

NordicAlps secures the top performance on five language pairs (Spanish to Asháninka, Chatino, Wizarika, Nahuatl, and Otomí), although they only exceed the baseline for Chatino. Similarly, the second-ranked team, DC_DMV, leads for four language pairs (Spanish to Aymara, Bribri, Shipibo-Konibo, and Rarámuri) but surpasses the baselines solely for Rarámuri. These results highlight the importance of meticulous pipeline design for data preprocessing and segmentation, as implemented by NordicAlps and the use of large multilingual models (NLLB) for finetuning, as employed by DC_DMV, for achieving robust results across most language pairs.

Finally, the BSC team, which participates for only two language pairs, Spanish to Guarani and Quechua, achieves the highest performance on both, surpassing the established baselines. Their strategic focus on finetuning a large multilingual model (NLLB) and gathering new data for these languages is key to their success.

6 Human Evaluation

Following prior AmericasNLP shared tasks (Mager et al., 2021; Ebrahimi et al., 2023), we also conduct a human evaluation of system outputs, focusing on Bribri.

6.1 Quantitative Analysis

As the test set has remained consistent across these competitions, we extend the prior evaluation using the best performing system from this year's shared task: Submission 4 by DC_DMV (DeGenaro and Lupicki, 2024). We consider the same 50 test inputs as in the prior analysis for this experiment, and a speaker of Bribri rates the system output on two axes: meaning and fluency. We consider a 5-point scale for evaluation, with a score of 5 being the best, and present results in Figure 1.

Similar to the pattern shown by the automatic metrics, we see a decrease in the perceived quality of translations from the best 2024 system as compared to the baseline (Gow-Smith and Villegas, 2023); i.e., scores suffer more, with a larger proportion being rated with a score of 1. For both metrics, scores of 5 are non-existent, showing a decrease in top-end performance as well. To further gain insights into the errors, we qualitatively look at the system outputs from the best 2024 system.

6.2 Qualitative Analysis

Table 5 shows examples of Bribri sentences translated by the best performing submission, organized by their score for meaning. The sentence with a score of 4 is readable and the original meaning is understandable, but there are parts that are not quite correct. In this example, "*Yes, you know she was great*", the hypothesis is very good, but it has at least one spelling mistake (**ujche̱n* instead of *ujché̱n* for "*it's known*"), and the word '*good*', *bua'*, is missing the intensifier {-ë} that it would need in order to become *bua'ë* '*great*'. In the case of the

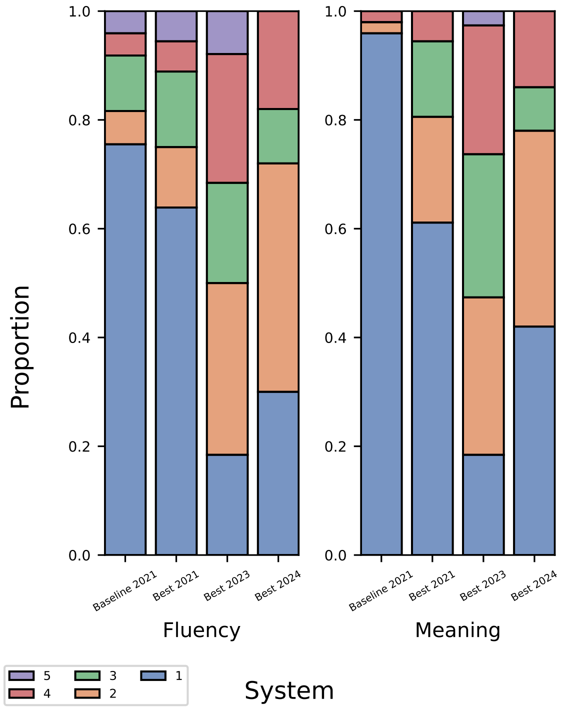

Figure 1: Results of the human evaluation for Bribri. Figure presents the proportion of evaluated example for each rating classification, with 1 representing the lowest quality and 5 representing the best. Values for the Baseline 2021, Best 2021, and Best 2023 systems are taken from (Ebrahimi et al., 2023).

sentence with the score 3, the words in the hypothesis still allow for an understanding of the meaning, but there are more mistakes. The example "*Hm, afterwards we moved to a new house*", has at least one spelling mistake: **pâ̱a̱li̱* instead of one of the other documented spellings of '*new*', for example *pà̱a̱li̱* or *pá̱li̱*. More importantly, it has a reflexive pronoun *e̱'* which does not belong in the sentence, and the verb is missing the plural marker {-yal} in the verb *mì̱ne̱yal* 'went.PL'.

The remaining hypotheses from Table 5 have more significant issues in their meanings. The example for meaning score 2, "*I spoke to Ramona again*", has some words correct, but there are errors and entire components missing. The translation is missing the postposition *ta̱* '*with*', which would be necessary to link the oblique argument '*Ramona*' to be verb *ujté* 'spoke'. It is possible that the system hallucinated the word *ta̱ma̱lé̱*, which resembles the word *ta̱má̱li* '*cuajiniquil fruit, Inga spuria*' because the word starts with the same letters as the postposition *ta̱*. But, in doing so, the system changed the meaning of the translation. A factor that might contribute to the hallucination is that there is an iterative morpheme, {-ma̱le̱}, which can mean '*again*' when it is attached to verbs (e.g. *ie déma̱le̱* '*he came again*' (Constenla et al., 2004, 119)). Unfortunately this morpheme is only found attached to verbs, not to postpositions,[4] and this makes the system hypothesis more difficult to understand.

Finally, the example for meaning score 1 can be translated, in its gold-standard version, as "*I am finishing with my project for next week*". The hypothesis produced by the system can be translated approximately as "*I am working*[sic], *finish, other*[sic] *weapon*". The verbs in the Bribri version are not connected properly, and the meaning of '*week*' is not present in the translation. Moreover, the system hallucinated the word *mó̱köl* '*weapon, rifle*', and it used the wrong numerical classifier to describe the rifle, **ië̀k* '*another* [round] *one*', when it should have used the classifier for long objects (e.g. *rifles*): *ië̀töm*. These errors combined make it so that the meaning of the original sentence cannot be inferred from the system's translation.

In summary, while we have made considerable progress as a community in the translation of Indigenous languages of the Americas, there is still much work ahead of us, both in terms of data collection and algorithm development.

7 Future Directions

In this section, we briefly discuss several possible future directions for the AmericasNLP shared task, given the results from the current as well as prior competitions.

Evaluation Data One bottleneck in the advancement of language technologies for low-resource, and particularly Indigenous, languages is the availability of evaluation data. High quality, gold standard data in target low-resource languages supports many important roles in the NLP research pipeline. First, and most importantly, it is the single resource which is necessary for experimentation; without held out data for evaluation, there cannot be any idea of how well a system performs for a given language. Second, the domain and source of data is important, as, over time, models are created to perform best on the data they are evaluated on. Particularly for low-resource languages, where there may not be great diversity in available data, it becomes vital to consider what data is used for evalu-

[4] There is a Bribri iterative morpheme, {-ne̱}, which can be attached to adverbs and verbs, but it has not been observed with postpositions either.

MS	Bribri	English translation
4	Tö́, be’ é̱n a̱ i̱à̱na̱ tö ie’ dör bua’ë. Tö́, be’ wa̱ i ujche̱n tö ie’ bák bua’.	Yes, you know she was great. Yes, you know[sic] that she was good.
3	Hum, ukö̀ki̱ sa’ mì̱ne̱yal ù pá̱li̱ a̱. Um, e’ ukö̀ki̱ sa’ e̱’ mì̱ne̱ ù pâ̱a̱li̱ a̱.	Hm, afterwards we moved to a new house. Hm, after that we went us to a new[sic] house.
2	Ramona ta̱ ye’ ujté skàne̱ Ye’ ujté Ramona ta̱ma̱lé̱.	I spoke to Ramona again. I spoke, Ramona, [cuajiniquil] fruit [sic].
1	Ye’ tso’ ka̱nè̱ ma̱ú̱k è̱we̱wa semana ië̀t wa. Ye’ tso’ kanè̱balök è̱nu̱k mó̱köl ië̀k.	I am finishing with my project for next week. I am working[sic]. Finish. Other[sic] weapon.

Table 5: Examples of Bribri sentences for each of the meaning scores (MS), accompanied by their translations in English. The first sentence is the gold standard, and the second sentence is the hypothesis by the best performing system.

ation. Future shared tasks should strive to continue creating new evaluation sets, both for currently supported languages (in order to increase diversity) as well as for new languages. Evaluation sets which contain data which is relevant to speakers and contain minimal biases increases the chances that good performance on the evaluation set is correlated with good real-world performance.

Additional Training Data This iteration of the shared task marks the first where performance did not increase for the majority of languages in the shared task. Of the four languages which did see improvements, two are relatively high-resource and have recently been included in large pretrained models (Costa-jussà et al., 2022). As such, additional data for training likely plays a large role in improving the performance for these languages. While teams continue to find new digital data for training, other non-digital sources may need to be considered for future systems.

Language Identification One of the main bottlenecks for gathering additional data is that every process of collecting resources from online sources starts with a good language identifier. Investing efforts into developing a language identification system for the shared task languages could boost the collection of additional training data.

New Language Pairs The performance of low-resource language pairs in multilingual MT models can benefit from incorporating additional data from other language pairs. Furthermore, our goal is to expand the scope of our shared task in future editions to include more underserved languages of the Americas. To achieve this, we plan to engage more researchers who have developed and published resources for the Indigenous languages of the Americas, both at our workshop and in other venues.

8 Conclusion

In this work, we present the results of the AmericasNLP 2024 Shared Task on Machine Translation. Overall, 6 teams participated in the shared task, and submitted a combined 157 submissions across all eleven supported languages. Prior to the start of the competition, the organizers provided two strong baselines and a training data set which includes data collected from prior submissions. While there were improvements for four languages in this year's shared task, the majority of languages did not see any performance gains over the baselines, which were the strongest systems from 2023.

Acknowledgments

We would like to thank all teams for participating in the shared task. Additionally, we would like to thank the University of Sheffield team for providing the necessary documentation for running the baseline system.

References

Joseph Attieh, Zachary William Hopton, Yves Scherrer, and Tanja Samardzic. 2024. System description of the nordicsalps submission to the americasnlp 2024 machine translation shared task. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, Mexico City, Mexico. Association for Computational Linguistics.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Adolfo Constenla, Feliciano Elizondo, and Francisco Pereira. 2004. *Curso Básico de Bribri*. Editorial de la Universidad de Costa Rica.

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*.

Ona De Gibert, Raúl Vázquez, Mikko Aulamo, Yves Scherrer, Sami Virpioja, and Jörg Tiedemann. 2023. Four approaches to low-resource multilingual nmt: The helsinki submission to the americasnlp 2023 shared task. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 177–191.

Dan DeGenaro and Tom Lupicki. 2024. Experiments in mamba sequence modeling and nllb-200 fine-tuning for low resource multilingual machine translation. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, Mexico City, Mexico. Association for Computational Linguistics.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2024. *Ethnologue: Languages of the world.*, 25 edition. SIL International, Dallas, Texas.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios Gonzales, Ivan Meza-Ruiz, et al. 2022. Americasnli: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6279–6299.

Abteen Ebrahimi, Manuel Mager, Shruti Rijhwani, Enora Rice, Arturo Oncevay, Claudia Baltazar, María Cortés, Cynthia Montaño, John E. Ortega, Rolando Coto-solano, Hilaria Cruz, Alexis Palmer, and Katharina Kann. 2023. Findings of the AmericasNLP 2023 Shared Task on Machine Translation into Indigenous Languages. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 206–219, Toronto, Canada. Association for Computational Linguistics.

Javier García Gilabert, Aleix Sant Savall Carlos Escolano, Francesca De Luca Fornaciari, Audrey Mash, and Maite Melero. 2024. Bsc submission to the americasnlp 2024 shared task. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, Mexico City, Mexico. Association for Computational Linguistics.

Edward Gow-Smith and Danae Sánchez Villegas. 2023. Sheffield's Submission to the AmericasNLP Shared Task on Machine Translation into Indigenous Languages.

Albert Gu and Tri Dao. 2023. Mamba: Linear-Time Sequence Modeling with Selective State Spaces.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Ayyoob ImaniGooghari, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling Multilingual Corpora and Language Models to 500 Languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.

Vivek Iyer, Bhavitvya Malik, Wenhao Zhu, Pavel Stepachev, Pinzhen Chen, Barry Haddow, and Alexandra Birch. 2024. Exploring very low-resource translation with llms: The university of edinburgh's submission to americasnlp 2024 translation task. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, Mexico City, Mexico. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A. Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. MADLAD-400: A Multilingual And Document-Level Large Audited Dataset.

Peiqin Lin, Shaoxiong Ji, Jörg Tiedemann, André F. T. Martins, and Hinrich Schütze. 2024. MaLA-500: Massive Language Adaptation of Large Language Models.

Manuel Mager, Arturo Oncevay, Abteen Ebrahimi, John Ortega, Annette Rios, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo Giménez-Lugo, Ricardo Ramos, Ivan Vladimir Meza Ruiz, Rolando Coto-Solano, Alexis Palmer, Elisabeth Mager-Hois, Vishrav Chaudhary, Graham Neubig, Ngoc Thang Vu, and Katharina Kann. 2021. Findings of the AmericasNLP 2021 shared task on open machine translation for indigenous languages of the Americas. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 202–217, Online. Association for Computational Linguistics.

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

Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, page 186. Association for Computational Linguistics.

Gema Ramırez-Sanchez and Jaume Zaragoza-Bernabeu. 2020. Bifixer and Bicleaner: Two open-source tools to clean your parallel data.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models.

Rodolfo Zevallos, John Ortega, William Chen, Richard Castro, Núria Bel, Cesar Toshio, Renzo Venturas, Hilario Aradiel, and Nelsi Melgarejo. 2022. Introducing QuBERT: A Large Monolingual Corpus and BERT Model for Southern Quechua. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 1–13, Hybrid. Association for Computational Linguistics.

Appendix

Lang.	Team	Ver.	BLEU	ChrF	ChrF++
aym	DC_DMV	2	3.49	35.43	30.97
aym	NordicsAlps	1	3.23	33.46	29.39
aym	DC_DMV	4	2.74	32.36	28.32
aym	DC_DMV	3	2.52	32.24	28.12
aym	DC_DMV	6	2.24	31.43	27.48
aym	DC_DMV	1	2.27	31.26	27.36
aym	DC_DMV	5	2.37	30.91	27.09
aym	NordicsAlps	2	1.99	30.37	26.37
aym	UEdin	3	1.13	25.14	21.89
aym	UEdin	1	1.14	24.94	21.77
aym	UEdin	2	1.06	24.56	21.37
aym	NordicsAlps	3	1.10	17.55	15.77
aym	Z-AGI_Labs	1	0.74	13.30	11.89
aym	DC_DMV	9	0.15	9.51	8.69
bzd	DC_DMV	4	4.84	22.23	23.47
bzd	DC_DMV	5	4.56	22.15	23.41
bzd	DC_DMV	3	4.63	22.02	23.32
bzd	NordicsAlps	1	5.00	22.27	23.32
bzd	DC_DMV	1	4.68	21.97	23.19
bzd	DC_DMV	6	4.75	21.99	23.15
bzd	DC_DMV	2	3.44	18.11	19.60
bzd	NordicsAlps	2	1.72	15.98	17.23
bzd	UEdin	1	2.21	15.43	16.54
bzd	UEdin	2	1.89	15.17	16.32
bzd	UEdin	3	1.75	14.53	15.56
bzd	NordicsAlps	3	1.39	13.17	12.24
bzd	DC_DMV	9	0.09	4.36	4.72
cni	NordicsAlps	1	2.41	27.76	23.20
cni	DC_DMV	6	3.49	26.15	22.98
cni	DC_DMV	3	3.56	26.05	22.87
cni	Z-AGI_Labs	1	3.22	26.75	22.65
cni	DC_DMV	5	3.41	25.63	22.53
cni	DC_DMV	4	3.51	25.53	22.46
cni	DC_DMV	1	3.56	25.48	22.44
cni	DC_DMV	2	3.52	22.13	19.89
cni	NordicsAlps	2	0.06	20.13	15.45
cni	NordicsAlps	3	1.68	17.30	15.23
cni	UEdin	1	0.41	17.54	14.82
cni	UEdin	3	0.43	17.08	14.50
cni	UEdin	2	0.37	16.26	13.68
cni	DC_DMV	9	0.14	11.83	9.81
ctp	NordicsAlps	1	13.44	40.37	37.38
ctp	NordicsAlps	2	4.65	26.61	23.64
ctp	UEdin	2	4.30	23.01	20.70
ctp	UEdin	1	3.35	19.50	17.66
ctp	UEdin	3	3.38	19.50	17.57
ctp	DC_DMV	1	1.73	20.58	16.52
ctp	DC_DMV	3	1.68	20.18	16.17
ctp	DC_DMV	5	1.68	20.06	16.11
ctp	DC_DMV	6	1.75	19.90	16.04
ctp	DC_DMV	4	1.74	19.59	15.78
ctp	NordicsAlps	3	1.78	14.97	12.96
ctp	DC_DMV	2	0.96	9.72	8.06
ctp	DC_DMV	9	0.00	3.38	2.62
gn	BSC	3	12.04	41.81	38.93
gn	BSC	4	11.28	40.66	37.64
gn	BSC	2	11.37	40.69	37.63
gn	BSC	1	11.04	40.38	37.42
gn	NordicsAlps	1	8.82	39.36	36.23
gn	BSC	5	11.53	38.37	35.73
gn	DC_DMV	2	5.46	36.78	33.31
gn	DC_DMV	3	6.30	35.72	32.58
gn	DC_DMV	4	6.42	35.51	32.44
gn	NordicsAlps	2	6.81	35.23	32.32
gn	DC_DMV	6	5.82	34.69	31.66
gn	DC_DMV	1	5.97	34.66	31.58
gn	DC_DMV	5	5.66	34.18	31.22
gn	UEdin	1	3.38	32.22	29.20
gn	UEdin	3	3.21	32.31	29.13
gn	UEdin	2	1.78	27.61	24.61
gn	NordicsAlps	3	1.60	16.11	14.80
gn	DC_DMV	9	0.32	10.10	8.91
hch	NordicsAlps	1	10.08	31.13	27.64
hch	DC_DMV	1	9.62	29.83	26.46
hch	DC_DMV	4	8.51	29.54	26.23
hch	DC_DMV	5	8.64	29.21	25.97
hch	DC_DMV	6	8.83	28.95	25.66
hch	DC_DMV	3	8.85	28.75	25.60
hch	UEdin	1	9.87	27.40	24.41
hch	UEdin	3	9.60	27.50	24.37
hch	NordicsAlps	2	6.46	26.92	23.47
hch	UEdin	2	7.03	24.51	22.03
hch	DC_DMV	2	3.29	22.36	19.56
hch	NordicsAlps	3	1.35	18.43	15.97
hch	DC_DMV	9	0.49	8.10	7.12
nah	NordicsAlps	1	2.30	26.91	22.87
nah	Z-AGI_Labs	1	1.09	26.29	21.71
nah	DC_DMV	1	1.79	25.58	21.63
nah	DC_DMV	4	1.73	25.41	21.44
nah	DC_DMV	5	1.86	25.35	21.43
nah	DC_DMV	6	1.78	25.24	21.41
nah	DC_DMV	3	1.85	24.84	21.07
nah	NordicsAlps	2	1.52	24.84	20.77
nah	UEdin	3	0.44	22.86	18.98
nah	DC_DMV	2	1.75	21.69	18.52
nah	UEdin	1	0.48	21.75	18.12
nah	UEdin	2	0.37	20.78	17.21
nah	NordicsAlps	3	1.64	17.08	14.57
nah	DC_DMV	9	0.12	13.14	10.46
oto	NordicsAlps	1	1.42	14.95	12.98
oto	DC_DMV	1	1.55	14.61	12.63
oto	DC_DMV	3	1.66	14.30	12.42
oto	DC_DMV	4	1.50	14.34	12.42
oto	DC_DMV	5	1.52	14.29	12.40
oto	DC_DMV	6	1.36	14.14	12.20
oto	NordicsAlps	2	0.20	13.80	11.63
oto	DC_DMV	2	1.46	13.05	11.50
oto	NordicsAlps	3	1.41	13.14	11.22
oto	UEdin	3	0.44	10.87	9.19
oto	UEdin	1	0.43	10.56	8.91
oto	UEdin	2	0.21	9.32	7.81
oto	DC_DMV	9	0.04	4.39	3.63
quy	BSC	1	4.85	44.04	38.21
quy	BSC	4	4.83	43.91	38.19
quy	BSC	2	4.72	43.87	38.10
quy	BSC	3	4.44	43.86	38.02
quy	DC_DMV	2	5.41	41.43	36.02
quy	DC_DMV	4	4.32	39.67	34.29
quy	DC_DMV	3	4.13	39.49	34.08
quy	DC_DMV	5	3.91	39.33	33.94
quy	DC_DMV	1	4.01	39.24	33.91
quy	DC_DMV	6	4.05	38.95	33.56
quy	NordicsAlps	1	4.08	37.92	32.98
quy	ND-NAIST	1	2.60	38.51	32.88
quy	Z-AGI_Labs	1	3.29	36.69	31.07

Lang.	Team	Ver.	BLEU	ChrF	ChrF++
quy	NordicsAlps	2	2.65	33.36	28.81
quy	UEdin	1	1.32	29.54	25.23
quy	NordicsAlps	3	2.77	28.99	25.15
quy	UEdin	3	1.31	29.37	25.04
quy	UEdin	2	0.94	26.69	22.77
quy	DC_DMV	9	0.40	13.08	11.42
shp	DC_DMV	2	4.45	32.95	29.37
shp	NordicsAlps	1	4.14	30.55	27.04
shp	DC_DMV	4	3.90	27.77	24.74
shp	DC_DMV	3	3.44	26.86	23.84
shp	DC_DMV	5	3.17	26.58	23.59
shp	DC_DMV	6	3.07	25.91	23.05
shp	UEdin	3	1.55	25.90	22.86
shp	UEdin	2	1.56	25.52	22.43
shp	DC_DMV	1	2.95	25.04	22.25
shp	NordicsAlps	2	1.09	25.68	22.20
shp	UEdin	1	1.34	25.08	22.04
shp	NordicsAlps	3	2.60	23.83	21.28
shp	DC_DMV	9	0.27	11.13	9.67
tar	DC_DMV	2	0.92	18.94	17.03
tar	DC_DMV	3	1.01	17.20	15.42
tar	DC_DMV	4	0.93	16.72	14.92
tar	DC_DMV	6	0.81	16.69	14.57
tar	NordicsAlps	1	0.55	17.03	14.57
tar	DC_DMV	5	1.04	16.58	14.51
tar	DC_DMV	1	0.86	16.41	14.39
tar	NordicsAlps	3	0.73	14.49	12.63
tar	NordicsAlps	2	0.12	12.54	10.53
tar	UEdin	1	0.11	11.46	9.65
tar	UEdin	2	0.11	11.07	9.49
tar	UEdin	3	0.15	11.32	9.48
tar	DC_DMV	9	0.07	7.65	6.64

Table 6: Full results of the shared task.