A Neural Machine Translator for Low-Resourced Luhya Language to Swahili

Richard Kimanzi[1,*] and Benson Kituku[1]

[1]Department of Computer Science, Dedan Kimathi University of Technology, P.O Box Private Bag - 10143, Nyeri, Kenya.

*Corresponding author(s). E-mail(s): rkimanzi@gmail.com;
Contributing authors: benson.kituku@dkut.ac.ke;

Abstract

The Luhya language, spoken by the Luhya people in Kenya represents a cluster of closely related Bantu languages that remain under-resourced in the field of computational linguistics. This study focuses on developing a neural machine translation (NMT) system to bridge the linguistic gap between Luhya and Swahili, another Bantu language widely spoken in East Africa. Leveraging the advances in deep learning, this study addresses the challenge of translating between these languages, despite limited available corpora. We used a parallel corpus for Luhya-Swahili particularly the Marachi and Bukusu dialects downloaded from the Maseno University Kencorpus repository. The corpus aligns bilingual texts collected from various sources such community contributions and local literature. The NMT model was trained using a Bidirectional LSTM model and then evaluated. Evaluation metrics such as BLEU scores were employed to assess the quality of the translations produced by the model. Initial results indicate promising translation accuracy despite the limited corpora, demonstrating the potential of neural networks in handling low-resourced languages. The success of this model not only facilitates communication and educational opportunities for Luhya speakers but also contributes to the preservation and digital representation of minority languages. Future work will focus on expanding the dataset, fine-tuning the model, and exploring multilingual NMT approaches to further improve translation quality and robustness. This research underscores the importance of technological inclusivity in linguistically diverse regions.

Keywords: Neural machine translation, Low-resourced languages, Bidirectional LSTM

1 Introduction

Machine translation (MT) has evolved significantly over the past few decades, from rule-based and statistical methods to the current state-of-the-art neural machine translation (NMT). NMT systems, particularly those employing sequence-to-sequence (Seq2Seq) architectures with attention mechanisms, have shown remarkable performance improvements in translating between high-resourced languages such as English and French [1]; [2]. These advancements offer a promising foundation for tackling translation challenges in low-resourced languages like Luhya.

The primary challenge in developing an NMT system for Luhya is the scarcity of parallel corpora, which are essential for training accurate and reliable translation models. Despite its cultural significance and a substantial number of speakers, Luhya lacks the extensive digital text collections and standardized resources necessary for traditional MT approaches. This gap hinders effective communication, information accessibility, and educational opportunities for Luhya speakers, contributing to their digital exclusion. Low-resourced languages often lack extensive digital text collections, standardized orthographies, and comprehensive linguistic resources [3]. To mitigate this, our research employs data augmentation as a strategy. By leveraging Swahili's extensive digital resources and its linguistic similarities to Luhya, we aim to enhance the NMT model's performance and reliability.

1.1 Objectives

This paper therefore intends to meet the following objectives;

1. Conduct a systematic literature review to identify the challenges of translating low-resourced languages using modern technologies and the best approaches for MT focusing on low-resourced languages.
2. Develop an NMT model for Luhya (Marachi and Bukusu dialects) to Swahili
3. Evaluate the performance of the NMT model using the metrics identified in the literature review.

This paper is organized in the following order, Section 2 handles the literature review, section 3 outlines the methodology used in this study, section 4 discusses the findings of this study while section **??** highlights the conclusion of this paper.**??**

2 Literature Review

This section reviews the literature relating to neural machine translation (NMT) and its application to low-resourced languages. We focus on approaches that may be used in translating the low-resourced Luhya language to Swahili. The review examines challenges unique to low-resourced languages, such as limited parallel corpora, and discusses innovative solutions like transfer learning, data augmentation, and community-driven data collection. Additionally, it evaluates prior research efforts in low-resourced language translation and their implications for the current study.

2.1 An Undertsnading of NMT

Neural Machine Translation (NMT), also referred to as Neural MT, Deep Neural Machine Translation, Deep NMT, or DNMT, represents the recent advancement in machine translation technology, employing neural network methodologies to estimate the probability of word sequences. This can apply to text fragments, complete sentences, or, with recent advancements, entire documents [4].

NMT employs a fundamentally different strategy for tackling language translation and localization issues by utilizing deep neural networks and artificial intelligence to train neural models. NMT has rapidly overtaken Statistical Machine Translation (SMT) as the leading method for machine translation. Compared to SMT, Neural Machine Translation typically yields significantly higher quality translations, offering improved fluency and adequacy [4].

2.2 Challenges Of Low-Resourced Languages

Studies in the field of natural language processing have identified various challenges associated with the translation of indigenous low-resourced languages. The subsections below discuss these challenges as highlighted by different researchers.

2.2.1 Lack of Parallel Corpora

One of the primary challenges for NMT in low-resourced languages is the scarcity of parallel corpora. A significant amount of data is required for NMT systems to learn mappings between languages. High-quality parallel datasets, where sentences in one language are directly translated into another, are crucial for training effective NMT models. Low-resourced languages often lack extensive parallel datasets, making it difficult to train models that can accurately capture the nuances of both languages [5]. Ranathunga et.al in their survey on Neural Machine Translation for low-resourced languages give guidelines on the possible NMT technique for a given LRL data setting based on their findings. [6] in their study on recent advances of low-resource neural machine translation, identify data augmentation, exploiting training data from other languages, and alternative learning strategies that use only monolingual data as some of the key strategies of building corpora for LRL.

2.2.2 Limited Monolingual Data

In addition to parallel corpora, having substantial monolingual data is beneficial for training language models that can assist in translation tasks. However, low-resourced languages often suffer from a lack of digital presence, resulting in limited monolingual datasets. This scarcity hampers the ability of NMT models to learn rich linguistic patterns and contexts [7].

2.2.3 Domain Specificity

NMT models trained on available data may not generalize well to different domains due to the specific nature of the training data. For low-resourced languages, the limited

data available is often domain-specific (e.g., religious texts, governmental documents), making it challenging for models to perform well across various contexts and topics [8].

2.2.4 Morphological Complexity

Many low-resourced languages exhibit high morphological complexity, meaning that words can have many forms depending on tense, case, gender, etc. The morphological variants of these words result in a large vocabulary. Most of these words share a common lemma, and contain regular morphological affixation; consequently much of the information required for translation is present, but not in an accessible form for models of neural MT. This complexity increases the difficulty of training NMT models, as the models need to account for a large variety of word forms and their meanings [9].

Low-resourced languages have been poorly translated due to their inherent morphological complexity and agglutination (a morphological process in which words are formed by stringing together morphemes, each of which corresponds to a single syntactic feature). Translation systems are still limited in their capacity to comprehend sentence and word structures for languages with rich morphology. Sub-word algorithms such as BPE, the incorporation of linguistically driven sub-word units into NMT systems [10] and the Complete Set of Endings (CSE) as proposed by [11] are some of the approaches that may be employed in solving these complexities.

The Luhya language as an indigenous and under-resourced language, faces the challenge of morphological complexity too. Most under-resourced languages often belong to language families that are typologically different from widely studied languages like English. Luhya is not an exception since it belongs to the Bantu language family. The cross-linguistic variability can pose additional challenges in transfer learning and model adaptation techniques that are commonly used to enhance NMT performance for low-resourced languages [12]. In addressing this challenge in the context of Bantu languages – Luhya being a Bantu language – Kituku. B. et. al. propose an approach that takes advantage of cross-linguistic similarities among Bantu languages using grammar engineering strategies such as grammar porting and grammar sharing [13]. Therefore, to address this challenge some of the proposed strategies identified from previous research as stated above may be applied when creating models for translating Luhya to other languages.

2.2.5 Lack of Standardization

The lack of standardized orthography or consistent spelling conventions in low-resourced languages further complicates data preparation and model training. Variations in spelling and writing can lead to inconsistencies in the dataset, affecting the model's ability to learn effectively [14]. Non-standardized languages may exhibit a high frequency of spelling variations for a single word. This challenge may vary based on different languages and their structure [15].

Millour and Fort in their paper on text corpora and the challenge of newly written language, propose collaborative lexicography i.e. the involvement of native speakers in developing lexical resources and building of text corpora with the involvement of native speakers as some of the ways in which the standardization challenge may be addressed [16].

2.3 Machine Translation Approaches For Low-Resourced Languages

[17] Comprehensively handle various machine translation approaches in their review paper. They tackle these approaches from a historical perspective of machine translation while highlighting the classical approaches that have been used as the building blocks of the current state-of-the-art Neural Machine Translation models. Continuous improvements, variations, and hybridization of these approaches have contributed greatly to the growth of natural language processing as we know it today. Kituku et. al further opine that each language differs on the best approach to translation based on the availability of language resources and its linguistic structure. The various approaches have been briefly discussed in below;

1. Rule-based Machine Translation - This is a classical approach to machine translation that relies on a set of linguistic rules to convert text from one language to another [18]. These rules encompass grammatical, syntactic, and semantic principles of both the source and target languages.
2. Data-driven Machine Translation - Data-driven Machine Translation (DDMT) refers to a family of machine translation approaches that rely on large datasets of bilingual text to learn how to translate from one language to another. Unlike rule-based methods that depend on handcrafted linguistic rules, DDMT methods use statistical and neural network techniques to model the translation process directly from data. [17] highlights Statistical Machine Translation and Neural Machine Translation as sub-types of Data-driven Machine Translation approaches
3. Hybrid Machine Translation - Hybrid Machine Translation is a machine translation method that integrates multiple translation approaches within a single system (Kituku et al., 2016)[17]. It may combine rule-based with data-driven machine translation approaches. Hybrid techniques are driven by the desire to achieve higher levels of translation accuracy that may not be possible to achieve using a single approach [4].

Considering that we obtained a parallel corpus for Luhya to Swahili, our focus onwards will be on the utilization of data-driven approaches specifically neural machine translation techniques in realizing the objectives of this study.

2.4 Neural Machine Translation Approaches Techniques

To understand the evolution of NMT techniques we studied a review paper by [19] which captures the various techniques and their application. Key highlights of the techniques discussed by Stahlberg include;

1. Ensembling - Ensembling in neural machine translation refers to the technique of combining multiple models to improve translation accuracy and robustness. Therefore, by aggregating the outputs of several models, the ensemble can produce better predictions than any individual model alone. This approach takes advantage of the strengths and mitigates the weaknesses of each model, leading to more reliable and accurate translations.

2. Hybrid NMT-SMT Systems - Integrating traditional Statistical Machine Translation (SMT) components with NMT models to utilize the strengths of both approaches
3. Rescoring and Reranking Methods - Using techniques like finite state transducer-based loose combination schemes and Minimum Bayes Risk (MBR) approaches to enhance translation quality.
4. Multimodal NMT - This involves ncorporating multiple modalities (e.g., text and images) into NMT systems for more context-aware translations.
5. Feature Integration - This involves incorporating SMT features like language models into NMT systems to address specific translation challenges.

[19] states that LSTM and bidirectional LSTM models have been successfully applied in various NMT architectures to improve translation quality and fluency. These models excel at capturing dependencies across words in a sentence and have shown promising results in generating accurate translations. Stahlberg further highlights that in more advanced NMT models, an attention mechanism is added to the basic LSTM encoder-decoder architecture. This helps the model to focus on different parts of the input sentence when generating each word in the output sentence, addressing the limitations of encoding the entire sentence into a single fixed-length vector.

Bi-Directional Long Short-Term Memory (Bi-LSTM), which is an extension of the traditional LSTM architecture enhances the capabilities of traditional LSTMs by processing input sequences in both forward and backward directions. In the context of neural machine translation, Bi-LSTMs process the input text in both directions, enabling the model to understand the context from both past and future words in the sequence. This bidirectional context is particularly beneficial for translating sentences where the meaning of a word can depend on both its preceding and succeeding words. The combined output captures context from both directions, which improves the model's understanding and translation of the sequence.

2.5 Deductions from Literature

Rule-based Machine Translation systems require extensive linguistic knowledge and resources, such as dictionaries, grammar rules, and language-specific heuristics. While they can produce high-quality translations for specific language pairs and domains, they are often less flexible and harder to scale compared to modern data-driven approaches like Neural Machine Translation (NMT) and Statistical Machine Translation (SMT)

Where parallel corpus is available the best translation mechanism to use is Neural Machine Translation. However for better results, one can implement a hybrid system that combines both neural and statistical machine translation techniques [19].

Transformers and attention mechanisms may also be applied to translation systems to improve the accuracy of translation models [2]. transformers have become a cornerstone in NMT research due to their ability to handle long-range dependencies, parallelize computations, and improve translation quality, marking a significant advancement in the field of machine translation.

3 Methodology

The neural machine translator developed in this study was based on two luhya dialects i.e Marachi and Bukusu. The evaluation was done on the datasets separately as well as the combined since the dialects are subsets of the main Luhya language. The evaluation was carried out to ascertain the performance of the model developed on each dataset. The following was undertaken in this stage of the study.

3.1 Data acquisition

The Luhya corpus was downloaded from the Maseno University Kencorpus repository. The data contained three dialects namely Bukusu, Logooli, and Marachi. Upon physical evaluation, we settled on the Marachi and Bukusu dialects, which had more data than the other dialects (Wanzare et al., 2021). As guided by literature, the more the data the higher the chances of attaining better translation results after training.

3.2 Text Pre-processing Techniques

[20] Emphasize the importance of text pre-processing for machine translation tasks. They confirm that preprocessing has a huge impact on how the model performs on the translation task. The Marachi dataset was organized into two hundred and fifty text files which contained numerous sentences and their corresponding translations. The following data modeling steps were undertaken.

1. Combined the text files into one file containing parallel corpora for Marachi and Swahili languages.
2. Filtering the data to ensure that each original Marachi sentence in the dataset had its corresponding translation.
3. Removing non-letter characters and numbers since they may not be useful in the translation process.
4. Converting all sentences into lowercase for uniformity during translation.
5. Converting the text file into a CSV file containing two columns Original (Marachi/Bukusu) and Translation (Swahili). The text file had 2473 records containing the source language (Marachi/Bukusu) and the corresponding translations or target language (Swahili). [21] asserts "that the development of training data is more critical than the choice of model or classification mechanism." Thus, any NLP model needs to have well pre-processed and adequate training data. In our view, our dataset was too small to generate satisfactory results. We therefore performed augmentation by introducing noise in the datasets. The Marachi dialect generated a new dataset with 7417 records after augmentation, While the Bukusu dialect generated 14, records after augmentation.
6. The augmented dataset in CSV format file was used as our input dataset to train the model.

3.3 Tokenization

We defined a function tokenize that creates and configures a text tokenizer using Keras. The Tokenizer object is initialized and then fitted on the provided text data. This is

meant to convert the input text into a sequence of tokens (numerical representations), which can be used as input for machine translation task. The fit_on_texts method updates the internal vocabulary based on the provided text data. We prepared the Luhya and Swahili text data for training a machine translation model. This involved tokenizing the text, converting it to sequences of integers, and padding the sequences to ensure they are of uniform length. We then printed out the vocabulary size in both Luhya and Swahili and the maximum length of all sequences (both Luhya and Swahili).

3.4 Training the model using Bi-directional LSTM

We defined a sequential model for neural machine translation. Which included; An embedding layer to convert words into vector representations (vector_length of 100), a bidirectional LSTM layer with 256 units, a RepeatVector layer to repeat the input sequence, another LSTM layer with 256 units that returns sequences, and a Time Distributed dense layer with a softmax activation to predict each word's probability distribution over the Luhya vocabulary.

The training was carried out using 30 Epochs and with K fold crossvalidation where K=10. An allocation of GPU memory was required to train both models. We utilized the free GPU allocation available on Kaggle for our experiment.

4 Results and Discussion

The Bidirectional NMT Model was tested using the Marachi and Bukusu dialects. The model was tested on the dialects separately then combined and the results were evaluated.

4.1 Bi-LSTM Results of the Marachi dialect dataset

For the Marachi Dialect, We observed a training accuracy of 89.83% and a training loss of 43.30%. The validation accuracy was 89.34% while the validation loss was 45.36% as indicated in the figures 1 and 2 below.

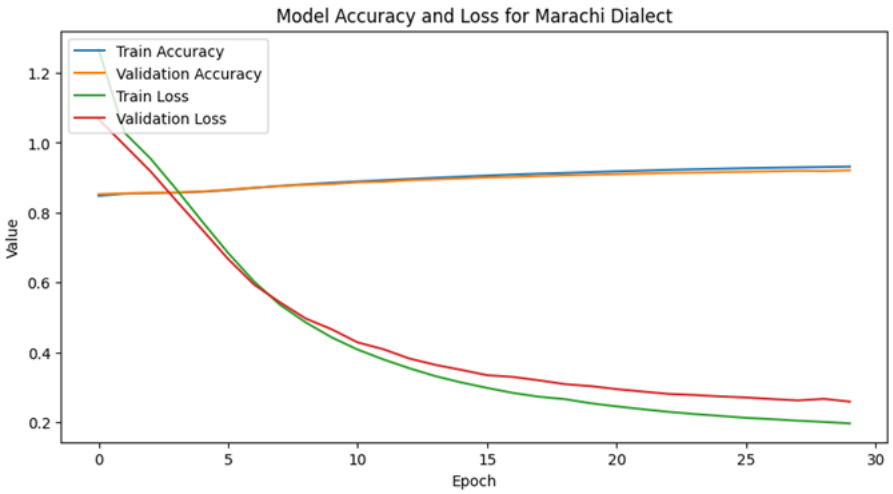

Fig. 1 Single Word Translations for Marachi Dialect

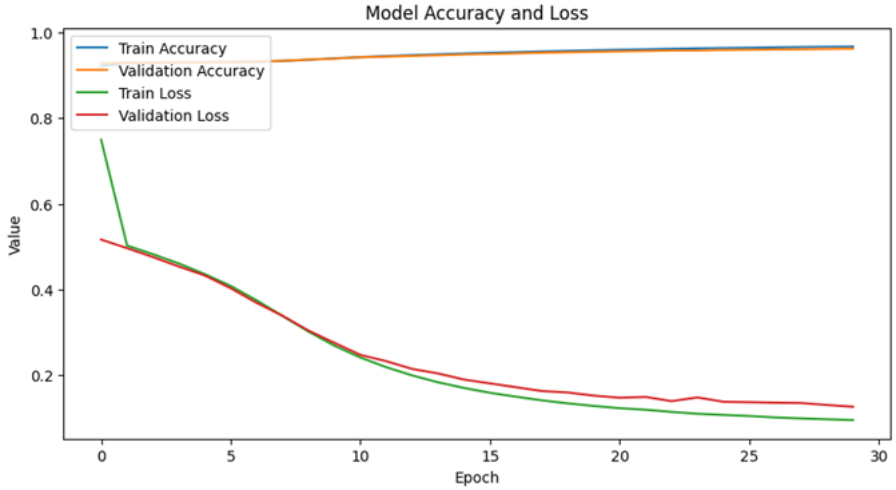

Fig. 2 Single Word Translations for Marachi Dialect

The translation accuracy was perfect with a BLEU Score of 100% for single-word sentences which were in the present in the Marachi dataset. This is exhibited in the figure below. However, the model accuracy using the BLUE Score measure reduced with short sentences. The model also performed poorly on for very long sentences as exhibitted in fig below. For an improved performance of the model, there is a need to fine-tune it to handle longer sentences. This could be done by adding more layers to the model and implementing attention and transformer mechanisms. Further, there is also

need to continuously build parallel corpora for all Luhya dialects since the available data is not enough to adequately train an efficient Neural Machine Translator model. As such better performing machine translators may be developed using the acquired data.

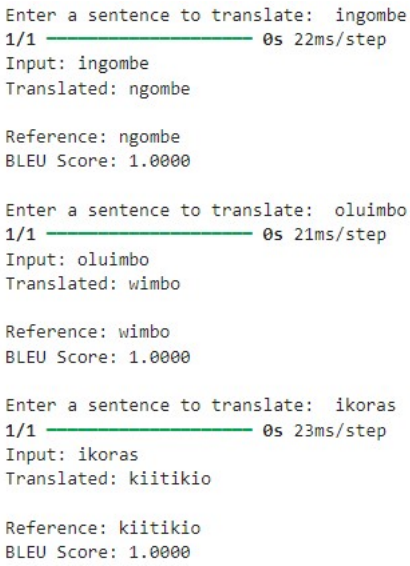

Fig. 3 Single Word Translations for Marachi Dialect

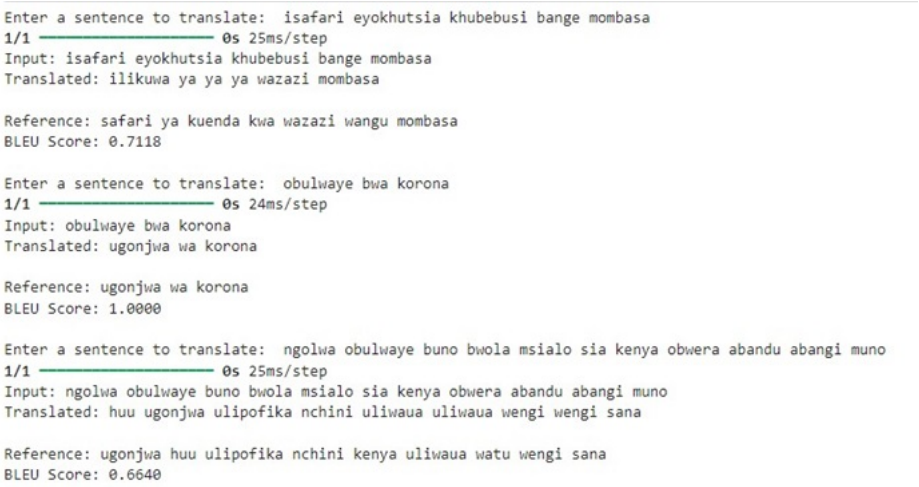

Fig. 4 Short Sentence Translations for Marachi Dialect

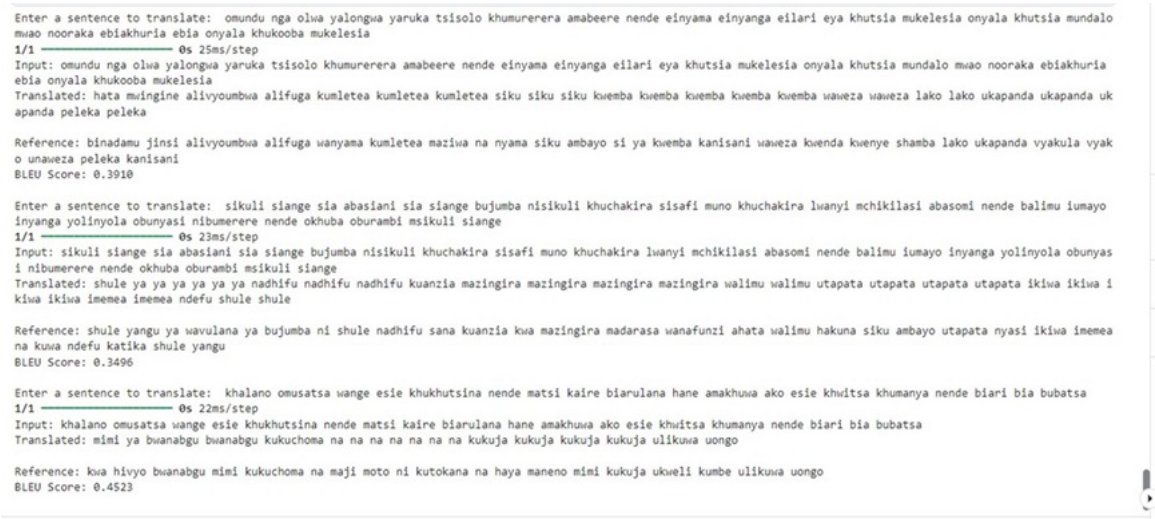

Fig. 5 Long Sentence Translations for Marachi Dialect[**?**])

As observed in translation results displayed in figures 3, 4, and 5. the accuracy in translation reduces as the sentences become longer.This is an indicator that the model does not perform well in learning more complex patterns in the Marachi input sentences.

4.2 Bi-LSTM Results of the Bukusu dialect dataset

Upon training the model using the Bukusu dialect dataset, the model had the following results. The training accuracy was 94.88%, the training loss was 23.76%, the Validation accuracy was 94.68% while the validation loss was 24.59%. The model performed better when trained on the Bukusu dataset based on the metrics discussed here. Figure 6 and 7 show the BLUE score using sample short and long sentences in the Bukusu dataset. The blue Scores for both short and long sentences in the Bukusu dialects were significantly higher as compared to the Marachi dataset.

```
Enter a sentence to translate:  khukhwama khu banuuli
1/1 ━━━━━━━━━━━━━━━━━━━━ 2s 2s/step
Input: khukhwama khu banuuli
Translated: kutoka kwa wanyanganyi

Reference: kutoka kwa wanyanganyi
BLEU Score: 1.0000

Enter a sentence to translate:  se baandire ese tawe
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 28ms/step
Input: se baandire ese tawe
Translated: hawajanitii mimi

Reference: hawajanitii mimi
BLEU Score: 1.0000

Enter a sentence to translate:  ekironi aambana ne baamori na kanaani
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 27ms/step
Input: ekironi aambana ne baamori na kanaani
Translated: ekironi ashirikiana na waamori wakanaani wakanaani

Reference: ekironi ashirikiana na waamori na wakanaani
BLEU Score: 0.5373
```

Fig. 6 Model accuracy for Marachi dialect

```
Enter a sentence to translate:  ne buli lwoosi omwimeelesi owa bioosi wele wabwe bamkholela ne afwiile babaandu bakaluukhananga bumayanu mala b
asaaya bawele baali bakobole khungila chikhale ne likhola ne ashitoreti
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 29ms/step
Input: ne buli lwoosi omwimeelesi owa bioosi wele wabwe bamkholela ne afwiile babaandu bakaluukhananga bumayanu mala basaaya bawele baali bakob
ole khungila chikhale ne likhola ne ashitoreti
Translated: na kila mara msimamizi msimamizi mungu mungu wao wamfanyie akiwa amefariki wakarejelea wakarejelea dhambi kisha wakaabudu baali baa
li wakarudi njia za za kitambo na na na

Reference: na kila mara msimamizi wa vyote mungu wao wamfanyie akiwa amefariki watu wakarejelea dhambi kisha wakaabudu miungu baali wakarudi ku
a njia za kitambo na matendo na ashitoreti
BLEU Score: 0.3825

Enter a sentence to translate:  babalekhwa mu sibala baba babaami baraano bafilisiti bakanaani boosi basidoni ne bahibi bamenyile mu bikulu via
lebanoni khukhwama khu sikulu sia bali herimoni khukhwola khu silibwa sie hamati
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 28ms/step
Input: babalekhwa mu sibala baba babaami baraano bafilisiti bakanaani boosi basidoni ne bahibi bamenyile mu bikulu via lebanoni khukhwama khu s
ikulu sia bali herimoni khukhwola khu silibwa sie hamati
Translated: walioachwa kwa nchi walikuwa watano watano wafilisti wakanaani wakanaani wasidoni na na wanaoishi wanaoishi kwa lebanoni lebanoni l
ebanoni mlima mlima wa baali baali kufika kufika lango lango hamati hamati

Reference: walioachwa kwa nchi walikuwa wafalme watano wafilisti wakanaani wote wasidoni na wahivi wanaoishi kwa vilima vya lebanoni kutoka kwa
mlima wa baali herimoni kufika kwa lango la hamati
BLEU Score: 0.1836

Enter a sentence to translate:  baabami kamakhumi  nenya chisebele che kameechi sikila alia babakhalwa chinjala chikhuumwa wamba kamaloba kali
khusiangalaamwe khumakhono ne chinjala chi sie ebunafubo owa kanaani
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 28ms/step
Input: baabami kamakhumi  nenya chisebele che kameechi sikila alia babakhalwa chinjala chikhuumwa wamba kamaloba kali khusiangalaamwe khumakhon
o ne chinjala chi sie ebunafubo owa kanaani
Translated: wafalme makumi nataka visima maji maji kwa alikula alikula vidole vidole vidole ukanipa mashamba mashamba mashamba kwa kwa kwa na n
a na vidole vidole kwa kaskazini kanaani kanaani

Reference: wafalme makumi nataka visima vya maji kwa sababu alikula waliokatwa vidole vya kusugua ukanipa mashamba yaliyo kwa tambarare na kwa
mikono na vidole vya kwa kaskazini mwa kanaani
BLEU Score: 0.1505
```

Fig. 7 Model accuracy for Bukusu Dialect

5 Conclusion

The research project on neural machine translation for Luhya (Marachi Dialect) to Swahili demonstrated that the Bidirectional LSTM (BiLSTM) model outperformed the unidirectional LSTM model in terms of translation accuracy and fluency. The bidirectional approach allowed the model to consider the context from both directions, providing a more comprehensive understanding of the input sequence and resulting in more accurate translations. This highlights the potential of advanced neural network architectures in addressing the complexities of low-resourced languages like Luhya.

Future research should explore the integration of transformer models and attention mechanisms for further improvement. With their self-attention mechanism, transformers can process entire sentences simultaneously, capturing long-range dependencies more effectively than RNN-based models. Attention mechanisms can further enhance translation quality by dynamically focusing on relevant parts of the input sentence during translation, addressing limitations seen in LSTM and BiLSTM models. By leveraging these advanced techniques, it is possible to significantly boost the performance of neural machine translation systems, making them more robust and capable of handling the intricacies of low-resourced languages. This direction promises to bridge the gap in translation quality, providing more accurate and reliable translations for Luhya to Swahili.

References

[1] Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)

[2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, , Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

[3] Guzmán, F., Chen, P.-J., Ott, M., Pino, J., Lample, G., Koehn, P., Chaudhary, V., Ranzato, M.: The flores evaluation datasets for low-resource machine translation: Nepali-english and sinhala-english. arXiv preprint arXiv:1902.01382 (2019)

[4] Wiggins, D.: What is Neural Machine Translation (NMT)? (2020). https://omniscien.com/faq/what-is-neural-machine-translation/ Accessed 2024-04-09

[5] Ranathunga, S., Lee, E.-S.A., Prifti Skenduli, M., Shekhar, R., Alam, M., Kaur, R.: Neural Machine Translation for Low-resource Languages: A Survey. ACM Comput. Surv. **55**(11) (2023) https://doi.org/10.1145/3567592 . Place: New York, NY, USA Publisher: Association for Computing Machinery

[6] Haque, R., Liu, C.-H., Way, A.: Recent advances of low-resource neural machine translation. Machine Translation **35**(4), 451–474 (2021) https://doi.org/10.1007/s10590-021-09281-1

[7] Sennrich, R., Haddow, B., Birch, A.: Improving neural machine translation models

with monolingual data. arXiv preprint arXiv:1511.06709 (2015)

[8] Zoph, B., Yuret, D., May, J., Knight, K.: Transfer learning for low-resource neural machine translation. arXiv preprint arXiv:1604.02201 (2016)

[9] Vylomova, E., Cohn, T., He, X., Haffari, G.: Word Representation Models for Morphologically Rich Languages in Neural Machine Translation. In: Proceedings of the First Workshop on Subword and Character Level Models in NLP. Association for Computational Linguistics, ??? (2017). https://doi.org/10.18653/v1/w17-4115 . http://dx.doi.org/10.18653/v1/w17-4115

[10] Chimalamarri, S., Sitaram, D.: Linguistically enhanced word segmentation for better neural machine translation of low resource agglutinative languages. International Journal of Speech Technology **24**(4), 1047–1053 (2021) https://doi.org/10.1007/s10772-021-09865-5

[11] Tukeyev, U., Karibayeva, A., Zhumanov, Z.h.: Morphological segmentation method for Turkic language neural machine translation. Cogent Engineering **7**(1), 1856500 (2020) https://doi.org/10.1080/23311916.2020.1856500 . Publisher: Informa UK Limited

[12] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised Cross-lingual Representation Learning at Scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, ??? (2020). https://doi.org/10.18653/v1/2020.acl-main.747 . http://dx.doi.org/10.18653/v1/2020.acl-main.747

[13] Kituku, B., Nganga, W., Muchemi, L.: Leveraging on Cross Linguistic Similarities to Reduce Grammar Development Effort for the Under-Resourced Languages: a Case of Kenyan Bantu Languages. In: 2021 International Conference on Information and Communication Technology for Development for Africa (ICT4DA), pp. 83–88 (2021). https://doi.org/10.1109/ICT4DA53266.2021.9672222

[14] Hovy, D., Spruit, S.L.: The social impact of natural language processing. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 591–598 (2016). https://doi.org/10.18653/v1/p16-2096

[15] Millour, A., Fort, K.: Unsupervised data augmentation for less-resourced languages with no standardized spelling. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), pp. 776–784 (2019). https://doi.org/10.26615/978-954-452-056-4_090

[16] Millour, A., Fort, K.: Text corpora and the challenge of newly written languages. In: Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU) and Collaboration and Computing for

Under-Resourced Languages (CCURL), pp. 111–120 (2020)

[17] Kituku, B., Muchemi, L., Nganga, W.: A Review on Machine Translation Approaches. Indonesian Journal of Electrical Engineering and Computer Science **1**(1), 182 (2016) https://doi.org/10.11591/ijeecs.v1.i1.pp182-190 . Publisher: Institute of Advanced Engineering and Science

[18] Shiwen, Y., Xiaojing, B.: Rule-based machine translation. In: Routledge Encyclopedia of Translation Technology, pp. 186–200. Routledge, ??? (2014)

[19] Stahlberg, F.: Neural machine translation: A review. Journal of Artificial Intelligence Research **69**, 343–418 (2020)

[20] Tabassum, A., Patil, R.R.: A survey on text pre-processing & feature extraction techniques in natural language processing. International Research Journal of Engineering and Technology (IRJET) **7**(06), 4864–4867 (2020)

[21] Bayer, M., Kaufhold, M.-A., Buchhold, B., Keller, M., Dallmeyer, J., Reuter, C.: Data augmentation in natural language processing: a novel text generation approach for long and short text classifiers. International journal of machine learning and cybernetics **14**(1), 135–150 (2023) https://doi.org/10.1007/s13042-022-01553-3