Neural Machine Translation for Mooré, a Low-Resource Language

Hamed Joseph Ouily[1,2], Aminata Sabané[1,2], Delwende Eliane Birba[1,3],
Rodrique Kafando[1], Abdoul Kader Kabore[1], Tégawendé F. Bissyandé[1,2]

Centre d'Excellence CITADEL, Université Virtuelle du Burkina Faso[1]
Département d'Informatique, UFR/SEA, Université Joseph KI-ZERBO[2]
AI-KING GROUP[3],
hamed.ouily@gmail.com, eliane.birba@aikinggroup.tech, rodrique.kafando@citadel.bf,
abdoulkader.kabore@citadel.bf, aminata.sabane@ujkz.bf, tegawende.bissyande@citadel.bf

Abstract. Natural Language Processing (NLP) is a field of artificial intelligence with the goal of enabling machines to understand human language. Neural Machine Translation (NMT) is one of the many applications of NLP and allows for the translation of a source language into a target language. NMT has made significant progress in recent years. However, most African languages, especially those in Burkina Faso, have received very little research attention in this context. In this article, we propose automated translation models for*Mooré* language to *French* based on Transformers. We obtained an average BLUE score of 44.82 for the model trained on all the data and 65.75 for the model trained only with the Jehovah's Witnesses Bible data for the machine translation task from *Mooré* to *French*. These encouraging results may evolve as the work is still in progress.

Keywords: Natural Language Processing, Neural Machine Translation, Low-ressource Language, Local language

1 Introduction

Linguistic diversity is one of the riches of Africa [1]. Languages hold strategic importance for both peoples and the planet, as they play a crucial role in the development process. They represent the wealth of cultural diversity and facilitate intercultural dialogue. Additionally, languages are an essential tool for ensuring quality education accessible to all. They encourage collaboration and contribute to the establishment of inclusive knowledge societies. They also preserve precious cultural heritage and stimulate political commitment to the beneficial application of science and technology for sustainable development. However, this diversity also presents a significant challenge in the form of linguistic barriers, given the importance of languages in communication.
The official language of Burkina Faso is French, and it has approximately 60 local languages. This

JRI 2023, December 18-20, Ouagadougou, Burkina Faso

DOI 10.4108/eai.18-12-2023.2348140

situation poses a challenge for communication and understanding among the different linguistic communities in the country, making it difficult for the majority of the population to access information. An effective solution to this situation is the automatic translation of local languages. Neural Machine Translation (NMT) is a rapidly evolving field fueled by advances in artificial intelligence (AI) and natural language processing (NLP). It is an architecture that allows machines to learn to translate between different languages [2]. However, Burkina Faso national languages have been underexplored in the field of neural machine translation, and the resources to do so are either non-existent or difficult to obtain, especially the data.

The overall objective of this study is to develop an efficient automatic translation system for Burkina Faso national languages, particularly ”Mooré”, to facilitate communication and understanding among speakers of these languages and other languages. To achieve this, we evaluated the effectiveness of various AI techniques for automatic translation of Burkina Faso national languages, collected and pre-processed a corpus of ”Mooré” texts, as well as their translation into French.

Our work aims to promote linguistic inclusion in administrative, educational, and media spheres, starting with Moore. The rest of the article is organized as follows: in section 2, we provided a state of the art of works related to our objectives. In section 3, we presented our methodology. In section 4, we present our results and challenges. We conclude in section 5.

2 Related work

Several recent works in the field of automatic language translation have been carried out, with the majority of them focusing on low-resource languages. In 2020, Dossou and Emezue [3] used an encoder-decoder architecture consisting of Gated Recurrent Units (GRU) to propose an automatic translation model for Fon, a language spoken in Benin, to French. They achieved a performance of 30.55 BLEU on the JW300 [4] and BeninLanguages datasets. The best results were obtained on data with diacritics (tonal marks). In the same year, Laura Martinus et al [5] proposed a translation model into English for six South African languages using the Transformer and achieved a score of 40 BLEU on the JW300 dataset. The authors demonstrated that the training data domain has an impact on model performance.

The Transformer is a neural network architecture based entirely on attention [6]. The authors of [6] introduced it in 2017 and showed that the Transformer outperforms encoder-decoder architectures based on recurrent neural networks for translation tasks on WMT2014 data. The absence of recurrent layers in the Transformer makes it faster to train.

In 2021, Hacheme [7] used the Transformer to propose a multilingual automatic translation model from English to Gbe (Fon and Ewe), known as English2GBE. The main goal was to demonstrate the benefits of a multilingual automatic translation model. They constructed three translation models: one for English to Ewe, one for English to Fon, and a multilingual model for English to Ewe and Fon (English2GBE). The results showed that the multilingual model outperformed the bilingual models. This was explained by the fact that the two languages are from the same family and share some characteristics, allowing them to learn from each other during model training.

The authors in [8] also demonstrated the effectiveness of the Transformer in translation tasks. They built two automatic translation models, one based on JoyNMT [9] and the other based on the Transformer. The Transformer-based models achieved better results. For their model training, they tested three data representation models and found that the Binary Pair Encoding (BPE) representation improved model performance. Tests were conducted using Bible data from YouVersion, JW300 data, and data provided by the South African government, Autshumato. The results were compared with the work of [5] and achieved a BLEU score at least 7 points higher.

Other researchers have used to pretrained models to train more powerful automatic translation models, despite the limited existing data. In 2019, the authors of [4] demonstrated that pretrained models have a significant impact on linguistic modeling, such as causal language modeling (CLM), masked language modeling (MLM), and translation language modeling (TLM). Pretraining multilingual language models leads to better results, especially in automatic translation tasks, where it achieved an average BLEU score of 75.1, compared to 71.5 for (Artetxe and Schwenk, 2018), which was the state of the art for the same language corpus translation task. Furthermore, they achieved a new state of the art with a BLEU score of 34.3 on WMT'16 German-English.

AfroLM [10] proposed a pretrained multilingual model on 23 African languages called AfroLM. This model is based on an active learning algorithm, which gives a model M the ability to query another model N to improve itself. In their case, they set M=N, making it a form of self-supervised active learning. They demonstrated that with 14 times less data, AfroLM is competitive with other pretrained language models, achieving an average F1-score of 80.13% with 0.73GB of data compared to 81.90% for AfroXLMR-Base with approximately 2.5 TB of data. Additionally, they proved that the model generalizes well for other NLP tasks. Their data was collected from news sources and covers various parts of the continent. They also used the BPE model for data representation.

3 Methodology

In this section, we have described the methodology used in this document. We begin by explaining how we collected and processed our data. Then, we describe how we have trained our language detection model, and finally, we present the results obtained in the next section. Figure 1 depicts our methodology.

3.1 Data collection and data processing

3.1.1 Data collection

We identified several data sources that contain text in ”Mooré” with their translations in French. The first data source we used is the Jehovah's Witnesses' Bible from the jw.org[1] website, which provides translations of the Bible in ”Mooré” and French. We collected both versions of the Bible directly from the website, treating each verse as a line in our dataset. We used web scraping for this

[1] https://www.jw.org/

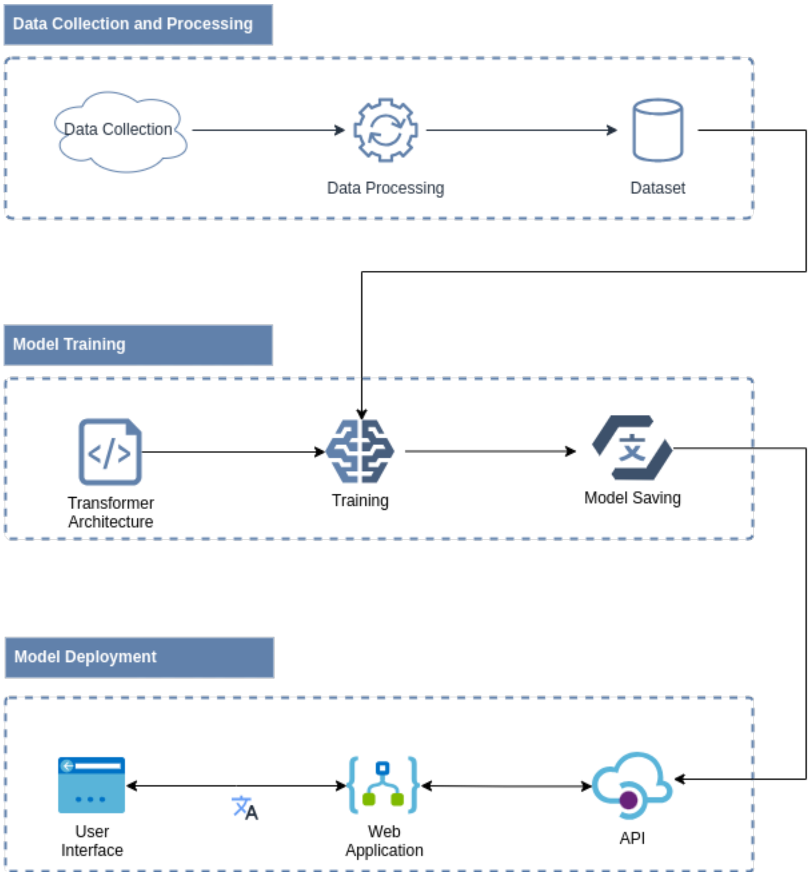


Fig. 1. Comprehensive methodology for developing a LOw-ressource language Translation system

task. Since we collected the entire Bible, we divided it into four (4) parts for each language, allowing us to run eight (8) tasks in parallel. This reduced the time required to approximately three (3) hours for both versions of the Bible, compared to six (6) hours for a single version without parallel processing.

Another data source is the ohchr website[2], which contains the Universal Declaration of Human Rights in both French and ”Mooré”. We also scraped this data, considering each sentence as a line in our dataset. Lastly, we utilized the ”Mooré”-French dictionary index [3] in PDF format. We extracted data from this PDF document using data extraction techniques, resulting in a total of 36,178 lines

[2] https://www.ohchr.org/fr

[3] https://www.webonary.org/moore/files/index-francais-moore.pdf

for our dataset. Table I provides the number of lines obtained for each source.

Table 1: The number of lines for each data source

Source	JW	ohchr	index
Number of lines	31078	64	5036
Number of words (mos - fr)	820817 - 757509	2033 - 1527	5036 - 5036

3.1.2 Data processing

We cleaned the data to obtain quality data. We performed data alignment for the Universal Declaration of Human Rights with the assistance of three (3) individuals who aligned and then verified that others had aligned correctly on their side. For the JW data, we conducted verse-level alignment during data collection and observed that some values were expressed in numbers in one verse and in words in its translation. This inconsistency could lead to model comprehension issues. We identified these lines (1415 lines) and removed them from the dataset. For the index, we did not need to make any modifications to the initially collected version. Our final dataset comprises 34,763 lines (”Mooré” : mos, French : fr). We present a data excerpt in Table 2.

Table 2: Data excerpt

mos	fr
Maam a Poll sn yaa ned ning Kirist Zeezi sn b...	De la part de Paul , appelé pour être apôtre de...
n gls sebkãngã n tool Wnnaam tiging ning sn ...	à l’assemblée de Dieu qui est à Corinthe , à vo...
B bark la laaf sn yit Wnnaam sn yaa tõnd B...	Que Dieu notre Père et le Seigneur Jésus Christ...
Mam psda Wnnaam bark wakat fãa yãmb yĩnga , ...	Je remercie toujours mon Dieu à votre sujet pou...
Bala yãmb sn be a pg wã , yãmb paamda bũmb f...	En effet , par votre union avec lui vous avez é...

3.1.3 Tokenization

We used the SentencePiece tokenizer [11], a language-agnostic subword tokenizer. SentencePiece is a widely used tokenizer in Natural Language Processing (NLP) due to its linguistic versatility, ability to handle compound and rare words, flexibility, and strong performance. It works well with many languages, including ”Mooré”, offers customization options, is supported by various

NLP frameworks, and is efficient for tokenization in various NLP tasks. We used the SentencePiece module implemented in TensorFlow[4] [5] with a vocabulary size (vocab_size) of 8,000 and set normalization to *false* to preserve accents, especially for ”Mooré”.

3.2 Transformer architecture

The Transformer was introduced in 2017 in [6], titled ”Attention Is All You Need.” The Transformer is a neural network architecture entirely based on the attention mechanism, unlike previous sequence-to-sequence architectures such as LSTM [12] or GRU [13], which combine RNNs and attention. This absence of recurrence helps reduce operations and provides better handling of long sequences [6]. This architecture has surpassed the state of the art in many fields, including Natural Language Processing (NLP). It is widely used today, particularly in Large Language Models (LLMs) such as OpenAI's GPT models[6]. Like all sequence-to-sequence architectures, the Transformer consists of an encoder and a decoder, as shown in the figure 2.

3.3 Model training

We divided our dataset as follows: 70% of the data for training, 20% for testing, and 10% for validation. We trained two machine translation models for ”Mooré”. The first model translates ”Mooré” to French, and the second one translates French to ”Mooré”. We implemented the Transformer architecture described in [14] using TensorFlow. Each model was trained for 80 iterations on the training data, with an average of 13.5 minutes per iteration, totaling approximately 18 hours to train a model.

For the configuration of our models, we used four (4) layers ($num\_layers = 4$) and four (4) attention heads ($num\_heads = 4$). Our models have a dimension of 128 ($d\_model = 128$), hidden layer dimension of 512 ($dff = 512$). The dropout rate, which is the probability that a neuron is deactivated, is set to 10% ($dropout\_rate = 0.1$).The workspace we have on the server has the following characteristics:

- Processor: 4 CPUs with an average frequency of 3.0GhzRAM
- Memory: 64GB
- Operating system: Debian 5.10.140-1

4 Initial results and challenges

For our work, we used the BLEU [15](bilingual evaluation understudy) score. This is an evaluation measure commonly used to assess the quality of automatically generated translations, such as

[4] https://www.tensorflow.org/?hl=fr

[5] https://www.tensorflow.org/text/api_docs/python/text/SentencepieceTokenizer

[6] https://openai.com/

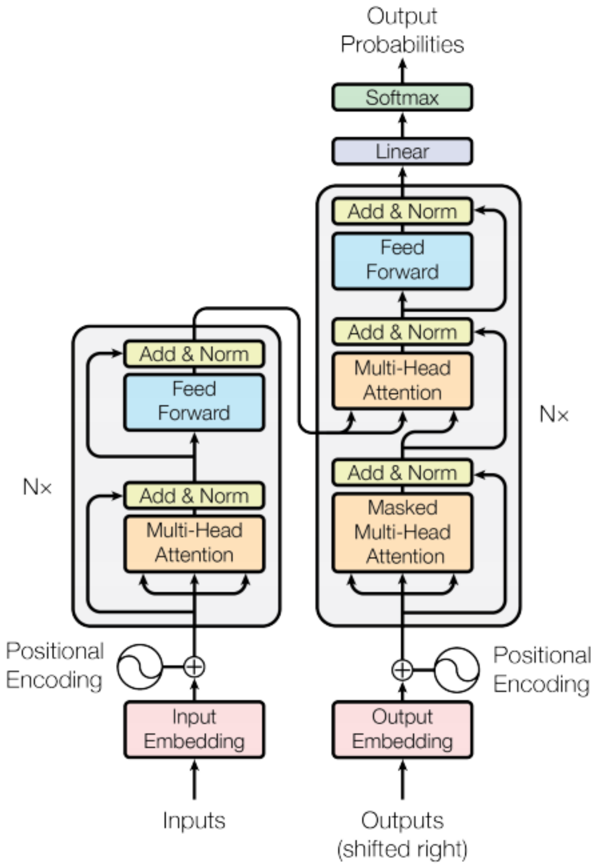


Fig. 2. Transformer Architecture

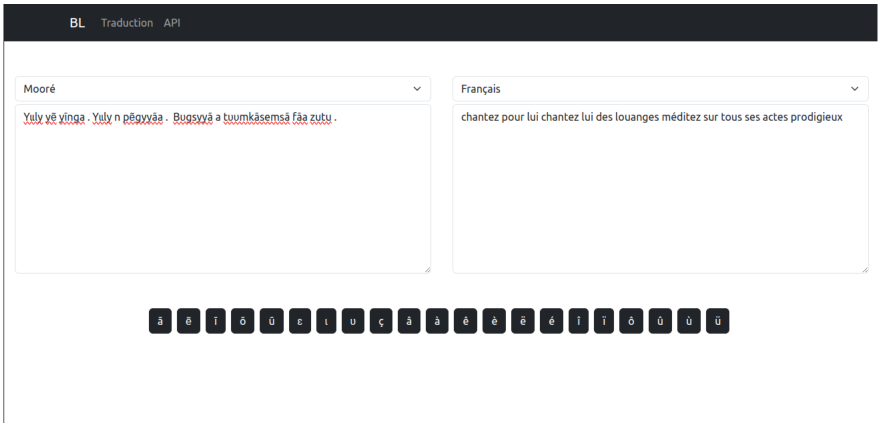
those produced by machine translation (MT) systems or text generation models. The BLEU score works by comparing the machine translation with a set of reference translations (often produced by humans) for the same source input.
Table 3 shows the results we have obtained after our initial training rounds. These results encourage us to continue our research work, even though we are severely limited in terms of linguistic resources, and the vast majority of our available data comes from biblical sources.

Mooré, a tonal language spoken in Burkina Faso, is characterized by words having the same spelling but different meanings depending on the tone, such as ’Saaga’ meaning ’balai’ or ’pluie’ (Saaga: broom or rain). The addition of diacritical marks can also change the meaning, as in ’sãaga’ for ’diarrhée’(sãaga: diarrhea). These nuances are more evident in spoken language than in written form, influencing the performance of linguistic models. Mooré also includes four main dialects, each associated with a specific region: yaadré from the Ouahigouya region, taooledé from the Koudougou region, saremdé from the Koupèla region, and lallweoogo, wubrweoogo, zudweoogo from the Central and Southern region [16]. These dialects reflect the linguistic and cultural diversity of Burkina Faso.

Table 3: Average BLUE score over training, test and validation BLUE scores

Data	mos-fr
JW	65.76 BLEU
JW + index	45.35 BLEU
JW + index + ohdh	44.82 BLEU

We have developed a web application for machine translation to make it easier for the public to use. Figure 3 shows the translation interface from ”Mooré” to French.



Mooré

Yʋly yẽ yĩnga . Yʋly n pẽgyyãa . Bʋgsyyã a tʋʋmkãsemsã fãa zutu .

Français

chantez pour lui chantez lui des louanges méditez sur tous ses actes prodigieux

Fig. 3. Translation from ”Mooré” to French

Table 4: Examples of ”Mooré” to French Translations

setence	ned kam fãa na n soga a babiig a ba zak n yeel yaa “ fo tara kobre d wã y tõnd naaba ges zĩkãngã sn yaa raboogã yelle ”
reference	chacun saisira son frère dans la maison de son père en lui disant tu as un manteau alors sois notre chef prends en charge ce tas de ruines
predict	chacun saisira son frère dans la maison de son père en lui disant tu as un manteau alors sois notre chef prends en charge ce tas de ruines
setence	kelgyyã bala mam sn yetã tara yõodo mam nobmsã togsda sn yaa trga
reference	écoutez car ce que je dis est important mes lèvres expriment ce qui est droit
predict	écoutez car ce que je dis est important mes lèvres expriment ce qui est droit
setence	tõnd na n paama pãng wnnaam maasem yĩnga a na n taba tõnd zabdntaasã
reference	par dieu nous obtiendrons de la force et il piétinera nos adversaires
predict	par dieu nous obtiendrons de la force et il piétinera nos adversaires

5 Conclusion

In this work, we trained two machine translation models based on the Transformer architecture. Our model facilitates automatic translation from ’Mooré’ to French. Our model obtained average BLUE scores of 44.82 on all the data collected. Ongoing research aims to refine these models by optimizing their parameters, and we have also developed a web application to make these machine translation capabilities accessible through a user-friendly interface.

In terms of perspectives, we first intend to propose a translation model from *French* to *Mooré* and then compile a comprehensive dataset that combines Mooré with several other languages, with the aim of developing multilingual machine translation models. In addition, we wish to address the unique challenges posed by the tonal nature of the ”mooré” language, such as differentiating words with identical spellings, but different meanings due to tonal variations. The aim will be to create tone-sensitive algorithms capable of efficiently interpreting and generating accurate translations in the face of such complexities.

Acknowledgments

This work was conducted as part of the Artificial Intelligence for Development in Africa (AI4D Africa) program, with the financial support of Canada’s International Development Research Centre (IDRC) and the Swedish International Development Cooperation Agency (Sida).

References

[1] Diversité culturelle et linguistique | African Declaration on Internet Rights and Freedoms;. Available from: https://africaninternetrights.org/fr/principles/diversit

[2] Luong T, Cho K, Manning CD. Neural Machine Translation. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts. Berlin, Germany: Association for Computational Linguistics; 2016. Available from: https://aclanthology.org/P16-5005.

[3] Emezue CC, Dossou FPB. FFR v1. 1: Fon-French neural machine translation. In: Proceedings of the The Fourth Widening Natural Language Processing Workshop;. p. 83-7.

[4] Agić Ž, Vulić I. JW300: A Wide-Coverage Parallel Corpus for Low-Resource Languages. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics; 2019. p. 3204-10. Available from: https://aclanthology.org/P19-1310.

[5] Martinus L, Webster J, Moonsamy J, Jnr MS, Moosa R, Fairon R. Neural machine translation for South Africa's official languages.

[6] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al.. Attention Is All You Need. arXiv;. Available from: http://arxiv.org/abs/1706.03762.

[7] Hacheme G. English2Gbe: A multilingual machine translation model for {Fon/Ewe}Gbe. arXiv;. Available from: http://arxiv.org/abs/2112.11482.

[8] Sefara TJ, Zwane SG, Gama N, Sibisi H, Senoamadi PN, Marivate V. Transformer-based Machine Translation for Low-resourced Languages embedded with Language Identification. In: 2021 Conference on Information Communications Technology and Society (ICTAS). IEEE;. p. 127-32. Available from: https://ieeexplore.ieee.org/document/9394996/.

[9] Kreutzer J, Bastings J, Riezler S. Joey NMT: A Minimalist NMT Toolkit for Novices. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations. Association for Computational Linguistics;. p. 109-14. Available from: https://aclanthology.org/D19-3019.

[10] Dossou BFP, Tonja AL, Yousuf O, Osei S, Oppong A, Shode I, et al.. AfroLM: A Self-Active Learning-based Multilingual Pretrained Language Model for 23 African Languages; 2022.

[11] Kudo T, Richardson J. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing; 2018.

[12] Hochreiter S, Schmidhuber J. Long Short-term Memory. Neural computation. 1997 12;9:1735-80.

[13] Chung J, Gulcehre C, Cho K, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:14123555. 2014.

[14] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Advances in neural information processing systems. 2017;30.

[15] Post M. A Call for Clarity in Reporting BLEU Scores; 2018.

[16] Zongo B. Petit manuel du mooré pratique - Bernard Zongo;. Available from: https://www.edilivre.com/petit-manuel-du-moore-pratique-langue-du-burkina-faso-bernard-zo.html/.