Findings from the Bambara - French Machine Translation Competition (BFMT 2023)

Ninoh Agostinho Da Silva[1], Tunde Oluwaseyi Ajayi[2], Alexander Antonov[3], Panga Azazia Kamate[4], Moussa Coulibaly[4], Mason Del Rio[5], Yacouba Diarra[4], Sebastian Diarra[4], Chris Emezue[6], Joel Hamilcaro[1], Christopher M. Homan[7], Alexander Most[8], Joseph Mwatukange[9], Peter Ohue[10], Michael Pham[11], Abdoulaye Sako[4], Sokhar Samb[12], Yaya Sy[1], Tharindu Cyril Weerasooriya[7], Yacine Zahidi[5], Sarah Luger[5*]

[1]Independent, previously Université Paris-Cité, [2]Insight Centre for Data Analytics, Masakhane, [3]Chuvash Language Laboratory, Yandex, [4]RobotsMali,[5]Orange Silicon Valley,[6]Mila Quebec AI Institute, Technical Universityof Munich, Lanfrica,[7]Rochester Institute of Technology,USA, [8]Montana State University,USA,[9]Meyabase Platforms,[10]University of Ibadan, Nigeria, [11]Swarthmore,[12]Dakar American University of Science & Technology, Senegal

[5*]sarahluger@gmail.com

Abstract

Orange Silicon Valley hosted a low-resource machine translation (MT) competition with monetary prizes. The goals of the competition were to raise awareness of the challenges in the low-resource MT domain, improve MT algorithms and data strategies, and support MT expertise development in the regions where people speak Bambara and other low-resource languages. The participants built Bambara to French and French to Bambara machine translation systems using data provided by the organizers and additional data resources shared amongst the competitors. This paper details each team's different approaches and motivation for ongoing work in Bambara and the broader low-resource machine translation domain.

1 BFMT 2023 - Competition Introduction

Orange Silicon Valley, hosted the "Bambara-French Machine Translation Competition 2023" (BFMT 2023) a low-resource machine translation (MT) competition that ended on February 15, 2023. The competition was launched on December 15, 2022. Participants had access to a Github repository with a training dataset of parallel French-Bambara aligned sentences[1]. The participants were also invited into a Slack community to share their approaches and data. An additional development dataset was provided to the teams and fewer than 48 hours before the submission deadline, a test dataset was released for generating text output to be sent to the competition organizers to evaluate translation performance using BLEU scores (Post, 2018).

The goals of the competition were to improve not only French to Bambara and Bambara to French automated translation systems, but also support a transparent and collaborative community to work on these and other language pairs, especially those (low-resource) languages spoken by West Africans. 50 people joined the online community and fourteen people competed in 6 teams. The teams contained participants from Mali, Senegal, Namibia, Nigeria, Ireland, Germany, Russia, Spain, France, the US, and the UK. Many of the participants speak or have working knowledge of a "low-resource language" or a language that does not have the digital resources that support highly accurate Natural Language Processing tool development.

Bambara is a tonal language with a rich morphology spoken by five million people as a first language and approximately 15 million people as a second language. Approximately 30–40 million people speak a language in the Mande language family, to which Bambara belongs (Lewis et al., 2014).

A predominately oral language, several competing writing systems have developed. A majority of Bambara speakers have not been taught to read or write in a standard format. Bambara's standardization is evolving and this poses challenges to automated text processing such as machine translation (Vydrin et al., 2022).

Additional contest information may be found in both French and English on the Orange Silicon Valley website[2].

[1]The dataset is available to share on request through the corresponding author.

[2]https://siliconvalley.orange.com/en/bambara-french-machine-translation-competition/

Proceedings of the The Sixth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2023), pages 110–122
May 6, 2023 ©2023 Association for Computational Linguistics

2 Background

Current state-of-the-art low-resource MT is surveyed in Haddow et al. (2022). Google Translate has integrated more low-resource languages into their language library sharing innovations as detailed in blog posts (Venugopal, 2010; Benjamin, 2019).

MT for the Bambara - French language pair has been explored in recent years in Akhbardeh et al. (2021); Tapo et al. (2020); Leventhal et al. (2020). This work is in part motivated by an increased financial and cultural focus on bringing machine learning to the Sahel region (Diarra and Leventhal, 2020).

2.1 Evaluation

MT can be evaluated by automated and manual methods. In this competition, we used automated tools to evaluate the closeness of translations to a gold standard. We use BLEU scores with sacreBLEU (Papineni et al., 2002; Post, 2018) for automated evaluation. Human evaluation would have been performed if the difference between the Team scores was less than 1 point in BLEU scale. The results were not close. Thus, we proceeded with using BLEU scores with sacreBLEU.

2.2 Datasets

The organizers provided a training dataset of aligned parallel Bambara - French sentences from the medical and dictionary domains as described in the original data collection (Akhbardeh et al., 2021). Each line in the dataset corresponds to a single sentence. The characteristics of the dataset provided by the organizers is shown in Table 1. In addition to the competition data, all participants were encouraged to gather, utilize, and share additional resources with other members of the competition community. The additional datasets used in the competition are shown in Table 2, with the Bayelemabaga (Vydrin et al., 2022) dataset being notable for the amount of additional data it gave to participants.

2.3 Baseline

The competition guidelines did not provide any baseline models nor baseline scores for the competition participants. The closest baseline to compare for this competition was from the findings of WMT21 (Akhbardeh et al., 2021), with BLEU scores of 1.32 for French to Bambara, and 3.62 for Bambara to French, using the Marian NMT (Junczys-Dowmunt et al., 2018) pre-trained model.

Data Split	Number of Sentences
Train	3,150
Dev	460
Test	460

Table 1: The characteristics of the dataset provided by the competition organizers.

Dataset	Teams
MAFAND (Adelani et al., 2022)	All Teams
NLLB-SEED (Team et al., 2022)	All Teams
FLORES (Goyal et al., 2022)	All Teams
BAYELEMABAGA (Vydrin et al., 2022)	All Teams
XP3 (Muennighoff et al., 2022a)	Yacine Zahidi
Wikipedia	Team Alpha

Table 2: Additional Bambara datasets used by the teams. Team Alpha use the Wikipedia dataset that is available through Wikimedia.org [3].

Technique	Reference
BART	(Lewis et al., 2019)
BLOOM-z 560M, mt0-small	(Muennighoff et al., 2022b)
byt5	(Xue et al., 2021a)
DeltaLM	(Ma et al., 2021)
HuggingFace	(Wolf et al., 2020)
LION optimizer	(Chen et al., 2023)
LoRA	(Hu et al., 2021)
M2M100 model	(Fan et al., 2020)
MarianNMT/Opus-MT	(Junczys-Dowmunt et al., 2018)
mt5	(Xue et al., 2021b)
NLLB model	Team et al., 2022
PEFT library	(Mangrulkar et al., 2022)
Sockeye	(Hieber et al., 2020)

Table 3: Techniques and models used by the teams.

2.4 Machine Translation Systems

Table 3 shows the different techniques and models used by the teams with transformer (Vaswani et al., 2017) and BERT models (Mishra et al., 2022; Sheshadri et al., 2023) inspiring much of the development.

3 Team-by-Team Machine Translation Findings from BFMT 2023

Six teams submitted system output that could be evaluated using sacreBLEU. Team Peter-Sokhar (Section 3.7) built an MT system but did not submit an output for scoring. Nonetheless, their findings from training and error analysis are included in this paper. In the following sections, each team first describes their methodology, then they describe their error analysis. See Table 2 for the datasets used by each team.

3.1 Team Alpha

We used an additional dataset from Wikipedia[4] which provided us with an extra 892 lines of data. Next, we made a list of MT models that contained Bambara and French in their dataset during pre-training. As a result, we started with the NLLB-200 (Team et al., 2022) pre-trained model. We fine-tuned both the 600M and the 1.3B (in order to test the impact of scaling on model capacity) parameter versions, from the Huggingface Hub. We found the NLLB model to be under-performing. Next, we switched to an M2M-100 (Fan et al., 2020) model after we discovered it had fine-tuned multilingual MT models separately for each language direction, which outperformed NLLB-200 (Adelani et al., 2022)

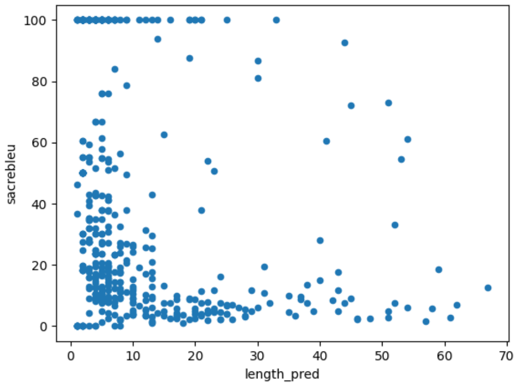

Figure 1: Scatterplot showing length of predicted sentences against sentence BLEU scores for FR → BAM.

To gain further insight into the challenges posed by certain sentence characteristics in our MT model, we conducted an analysis of the per-sentence BLEU scores plotted against the length of the predicted sentences. Initially, we postulated that our MT model would perform better with shorter sentences and perform worse with longer sentences. However, as illustrated in Figure 1, which presents a scatterplot of the lengths of the predicted sentence against their sentence BLEU scores, our model struggled even with shorter sentences. This led us to reconsider our hypothesis and explore the possibility that our model was underfitting. Next, we decided to investigate the potential benefits of implementing backtranslation.

[4] https://dumps.wikimedia.org

Algorithm 1 Team Alpha's Backtranslation Approach

$n\_epochs \leftarrow$ number of fine-tuning epochs
$D_{train} \leftarrow$ training dataset of French- Bambara parallel sentences
$D_{bam}^{wiki} \leftarrow$ 892 monolingual cleaned sentences from Wikipedia.
$D_{fr} \leftarrow$ dataset of French sentences only. For our case it was gathered by taking the French instances of D_{train}
$D_{bam} \leftarrow$ dataset of Bambara sentences only. For our case it was gathered by taking the Bambara instances of D_{train} and additional monolingual sentences from D_{bam}^{wiki}

$M_{fr\longrightarrow bam}^{0} \leftarrow$ fine-tuned MT model of (Adelani et al., 2022) for French $\longrightarrow$ Bambara

$M_{bam\longrightarrow fr}^{0} \leftarrow$ fine-tuned MT model of (Adelani et al., 2022) for Bambara $\longrightarrow$ French.

$D_{train}^{0} \leftarrow D_{train}$.
for $k \leftarrow [0, 1, 2...n]$ **do**
  $M_{fr\longrightarrow bam}^{k+1} \leftarrow$ fine-tune $M_{fr\longrightarrow bam}^{k}$ on D_{train}^{k} for n_epochs epochs.

  $M_{bam\longrightarrow fr}^{k+1} \leftarrow$ fine-tune $M_{bam\longrightarrow fr}^{k}$ on D_{train}^{k} for n_epochs epochs.

  $D_{bam}^{k} \leftarrow$ generated synthetic translations to Bambara from D_{fr} using $M_{fr\longrightarrow bam}^{k+1}$.

  $D_{fr}^{k} \leftarrow$ generated synthetic translations to French from D_{bam} using $M_{bam\longrightarrow fr}^{k+1}$.

  $D_{train}^{k+1} \leftarrow$ concatenated training dataset gotten from $D_{train}^{0} \cup \{D_{bam}^{k} \leftrightarrow D_{fr}\} \cup \{D_{fr}^{k} \leftrightarrow D_{bam}\}$
end for

3.1.1 Team Alpha's Backtranslation Approach

Several papers have highlighted the positive effect of backtranslation (Sennrich et al., 2016a; Poncelas et al., 2018; Zhang et al., 2020; Dossou and Emezue, 2020; Fan et al., 2020; Emezue and Dossou, 2021; Adelani et al., 2022; Team et al., 2022). Inspired by random online backtranslation (Zhang et al., 2020), we created our version, explained in Algorithm 1, to help our model better utilize the training dataset, and the 892 monolingual Bambara sentences from Wikipedia. Our approach, dubbed *Cyclic backtranslation* (Lam et al., 2021), would theoretically enable the model to leverage the available training and monolingual dataset by compelling the MT model for each direction, at each step k, to learn from a concatenation of the original training dataset, its synthetically generated sentences, and those generated by the MT model of the opposite direction in the previous step.

Despite its potential benefits, implementing backtranslation presented several challenges. First, it was a difficult process to set up, particularly in achieving a high degree of automation and reducing the need for human intervention. Secondly, it was computationally expensive and time-consuming, as each iteration of the backtranslation process involved working with three times more data than the previous iteration. Consequently, we were only able to complete one backtranslation successfully.

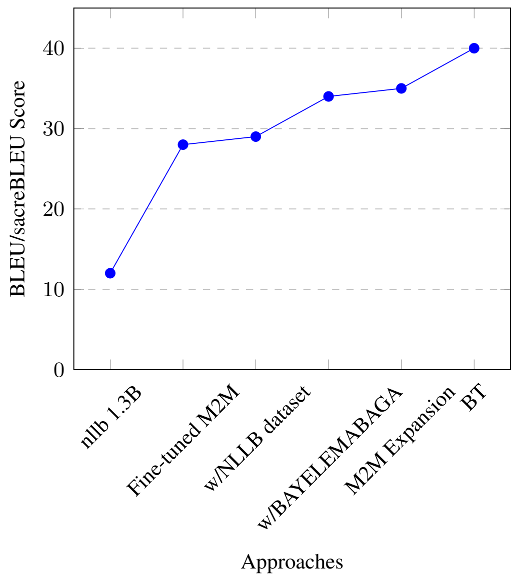

Figure 2: Timeline of Team Alpha efforts and BLEU score on dev set. The chart begins with our use of NLLB, switches to fine-tuned M2M, incorporates NLLB Seed dataset, then includes the BAYELEMABAGA dataset, and ends in our hypothetical performance using our cyclic backtranslation approach. The scores reported are for doing French → Bambara translation.

We included a potential impact in Figure 2 which shows the timeline of our activities and their corresponding evaluation results on the French → Bambara direction.

One of the major challenges facing machine translation for African languages is the limited availability of high-quality datasets (Nekoto et al., 2020; Caswell et al., 2021; Adelani et al., 2022). This became apparent in our study, where the use of the BAYELEMABAGA dataset resulted in a significant increase in the performance of our MT model. The scarcity of such resources highlights the need for continued efforts to develop and curate datasets for African languages, which could significantly improve the performance of machine translation models for African languages.

3.2 Team Most-Pham

We used a pre-trained MarianMT transformer model (Junczys-Dowmunt et al., 2018) which was pre-trained for Romance languages to English due to the non-existence of Bambara-French pre-trained weights for the MarianMT model. The model was then trained using a set of hyperparameters which were inspired by findings from Araabi and Monz (2020); Van Biljon et al. (2020) where the authors found the hyperparameters that would achieve the highest BLEU scores when dealing with low-resource languages. Our implementation was limited due to insufficient computing power (we were not able to increase attention heads without the GPU crashing during training).

We use the following set of hyperparameters; optimizer: adam, learning rate: $2e^{-5}$, beta 1:0.9, beta 2: 0.999, epsilon: $1e^{7}$, batch size: 64, and attention heads: 8.

3.2.1 Error analysis

Due to limited computing power, we were not able to fully train our MT model until convergence. It is plausible our model could have achieved higher accuracy or lower bias with more iterations of gradient descent. We also were not able to fine-tune our hyper-parameters as much as we would have liked.

In the seq2seq translation output, one word would get repeated multiple times back-to-back. This hallucination could be reduced by using a model that was pre-trained in French, so it would know from experience that French sentences do not normally include back-to-back repeated words.

There were words that appeared infrequently in the training set and were frequently mistranslated. With more time in this competition, this could have been alleviated with Byte Pair Encoding (BPE).

3.2.2 Discussion

While the existing literature suggests that Transformer models typically need a large training corpus to do well, our model suggests otherwise. With minor (out-of-the-box) modifications made to the architecture, the Transformer seq2seq model was still able to achieve a BLEU of 14.81 despite a limited training corpus, lack of a pre-trained Bambara model, computing power, and hyper-parameter tuning. In hindsight, we should have used a model that was pre-trained for Bambara to any Romance language, because it would be easier to learn Bambara to French if it had been pre-trained in Bambara to English, for example. We hypothesize that the difference between Bambara and the pre-trained data is very large, thereby making the model struggle to learn a different language with such a small dataset.

3.3 Team JYN

Our team had previously worked on MT tasks on languages such as French, Reunionese Creole, Portuguese, Umbundu, and Kimbundu, where we observed sub-optimal outcomes when training an autoregressive generative transformer model, either encoder-only or decoder-only, starting from scratch. Hence, for the given task, we wanted to use a Sequence to Sequence (seq2seq) model with prior training on the Bambara language. We evaluated different models of different sizes and with different number of training steps. We evaluated the following models on the development datasets: mt0-small, BLOOM-z 560M (Muennighoff et al., 2022b), NLLB 600M distilled, NLLB 1.3B, NLLB 1.3B distilled, and NLLB 3.3B (Team et al., 2022).

Upon evaluating the dev dataset, NLLB 600M distilled and NLLB 1.3B distilled exhibited superior performance. However, due to computational limitations even with our optimizations, training the NLLB 3B version would have been impossible. For an auto-regressive/instruction model, BLOOM-z exhibited more potential than mt0-small, and after two epochs, it produced acceptable scores. Nevertheless, it appears that general-purpose models of such small sizes do not rival specialized seq2seq models of similar dimensions, especially in a low-resource scenario.

We focused our scarce GPU hours to the two most promising models (NLLB 600M and NLLB 1.3B, which are both distilled models) and fine-tune them until the competition deadline. This provided an avenue to utilize and fine-tune distilled models. 1.3B distilled was better than not distilled models. Without fine-tuning, by using the default HuggingFace *generate* method, the 600M distilled model had a BLEU score of 19.8157 and 17.9217 for BAM to FR and FR to BAM, respectively. And the non-fine-tuned distilled 1.3B model had 24.5496 and 25.5610 for BAM to FR and FR to BAM, respectively. Both were tested on the dev corpus provided by the competition organizers. Table 4 shows the BLEU scores using different models and training steps, the latter indicating the amount of training a model should undergo.

Model size/Training steps	BAM → FR	FR → BAM
600M/3000 steps	21.7641	18.8674
600M/6000 steps	21.5270	**21.3773**
600M/9000 steps	21.3773	17.8374
1.3B/1500 steps	20.3349	17.8032
1.3B/3000 steps	18.6542	17.6243
1.3B/4500 steps	24.2556	19.3324
1.3B/6000 steps	25.3816	18.7743
1.3B/7500 steps	**26.0991**	18.1205

Table 4: BLEU Scores on development set (Team JYN), with increasing training steps showing a constant increase in translation for Bambara to French.

The hyperparameters used for fine-tuning the NLLB models are: Optimizer: Adafactor; Learning rate: $1e^{-04}$; Batch size (1.3B model): 4; Batch size (600M model): 10; Gradient acc. (1.3B model): 16; and Gradient acc. (600M model): 10.

3.3.1 Error Analysis

We made a challenging discovery during this competition. In the NLLB paper, the source and target sequences are fed to the model with this scheme: (src_sequence, src_lang) for the source sequence and (tgt_lang, tgt_sequence) for the target sequence. On the other hand, the NLLB tokenizer in the HuggingFace transformer tokenizes the pair of sequences as (src_squence, src_lang) and (tgt_sequence, tgt_lang). Once we fixed this issue, the sacreBLEU scores of our finetuned NLLB models started to improve, consistently with the decrease of the loss, and with the quality differences that we could observe. However, we discovered and fixed this issue less than 24 hours before the deadline, and we had lost quite a bit of time by trying other fixes. Considering French is our native language, and a member of our group has some understanding of Bambara, we were able to compare the outputs of the model to the targets of the development set. Prior this discovery, the BLEU scores of our fine-tuned models were not impressive and inconsistent with the steadily decreasing loss on the dev set, and our observations of the outputs. After this fix, the BLEU scores showed improvements, even when we did not resolve the difference in behaviour between the two translation directions. The Bambara to French translations got marginally better in terms of BLEU scores compared to the French to Bambara, which was dramatically worse than the base performance.

3.3.2 Discussion

For our next MT project, we would explore large language models (LLM). We believe it would be a good idea to investigate the performance of few-shot prompting on these LLMs, because we have seen that the most promising model is still very limited for languages like Bambara.

Since Bambara, like many languages, is primarily spoken, we will try speech-based approaches in future work. These approaches will potentially have more impact and be more useful to these communities, especially to those who cannot write in their languages.

3.4 Yacine Zahidi

For pre-trained models, we explored several models available on the HuggingFace Hub, including M2M-100 (Fan et al., 2020), NLLB (Team et al., 2022), mT5 (Xue et al., 2021b) and byt5 (Xue et al., 2021a) models each pre-trained by the Masakhane Organization (Nekoto et al., 2020). Each model was evaluated on the dev set provided by the organizers with respect to the BLEU score. The M2M-100 (Fan et al., 2020) was chosen as a starting point since it scored the highest. It is a 483 million parameters distilled version of the original 1.2 billion parameters encoder-decoder transformer model.

Fine-tuning on the challenge dataset was promising, but the model validation loss curves showed overfitting despite fine-tuning for weight decay, small learning rate with decreasing linear schedule, warmup, and dropout. In addition, the BLEU score would not exceed 15 on the dev dataset, but upon manual investigation, the produced translations were shallow and sometimes semantically unrelated to the ground truth.

3.4.1 Error analysis

We examined the generated translations for common issues such as mistranslations, omissions, and word order errors. The resulting training process consisted of two steps: fine-tuning on the additional dataset in Table 2 and a step involving the challenge data. Yielding a BLEU score of **27** on the dev set, this approach produced a better result than fine-tuning on a mix of both extended and challenge data. The challenge data would then be under-represented, which would allow for a low BLEU score since the model is evaluated on a dev set from the challenge data distribution and not the additional data in Table 2.

The score was further improved by changing the generation algorithm and number of beams, resulting in the final dev BLEU score of **28.93** seen in Figure 3. This improved the score by 2 points.

Error analysis showed the gap in BLEU score between the dev set medical data and dictionary data. An average of 10 points difference was reported from one distribution to the other, which could be explained by two main differences: that in sequence length (the dictionary data was notably shorter) and in vocabulary distribution (the medical data was more domain-specific).

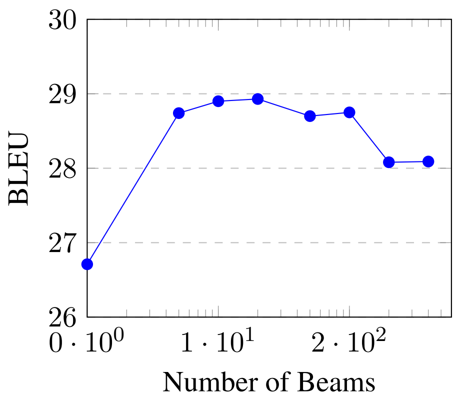

Figure 3: BLEU as a function of the number of beams. A value of one implies greedy decoding while bigger values correspond to the beam-search algorithm. Not surprisingly, the score dramatically improves before plateauing around 10 and reaching diminishing returns. Notably, the optimum is reached at 15 and increasing the number of beams further has a negative impact on the score.

3.4.2 Discussion

In addition to the data in Table 2, we extended our training data by processing a many-to-Bambara dataset from BigScience: the Bambara split of XP3-all (Muennighoff et al., 2022a). XP3-all contains 265,180 many-to-Bambara lines, but we only included the French-to-Bambara subset, and enriched it with the English-to-Bambara subset that was translated with the opus-mt-en-fr model from Helsinki-NLP (Tiedemann and Thottingal, 2020) resulting in 8,377 additional lines of training data.

In the future, we would spend more time automating tasks, including hyper-parameter tuning, to improve the efficiency of the system. Notably, the cross-entropy loss function is only a differentiable proxy for the metric we are trying to optimize i.e. the BLEU score (which is not differentiable). With the recent success of Reinforcement Learning techniques in natural language generation tasks (Stiennon et al., 2020), we plan to further fine-tune the model using the BLEU metric as a task reward, similar to Pinto et al. (2023).

In the future we will explore techniques, such as the recently introduced PEFT (Mangrulkar et al., 2022), which allows for fine-tuning of LLM on very small datasets using parameter efficient fine-tuning methods. IA3 (Liu et al., 2022), Prompt-Tuning (Lester et al., 2021), Prefix-Tuning (Liu et al., 2021), and Low Rank Adaptation (LoRA) (Hu et al., 2021) methods are currently leveraged to train large models efficiently on as few as 10 examples. In comparison to classic fine-tuning that involves training all the weigths of the model, these methods have the added advantage of achieving similar (sometimes even better) results by training only a small subset of the weights (by freezing the pre-trained weights and adding trainable adapter weights as seen in the case of LoRA and IA3). We therefore expect these methods to be increasingly used for any low-resource task in the near future.

Moreover, it seems that the Adam optimizer has finally found a worthy, artificially evolved rival (Chen et al., 2023). We look forward to testing it using the parameters of this task.

Finally, we would suggest the use of learned metrics for the evaluation of the translations instead of the BLEU metric (that ignores synonyms and idioms) building on the works of (Zhang* et al., 2020). Although such models are not yet trained on Bambara, Eddine et al. (2021) seems to offer part of the solution, and an alternative would simply be computing the cosine-distance between the embedding representation of the produced translation and that of the reference (Reimers and Gurevych, 2020).

3.5 Alexander Antonov

All of our models were trained using Sockeye (Hieber et al., 2020). In this task, we focused on building models *from scratch* and utilized 4 checkpoints averaging model parameters in our system. We averaged the parameters of the best 4 checkpoints, which helped to improve results. In addition we used BPE for word segmentation (Sennrich et al., 2016b).

3.5.1 Error analysis

We performed error analysis based on the BLEU metric, and used it as an optimized metric while training. We also used the sacreBLEU (Post, 2018).

3.5.2 Discussion

There are other extended techniques, such as back translation and pre-trained models that we intend to explore in future research. In addition, we also plan to add additional training datasets that were provided and used by the other teams.

3.6 Team Mali

The team attempted multiple approaches concurrently, first pre-training a bilingual Bambara-French denoising Seq2Seq-based foundational model with a lower quality dataset, inspired by Lewis et al. (2019), then fine-tuning it with a higher-quality dataset. This approach yielded non-optimal translations and performance, with all the scores being sub-8 BLEU (it was also resource-heavy and time-consuming). We fine-tuned with DeltaLM (Ma et al., 2021), the training failed to converge with both the base checkpoint and large checkpoint. The problem could be attributed primarily to limited compute resources.

We were able to double our performance from the previous approaches when we re-trained with the NLLB-200 (Team et al., 2022) 600M parameters pre-trained model, with a learning rate of 2, batch size of 512, and training steps of 20k with the lower-quality dataset. Using both DABA-assisted and non-Daba-assited pre-processing[5].

Furthermore, we obtained another peak in performance when we unfreeze the model and then tuned it with the competition dataset with the same configuration, for an understanding of the type of text used for the competition (although we suspected over-fitting). We have seen similar results from both directions, Bambara to French and French to Bambara.

3.6.1 Error analysis

We knew that Bambara is a complex and morphologically sophisticated language. Bambara and French have a one sentence to many translation

[5] https://github.com/maslinych/daba

scheme, where one sentence can have multiple interpretations in the other language, in a polysemous phrasal relationship. Additionally, with Bambara being predominantly a spoken language, there are many fluidities that only native speakers can pick up from translations, compared to a more structured language. We chose to weigh human evaluation higher than automated metrics. Both evaluation techniques gave an insight into the overall performance of our models.

Human Evaluation We came up with our own defined method for manual evaluation, described as follows: For every model trained, we sampled 50 lines from our test set and classified each line into three classes manually *BAD*, *ACCEPTABLE*, and *GOOD*. Where *BAD* was given a value of 0; it is chosen when the hypothesis does not relay any information from the source or is a bad translation. *ACCEPTABLE* was given a value of 1; it is chosen when the hypothesis is a literal translation of the source without context. *GOOD* was given a value of 2; it is chosen when the hypothesis is an accurate translation of the source with context.

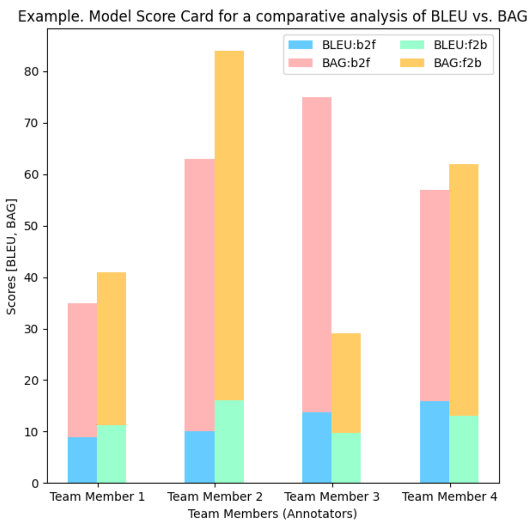

Figure 4: Example model score card analysis comparing human-evaluation vs BLEU. where **b2f**: Bambara-French, **f2b**: French to Bambara. **BAG**: Bad, Acceptable, Good

Each member of the team evaluated a batch of 50 lines per model trained, given the source text, a reference translation, and the hypothesis generated by the model. They were tasked to evaluate the manual score and to compute the BLEU score of the batch, for a comparative analysis of the two results, an example evaluation is shown in Figure 4.

Acknowledging the subjective nature of human evaluation, we should state that while the human evaluations was used to guide our analysis of the performance of our models for the competition, further investigations are needed to validate its viability.

3.6.2 Discussion

Bambara's complexity made it challenging to find the best possible approach, as each aspect of the training required analysis. From pre-processing to evaluation, we found that fine-tuning with the NLLB200 600M model to be more performant. The most significant aspect in our method was the human-in-the loop approach, where coupling human annotation and automated metrics was the primary indicator that informed our decisions during the competition.

3.7 Team Peter-Sokhar

We experimented with transformer-based models and utilized the attention mechanism, which enables one component of the model to concentrate on another part of the model. Due to the issue of vanishing gradient and the weakness of limited levels of parallelization, respectively, both recurrent neural networks (RNNs) and Long Short Term Memory (LSTM) were not considered (Vaswani et al., 2017). The selected transformer model was Facebook/nllb-200-distilled-600M (Team et al., 2022), which was fine-tuned on the training dataset, which allowed for the design of the encoder, latent representation, and decoder. By using semi-supervised learning, the decoder fed features to the model. The team explored training the model for 100 epochs.

3.7.1 Error Analysis

By using Google Translate, the team was able to avoid having a native speaker as a teammate. In the future, a native speaker will be a part of the team.

3.7.2 Discussion

Beyond needing additional compute and a powerful internet connection, we would like to consider other alternative models for cross-validation.

Team Name	BLEU Score (BAM to FR)	BLEU Score (FR to BAM)
Team Alpha	16.31	17.45
Team JYN	13.12	11.1
Yacine Zahidi	19.05	N/A
Alexander Antonov	7.54	8.06
Team Most-Pham	14.81	N/A
Team Mali	5.82	N/A

Table 5: BLEU score results by team for Bambara - French and French - Bambara, with placement ordering.

4 BFMT 2023 Results and Discussion

Table 5 shows the BLEU scores for both Bambara to French and French to Bambara translations. Not all of the teams attempted both translation directions and the scores were averaged across both language pairs to determine the winners.

The BFMT 2023 competition aimed to increase research in low-resource language machine translation by providing training and evaluation data and supporting community-building around scientific transparency. Community-building included teams being constructed from individuals with complementary skills and all relevant training data discovered by one team being shared amongst the teams.

Nonetheless, there were key themes to the submissions. All of the teams used the same core datasets, with two teams bootstrapping alternatives as shown in Table 2. Additional data provided a significant advantage in this low-resource situation. From a machine learning perspective, many of the teams shared similar approaches with effectively utilizing the M2M-100 model (Fan et al., 2020) as the differentiator between the top performing teams. Notably, the NLLB-200 (Adelani et al., 2022) model comparatively under-performed. We believe this is because the M2M-100 model had fine-tuned MT models separately for each language direction.

Subsequent insights were that the winning team used a backtranslation approach, *cyclic backtranslation*, and another successful team used a beam search optimization. Also, we learned that smaller distilled models could beat larger models with limited amounts of data (i.e., fine-tuning distilled models yields more accurate results).

Only one team had members that spoke Bambara but many participants are speakers of other low-resource languages and hope to extend their experience with MT system development to languages that their families and friends speak.

5 Conclusion and Future Work

Because of BFMT 2023, researchers have successfully implemented innovative low-resource machine translation systems. These implementations are extensible to other language pairs, which is helpful since low-resource languages continue to face numerous challenges in terms of research focus and funding. We believe BFMT 2023 has not only supported increased visibility of the Bambara language, but it has also showcased the talent that is working on using creative techniques to address these technical challenges globally.

The BFMT 2023 competition community would like to extend this work by holding other competitions. Ideally, the next competition will utilize automatic speech recognition data. Including spoken data in MT might circumvent a challenge in low-resource language, where only a few online datasets support predominately oral language text processing.

The output of BFMT 2023 is a viable baseline for French - Bambara and Bambara - French machine translation. In addition, the competition dataset is now available to researchers seeking to exceed this baseline or evaluate their translation systems. Similar to the practice in some Kaggle competitions, we can also provide a baseline model in the next competition iteration that is based on the top scoring competition submission [6].

Finally, we would like to provide greater financial support to the participating teams by sponsoring equal and standard access to computational resources. This could better illuminate which machine learning models are the highest performers.

Limitations

There are several limitations we observed during the BFMT 2023 competition. We hope these limitations and findings help researchers to understand the challenges of organizing an MT shared task and use them to improve their competitions.

1. Bambara is a low-resource language and the amount of data needed to significantly improve MT is very large. Inconsistent Bambara orthographies might mitigate translation quality improvement even with additional data collection. There are very high rates of illiteracy for Malians (35%, the 5th highest in the world (Diarra and Leventhal, 2020)) and

[6] https://www.kaggle.com/competitions

Bambara speakers. We would like to gather and translate spoken Bambara audio data to counter these challenges.

2. The test set used for BLEU score evaluation was data previously used in WMT21 (Akhbardeh et al., 2021). It contained transcripts of conversations between translators and Bambara speakers, and translations of medical information[7]. Nonetheless, this dataset was extensively re-aligned and post-processed to remove encoding errors. Due to this additional data cleaning, the processed, competition dataset is of higher quality and thus has no exact baseline for comparison. Further, many competitors trained models with additional data, potentially leading to over-fitting of models to a different format of Bambara-French translations, rather than the original dataset.

3. BLEU has known limitations for meaningful evaluation including how well it corresponds to human evaluation of language correctness and naturalness. In the future we would like to conduct human evaluation of the MT competition output. Many of the diverse competition participants speak other low-resource languages, but only Team Mali had Bambara speakers. Team Mali performed human evaluation and gave human results more weight than automated ones. Human evaluation was used to guide the analysis of the performance of their models. They would like to extend this work but were limited due to the time constraints required for a competition. Finally, the participants' BLEU scores did not meet the closeness threshold (within 1 point) the judges deemed necessary for supplementary human evaluation.

4. We understand human evaluation of the translation predictions can be a strategic piece for judging translation quality and naturalness. Human evaluation can give insight on how systems actually perform and direct focus for improvement based on linguistic analysis. As a low-resource language, it is difficult to find human evaluators with translator-level written French and Bambara skills on the data annotation platforms used in conducting and collecting supplemental human evaluation. We hope these observations will help future MT competition organizers to plan and allocate resources for human evaluation for judging.

5. The importance of compute power was also evident in this competition but the MT systems were not compared in regards to computational resources. In future work we will support equal computational resources for all teams.

[7]The dataset is available to share on request through the corresponding author.

Ethics Statement

Any evaluation system that incorporates human workers motivates reflection on the ethical implications of their contribution. Two of the teams competing in the competition had members that were able to annotate their system's output for translation quality due to their Bambara knowledge. This was part of their team's evaluation efforts and all the team members had already consented to participate in the competition.

In addition to considering how participating in the competition affected the team members, this work also affects the many millions of Bambara speakers who have not historically had access to technology. A recent focus on Machine Learning by the Malian government aims to change that (Diarra and Leventhal, 2020). As a consequence, increasing awareness and access to MT data, tasks, and their applications has wide global impact.

Finally, due to the BLEU scores the competing teams produced, these current translation systems should not be used in critical situations where inaccurate translations could lead to harm.

Acknowledgements

The data used in this work was supported by funding from the Artificial Intelligence Journal promoting AI research and previously utilized for *WMT* (Akhbardeh et al., 2021). We would like to thank Marcos Zampieri, David Talbot, and Francois Lefevre, Alexandra Mephon, and Douglas Cramer for their extremely helpful insight and time they contributed to the competition. We would like to thank Orange Silicon Valley for their support of the machine translation competition including the monetary awards for the winners.

References

David Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen Muhammad, Guyo Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ajibade, Tunde Ajayi, Yvonne Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. 2022. A few thousand translations go a long way! leveraging pre-trained models for African news translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3053–3070, Seattle, United States. Association for Computational Linguistics.

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondrej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussà, Cristina España-Bonet, Angela Fan, Christian Federman, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher M. Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (wmt21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online.

Ali Araabi and Christof Monz. 2020. Optimizing transformer for low-resource neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3429–3435.

Martin Benjamin. 2019. Teach You Backwards: An In-Depth Study of Google Translate for 108 Languages — teachyoubackwards.com. `https://www.teachyoubackwards.com/`. [Accessed 04-Apr-2023].

Isaac Caswell, Julia Kreutzer, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Auguste Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios Gonzales, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Muller, Andre Matthias Muller, Shamsuddeen Hassan Muhammad, Nanda Firdausi Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, M. Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine cCabuk Balli, Stella Rose Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi N. Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2021. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Yao Liu, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, and Quoc V. Le. 2023. Symbolic discovery of optimization algorithms.

Haby Sanou Diarra and Michael Leventhal. 2020. Developing machine learning competence in africa in the francophone sahel region.

Bonaventure F. P. Dossou and Chris C. Emezue. 2020. Ffr v1.0: Fon-french neural machine translation. *arXiv preprint arXiv: Arxiv-2003.12111*.

Moussa Kamal Eddine, Guokan Shang, Antoine J. P. Tixier, and Michalis Vazirgiannis. 2021. Frugalscore: Learning cheaper, lighter and faster evaluation metricsfor automatic text generation.

Chris Chinenye Emezue and Bonaventure F. P. Dossou. 2021. MMTAfrica: Multilingual machine translation for African languages. In *Proceedings of the Sixth Conference on Machine Translation*, pages 398–411, Online. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation. *arXiv preprint*.

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.

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

Felix Hieber, Tobias Domhan, Michael Denkowski, and David Vilar. 2020. Sockeye 2: A toolkit for neural machine translation. In *Proceedings of the 22nd Annual Conference of the European Association for*

Machine Translation, pages 457–458, Lisboa, Portugal. European Association for Machine Translation.

Edward Hu, Yelong Shen, Phil Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

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, Melbourne, Australia.

Tsz Kin Lam, Shigehiko Schamoni, and Stefan Riezler. 2021. Cascaded models with cyclic feedback for direct speech translation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7508–7512. IEEE.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *CoRR*, abs/2104.08691.

Michael Leventhal, Allahsera Tapo, Sarah Luger, Marcos Zampieri, and Christopher M Homan. 2020. Assessing human translations from french to bambara for machine learning: a pilot study. *arXiv preprint arXiv:2004.00068*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

M.P. Lewis, G.F. Simons, and C.D. Fennig. 2014. *Ethnologue: Languages of Africa and Europe*. SIL International.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2021. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *CoRR*, abs/2110.07602.

Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, Alexandre Muzio, Saksham Singhal, Hany Hassan Awadalla, Xia Song, and Furu Wei. 2021. Deltalm: Encoder-decoder pre-training for language generation and translation by augmenting pretrained multilingual encoders. *arXiv preprint arXiv:2106.13736*.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, and Sayak Paul. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. urlhttps://github.com/huggingface/peft.

Swaroop Mishra, Matthew Finlayson, Pan Lu, Leonard Tang, Sean Welleck, Chitta Baral, Tanmay Rajpurohit, Oyvind Tafjord, Ashish Sabharwal, Peter Clark, and Ashwin Kalyan. 2022. Lila: A unified benchmark for mathematical reasoning.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2022a. Crosslingual generalization through multitask finetuning.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2022b. Crosslingual generalization through multitask finetuning.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

André Susano Pinto, Alexander Kolesnikov, Yuge Shi, Lucas Beyer, and Xiaohua Zhai. 2023. Tuning computer vision models with task rewards.

Alberto Poncelas, Dimitar Shterionov, Andy Way, Gideon Maillette de Buy Wenniger, and Peyman Passban. 2018. Investigating backtranslation in neural machine translation. *arXiv preprint arXiv: Arxiv-1804.06189*.

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.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. *ACL*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Shailashree K Sheshadri, Deepa Gupta, and Marta R Costa-Jussà. 2023. A voyage on neural machine translation for indic languages. *Procedia Computer Science*, 218:2694–2712.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021. Curran Associates, Inc.

Allahsera Auguste Tapo, Bakary Coulibaly, Sébastien Diarra, Christopher Homan, Julia Kreutzer, Sarah Luger, Arthur Nagashima, Marcos Zampieri, and Michael Leventhal. 2020. Neural machine translation for extremely low-resource african languages: A case study on bambara. *arXiv preprint arXiv:2011.05284*.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Anna Sun Jean Maillard, 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.

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

Elan Van Biljon, Arnu Pretorius, and Julia Kreutzer. 2020. On optimal transformer depth for low-resource language translation. *arXiv preprint arXiv:2004.04418*.

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.

Ashish Venugopal. 2010. Five more languages on translate.google.com — translate.googleblog.com. `https://translate.googleblog.com/2010/05/five-more-languages-on.html`. [Accessed 04-Apr-2023].

Valentin Vydrin, Jean-Jacques Meric, Kirill Maslinsky, Andrij Rovenchak, Allahsera Auguste Tapo, Sebastien Diarra, Christopher Homan, Marco Zampieri, and Michael Leventhal. 2022. Machine learning dataset development for manding languages. urlhttps://github.com/robotsmali-ai/datasets.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2021a. Byt5: Towards a token-free future with pre-trained byte-to-byte models.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021b. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.