Development of a Recurrent Neural Network Model for English to Yorùbá Machine Translation

Adebimpe Esan[1]*, John Oladosu[2]*, Christopher Oyeleye[3]*, Ibrahim Adeyanju[4]
Olatayo Olaniyan[5], Nnamdi Okomba[6], Bolaji Omodunbi[7], Opeyemi Adanigbo[8]
Department of Computer Engineering, Federal University Oye-Ekiti, Ekiti state, Nigeria[1,4,5,6,7,8]
Ladoke Akintola University of Technology, Ogbomoso[2,3]

***Abstract*—This research developed a recurrent neural network model for English to Yoruba machine translation. Parallel corpus was obtained from the English and Yoruba bible corpus. The developed model was tested and evaluated using both manual and automatic evaluation techniques. Results from manual evaluation by ten human evaluators show that the system is adequate and fluent. Also, results from automatic evaluation shows that the developed model has decent and good translation as well as higher accuracy because it has better correlation with human judgment.**

***Keywords*—Recurrent; tokenizer; corpus; translation; evaluation; correlation**

I. Introduction

The demand for translation and translation tools currently exceeds the capacity of available solution [1], hence, the need to intensify research in the field of machine translation [2]. Machine Translators (MT) accept characters of source language and map to the characters of the target language to generate the words with the help of various rules and other learning process techniques [3]. Previous researchers have employed various approaches to develop machine translators and the approaches were categorized into two by [4], namely; single and hybrid approaches. Single approaches include: rule-based, knowledge-based, statistical and direct approaches while Hybrid approaches are: word-based, phrase-based, syntax-based, forest-based and neural machine translation models.

Neural Machine Translation (NMT) is an improvement in the field of machine translation where a large neural network is built and trained to read a sentence and output a correct translation [5]. The approach consists of the encoder and the decoder for encoding a source sentence and decoding it to a target sentence [6]. Neural machine translators have shown promising results than previous MT approaches through the incorporation of some neural components to existing translation systems like phrase-based systems [7]. In addition, research revealed that NMT produces automatic translations that are significantly preferred by humans when compared to other machine translation approaches.

However, the most widely used model for NMT is the Recurrent Neural Network model which is a supervised machine learning model that is made of artificial neurons with one or more feedback loops. In order to train a RNN, a parallel corpus is trained so as to minimize the difference between the output and target pairs by optimizing the weights of the network [8]. In addition, a portion of the corpus is used as the validation dataset [9] to watch the procedure during training and prevent the network from underfitting or overfitting. RNNs have distributed hidden states used for storing information about the past efficiently and non-linear dynamics for updating their hidden state [10]. Hence, this research developed a recurrent neural network model for English to Yoruba machine translation.

*Corresponding Author

II. Related Works

Neural Machine Translation (NMT) is an improvement in the field of machine translation and it is based purely on deep neural networks. The encoder–decoder architecture [5] which is a conventional approach to neural machine translation, encodes a whole input sentence into a fixed-length vector from which a translation was decoded. Research show that the use of a fixed-length context vector is a challenge for the translation of longer sentences, hence, the research was extended by developing a model that soft- search for a set of input words, or their annotations computed by an encoder, when generating each target word [11]. The method prevents the model from encoding all the source sentences into a fixed-length vector but focuses only on relevant information that will help to generate target word. This approach outperformed the conventional encoder-decoder model significantly.

However, as training and decoding complexities increase proportionally to the number of target words in previous NMT systems, the size of the target vocabulary was extended by using an approach that enables training a model with much larger target vocabulary without substantial increase in computational complexity [12]. Decoding was efficiently done using a very large target vocabulary by selecting a small portion of the target vocabulary. Research show that the models trained outperformed the baseline models with a small vocabulary size. Though, it is unable to translate words which could not be found in the vocabulary. Therefore, alignment-based technique was used by [13] to mitigate this problem. The technique was carried out by training the model on data that is augmented by the output of a word alignment algorithm, allowing the NMT system to emit, for each Out of Vocabulary (OOV) word in the target sentence, the position of its corresponding word in the source sentence.

Moreover, [14] developed a multi-task learning model by training a unified neural machine translation model. In the research, an encoder is shared across different language pairs and each target language has a separate decoder. The

challenge with this model is the inability to address the data scarcity problem of some resource-poor language pairs. Thus, attention mechanism was incorporated in the models by [15] to overcome the problem. The attention mechanism was incorporated into the multi-task neural machine translation model and this method helps eliminate the data scarcity problem of the baseline model. Despite this achievement, the model still relies on word-level modeling.

Therefore, to reduce reliance of MT systems on word-level modeling, an attention-based encoder– decoder with a sub word-level encoder and a character-level decoder were developed for NMT [16]. The approach focused on the target side, in which a decoder generated one character at a time, while soft-aligning between a target character and a source sub-word. Research showed that the character-level decoder outperformed the sub-word-level decoder. Finally, [17] also addressed the data scarcity problem by developing a multi-task learning model by training a unified neural machine translation model with the decoder shared over all language pairs and each source language has a separate encoder. Research showed that given small parallel training data, the model was effective in learning the predictive structure of multiple targets.

III. Methodology

A. Design of the Recurrent Neural Network Model

The RNN model was designed to include three layers: input layer, hidden layer and output layer. The input layer was designed to have N input units and the inputs to this layer is a sequence of vectors through time *t* such that $\{x_{t-1}, x_t, \dots x_{t+1}\}$, where $x_t = (x_1, x_2, \dots x_N)$. The encoder RNN reads the input sentence which is a sequence of vectors $x = (x_1, \dots xT_x)$ into a vector c as in equations 1 and 2:

$$h_t = f(x_t, h_{t-1}) \quad (1)$$

and

$$c = g(h_1 \dots . h_{Tx}) \quad (2)$$

Where $h_t \in \mathrm{R}^n$ is a hidden state at time *t*, and *c* is a vector generated from the sequence of the hidden states f and g are non-linear functions. The input units were connected to the hidden units in the hidden layer, where the connections are defined with a weight matrix W_c.

In addition, at the hidden layer, this research modified the recurrent neural network of [5] by estimating the distribution with an attention mechanism [13] to overcome the shortcoming of previous research [7]. The hidden layer has *M* hidden units $h_t = (h_1, h_2, \dots, h_M)$. The source encoder recurrent neural network (RNN) maps each source word from the input unit to a word vector and processes these to a sequence of hidden vectors $h_1 \dots . h_t$ as shown in equation 1. The source hidden vectors influence the distribution through an attention pooling layer h_t that weighs each source word relative to its expected contribution to the target prediction as shown in equation 3.

$$h_{t=}^2 \tan h(W_c\,[C_t\,;\,h_t]) \quad (3)$$

From equation 3, *tan h* is the activation function, W_c is the weight matrix, C_t is the context vector and h_t are the hidden states. This research used *tanh* activation function at the hidden layer to overcome the shortcoming of the sigmoid function in previous work [5].

Moreover, at the output stage, previous model [7](Cho et al. 2014) mapped the input vector to the target sequence with another RNN during sequence learning and this prevented the model from learning long term dependencies while training the RNNs. Therefore, this research computed the output layer by combining the RNN (GRU) hidden representation of previously generated words $(w_1, \dots \dots . w_{t-1})$ with source hidden vectors to predict scores for each possible next word as shown in equation 4. The activation in the GRU was modeled as equation 4:

$$y_t = W_t h_t \quad (4)$$

where

$$h_t = (1 - Z_t)h_{t-1} + Z_t \hbar_t \quad (5)$$

From equation 5, Z_t is the update gate which controls the update value of the activation as shown in equation 6.

$$Z_t = (W_z X_t + U_z h_{t-1}) \quad (6)$$

From equation 3.6, *W* and *U* are weight matrices to be learnt.

The candidate activation is shown in equation 7;

$$\hbar_t = tanh(W_h X_t + U_h(r_t \odot h_{t-1}) \quad (7)$$

Where r_t is a set of reset gates defined as equation 8:

$$r_t = \sigma(W_r X_t + U_r h_{t-1}) \quad (8)$$

The diagram of the designed RNN model is shown in Fig. 1.

B. Method and Size of Data Collection

Fourteen thousand two hundred and forty five sentences were extracted from the bible parallel corpus for training the system and five thousand sentences extracted from the bible parallel corpus were used in validating the system. The developed system was tested automatically using two different data sets from two literatures: five hundred and eighty eigthy (588) sentences and one thousand (1,000) sentences respectively. Manual testing was also done using one two hundred (200) sentences from the third literature.

C. Text Preprocessing

Pre-processing of this corpus was carried out in three phases, namely; data loading, tokenization and vocabulary building.

1) Data loading: The parallel corpus was loaded as strings into memory. Every English sentence is placed on a line with its corresponding Yoruba translation and separated by a TAB. Cases were ignored and spaces were added between words and punctuation marks.

2) Tokenization: Morphology-based and frequency based tokenization approaches were used in tokenizing the coprus

used. Morphology based tokenization was employed to split off punctuation and numbers. Frequency-based tokenization was carried out using byte-pair encoding (BPE) [18].

3) Vocabulary building: Tokens that rarely appeared were mapped into a special unknown ("<unk>") token. Special tokens like: "<pad>", "<bos>" and "<eos>" were added for padding, beginning of sentence and end of sentence respectively.

D. Training of the Model

Training of data was done at every 1000 checkpoint. The model was validated on a dataset of 5000 sentences with an accuracy of 60.051. The sequence diagram of the designed model is shown in Fig. 2 while the class diagram of the designed model is shown in Fig. 3.

E. The Developed RNN Model

The developed model is shown in Fig. 4. From the diagram, source words were depicted by colour yellow while target words were depicted by blue colour. The source words were first mapped to word vectors and then fed into a recurrent neural network (RNN). At the end of sentence <eos> symbol was displayed and the final time step initializes a target RNN. At each target time step, attention was applied over the source RNN and combined with the current hidden state to produce a prediction of the next word $p(w_t|w_{1:t-1}, x)$. This prediction was then fed back into the target RNN. The developed system was tested and implemented for English to Yoruba translation. Fig. 5 shows a sample page of the developed system and Fig. 6 shows a sample page that confirms the ability of the system to translate long sentences.

F. Evaluation of the Developed Model

The Modified Recurrent Neural Network model was evaluated using Human judgment and Bilingual Evaluation Understudy (BLEU). Ten human evaluators evaluated the developed model using adequacy and fluency metrics [19] on a 5 point likert scale (over 0 to 4). The guidelines for evaluation required that the following score be given to a sentence by looking at each output sentence on a 5 point Likert scale (over 0- 4): 4=: All Meaning,3 =: Most meaning, 2 =: Much meaning, 1=: Little meaning and 0=: No meaning. The overall adequacy of the system was computed using the formula by [20] where the total number of sentences with scores 2, 3 and 4 were added and divided by the total number of sentences N as shown in equation 9.

$$Adequacy = \frac{Scores\ 2,3,4}{N} \quad (9)$$

The correct grammatical constructions present in the translated sentences were evaluated using fluency metric based on the research by [21]. The guidelines for evaluation required that the following scores be given to a sentence by looking at each output sentence on a 5 point Likert scale (over 0- 4): 4 := for Perfect:, 3 := for Good, 2 := for Non-native, 1 := for Diffluent, 0 := for Incomprehensible. The overall fluency of the system was computed using the formula by [22] as shown in equation 10. Scores above 2 were considered. Scores 3 and 4 were penalized by multiplying their count by 0.8 and 0.6 respectively so as to make the estimated score better.

$$(F = 100 * ((S4 * 0.8 + S3 * 0.6 + S2))/N) \quad (10)$$

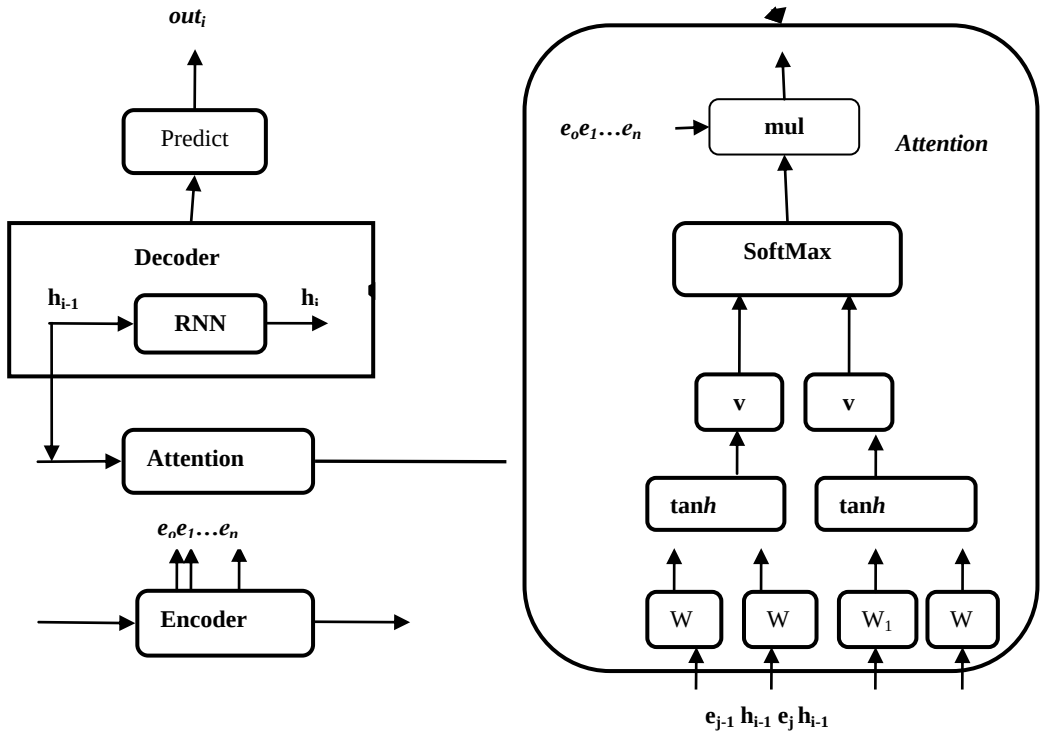

Fig. 1. Design of the RNN Model.

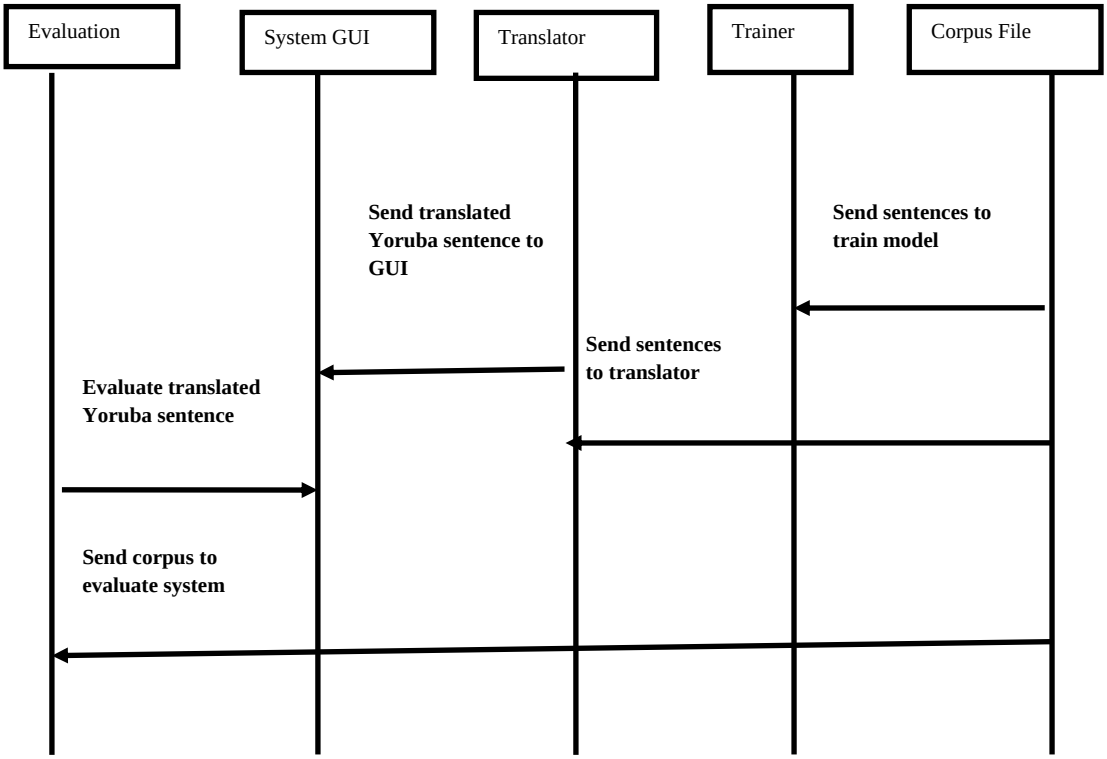

Fig. 2. Sequence Diagram of the Developed Model.

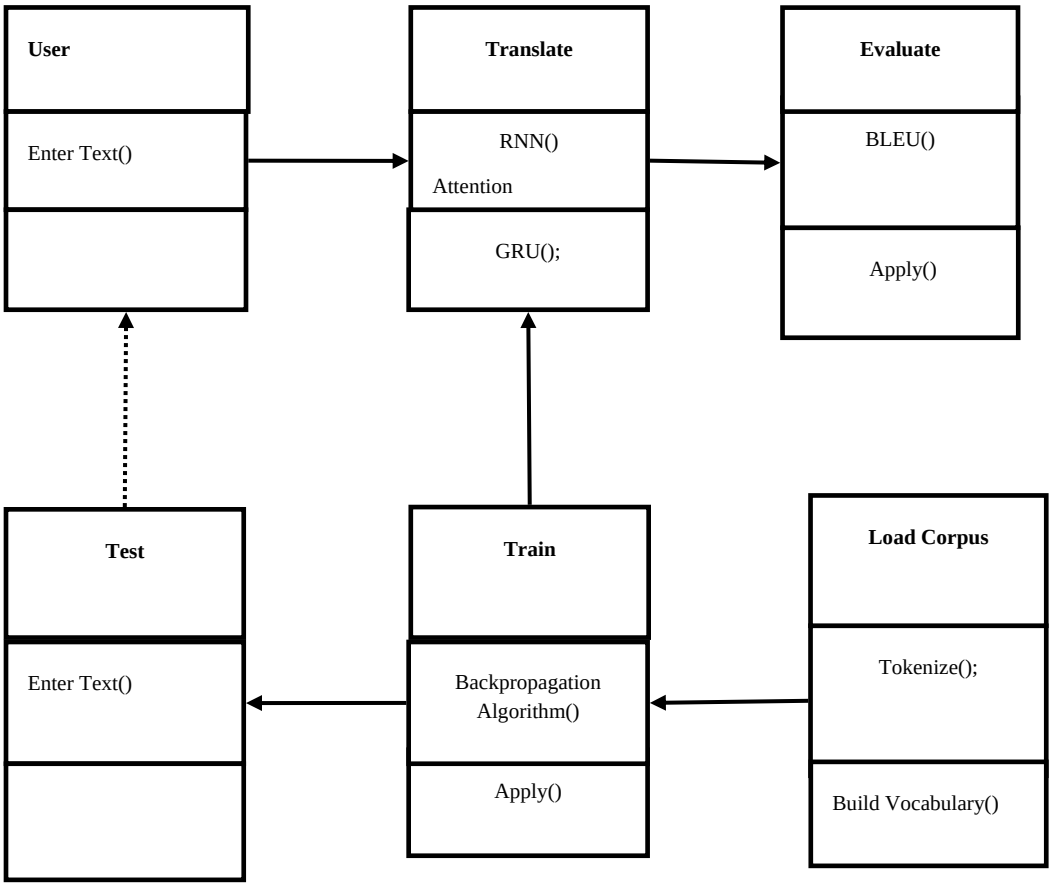

Fig. 3. Class Diagram of the Developed Model.

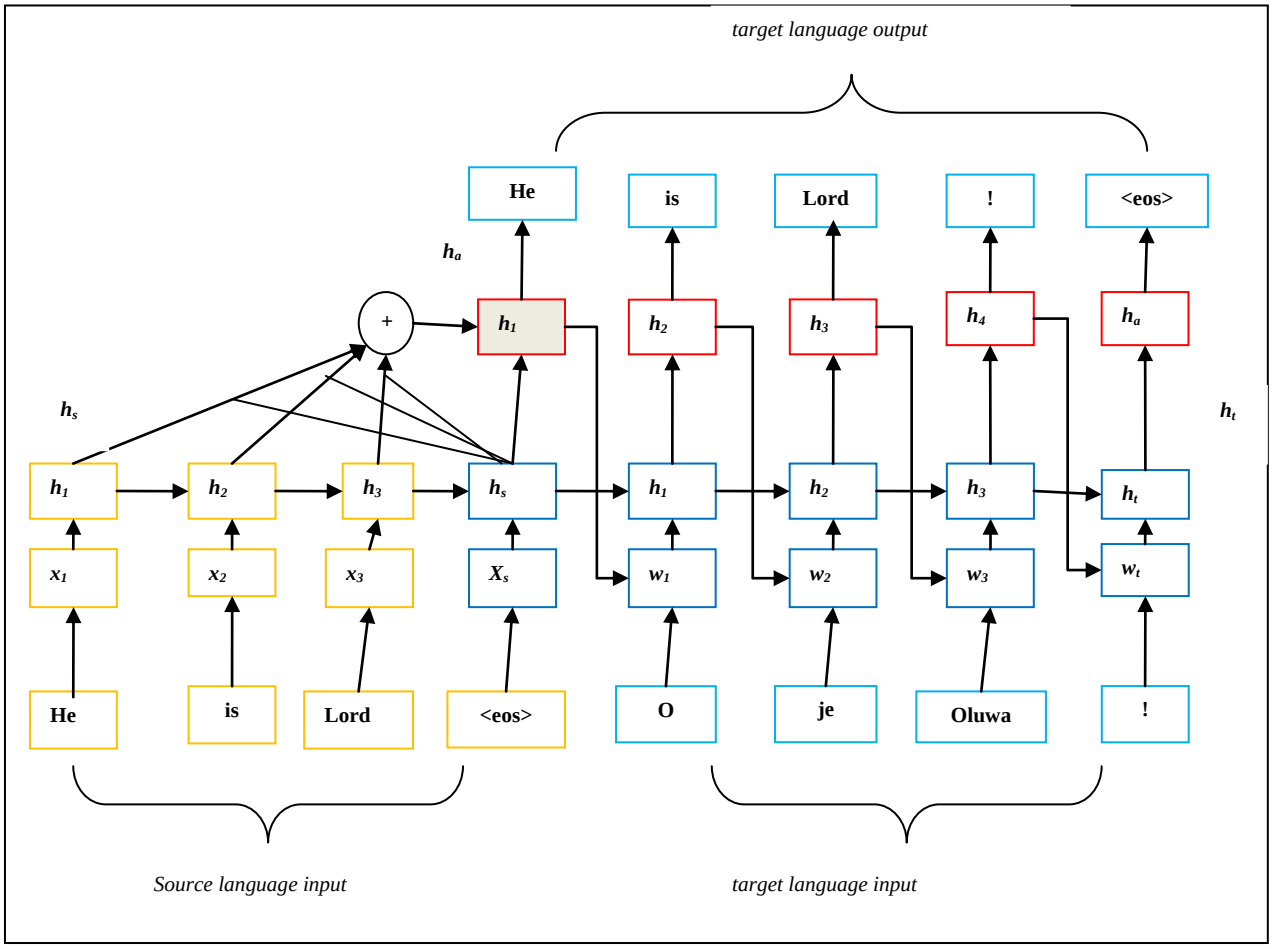

Fig. 4. The Developed RNN Model.

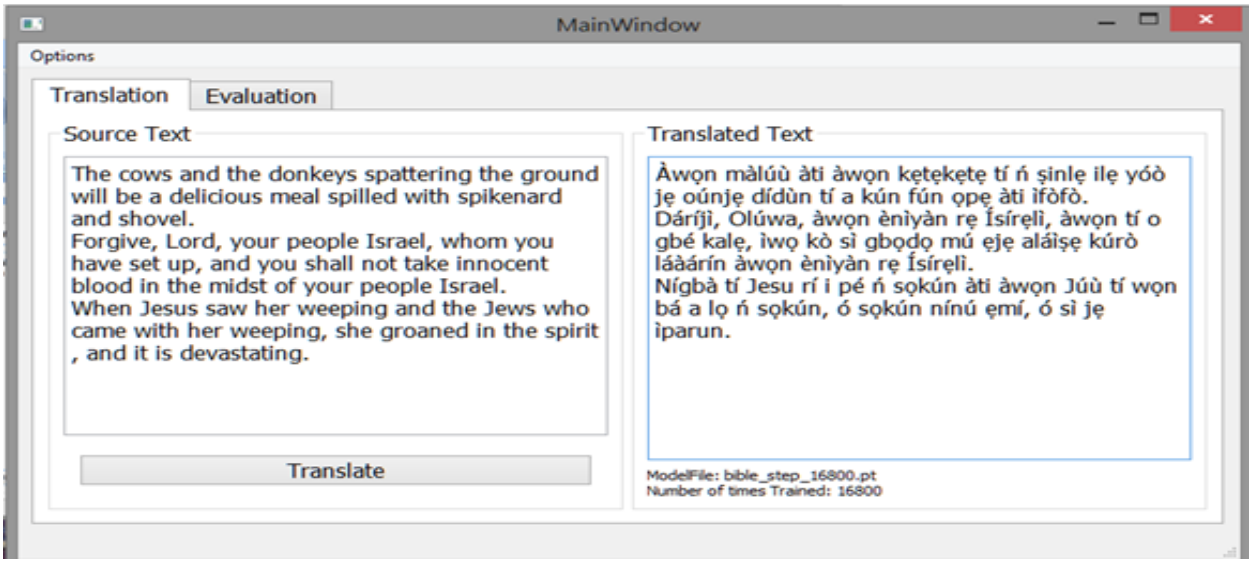

Fig. 5. Sample Page of the Developed NMT Model.

Fig. 6. Sample Page that Shows the Ability of the Model to Translate Long Sentences.

IV. Results

Results from ten human evaluators were computed using the adequacy formula from [20] as shown in Table I and results from ten human evaluators were computed using the fluency formula from [22] and the results obtained are shown in Table II. The average scores for adequacy and fluency metrics for the developed system were computed and recorded in Table III. The overall average for adequacy and fluency metrics are 86.65 and 70.72, respectively.

TABLE I. Adequacy Metric Scores for the Developed System

SCALE 0-4	**RNMT**	**ADEQUACY**	**METRIC**	**SCORES**						
	User 1	User 2	User 3	User 4	User 5	User 6	User 7	User 8	User 9	User 10
0	9	8	10	7	6	4	5	7	5	6
1	25	21	23	18	19	20	17	20	19	18
2	42	38	40	35	37	32	30	31	36	34
3	50	52	49	54	53	53	50	51	54	51
4	74	81	78	86	85	91	98	91	86	91

TABLE II. Fluency Metric Scores for the Developed System

SCALE 0-4	**RNMT**	**FLUENCY**	**METRIC**	**SCORES**						
	User 1	User 2	User 3	User 4	User 5	User 6	User 7	User 8	User 9	User 10
0	6	10	7	3	8	9	9	5	6	4
1	12	15	13	7	10	10	11	10	11	9
2	49	42	45	47	40	43	41	43	42	46
3	60	62	67	69	70	68	66	65	72	71
4	73	71	68	74	72	70	73	77	69	70

TABLE III. Computed Scores for Adequacy and Fluency Metrics

METRIC	**CALCULATED**	**METRIC**	**SCORES**							
	User 1	User 2	User 3	User 4	User 5	User 6	User 7	User 8	User 9	User 10
Adequacy	83.0	85.5	83.5	87.5	87.5	88.0	89.0	86.5	88.0	88.0
Fluency	71.7	68.4	69.8	73.8	69.8	69.9	69.5	71.8	70.2	72.3

V. Discussion

Results from adequacy and fluency metrics of the developed system were compared and it was discovered that the system's adequacy score is higher than the fluency score. This is in line with the research by [20] where Google translate' Comprehensibility (adequacy) score was found to be higher than its fluency score. Also, the overall adequacy score of the developed system show that neural machine translators are more adequate than other MT approaches. This is according to the research by [22] where English language was translated to Hindi using Rule based and Statistical approaches and percentage adequacy scores recorded was significantly lower than the developed system. Moreover, the developed system has an improved quality output when compared to previous systems (Google and Bing SMT systems) [20] because its adequacy and fluency scores are higher than the previous systems. The results obtained in this research also confirm that NMT systems give relatively high accuracy when trained on a larger dataset and can yield good predictions as well [23]. The results from manual evaluation also affirms that GRU based RNN provides a stronger and more robust translation model with high resourced languages [24].

Automatic evaluation of the developed system gives a percentage BLEU score of 54.8 when tested on a dataset of five hundred and eighty eight sentences and 57.3 when tested on a data set of one thousand sentences. The average BLEU score obtained from the two datasets is fifty-six percent (56%). The research by [25] reveals that BLEU scores up to 50 and above generally reflect good and fluent translations. Author in [26] also confirms that a BLEU score of 50 implies a decent translation. Hence, the result shows that the system has good and fluent translation.

In addition, research by [6] revealed that neural machine translators have higher quality output than phrase-based machine translators. Hence, the results obtained from automatic evaluation of the developed system were compared to the results from previous system [1] and the BLEU score is significantly higher than the previous system. The performance of neural machine translators was also confirmed to be higher than statistical machine translators by [27] when English was translated to Arabic language. The NMT model outperforms SMT model by 1.5 BLEU in the out-of-domain testing.

In other words, this research establishes that NMT systems have higher quality output than previous rule based and statistical MT sytems. This was established by comparing the result obtained from the developed system to the research by [22] and a significant improvement in BLEU score was recorded. The result also revealed that Recurrent Neural network Models coupled with an attention mechanism produces higher quality output than previous neural machine translators. This was confirmed by [28] where different neural machine translators were compared and the recurrent neural MT outperformed other approaches considered.

VI. Conclusion

This research developed a recurrent neural network model for English to Yoruba machine translation. The model was tested and evaluated using both human and automatic evaluation techniques and the system was found adequate and fluent with decent and good translation. Hence, this research established that neural machine translators outperformed previous machine translation approaches and affirms that the addition of attention mechanism and gated recurrent units improves the quality of translation. It is recommended that future work extends the vocabulary size of this research and modify the model to handle multiple tasks for translation into different languages.

References

[1] Ayogu, I.I., Adetunmbi, A.O. and Ojokoh, B.O. (2018). Developing Statistical Machine Translation System for English and Nigerian Languages. Asian Journal of Research in Computer Science; 1(4):1-8.

[2] Esan,A.O, Omodunbi, B. Olaniyan, O. Odiase, P. and Olaleye, T. (2018). Development of Adjectival Phrase-Based English to Yorùbá Machine Translator. International Digital Organization for Scientific Research Journal of Applied Sciences, 3(1) 29-42.

[3] Pankaj, K., and Er.Vinod, K. (2013). Statistical Machine Translation Based Punjabi to English Transliteration System for Proper Nouns. International Journal of Application or Innovation in Engineering & Management (IJAIEM), 318-321.

[4] Oladosu J.B., Esan A.O., Adeyanju I. A., Adegoke B.O, Olaniyan O.M. and Omodunbi B.A. (2016). Approaches to Machine Translation: A Review. FUOYE Journal of Engineering and Technology,1 (1), 120-126.

[5] Sutskever, I., Vinyals,O. and Le, Q. (2014). Sequence to sequence learning with neural networks. Proceedings of International conference in Advances in neural information processing systems, (3104–3112).

[6] Bentivogli, L., Bisazza, A., Cettolo, M. and Federico, M. (2016). Neural versus Phrase-Based Machine Translation Quality: a Case Study.

[7] Cho, K. Bahdanau, D. Bougares, F. Schwenk, H. and Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical MachineTranslation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, pp. 1724–1734.

[8] Agrawal, R. and Sharma, D. M. (2017). Building an Effective MT System for English -Hindi Using RNN 'S. International Journal of Artificial Intelligence and Applications (IJAIA), 8(5).

[9] Salehinejad, H., Sankar, S., Barfett, J., Colak, E. and Valaee, S. (2018). Recent Advances in Recurrent Neural Networks. http://arxiv.org /abs/1801.01078v3 pp. 1-20.

[10] Mahata and Das, (2018). Machine Translation Using Recurrent Neural Network on Statistical Machine Translation. Journal of Intelligent Systems, 34-43.

[11] Bahdanau, D., Cho, K. and Bengio, Y. (2015). Neural Machine Translation By Jointly Learning To Align And Translate. In proceedings of ICLR.

[12] Jean,S., Cho, K., Memisevic, R., and Bengio, Y. (2015). On using very large target vocabulary for neural machine translation. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Beijing, China, 1–10.

[13] Luong, M., Pham, H. and Manning, C.D. (2015a). Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025.

[14] Dong, D., Wu, H., He, W.,Yu, D. and Wang, H. (2015). Multi-Task Learning for Multiple Language Translation. In Proceedings of ACL, Beijing, China.

[15] Firat,O. Cho, K. and Bengio, Y. (2016). Multi-Way, Multilingual Neural Machine Translation with a Shared Attention Mechanism. arXiv:1601.01073v1 [cs.CL].

[16] Chung, J., Cho, K. and Bengio, Y., (2016). A Character-Level Decoder without Explicit Segmentation for Neural Machine Translation.

[17] Baniata, L.H., Park, S. and Park S. (2018). A Neural Machine Translation Model for Arabic Dialects that Utilizes Multitask Learning (MTL). Computational Intelligence and Neuro-Science; (2018):1-10.

[18] Sennrich, R., Haddow, B. and Birch. A., (2015). Improving neural machine translation models with monolingual data. arXiv preprint arXiv:1511.06709. [19]Papineni, K., Roukos, S., Ward, T. and Zhu, W. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 311-318.

[19] Papineni, K., Roukos, S., Ward, T. and Zhu, W. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 311-318.

[20] Ojha, A., Bansal, A., Hadke, S. and Jha, G. (2014). Evaluation of Hindi-EnglishMT Systems. Proceedings of second workshop onIndian Language Data: Resources and Evaluation, Harpa Conference Centre, Reykjavik, Iceland.

[21] Pathak, S., Ahmad, R. and Ojha, A. (2012): A Language Engineering Approach to Enhance the Accuracy of Machine Translation Systems, 2(1):205-214.

[22] Sreelekha, S., (2016). Statistical Vs Rule Based Machine Translation; A Case Study on Indian Language Perspective. World Journal of Computer Application and Technology. 4(4): 46-57.

[23] Greenstein, E. and Penner, D. (2015). Japanese-to-English Machine Translation Using Recurrent Neural Networks. pp. 1-7.

[24] Wang, R., Panju, M.and Gohari, M. (2017). Classification-Based RNN Machine Translation using GRUs. arXiv:1703.07841v1 [cs.NE].

[25] Klein, G. Kim, Y. Deng, Y. Senellart, J. and Rush, A.M. (2017). OpenNMT: Open-Source Toolkit for Neural MachineTranslation. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics-System Demonstrations, Vancouver, Canada, 67–72.

[26] Lavie, A. (2011). Evaluating the Output of Machine Translation Systems. In Proceedings of 13th Machine Translation Summit Tutorial, Xiamen, China.

[27] Oudah,M., Almahairi, A. and Habash, N. (2019). The Impact of Pre-processing on Arabic-English Statistical and Neural Machine Translation, arXiv:1906.11751v1 [cs.CL].

[28] Chen, M., Firat, O., Bapna, A., Johnson, A., Macherey, A., Foster, G., Jones, L., Parmar, N., Shazeer, N., Vaswani, A., Uszkoreit, J., Kaiser, L., Wu, Y., Hughes, M., and Zhifeng, M. (2018). The Best of Both Worlds: Combining Recent Advances in Neural Machine Translation. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers) (pp. 76–86), Melbourne, Australia.