Arewa NLP's Participation at WMT24

Mahmoud Said Ahmad[1], **Auwal Abubakar Khalid**[2], **Lukman Jibril Aliyu**[3],
Babangida Sani[4], **Mariya Sunusi Abdullahi**[5]

[1]*Federal University of Technology Babura (FUTB)*, [2]*Bayero University Kano (BUK)*, [3]*Arewa Data Science Academy*,
[4]*Arewa Data Science Academy*, [5]*Bayero University Kano (BUK)*
msahmad.cs@futb.edu.ng, aka2000078.mcs@buk.edu.ng, lukman.j.aliyu@gmail.com,
sunusimariya@gmail.com, bsani480@gmail.com

Abstract

This paper presents the work of our team, "ArewaNLP," for the WMT 2024 shared task. The paper describes the system submitted to the Ninth Conference on Machine Translation (WMT24). We participated in the English-Hausa text-only translation task. We fine-tuned the OPUS-MT-en-ha transformer model and our submission achieved competitive results in this task. We achieve a BLUE score of 27.76, 40.31 and 15.85 on the Development Test, Evaluation Test and Challenge Test respectively.

1 Introduction

Machine translation (MT) is widely regarded as one of the most successful applications of natural language processing (NLP). It has seen significant advancements, particularly in the accuracy of its results. While MT has achieved near-human performance for several language pairs, it still faces challenges when dealing with low-resource languages or when incorporating other modalities (such as images.(Parida et al., 2021).

In the broader field of machine learning and deep learning, multimodal processing involves training models using a combination of different information sources such as images, audio, text, or video. By incorporating multimodal data, models can learn features from various subsets of these sources (depending on the data modality), leading to improved prediction accuracy. Multimodal machine translation leverages information from multiple modalities, with the expectation that these additional modalities will offer valuable alternative perspectives on the input data. Despite machine translation's near-human performance for several language pairs, it still faces difficulties in translating low-resource languages and effectively utilizing other modalities. (Sen et al., 2022).

WMT is a workshop on Machine Translation. WMT24 features the English-to-Low-Resource Multimodal Translation Shared Task, which involves Bengali, Hausa, Hindi, and Malayalam datasets from the Visual Genome project. These datasets include both text and images, providing a rich resource for research in English-to-[Hindi, Bengali, Malayalam, Hausa] Machine Translation and Multimodal studies.(Parida et al., 2024; Scientist, 2024).

In this system description paper, we outline our approach to the English-Hausa text-only translation task.

2 Dataset

We utilized the Hausa Visual Genome (HaVG) dataset (Abdulmumin et al., 2022) provided by the organizers. This dataset comprises 32,923 images with corresponding descriptions, divided into training, development, test, and challenge-test sets. The training set includes 28,930 English and Hausa sentence pairs, while the development set contains 998 sentences, the evaluation test set has 1,595 sentences, and the challenge test set consists of 1,400 sentences. A summary of the sentence statistics is provided in Table 1.

3 Experimental Details

The experimental setup involved fine-tuning a pre-trained sequence-to-sequence language model, specifically the OPUS-MT-en-ha model, which was pre-trained on English-Hausa data. Fine-tuning was performed using PyTorch and Hugging Face Transformers. For the English-Hausa text-only translation task, we fine-tuned the OPUS-MT-en-ha model[1], a translation model pre-trained on English-Hausa data by the Language Technology Research Group at the University of Helsinki[2] .

[1]https://huggingface.co/Helsinki-NLP/opus-mt-ha-en
[2]https://github.com/Helsinki-NLP

Proceedings of the Ninth Conference on Machine Translation (WMT), pages 829–832
November 15–16, 2024.

Set	Sentences	Tokens	
		English	Hausa
Training set	28,930	147,219	144,864
Development test	998	5,068	4,978
Evaluation test	1,595	8,079	7,952
Challenge test	1,400	8,411	9,514
Total	32,923	-	-

Table 1: Statistics of data used in the English-Hausa text-only translation: the number of sentences and tokens.

3.1 Preprocessing

The Hausa Visual Genome dataset was prepared to train the translation model. The preprocessing phase involved preparing the Hausa Visual Genome (HaVG) dataset for training the translation model, The data was loaded using 'pandas' and converted into Hugging Face 'Dataset' objects for both English and Hausa texts. We employed the 'Helsinki-NLP/opus-mt-en-ha' tokenizer to tokenize the text, truncating or padding sequences to a maximum length of 128 tokens. The tokenized data was then formatted for PyTorch, including input IDs, attention masks, and labels, to ready it for training.

3.2 Model Fine-Tuning

Model fine-tuning is a crucial step in which the pre-trained model is adapted to the specific task of English-Hausa translation. We fine-tuned a pre-trained sequence-to-sequence language model using PyTorch and Hugging Face Transformers. The model was trained for 3 epochs with an AdamW[3] optimizer and a linear learning rate scheduler. Training was conducted on a GPU in batches of 8, with evaluation performed after each epoch. Upon completion, the fine-tuned model and tokenizer were saved. Fine-tuning not only enhanced the model's translation accuracy but also allowed it to perform well on different test sets, although it faced challenges with more difficult content as seen in the Challenge Test results.

This methodology enabled the model to achieve competitive BLEU scores on the various test sets, demonstrating its effectiveness in translating between English and Hausa, albeit with some room for improvement in handling more complex or less familiar content

[3] https://keras.io/api/optimizers/adamw

4 Results

Table 4 presents the results of automatic evaluation of our model.

Development Test (D-Test BLEU: 27.76): The model scored 27.76 on the Development Test set. This is a solid result, indicating that the model produces translations that are reasonably accurate, though there's some room for improvement. This test set is typically used during the model's development phase to fine-tune its performance.

Evaluation Test (E-Test BLEU: 40.31): On the Evaluation Test set, the model achieved a BLEU score of 40.31, which is quite a bit higher than on the Development Test set. This suggests that the model is particularly good at translating the kinds of sentences found in this set, perhaps because they are similar to what the model has seen during training.

Challenge Test (C-Test BLEU: 15.85): The model scored 15.85 on the Challenge Test set, which is significantly lower than the other two scores. This suggests that the Challenge Test set contains more difficult or unfamiliar content, making it harder for the model to produce accurate translations.

Zero-shot vs. Finetuned Scenarios

The zero-shot evaluation BLEU scores (table 3) are very low compared to the fine-tuned results (table 4). This demonstrates that without prior exposure or training on this specific data, the model struggles to perform accurate translations. These low BLEU scores suggest that the model's ability to generalize to completely unseen data (zero-shot scenario) is limited.

The significant difference between fine-tuned and zero-shot BLEU scores across all sets illustrates the importance of HaVG data. Fine-tuning has allowed the model to learn the translation patterns within the datasets, leading to far superior performance compared to the zero-shot setting.

English-Hausa Translation Examples

Table 2 presents sample English sentences alongside their Hausa translations, sourced from the challenge test set. Some examples are straightforward, where the model successfully translated simple, clear sentence structures. However, other examples are more challenging, showcasing the model's ability to handle complex or ambiguous translations. For instance, in examples 7 and 8, the word "cross" appears, which can refer to either a cruciform symbol or the act of crossing a street. The model accurately interpreted the context in both cases, delivering correct translations for each meaning. These more difficult examples illustrate the differences between the Dev, Eval, and Challenge sets, with the Challenge set specifically designed to test the model's performance by including context-dependent and nuanced sentences. The model's ability to navigate these complexities demonstrates its overall effectiveness.

S/N	English	Hausa Translation
1	A second pizza in a pan.	Pizza na biyu a cikin kwanon suya.
2	A girl on the tennis court is preparing to hit the ball.	Wata yarinya a filin wasan tanis tana shirin buga kwallon.
3	Knife block sitting on counter with knives in it.	Sandar wuka zaune akan kan tebur tare da wukake a ciki.
4	The players' socks are blue.	Yan wasan safa sune shui.
5	Balconies on the second story of the buildings.	Baranda akan bene na biyu na gine-ginen.
6	Beige stairway going to second level.	Matakala na beige zuwa bene na biyu.
7	The woman is waiting to cross the street.	Matar tana jira ta tsallaka titi.
8	A black cross on a vertical stabilizer.	Gicciye mai baar fata akan mai tsaye tsaye.
9	Man cross country skiing.	Mutum ya tsallaka kan asa a lokacin tsere.

Table 2: Sample of English to Hausa translations generated by our model.

D-Test BLEU	E-Test BLEU	C-Test BLEU
1.87	1.95	2.56

Table 3: Results of text-only translation task: Zero-shot

D-Test BLEU	E-Test BLEU	C-Test BLEU
27.76	40.31	15.85

Table 4: Results of text-only translation task: Fine-tuned model

5 Conclusion

This paper describes our system for English-to-Hausa text-only translation. The system performs well on more standard test sets (especially the Evaluation Test) but struggles with more challenging or unusual content, as seen in the Challenge Test results. This indicates that while the system is effective in many scenarios, it may need further training to handle more complex translation tasks. We plan to extend our work to include English-Hausa multimodal translation and image captioning tasks in the future.

Ethics Statement

In our work on the English-to-Hausa text-only translation task, we adhered to the highest standards of ethical research and data use. The datasets employed, including the Hausa Visual Genome dataset, were provided under appropriate licenses, and we ensured that all data used was handled in accordance with the terms specified by the providers. Our research also followed guidelines for responsible AI development, including fairness, transparency, and privacy considerations. We took particular care to avoid biases in our models that could negatively impact the communities whose languages we are working with. Additionally, we acknowledge the potential risks of deploying machine translation systems in sensitive contexts and emphasize the importance of human oversight in such applications.

Acknowledgements

We would like to express our gratitude to the organizers of WMT24 for their support and guidance throughout this project. Special thanks to Shantipriya Parida, Ondřej Bojar, Idris Abdulmumin, and Shamsuddeen Hassan Muhammad for providing the necessary datasets and for their insightful feedback. We also acknowledge the Language

Technology Research Group at the University of Helsinki for the OPUS-MT-en-ha model. Computational resources provided by our institutions were crucial for this research.

References

Idris Abdulmumin, Satya Ranjan Dash, Musa Abdullahi Dawud, Shantipriya Parida, Shamsuddeen Muhammad, Ibrahim Sa'id Ahmad, Subhadarshi Panda, Ondřej Bojar, Bashir Shehu Galadanci, and Bello Shehu Bello. 2022. Hausa visual genome: A dataset for multi-modal English to Hausa machine translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6471–6479, Marseille, France. European Language Resources Association.

Shantipriya Parida, Ondřej Bojar, Idris Abdulmumin, and Shamsuddeen Hassan Muhammad. 2024. Wat2024 english-to-lowres multi-modal translation task. https://ufal.mff.cuni.cz/wat2024-multimodal. Accessed: 2024-08-29.

Shantipriya Parida, Subhadarshi Panda, Ketan Kotwal, Amulya Ratna Dash, Satya Ranjan Dash, Yashvardhan Sharma, Petr Motlicek, and Ondřej Bojar. 2021. NLPHut's participation at WAT2021. In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 146–154, Online. Association for Computational Linguistics.

ML Scientist. 2024. Join the wmt2024: English-to-lowres multi-modal translation task! https://mlscientist.com/join-the-wmt2024-english-to-lowres-multi-modal-translation-task/. Accessed: 2024-08-29.

Arghyadeep Sen, Shantipriya Parida, Ketan Kotwal, Subhadarshi Panda, Ondřej Bojar, and Satya Ranjan Dash. 2022. Bengali visual genome: A multimodal dataset for machine translation and image captioning. In *Intelligent Data Engineering and Analytics*, pages 63–70, Singapore. Springer Nature Singapore.