Extending new language in NLLB-200: language informal Malagasy

Francis Rakotomalala
Laboratory for Mathematical and Computer Applied to the Development Systems (LIMAD)
University of Fianarantsoa
Fianarantsoa, Madagascar
francis_rakotomalala@ymail.com

Aimé Richard Hajalalaina
Laboratory for Mathematical and Computer Applied to the Development Systems (LIMAD)
University of Fianarantsoa
Fianarantsoa, Madagascar
arhajalalaina@yahoo.fr

Ndaohialy Manda Vy Ravonimanantsoa
Engineering and Innovation Sciences and Techniques (STII)
University of Antananarivo
Antananarivo, Madagascar
ndaohialy@gmail.com

***Abstract*—This study focuses on integrating informal Malagasy into the NLLB-200 model for machine translation. The model underwent supervised pretraining, which quickly led to improved performance, marked by a significant reduction in both loss and perplexity. This step allowed the model to effectively adapt to the unique linguistic structures of Malagasy. The evaluation of key translation metrics such as BLEU, ROUGE, and BertScore showed that the model produces high-quality translations, combining fluency with semantic coherence. Although the BLEU score was moderate, the ROUGE and BertScore results revealed a remarkable level of lexical and semantic fidelity. This work highlights the importance of developing translation systems that can handle low-resource languages, which are often overlooked by traditional technologies. The study also demonstrates the model's ability to grasp the nuances of informal Malagasy, resulting in significant improvements over existing translation tools. In conclusion, this approach emphasizes the need to include informal languages in translation systems, paving the way for more inclusive and linguistically tailored applications.**

Keywords*—Language Informal Malagasy, Machine translation, NLLB-200

I. INTRODUCTION

In today's globalized and interconnected world, the ability to process and comprehend multilingual texts has become a critical requirement. Recent developments in neural machine translation (NMT) have yielded significant advancements, particularly through models such as mBART [1] and NLLB [2], which have demonstrated notable robustness across a wide array of languages. Nevertheless, the majority of existing translation models are predominantly trained on a limited set of high-resource languages, thereby neglecting a substantial number of low-resource languages [3].

Informal language varieties especially those emerging in digital environments such as social media, online forums, and text messaging remain largely overlooked by mainstream machine translation systems. Despite their pivotal role in reflecting linguistic diversity, such varieties are still insufficiently represented in current language modeling efforts [4]. Most systems are optimized to process standardized language forms, often disregarding informal registers, dialectal variations, and the non-standard orthographic and syntactic patterns that characterize everyday communication. This exclusion reveals a major shortcoming in language technologies, which struggle to accommodate the linguistic richness inherent in informal discourse.

Against this backdrop, the integration of informal language varieties into machine translation systems has become essential for promoting linguistic equity, particularly with respect to underrepresented and low-resource languages such as Malagasy. As the proliferation of digital platforms continues to foster the evolution and diversification of informal language use, there is an increasing need to extend the capabilities of translation models accordingly [5]. The adaptation of models like NLLB-200 to incorporate informal Malagasy thus constitutes a significant advancement towards the equitable inclusion of low-resource languages and the enhancement of language technology accessibility for marginalized linguistic communities.

II. RELATED WORKS

Previous research has extensively examined the use of translation models to address the challenges posed by low-resource languages, with a particular focus on bilingual models. While such approaches have yielded promising results, recent advances have shifted attention towards multilingual models. Unlike bilingual systems, multilingual models are trained on corpora encompassing multiple languages, thereby enabling translation across a broader range of language pairs. However, despite the occasional inclusion of low-resource languages, these models continue to favor dominant languages and demonstrate superior performance on them.

With the rise of large language models (LLMs), the sequence-to-sequence (Seq2Seq) paradigm, widely employed in machine translation, has proven particularly effective. Seq2Seq models based on the Transformer architecture such as mBART [1] and T5 [6] are capable of capturing long-range dependencies and producing grammatically accurate, contextually coherent translations. These models exhibit enhanced generalization capabilities for low-resource languages through transfer learning, leveraging cross-lingual similarities.

Nonetheless, the effectiveness of multilingual models remains limited when it comes to handling informal registers, especially those found in digital communication. It is in this

Cite as: F. Rakotomalala, A. R. . Hajalalaina, and N. M. V. Ravonimanantsoa, "Extending new language in NLLB-200: language informal Malagasy", *Syst. Theor. Control Comput. J. ,* vol. 5, no. 1, pp. 41–50, Dec. 2025.
DOI: 10.52846/stccj.2025.5.1.70

context that NLLB-200 [2] was developed to encompass a wide spectrum of languages. However, its performance on informal corpora remains suboptimal. To address these limitations, techniques such as back-translation have been widely explored. Back-translation [7] involves generating additional training data by translating text from a target language into a source language and then back into the target language, thereby expanding the training set and improving model performance.

Other methods, such as lexical noise injection [8] and partial linguistic normalization, have shown promising results in handling non-standard language varieties. These techniques are particularly relevant for languages like informal Malagasy, which exhibit high levels of orthographic and stylistic variability.

The work of [9] on SentencePiece has also demonstrated that adaptive tokenizers, capable of accommodating diverse linguistic variations, are essential for processing informal language. Such tokenizers facilitate vocabulary enrichment by incorporating register-specific or language-specific tokens, as in the case of integrating informal Malagasy.

Moreover, model robustness in the face of linguistic variation has been a key topic of investigation. Notably, [10] emphasize that for multilingual models to be truly effective, they must not only provide accurate translations but also capture the nuances of informal language. The inclusion of diverse corpora particularly those composed of digital dialogues and informal exchanges can substantially enhance translation quality.

Although recent studies have investigated models such as MarianMT and NLLB for specific languages, few have explored their application to informal registers like informal Malagasy. Furthermore, the integration of informal language into machine translation systems remains an emerging area of research. While existing studies have laid a solid foundation for the inclusion of low-resource languages, the application of innovative techniques including corpus enrichment, vocabulary extension, and back-translation remains crucial to addressing the challenges posed by informal Malagasy.

III. MODEL DESCRIPTION

In this study, we have chosen to use the NLLB-200 model as our main solution. It relies on multi-task learning and masking techniques to improve translation accuracy and fluency, while being pre-trained on large and varied multilingual corpora.

A. General Overview

NLLB-200 is a multilingual neural machine translation model developed by Meta AI to enhance translation for low-resource languages. The model was designed to promote broader linguistic coverage by including languages that are rarely represented in traditional translation systems.

B. Transformer-based Seq2Seq model formulation

For the Transformer-based translation model on which NLLB-200 is built, the mathematical formulation can be presented as follows.

Let $x = (x_1, x_2, \dots, x_n)$ be a source sentence and $y = (y_1, y_2, \dots, y_m)$ a target sentence. The objective of the model is to estimate the conditional probability of the target sequence given the source sequence:

$$P(y \mid x) = \prod_{t=1}^{m} P(y_t \mid y_{<t}, x) \quad (1)$$

where $y_{<t} = (y_1, \dots, y_{t-1})$ denotes the previously generated tokens.

1) Input encoding

Each source token is transformed into a vector representation by combining token embedding and positional encoding:

$$h_i^{(0)} = E(x_i) + p_i \quad (2)$$

where $E(x_i)$ is the embedding of token x_i, and $p_i$is its positional encoding.

2) Attention mechanism

For a given layer, the query, key, and value matrices are obtained through linear projections:

$$Q = HW_Q, K = HW_K, V = HW_V \quad (3)$$

where $H \in \mathbb{R}^{n \times d}$ represents the input activations of the layer. The scaled dot-product attention is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \quad (4)$$

This operation allows each token to assign different weights to the other tokens in the sequence.

3) Multi-head attention

The Transformer applies several attention mechanisms in parallel:

$$\text{head}_j = \text{Attention}(Q_j, K_j, V_j) \quad (5)$$

and then combines them as:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W_O \quad (6)$$

where h denotes the number of attention heads.

4) Encoder

In the encoder, each layer combines multi-head attention, residual connections, normalization, and a feed-forward network:

$$H^{(l)} = \text{FFN}(\text{MultiHead}(H^{(l-1)})) \quad (7)$$

The final encoder representations provide the contextual representation of the source sentence.

5) Decoder

The decoder receives the previously generated target tokens and applies:

- masked self-attention, which prevents access to future tokens;
- cross-attention between the decoder representations and the encoder outputs.

The probability of the next token is then given by:

$$P(y_t \mid y_{<t}, x) = \text{softmax}(W_o h_t^{(L)} + b_o) \quad (8)$$

where $h_t^{(L)}$ denotes the final decoder representation at time step t.

6) Loss function

Training consists of minimizing the negative cross-entropy loss over the training corpus:

$$\mathcal{L}(\theta) = -\sum_{t=1}^{m} \log P(y_t \mid y_{<t}, x) \quad (9)$$

The optimization problem therefore consists of finding the model parameters θ that minimize this objective:

$$\theta^* = \arg\min_\theta \mathcal{L}(\theta) \quad (10)$$

C. Architecture

The architecture of NLLB-200 is based on the Transformer model, a foundational architecture widely used in many translation systems. Its architectural components include:

- Number of layers: The model employs 12 encoder layers and 12 decoder layers, consistent with standard Transformer architectures.
- Embedding size: Each word or subword is represented by a 1024-dimensional vector.
- Number of attention heads: Each layer uses multi-head attention with 16 attention heads, enabling the model to capture complex relationships between words.
- Vocabulary size: The model supports a large vocabulary, managed through subword segmentation using Byte-Pair Encoding (BPE), which allows efficient processing of languages with diverse morphological structures.

B. Specificity

Meta designed the NLLB-200 model to address the challenge of large-scale multilingual translation, particularly for low-resource languages. The following key components distinguish this model:

1) Language Identification and Dataset Construction

Meta first employed a language identification system based on an automatic detection model to build the multilingual datasets required for training NLLB-200. Using a Transformer-based language model, they were able to identify sentence pairs in various languages to feed into the training process.

2) Multilingual Architecture and Specific Adaptations

To handle a wide range of languages, a language identification token is added to the input of both the encoder and decoder. This allows the model to better understand the source language and accurately target the output language. This conditioned approach optimizes the handling of low-resource languages, enabling the model to produce more coherent and contextually appropriate translations.

3) Mixture of Experts

A notable feature of NLLB-200 is its use of a Sparsely Gated Mixture of Experts model. Unlike traditional models that activate most of their parameters for each input, this approach selectively activates only a subset of parameters. This reduces cross-lingual interference and enhances interlingual transfer, which is crucial for low-resource languages such as informal Malagasy. The mechanism relies on multiple feed-forward subnetworks, and during training, the model learns to activate the most suitable subnetwork for each language, thereby optimizing translation performance.

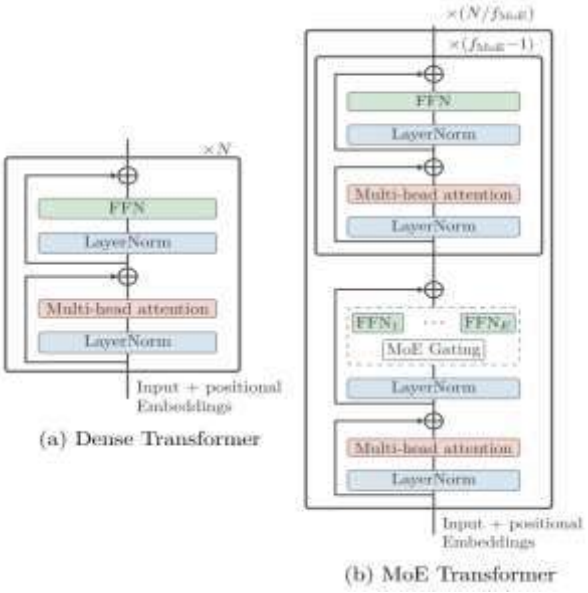

Fig. 1. Traditionnal Dense Transformer VS MoE Transformer [2]

IV. EXTENSION TO INFORMAL MALAGASY

The process of adapting the NLLB-200 model for informal Malagasy translation consists of several key steps:

- Pre-training: The NLLB-200 model will first be trained on existing sentence pairs in Malagasy.
- Back-translation: The back-translation technique will be used to generate additional sentence pairs.
- Post-training: The model will then be globally trained by combining the existing real data with the data generated through back-translation.

A. Supervised Pre-training

1) Corpus

The dataset [11] initially used for supervised training typically contains formal texts, with sentence pairs in Malagasy and another language. This data is essential for establishing a solid foundation for standard translation before enriching it with informal texts through techniques such as back-translation.

2) Normalization

The texts are first normalized, which involves standardizing punctuation, spaces, and symbols used in the texts to ensure that the data is clean and ready for the tokenization step.

3) Tokenization

The sentences in the corpus are then converted into tokens. The NLLB-200 model uses a vocabulary of 256,204 tokens, enabling extensive language coverage. An unknown word in the vocabulary will be broken down into several sub-words or tokens to ensure full coverage.

4) Embeddings

Each token is converted into a numeric embedding vector. These vectors are 1024-dimensional representations, where each token is mapped to a position in the latent space. This step enables the model to understand semantic relationships between words.

5) Preparation of Training Data

The formal corpus is divided into source-target sentence pairs and prepared using techniques such as Byte Pair Encoding (BPE) to maximize model efficiency. The sentences are then binarized to enable more efficient memory management during training.

6) Training Data Preparation

The environment is configured for training on three GPUs, with a batch size of 9, and a total of over 500,000 training examples. This ensures optimal resource usage while maintaining stable training performance.

7) Training

Finally, the model training is carried out using the Hugging Face Transformers API. It is based on stochastic optimization of the loss function defined in equation (9), through backpropagation.

The optimization process aims to update the model parameters defined in equation (10). The optimizer used is AdamFactor, a variant of adaptive gradient descent methods, well-suited for training large-scale models such as NLLB-200. It enables stable parameter updates while incorporating normalization mechanisms and dynamic learning rate reduction.

The training process is performed in two stages: an initial phase on formal data, followed by an adaptation phase using informal data through back-translation, in order to improve the model's robustness to non-standard Malagasy.

B. Back-translation

Back-translating the informal Malagasy corpus is a crucial step to enrich the data and improve translation performance.

1) Preparation of the Informal Corpus

For this study, the informal corpus used comes from monolingual dataset, FLICs[1], which contains a variety of informal Malagasy texts, such as comments, conversations, and social media posts.

2) Tokenization of the Corpus

After preparation, the corpus is tokenized using the appropriate tokenizer for the NLLB-200 model. This phase transforms the informal text into a sequence of tokens, which are text units that can be processed by the model as numerical vectors. Tokenization is essential to ensure that the model can correctly interpret Malagasy sentences, enabling effective translation.

3) Translation

The back-translation process takes place in two sub-steps. First, the informal sentences are translated into the target language, for example, French. For each Malagasy sentence, the fine-tuned NLLB-200 model is used to generate the translated version. Then, the produced translations are translated back into Malagasy. This second step is crucial, as it allows enriching the initial corpus with alternative formulations that better capture the nuances of informal language. By integrating these new formulations, not only is the diversity of the corpus increased, but the model's ability to handle linguistic variations is also enhanced.

These generated data will then be used for the final training of the model, thereby strengthening its effectiveness in translating informal texts.

C. Integration of the New Language

1) Vocabulary Expansion

In this phase, it is crucial to ensure that the model's vocabulary fully represents the characters present in informal texts. Since these texts do not need to be aligned with another language, the goal is to collect and preprocess all available informal texts. To improve the tokenizer's learning capability, only characters that appear at least five times in the corpus were retained, while rarer occurrences, appearing once or twice, were excluded to optimize vocabulary size and performance. In our work, the original vocabulary was enriched with more than 12,000 new tokens, ensuring broader coverage and better adaptation to the specifics of the data.

2) Adding a New Language Tag

When adapting the NLLB model to handle informal Malagasy texts, it is essential to add a new language tag to the tokenizer. This step is crucial as language tags play a fundamental role in identifying the source and target languages for the model.

a) Function of Language Tags

In an NLLB tokenizer, language tags are special tokens inserted at the beginning of source and target texts. These tags inform the model about the language it should process, facilitating accurate and contextually appropriate translation.

b) Process of Adding the Tag

During the fine-tuning phase of the NLLB model, it is necessary to integrate a new language tag specifically designed for Malagasy. This process requires modifications both in the tokenizer and the model's architecture. The tag is typically added as a unique token representing the target language, which will be recognized and used by the model during inference.

c) Impact on Model Performance

By adding this new tag, the model can better handle linguistic variations and nuances of informal language, improving its ability to generate more natural and contextually adapted translations. This conditioned encoding mechanism also allows the model to leverage similarities between Malagasy and other languages, contributing to the generalization of translations.

3) Training

Just like in pre-training, the new model is trained using the Hugging Face library. However, this time, the model is trained on a dataset of over 770,000 example pairs, mostly composed of informal texts. This adaptation allows the model to specifically learn the nuances and characteristics of informal Malagasy, thus enhancing its ability to translate effectively in this context.

V. RESULTS

In the following section, we evaluate the performance of the NLLB-200 translation model, fine-tuned for a new target language namely informal Malagasy by examining the results obtained during the supervised pretraining and generalization phases.

A. Supervised Pretraining

1) Loss

At the beginning of the pretraining phase, the model shows a relatively high loss, with values of 10.96 on the training set and 8.75 on the test set. These values indicate that the model is still not well adjusted to the task, with a significant gap

[1] F. Rakotomalala, A. R. Hajalalaina, M. V. Ravonimanantsoa Ndaohialy, A. Andriavelonera Alexandre, et A. H. Ranaivoson, « FLICs (Facebook Language Informal Corpus): a novel dataset for informal language », Int J Data Sci Anal, vol. 18, no 4, p. 393-403, oct. 2024, doi: 10.1007/s41060-023-00460-2.

between predictions and actual labels. However, after just one epoch of training, the loss drops sharply, reaching 0.21 for training and 0.23 for testing. This rapid improvement shows that the model effectively integrated the fundamental information required for the translation task, which reflects the quality of the dataset and the robustness of the model's architecture.

The figure 2 illustrates the evolution of the loss value for NLLB-200 600M, recorded at every 1,000-step interval:

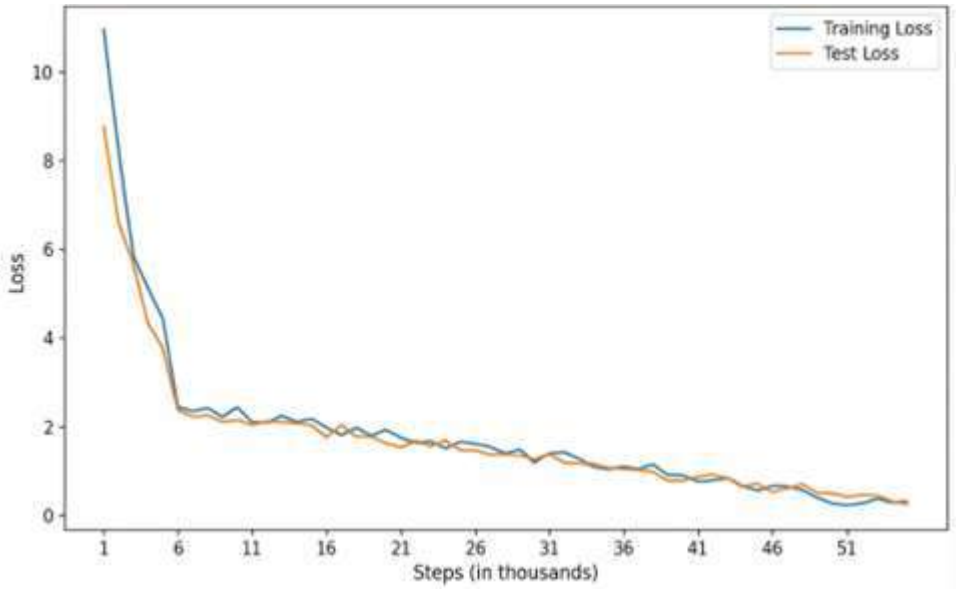

Fig. 2. Evolution of the loss function during pretraining by 1k steps.

2) Perplexity

The initial perplexity is extremely high, at 6310, reflecting a great degree of uncertainty in the model's initial predictions. However, after just one epoch, the perplexity drops to 1.26, indicating increased confidence in the predictions. This final score is very low, showing that the model can predict the next sequence with a high level of accuracy and is thus well-suited to the translation task on this dataset.

B. Generalization to the New Language

1) Loss

During generalization to the new language, the initial training loss is 0.25, then drops to 0.056 after one epoch. This rapid decrease indicates that the model quickly adapted to the linguistic particularities of the new language. On the test set, the loss follows a similar trajectory, dropping from 0.23 to 0.061, which confirms the effectiveness of the generalization. These low values demonstrate that the model maintains high accuracy even when faced with previously unseen data.

The figure 3 illustrates the evolution of the loss value during generalization to the new language, recorded at every 1,000-step interval:

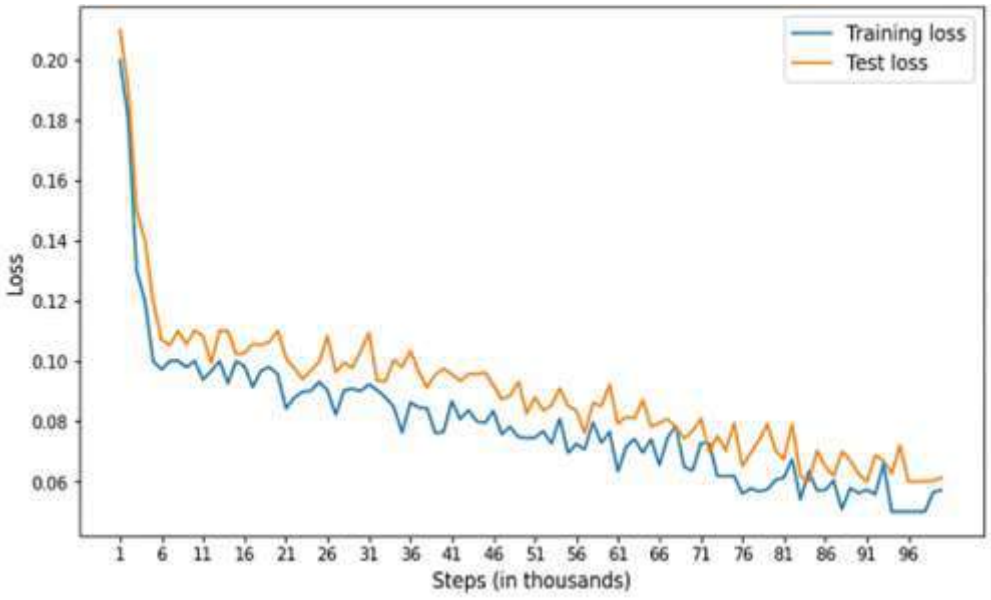

Fig. 3. Evolution of the loss function during generation on the new language by 1k steps.

2) Perplexity

The measured perplexity is 1.06, which is remarkably low. This indicates that the model is very confident in its predictions for this specific language, and that it has successfully assimilated the linguistic structures and characteristics of the new language.

3) Evaluation with Additional Metrics

The table 1 presents excerpts of translations from Malagasy to French, generated by the fine-tuned NLLB-200 model. Each translation is accompanied by additional metric scores, providing a quantitative evaluation of translation quality.

a) Average BLEU Score: 23.41

The BLEU score is a metric that compares the generated translation with a human reference by measuring n-gram overlap. A score of 23.41 is moderate for a translation model working with an informal language, suggesting that the model reproduces reference sentence segments fairly well, even though some phrasing differences remain. This score may be impacted by reformulations introduced by the model to improve fluency, particularly in back-translated sentences, which reduce exact n-gram matches and thus the BLEU score.

b) Average ROUGE Scores – ROUGE-1: 0.47, ROUGE-2: 0.29, ROUGE-L: 0.46

ROUGE scores assess translation quality based on word overlap (ROUGE-1), bigrams (ROUGE-2), and the longest common subsequence (ROUGE-L) between the generated translation and the reference. With ROUGE-1 at 0.47, ROUGE-2 at 0.29, and ROUGE-L at 0.46, the model demonstrates solid lexical and syntactic coverage, although it may introduce new expressions. This supports the observation that the model adapts certain structures to make translations more fluid, which can slightly deviate from the original text, especially in back-translation cases.

c) Average BERTScore: 0.88

BERTScore uses embeddings from language models to compare the semantic meaning of generated translations with human references. A score of 0.88 is excellent, indicating that the model maintains a high level of semantic consistency, even when the wording differs. This shows that the reformulations introduced do not alter the overall meaning and suggests strong adaptation of the model to informal Malagasy.

These scores indicate that the model produces fluent and coherent translations, even if some reformulations differ from the reference text. While this may slightly reduce exact match metrics like BLEU and ROUGE, it can enhance readability in an informal context.

TABLE I. EXCERPTS OF TRANSLATIONS WITH ADDITIONAL METRICS SCORES

	mg	fr	pred	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	BertScore
642428	makasitrak andriamatoa filoha hajaina mino iza...	nous croyons que nous serons bientôt ici	nous croyons que nous sommes bientôt dans la r...	31.559845	0.631579	0.470588	0.631579	0.927010
713111	ts azoko ko zani, ko manin mali nteni oe eto f...	je ne comprends pas, mais je suis mal à l'aise...	je ne comprends pas, mais je suis ici pour finir.	35.243002	0.518519	0.480000	0.518519	0.905588
630961	al do nu nu harivony jonah daniel lelahy vonon...	al do nu nu l'après-midi jonah daniel le gars ...	al do nu nu harivony jonah daniel le gars prêt...	36.619264	0.705882	0.500000	0.705882	0.896920
751289	lasa le pays bas 2 1	est devenu le pays bas 2 1	devient le pays bas 2 1	64.318702	0.769231	0.727273	0.769231	0.907934
509312	rningd luxmina training institute officiel 02 ...	instituts de formation rnedd luxmina officiel ...	rningd luxmina formation institut officiel 02 ...	43.443713	0.842105	0.470588	0.631579	0.907908
427130	iz iz daoli naita anle film.	il est daoli naita anle film.	iz iz daoli naita anle film.	61.478815	0.666667	0.600000	0.666667	0.920408
711125	ko efa nteneniko izy fa ia mo mahita ze ataony...	je lui ai dit que tu voyais ce qu'il faisait, ...	je lui ai dit que vous voyez ce qu'il fait, ce...	34.642262	0.709677	0.551724	0.709677	0.938177
530928	le mod ariv zao problème fa le lalan makani tap	le mod ariv zao problème mais le chemin makiya...	le mod ariv zao problème mais le lalan makani tap	66.063286	0.818182	0.700000	0.818182	0.962287
525137	kery introduit la. au tubydi frere on en a bes...	l'inotroduit la.au tubydi frere sur en un beso...	qui introduit la. au tubydi frere on en a besoin.	37.700638	0.571429	0.315789	0.571429	0.888604
768823	teto orinasa malagasy tiko nopotehana noho fia...	dans cette société malgache, les gens sont dét...	dans une entreprise malgache, les gens ont été...	64.870669	0.850000	0.684211	0.800000	0.960970

C. Model Usage

The NLLB-200 model can be used to translate informal Malagasy phrases into a target language—French in this example—with a particular focus on the accuracy of common expressions.

Let's take the phrase *"namako mandeh makati"*:

- Translation by our model: *"Mon ami va à la maison"* ("My friend is going home") — a fairly accurate translation.
- Translation by Google Translate: *"Mon ami mandeh makati"* — incorrect translation, lacking proper adjustment.
- Translation by ChatGPT: *"Mon ami va bien"* ("My friend is doing well") — incorrect translation, changing the meaning entirely.

This example illustrates the NLLB-200 model's ability to grasp the meaning of informal Malagasy phrases and to provide contextually appropriate translations, outperforming other state-of-the-art translation systems.

VI. DISCUSSIONS AND RECOMMENDATIONS

A. Self-Generalization Capacity to the New Language

The fine-tuned NLLB-200 model demonstrates a strong ability to generalize to Malagasy, an informal and low-resource language. At the beginning of pretraining, the high loss values and initial perplexity indicate significant uncertainty in its predictions. However, after just one epoch, the loss drops to 0.21 and 0.23 respectively, and perplexity to 1.26. This rapid decline reflects a fast integration of linguistic characteristics and efficient learning of Malagasy language structures.

These results show that NLLB-200 is well-designed to adapt to underrepresented languages, a valuable asset for users aiming to expand machine translation applications to new languages with limited resources.

B. Translation Quality

The additional metrics provide an overall view of the translation quality produced by the model, analyzed from various perspectives.

The BLEU score, although relatively moderate, remains encouraging for an informal language. This score, which measures n-gram overlap with the reference translation, is affected by the reformulations the model uses to preserve fluency, especially during back-translation. These reformulations, while enhancing readability, can lower exact n-gram matches.

ROUGE scores, on the other hand, indicate good lexical and syntactic similarity. This means the model retains adequate coverage of key lexical elements, even if some reformulations lead to slight differences in the structure of n-grams. This balance between fidelity and fluency allows for clear and readable translations, with a tendency to paraphrase for better comprehension.

Finally, the high BertScore shows that the model effectively captures the overall meaning and nuances of the text, even with formulation variations compared to the reference. This means that reformulations do not compromise the original meaning, ensuring semantically faithful translations.

Overall, these results reveal the model's ability to generate natural and coherent translations suited to informal communication contexts, where readability is especially valued.

C. Advantages

1) Natural and Coherent Translations

The model is particularly effective at producing translations that sound natural, a crucial aspect when translating informal languages like Malagasy. These translations are suitable for spontaneous communication, where the final text needs to be fluid and easy to understand. This is especially relevant for

users requiring quick and readable translations, notably in informal exchanges or daily conversations. Such translation improves overall comprehension while preserving the original intent of the text.

2) Ability to Normalize Informal Text

One notable feature of the model is its ability to normalize informal text. Informal Malagasy often contains abbreviations, spelling mistakes, and other non-standard forms, commonly found in online messages or casual conversation. The model can convert these into a more standardized form without altering the meaning. This feature can be particularly useful for tasks such as processing social media data or generating text that must comply with stricter language norms.

3) High Semantic Fidelity Despite Reformulations

While the model may introduce reformulations to improve translation fluency, it maintains high semantic fidelity. This means that even when the structure of sentences is changed for a more natural rendering, the overall meaning remains intact. This is crucial in translation contexts where accuracy is essential—such as technical or scientific translations—as well as in dialogues where understanding the message is a priority.

4) Extended Applications

Due to its ability to smoothly reformulate sentences, the model is also well-suited for other language tasks such as text cleaning or online content moderation. For example, it can transform informal text into a more formal version or detect and correct errors in informal Malagasy corpora. Additionally, its ability to handle informal data makes it useful in digital communication contexts, such as social media or online forums.

D. Limitations

1) Reduced Accuracy in Highly Informal or Non-Standard Contexts

When the source text contains extreme forms of informal language, the model may still struggle to produce perfectly fluent translations. Although it is adapted to a certain degree of informality, the translation may lose precision or coherence when the variation in language style becomes too pronounced. This issue is especially noticeable when the target language requires a higher level of formality than that used in the informal Malagasy source text.

2) Difficulties with Long or Complex Sentences

Like many neural machine translation models, NLLB-200 can face challenges with long or complex sentences, where maintaining syntactic structure is more difficult. Specifically, in sentences with multiple clauses or complex grammatical dependencies, the model may generate translations that lack clarity or introduce structural errors.

3) Issues Translating Ambiguous Sentences

Due to the nature of certain idiomatic expressions or language-specific constructs in Malagasy, the model may not successfully capture all semantic ambiguities. Some phrases may be translated too literally, leading to loss of meaning or inconsistencies. Linguistic ambiguity remains a challenge for all automatic translation systems, and while NLLB-200 performs excellently in most cases, these situations continue to pose problems.

4) Lower BLEU and ROUGE Scores Due to Reformulations

Although reformulations enhance the fluency of the translation, they can also negatively affect exact-match metrics such as BLEU and ROUGE. These reformulations, aimed at improving readability of the target text, may reduce the exact n-gram overlap between the translation and the human reference. This lowers the precision of these metrics, even though the overall semantic quality—measured by tools like BertScore—remains high. This means that while the translation is generally of good quality, it does not always perfectly match the reference, which could be problematic in cases requiring exact fidelity.

5) Limitations in Handling Low-Resource Languages

Despite promising results, the model remains constrained by the availability of Malagasy data—a language considered low-resource in the field of machine translation. The model's quality is therefore heavily influenced by the size and diversity of the training corpus. If additional Malagasy resources become available, performance could improve; for now, however, the model may not fully cover all varieties and nuances of the language.

E. Recommendations

To further improve the performance of the translation model, the following recommendations are proposed:

1) Enrich the Training Corpus

To better capture variations in the Malagasy language and overcome current limitations due to a limited dataset, it is recommended to expand the corpus with more texts from various registers and contexts, including more formal texts and specialized domains. This would enhance the model's flexibility in handling more complex translations and improve its ability to manage ambiguity or context-specific expressions. Integrating data from social media, informal dialogues, and academic sources could also help diversify the output.

2) Use Ambiguity Detection and Management Techniques

For ambiguous sentences, it would be beneficial to incorporate additional disambiguation methods into the model, allowing it

to better choose between multiple possible meanings of a word or phrase depending on the context. This could involve introducing auxiliary models or context-based rules to improve the accuracy of translations involving semantic ambiguity.

3) Improve Handling of Highly Informal Texts

To better manage very informal language (slang, abbreviations, etc.), it is advisable to apply specific text normalization techniques prior to translation. This could include converting informal terms into their more standardized equivalents before performing the translation, to ensure greater coherence and understanding in the target language.

4) Optimize the Handling of Long and Complex Sentences

To address the limitations in translating long or syntactically complex sentences, one approach could be to use sentence segmentation or chunk-based translation methods, allowing the model to process each segment independently before reassembling them. This strategy may improve translation clarity without losing the overall meaning of the text.

5) Use a Hybrid Approach for Complex Expressions

In cases where the model struggles with idioms or complex expressions, a hybrid approach combining neural translation techniques with rule-based translation could be helpful. Predefined linguistic rules can support more accurate translations in highly specialized contexts or when dealing with rare expressions.

6) Enhance Low-Resource Language Support

Given that Malagasy is a low-resource language, expanding the model to include multilingual pretraining approaches could Ultimately, this study highlights the considerable potential of NLLB-200 to expand the capabilities of machine translation models by incorporating low-resource and informal languages. It also paves the way for more inclusive applications tailored to the specificities of global languages, and shows how such models can serve as powerful tools to enhance linguistic diversity and global communication.

ACKNOWLEDGMENT

The author would like to express sincere gratitude to the Politehnica University of Timișoara (UPT)[2] for the warm welcome and the logistical support provided during the research stay. Special thanks are also extended to the research supervisor for valuable guidance and support, as well as to the research teams for their collaboration and availability throughout this period.

strengthen its translation capabilities by leveraging related or similar languages. This includes employing techniques such as transfer learning to improve the model's adaptability to this specific language.

VII. CONCLUSION

The evaluation of the fine-tuned NLLB-200 model for translating informal Malagasy into other languages—specifically French—reveals particularly promising results. During supervised pretraining, the model demonstrated rapid performance improvements, with a notable decrease in loss and perplexity, indicating its ability to effectively grasp the complex linguistic structures of Malagasy. As it was extended to this new language, the results remained robust, with exceptionally low loss and perplexity levels, highlighting the model's strong adaptation to the specificities of informal Malagasy.

Translation metrics such as BLEU, ROUGE, and BertScore confirm the model's capacity to generate translations that are both fluent and semantically coherent. While the BLEU score is moderate—mainly due to necessary reformulations in informal contexts—the ROUGE and BertScore metrics show that the model preserves high quality in both lexical and semantic terms. These results are particularly significant when dealing with the translation of informal, low-resource languages, a domain often overlooked by traditional translation systems.

A concrete example of translation highlights the model's effectiveness in handling the subtleties of informal Malagasy, demonstrating its advantage over other translation tools such as Google Translate or ChatGPT, which sometimes struggle to accurately convey the meaning of informal phrases. This ability to understand and translate informal nuances gives the model a real edge when working with underrepresented languages.

REFERENCES

[1] Y. Liu et M. Lapata, « mBART: multilingual denoising pre-training for neural machine translation », in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, p. 7871-7880.

[2] N. Team *et al.*, « No Language Left Behind: Scaling Human-Centered Machine Translation », 25 août 2022, *arXiv*: arXiv:2207.04672. doi: 10.48550/arXiv.2207.04672.

[3] E. M. Bender, T. Gebru, A. McMillan-Major, et S. Shmitchell, « On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜 », in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, Virtual Event Canada: ACM, mars 2021, p. 610-623. doi: 10.1145/3442188.3445922.

[4] S. Jin, A. P. de Vries, A. Szuba, et D. Hiemstra, « Classification and Interchange of Informal and Formal English Text », 2022.theses/2022/Seraph_Jin___1032019___Classification_and_Interchange_of_Informal_and_Formal_English_Text.pdf

[5] C. Zhao *et al.*, « A Systematic Review of Cross-Lingual Sentiment Analysis: Tasks, Strategies, and Prospects », *ACM Comput. Surv.*, vol. 56, nº 7, p. 1-37, juill. 2024, doi: 10.1145/3645106.

[6] C. Raffel *et al.*, « Exploring the limits of transfer learning with a unified text-to-text transformer », *Journal of machine learning research*, vol. 21, nº 140, p. 1-67, 2020.

[2] https://www.upt.ro/

[7] R. Sennrich, B. Haddow, et A. Birch, « Improving Neural Machine Translation Models with Monolingual Data », 3 juin 2016, *arXiv*: arXiv:1511.06709. doi: 10.48550/arXiv.1511.06709.

[8] S. Edunov, M. Ott, M. Auli, et D. Grangier, « Understanding Back-Translation at Scale », 3 octobre 2018, *arXiv*: arXiv:1808.09381. doi: 10.48550/arXiv.1808.09381.

[9] T. Kudo et J. Richardson, « SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing », 19 août 2018, *arXiv*: arXiv:1808.06226. doi: 10.48550/arXiv.1808.06226.

[10] P. Michel et G. Neubig, « Extreme Adaptation for Personalized Neural Machine Translation », 4 mai 2018, *arXiv*: arXiv:1805.01817. doi: 10.48550/arXiv.1805.01817.

[11] J. Tiedemann, « OPUS-Parallel Corpora for Everyone. », *Baltic Journal of Modern Computing*, vol. 4, nº 2, 2016