A Deep Learning-based English to Yoruba Neural Translation Model

Olawale Timothy Adeboje
Department of Mathematical and Computing,
Koladaisi University Ibadan, Nigeria

Gabriel Junior Arome
Department of Cybersecurity,
Federal University of Technology, Akure. Nigeria

Olusola Adebayo Adetunmbi
Department of Computer Science,
Federal University of Technology, Akure. Nigeria

Raphael Olufemi Akinyede
Department of Information Systems,
Federal University of Technology, Akure. Nigeria

ABSTRACT
Yoruba is one of the most widely spoken languages in Africa and certain parts of the world. However, the dominance of the English language in Nigeria has contributed to the gradual decline of Yoruba language. This research was motivated by the need to address this by developing a Yoruba neural translation that aim to translate text written in English Language to Yoruba language. The research utilized a parallel corpus obtained from MENYO-20k and odunola/yoruba-english pairs (https://huggingface.co/datasets/odunola/yoruba-englishpairs). Transformer model was used, Bidirectional Encoder Representations from Transformers (BERT) was used for the encoder and T5-Base model on the decoder side of the transformer. The developed model achieved the BLUE score of 72%, which means a strong alignment between the translated outputs and reference texts, reflecting the model's capability to maintain the integrity of the original message.

Keywords
Deep Learning, Neural machine translation, BERT, T5 Model, NLP, Encoder, Decoder.

1. INTRODUCTION
The ability to effectively translate languages is a crucial aspect of global communication and cultural exchange. The process by which people or groups exchange ideas, thoughts, feelings, and information is known as communication. It entails sending and receiving signals via a variety of mediums, such as writing, speaking, body language, gestures, and contemporary technology tools like social media and emails[1]. In Nigeria, English language has been the official language since the British colony. The dominance of English language in Nigeria is gradually making the indigenous languages to be going extinct. This is because English language is used to teach children at school, English language is used in all sectors, judiciary, political, medical and even religion organization. This has brought significant cultural and linguistic consequences for national identity, leaving many children unable to communicate in their native language. Among the languages in Nigeria are Yoruba, Igbo, Hausa and so on.

Yoruba language is celebrated for its rich linguistic and cultural heritage, encompassing a longstanding tradition of written and oral literature, arts, religion, and philosophy. The language boasts a wide variety of dialects, reflecting the geographical and historical diversity of the Yoruba people. As one of the twelve languages of the Edekiri sub-branch of the Niger-Congo language family, Yoruba, which emphasizes the first syllable, is a member of the Kwa branch and one of Nigeria's three major languages, Yoruba features approximately twenty dialects, which differ in phonology and vocabulary. These dialects are also spoken in regions beyond Nigeria with linguistic and cultural ties to the Yoruba people, such as the Republic of Benin, Togo, and Sierra Leone [2]. As language and culture are deeply interconnected, the decline in Yoruba usage among younger generations has accelerated the erosion of our cultural heritage. Consequently, the younger population is increasingly disconnected from the fundamental values and traditions of their culture. This shift is further reflected in their fashion choices, which increasingly align with those of dominant language groups [1]. Therefore, this study focuses on the translation of Yoruba text to English text using a transformer-based model, a technique that has shown promising results in various language translation tasks.

The Transformer model is a type of neural network [3] initially recognized for its exceptional performance in machine translation. Today, it has become the standard for constructing extensive self-supervised learning systems. In recent years, Transformers have surged in prominence not only within natural language processing (NLP) but also across diverse domains like computer vision and multi-modal processing. As Transformers advance, they are progressively assuming a pivotal role in advancing both the research and practical applications of artificial intelligence (AI) [4].

2. LITERATURE REVIEW
In [5], the researchers aimed to achieve English-to-Yoruba text translation using a rule-based method. Their approach primarily utilized a dictionary-based rule-based system, considered the most practical among various types of machine translation methods. However, a limitation of this approach was its inability to accurately translate English words with multiple meanings, influenced by their grammatical context.

In [6], the researchers aimed to create a machine translator that converts English text into Yorùbá, thereby making the Yorùbá language more accessible to a wider audience. The translation method used rewrite rules and phrase structure grammar. Utilising Natural Language Tool Kits (NLTKs), these rules were created and assessed. Parse tree and Automata theory-based analytical methods were used for the analysis. However, the study noted that accuracy decreases as the length of sentences increases.

In [7], the research aims to create a translation system between English and two Nigerian languages: Igbo and Yorùbá. The study employs a phrase-based Statistical Machine Translation approach, which involves grouping language words into

sequences, translating each phrase into the target language, and optionally reordering them based on target language models and distortion probabilities. However, the research faces challenges such as insufficient data, orthographic errors, and high error rates during system compilation.

In [8] titled Japanese-to-English Machine Translation Using Recurrent Neural Networks system. The research objective is to translate Japanese language to English language. Bidirectional recurrent neural networks (BiRNNs) were employed as the encoder. They have two hidden states: one that reads the source phrase in reverse order, and the other that reads it in order. However, there were no complex sentence translations.

In [2], the authors presented the development of a POS tagger tailored for the Yoruba language text using deep neural network technique, aimed at enhancing the performance of natural NLP applications and high-level tasks. To generate a sequence of tags, the current POS taggers for Yoruba either use stochastic approaches, which are highly redundant, or rule-based systems, which are constrained by the correctness and comprehensiveness of the stated rules. This study is therefore driven by the need to promote the use of machine learning models to create reliable and extremely efficient POS taggers that are customised for Yoruba text. Data collection encompassed sources such as religious texts, newspapers, and literature books, resulting in the extraction of 621 sentences, which were subsequently preprocessed and tokenized, yielding a corpus of 20,795 words. The tokenized sentences underwent manual tagging with the correct POS labels, providing labeled data for training the deep neural network model. This model was then employed to predict the POS tags for untagged words, leveraging the learned associations between words and their respective parts of speech. However, there is a need to expand the number of POS categories considered beyond the initial eight.

3. METHODOLOGY

This research aims to translate English language text into Yoruba language text using transformer model. The neural machine translation module converts text written in English language to Yoruba language text. Transformer model has been proven to solve the problems of Recurrent Neural Network. To illustrate global interdependence between input and output, the Transformer model architecture completely relies on an attention mechanism rather than recurrence. The Transformer model is divided into two macro blocks, which are : Encoder and Decoder. Figure 1 shows the architecture of the developed model.

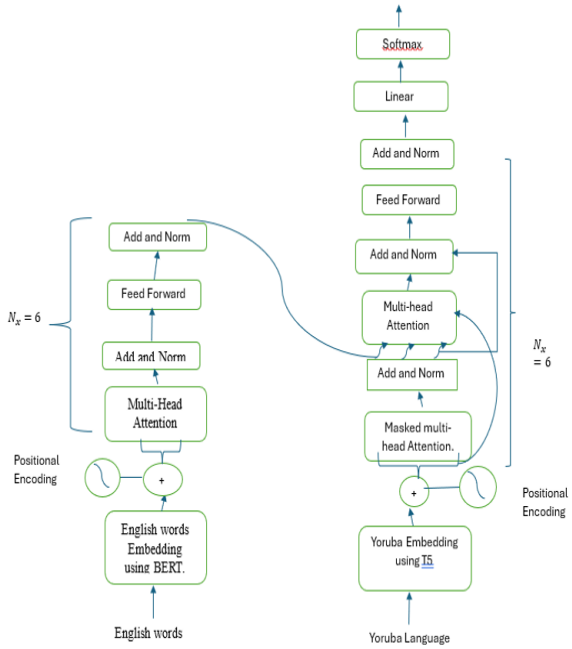

Figure 1: The Architecture of the developed model

3.1 Dataset

The corpus text was obtained from two sources; MENYO-20k and odunola/yoruba-English-pairs (https://huggingface.co/datasets/odunola/yoruba-english-pairs) because they have a nonrestrictive license, the Yorùbá sentences have been further verified for quality issues and to have a robust Yoruba corpus. With texts gathered from news stories, TED presentations, movie and radio transcripts, science and technology texts, and other brief pieces selected from the internet and expert translators, MENYO-20k is an open source, multi-domain parallel dataset. There are 20,100 parallel sentences in the dataset, which are divided into 6,633 test sentences, 3,397 development sentences, and 10,070 training sentences (3,419 multi-domain, 1,714 news domains, and 1,500 ted talks speech transcript domains).

3.2 Encoder

The encoder encodes the English language input sequence and passes it to the decoder. It is responsible for language learning and identification. A stack of N = 6 identical layers makes up the encoder block. There are two sub-layers in every layer. One is a fully connected feed-forward network, and the other is a multi-head self-attention mechanism. Bidirectional Encoder Representations from Transformers (BERT) were employed in the study as the encoder. The input (English language text) was transformed into embedding vector that can easily be understood by the BERT algorithm. The purpose of the language representation model BERT is to extract context-sensitive characteristics from the input text by pre-training deep bidirectional representations. It is a neural network-based method for pre-training language processing. The input (string) gets converted into sets of tokens; the token goes through the token embedding where it is vectorize which has contextual meaning. The BERT embedding text is based on token, segment and position. Consider the embedding of the $i-th$ word in the sequence, denoted by

$$word_{input} = x_i \in R^d \quad (1)$$

where x_i is the word in the sequence, R^d is the transformation matrix and $word_{input}$ is the transformed matrix.

Positional Embedding was calculated for time set to even and odd respectively in equations 2 and 3:

$$PE\ (POS,\ 2i) = \sin \frac{POS}{1000^{2i/dmodel}} \quad (2)$$

$$PE\ (POS,\ 2i+1) = \cos \frac{POS}{1000^{2i/dmodel}} \quad (3)$$

where PE = position encoding, POS = position of the sentence, i = index position in the input sequence, d_{model}= dimension of the vector.

After, the sum of position segment and token embedding to form the input for the BERT model

$$BERT_{input} = PE + \text{TE} + \text{SE} \quad (4)$$

where PE is the positional encoding, TE is the Token encoding and SE is the segment Encoding, $BERT_{input}(\ seq,\ d_{model})$ where seq= sequence of length and d_{model} is the size of the embedding vector.

At the Multi head attention layer, BERT embedded input $BERT_{input}(\ seq,\ d_{model})$ will be divided into four, one will be sent to the Add and Norm phase and three will be sent to the Multi-Head Attention. The three will be Query (Q), Key (K) and Value (V).

The attention between the Q,K and V is calculated by

$$Attention\ (Q,K,V) = \ softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)v \quad (5)$$

This enables information from several subspaces at various places to be concurrently attended to by the model. This is inhibited by averaging with a single attention head.

$$head_i = Attention(QW_i^Q\ , KW_i^k\ , VW_i^v\ ) \quad (6)$$

with the new head, the next is to concatenate the heads by :

$$MultiHead(Q,\ K,\ V\ ) \ = \ Concat(head_i,\ \ldots, head_h)W^O \quad (7)$$

where Q=Query, K= Key and V=Value

The residual connection is done to ensure that there is a strong information signal that flows through deep networks. This is required because during back propagation, there is vanishing gradient, so to prevent that, we induce small strong signal from the input to different parts of the network.
After the Multi Head attention, the result is added to the embedded input using the residual connection.

$$v_{(seq,\ d\ model)} = x + MultiHead\ _{(Q,V,K)}(x) \quad (8)$$

where x is the input embedded and $MultiHead\ _{(Q,V,K)}(x)$ is the resultant multi head attention process

The next is to perform a layer normalization and this is calculated by :

$$\text{Layer Norm}\ (v_{(seq,\ d\ model)}) = \gamma\frac{v-\mu}{\sigma} + \beta \quad (9)$$

where μ= mean, σ = standard deviation, γ= gamma (multiplicative), β =beta (additive) and will initially set to 1 and 0.

The Feed Forward Network (FFN) comprises two dense layers that are individually and uniformly applied to every position. The Feed Forward layer is primarily used to transform the representation of the input sequence into a more suitable form for the task at hand. This is achieved by applying a linear transformation followed by a non-linear activation function. The output of the Feed Forward layer has the same shape as the input, which is then added to the original input.

3.3 Decoder

The output of the encoder will be passed to the next encoder and the adjacent decoder. The research utilized Text-to-Text Transfer Transformer (T5) model on the decoder side of the transformer. The baseline or standard form of T5 is called T5-Base. It is popular and suitable for a wide range of NLP jobs since it finds a balance between model complexity and efficiency. It is a decent place to start for most applications and provides a good trade-off between computational cost and performance. The goal of the decoder is to translate English language to Yoruba language. The output of the Encoder (value and key) is joined to the output (Query) of the masked multi head attention in the decoder phase.

4. RESULT

The integrity of a system rests on the outcome of its evaluation. As part of the effort to examine the efficiency of the developed system, standard metrics such as: Bilingual Evaluation Understudy (BLEU score), Recall, F1 score and Precision. Table 1 shows the result of the evaluation. The BLEU score evaluates the quality of the text output that is received via a machine translation. The precision defines the ratio of relevant information that has been retrieved from the model versus all the information that has been retrieved by the model. It is calculated using equations

$$precision = \frac{correct}{length_o} \quad (10)$$

where correct is the count of words that are translated correctly, and the length is the length of the output of the translation, in other words, it is the total number of words in the output.

Recall measures are defined as the ratio of the relevant information that is retrieved versus all the relevant information that exists in the input.

$$\text{recall} = \frac{correct}{length_r} \quad (11)$$

where correct is the count of words that are translated correctly, and length is the length of the reference of the translation, in other words, it is the total number of words in the desired or reference output.

The F1 score is a better measurement tool as it allows for the balance between precision and recall measures.

$$F1\ score = \frac{2*(precision)(recall)}{precision+recall} \quad (12)$$

Table 1 shows the results of the system when being evaluated using Bleu score, recall, precision and F1 score.

Table 1 : The result of the system evaluation

Metric	Value	Percentage
BLEU Score	0.7200	72
Precision	0.8500	85
Recall	0.8200	82
F1 score	0.8300	83

The figure below depicts the graphical representation. The percentage on the y axis and the evaluation metrics on the x axis.

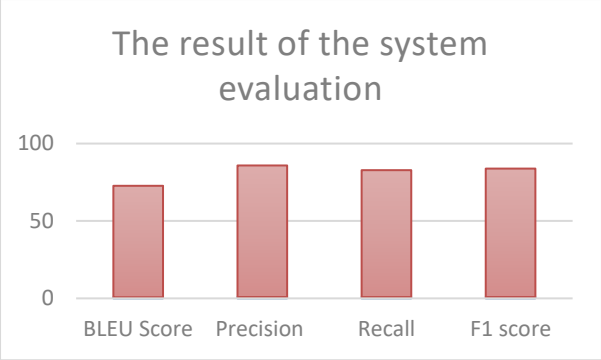

Figure 2: The graphical result of the system evaluation

The developed model achieved the BLEU score of 72%, which means a strong alignment between the translated outputs and reference texts, reflecting the model's capability to maintain the integrity of the original message. The word achieved the precision, recall and F1 Score of 85%, 82% and 83% respectively. Precision quantifies the accuracy of the positive prediction, recall quantifies the percentage of real positive cases that the model correctly recognized, and the F1 Score evaluates the model's accuracy by combining precision and recall.

5. CONCLUSION

Few researchers have worked on neural machine translation, particularly focusing on English-to-Yoruba translation. There is need to address the emerging concerns and limitations which includes but not limited to the decrease in translation accuracy as the sentence increases and the issue of ambiguities in translation, inadequacies of data, orthographic errors and high error rates at the compilation level of the system, handling small vocabularies, inadequacies in the areas of missing words, wrong words or spellings, semantic and syntax errors as well as grammatical and word order issues, low pronunciation accuracy, using a deep learning-based text-to-speech translation model.

The developed model achieved the BLUE score of 72%, which means a strong alignment between the translated outputs and reference texts, reflecting the model's capability to maintain the integrity of the original message. The developed system achieved high percentage of precision, recall and F1score.

The research is highly needed for individuals who are learning Yoruba and need to practice or translate English texts into Yoruba to understand and improve their skills. Also, for travelers and tourists visiting Yoruba-speaking regions, who may need to communicate with locals in their native language.

Future work should include the implementation of Yoruba speech on machines such as Automated Teller Machine, Cell Phones, Smart televisions and so on, to accelerate the development of technology for the Yorùbá language

6. REFERENCES

[1] Timothy, A. O., Adebayo, A. O., Junior, A. G., & Olufemi, A. R. Bilingual Neural Machine Translation from English To Yoruba Using A Transformer Model.

[2] UGWU, C. C., OYEWOLE, A. R., POPOOLA, O. S., ADETUNMBI, A. O., & ELEBUTE, A. (2024). A Part of Speech Tagger for Yoruba Language Text using Deep Neural Network. *Franklin Open*, 100185.

[3] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention Is All You Need.(Nips), 2017. arXiv preprint arXiv:1706.03762, 10, S0140525X16001837.

[4] Xiao, T., & Zhu, J. (2023). Introduction to Transformers: an NLP Perspective. arXiv preprint arXiv:2311.17633.

[5] Mishina, U. L., & Iskandar, I. (2019). The role of English language in Nigerian development. GNOSI: An Interdisciplinary Journal of Human Theory and Praxis, 2(2), 47-54.

[6] Eludiora, S. I., & Odejobi, O. A. (2016). Development of an English to Yorùbá Machine Translator. International Journal of Modern Education and Computer Science, 8(11), 8.

[7] Ayogu, I. I., Adetunmbi, A. O., & Ojokoh, B. A. (2018). Developing statistical machine translation system for english and nigerian languages. Asian Journal of Research in Computer Science, 1(4), 1-8.

[8] Greenstein, E., & Penner, D. (2015). Japanese-toenglish machine translation using recurrent neural networks. Retrieved Aug, 19, 2019.