IST-Africa 2025 Conference Proceedings
Miriam Cunningham and Paul Cunningham (Eds)
IST-Africa Institute and IIMC, 2025
ISBN: 978-1-905824-75-5

Neural Machine Translation in Low-Resourced Languages: Case of Dholuo – Swahili Translation

Peter KIMANGA, Benson KITUKU, Juliet MOSO, Patrick GIKUNDA
Department of Computer Science, Dedan Kimathi University of Technology, Nyeri, Kenya
Email:peterkimanga.m@gmail.com, benson.kituku@dkut.ac.ke, juliet.moso@dkut.ac.ke, patrick.gikunda@dkut.ac.ke

Abstract: This study developed a Neural Machine Translation (NMT) model for Dholuo-Swahili translation, addressing the lack of translation resources for low-resource languages in Kenya. Using the OpenNMT-py framework, the model was trained on a parallel dataset of 27,415 sentence pairs from two datasets to improve model performance. Performance was evaluated using BLEU (Bilingual Evaluation Understudy), METEOR (Metric for Evaluation of Translation with Explicit Ordering), perplexity (PPL), and prediction scores, with results showing a BLEU score of 16.29, METEOR score of 0.27, and PPL of 1.88, indicating moderate translation accuracy. Despite achieving low PPL, fluency and lexical accuracy challenges persisted due to limited training data and linguistic differences. Future research will focus on expanding datasets, leveraging multilingual models like LLaMA 3.2, and integrating hybrid approaches to improve translation quality. This work contributes to digital inclusion, language preservation, and enhanced cross-linguistic communication for underrepresented languages in Kenya.

Keywords: Neural Machine Translation, OpenNMT-py, Kencorpus, Low-Resource Languages

1. Introduction

Machine translation (MT) is one of the most successful applications of Natural Language Processing, with numerous practical applications plus research papers at leading machine learning and NLP conferences [1]. Furthermore, it is a discipline of computational linguistics concerned with the computational analysis of text or speech for translation purposes. Consequently, it is defined as converting source language text to target language text using a computerized system with human in-loop or not [2] to create a sound MT system that translates human languages with acceptable accuracy [3]. The field of MT is an intersection of several fields, such as linguistics, computation, artificial intelligence, translation theory, and statistics. Machine translation has main branches, namely, rules-based translation and Data-driven approaches. In rule-based approaches, rules are crafted by linguists and usually have high precision but low coverage, whereas data-driven approaches do have high coverage and need an annotated parallel corpus [4]; in our case, the research shall apply the second approach using neural networks.

The need for MT translation resources for low-resourced languages cannot be over-emphasized in the current knowledge technology-driven economy [5]. Low-resourced languages have little or no substantial presence on the Internet, have very little or no digitized corpora or digital text, no commercial interest, and are technologically marginalized or disadvantaged [6], [7], [8]. However, as shown in the subsequent section, most of the global population still speaks low-resource languages. Thus, they stand to gain from enhanced translation quality in their language [9]. Consequently, there is a lot of concern for low-resource languages in order to achieve language equity and inclusion

Authorized licensed use limited to: Dedan Kimathi University of Technology. Downloaded on July 23,2025 at 09:40:42 UTC from IEEE Xplore. Restrictions apply.

around the globe. These tend to support a few low-resource languages with deficient levels of coverage. Therefore, creating the neural network machine translation between these low-resourced languages is a step toward a basic language toolkit [10] besides revitalizing, preserving, and documenting the languages.

Several research efforts in Neural Machine Translation (NMT) have aimed at improving translation quality, accessibility, and effectiveness. The paper [11] addresses the lack of translation resources for Ometo languages (Wolaita, Gamo, Gofa, and Dawuro) by developing a multilingual Neural Machine Translation (NMT) system for Ometo-English translation. The study utilized a parallel corpus collected from religious texts, which was preprocessed and used for training different combinations of Ometo-English translation pairs in OpenNMT. A multilingual training approach was applied, where multiple related Ometo languages were trained to enhance translation accuracy. The best results were achieved when Wolaita, Dawuro, and Gamo were used for training, while Gofa was used for testing, with a BLEU score of 4.5. Despite the complexity of Ometo languages, multilingual NMT showed promising results, and future research aims to address the morphological challenges affecting translation quality.

The study [12] introduced adaptNMT, an open-source toolkit built on OpenNMT, tailored to simplify NMT development for technical and non-technical users. It features a user-friendly interface, built-in visualization, SentencePiece subword segmentation, and an automated training setup supporting RNN and Transformer models. Additionally, it can track energy consumption. AdaptNMT lowers the learning curve, allows easy hyperparameter tuning, and benefits researchers working with low-resource languages by enabling quick model development with minimal expertise.

Another study [13] focused on preserving the endangered Gayo language by developing an Indonesian-to-Gayo NMT model using OpenNMT and Transformer architecture. A parallel corpus was created from an Indonesian-Gayo dictionary, preprocessed with case folding and punctuation removal, and expanded through data augmentation. The model achieved 90% accuracy on augmented data and a BLEU score of 79.90, proving Transformer-based NMT's effectiveness. However, accuracy dropped for unseen words, highlighting the need for larger datasets.

Additionally, research [14] explored translation between Literary and Vernacular Chinese, focusing on segmentation techniques. It compared character-level segmentation to word-level segmentation using Jiayan (BiLSTM+CRF) and Jieba (dictionary-based methods). Tested on 2-layer LSTM and Transformer models in OpenNMT, the Transformer with word-level segmentation produced the most accurate and natural translations. The study emphasized the importance of segmentation techniques in NMT and suggested further research into hybrid segmentation models for improved performance.

In this research, we established that the Luo community in Kenya, which widely uses the Dholuo language, has a challenge in the digital language field and, more significantly, in the translation from Dholuo to Swahili. Although Swahili is one of the Kenyan national languages [15], translating from Dholuo is poorly supported by the existing translation tools, leading to communication barriers in regard to socio-economic and education integration. In the case of this problem, there is a clear need for the development of a Neural Machine Translator model to translate Dholuo to Swahili in order to improve understanding between people using the two languages.

The training of an NMT model for translating Dholuo to Swahili has a major importance in the following ways. First, it creates linguistic inclusion through a digital interface that allows Dholuo-spoken individuals to engage in the digital services available in Swahili. Also, it helps support culture in the following ways: The study facilitates putting the Dholuo language into digital form and translation to other languages, which, in the long run, will be of extreme importance in the preservation and documentation of the said language

Authorized licensed use limited to: Dedan Kimathi University of Technology. Downloaded on July 23,2025 at 09:40:42 UTC from IEEE Xplore. Restrictions apply.

[16]. This research also builds education by offering materials that enable Dholuo-speaking students to learn Swahili and help boost their academic potential. As an economic advantage, the translation model enhances cross-linguistic complexities between Dholuo and Swahili speakers, enhancing regional business, trade, and employment [1]. Lastly, this research also contributes significantly to the NMT research by exploring the issue of low-resource language translation and opening a new avenue for subsequent technology development [17].

The focus of this paper was to build an open-source Neural Machine Translator using OpenNMT–py that can be used to translate from Dholuo to Swahili. OpenNMT is an open-source project under active development for neural machine translation and related tasks in neural sequence learning [18].

The remainder of this paper is structured into the following sections: **Objectives**: highlighting the aims of this research. Methodology: covering datasets, preprocessing, and model architecture; **Technology Description**: detailing tools and libraries used; **Development**: detailing the experimental setup; **Business Benefits**: Exploring the benefits the model will have; **Results**: discussing findings and implications; and **Conclusion**s; summarizing key insights, limitations, and future directions.

2. Objectives

Main and specific objectives guided this research.

2.1 Main Objective

The primary aim of this study was to build an open-source NMT model to translate the Dholuo language to Swahili to facilitate interaction or access to Swahili among the Dholuo language speakers.

2.2 Specific Objectives

1. To gather and preprocess a parallel corpus of Dholuo and Swahili texts, ensuring the data is clean, aligned, and suitable for training the NMT model.
2. To train the Dholuo-Swahili NMT model using the OpenNMT–py framework.
3. To evaluate the NMT model using BLEU, METEOR, Prediction, and Predicted Perplexity scores.

3. Methodology

In this research, we utilized two datasets. The first dataset we used is the Kencorpus dataset, which serves as a text and speech corpus for three languages predominantly spoken in Kenya: Swahili, Dholuo, and Luhya [19]. Our focus in this dataset was on the Dholuo to Swahili corpus. The dataset contained 99 small text files that must be combined into a single reference document for more accessible work. While merging the files, ensuring each source sentence had its corresponding translated text was critical. In cases where no translation was found for a given sentence, the source sentence was removed. This step was necessary so all the original sentences produced while collecting the data had the correct translations. After merging all the 99 files, we had 4,033 Dholuo sentences with their Swahili translations.

The second dataset that we used is “Low-Resource Language Data: Parallel Corpora for Kiswahili and Kidaw'ida, Kalenjin, and Dholuo” from Tech Innovators Network(THiNK) [22]. This dataset consists of three parallel corpora: Kidaw'ida-Kiswahili, Kalenjin-Kiswahili, and Dholuo-Kiswahili. We chose the Dholuo-Kiswahili corpus, which contains 23,446 training sentences and 5,862 test sentences. In this, we stepped toward ensuring the source sentences had their corresponding target sentences. We found 64 sentences, and 14 sentences were missing their target sentences. We removed them and remained with 23,382

Authorized licensed use limited to: Dedan Kimathi University of Technology. Downloaded on July 23,2025 at 09:40:42 UTC from IEEE Xplore. Restrictions apply.

training sentences and 5,848 for testing. The training sentences were combined with the first dataset to form a new dataset with 27,415 sentences.

Preprocessing played a key role in preparing the dataset for training and testing, ensuring it was cleaned and ready for use. Several steps were taken to enhance translation accuracy. First, the sentences were cleaned to remove unnecessary characters, punctuation, special symbols, and extra spaces. This helped standardize the input and reduce noise. Next, all the texts were converted to lowercase to maintain consistency, ensuring that words with different capitalizations were not treated as separate entities. These steps streamlined the data, making the model's learning easier.

Table 1: Sample of Parallel Corpus Dholuo- Swahili

Dholuo(Source)	**Swahili(Target)**
adhi tiyo ne kambi cha	naenda kufanya kazi kwa ile kambi
kane jotend siasa ariyo mopogore oromo ne gigore kendo hinyo ngato gi nyawadgi kendo josiasa moko ne omaki nikech wechego	mawili yenye miengo tofauti ya kisiasa yalipokutana yakapigana na kujeruhiana baadhi ya wanasiasa walitiwa mbaroni kwa tuhuma hizi
nyithindo go ne oherore ainya maonge gima ne pogogi	watoto hao walipendana sana hakuna chochote kiliwatenganisha
e alworawa nitie kuonde matinie mogik ma thieth yudoree	huduma ya matibabu katika wilaya yetu haitoshi ikilinganishwa na maeneo mengine
jokuoye nomaki kakualo od pesa	wezi walishikwa wakiiba benki
kenya noyudo ohuru higni mangeny mokadho	kenya ilipata huru miaka mingi iliopita

To train and evaluate our model, the dataset was split into 80% (21,932 samples) for training, and 20% (5,483 samples) for validation, and a separate test set was used for testing purposes. This structured approach ensured the model learned effectively and unbiasedly assessed unseen data.

The training set was used to teach the model to map Dholuo-Swahili sentence pairs, making it possible to learn sentence structure, grammar, and vocabulary for improved translation accuracy. The validation set monitored the model performance during training, with evaluations happening every 500 steps using metrics like validation accuracy and perplexity.

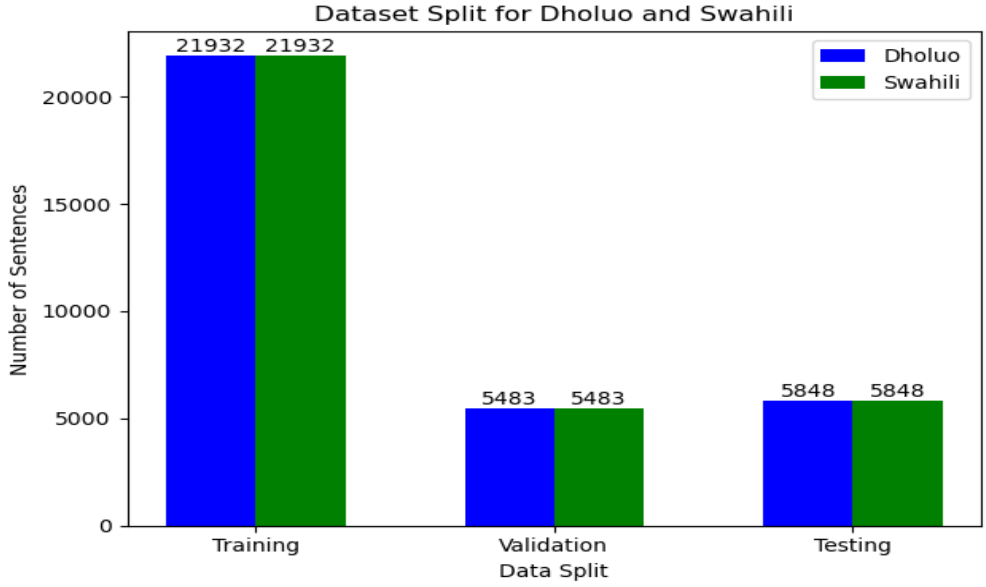

Figure 1: Data split Counts for Training, Validation, and Testing

The Transformer model, implemented using OpenNMT-py, was chosen for this task due to its efficiency in handling sequence-to-sequence translation and ability to capture long-range dependencies using self-attention mechanisms. Unlike Recurrent Neural Networks

Authorized licensed use limited to: Dedan Kimathi University of Technology. Downloaded on July 23,2025 at 09:40:42 UTC from IEEE Xplore. Restrictions apply.

(RNNs), which process words sequentially, the Transformer processes words in parallel, making it significantly faster and more effective for translation tasks.

The model was designed with six encoder layers and six decoder layers, following the standard Transformer architecture. The encoder processes the Dholuo text, while the decoder generates the corresponding Kiswahili translations. Each layer within the encoder and decoder incorporates multi-head self-attention mechanisms, which allow the model to focus on different parts of the sentence during translation.

Feed-forward networks and self-attention were combined to learn meaningful patterns using Transformer architecture. Since Transformers lack recurrence, positional encoding preserved word order. Word embeddings were set to 512 dimensions for similarity and shared between Dholuo and Swahili. The learning rate for the Adam optimizer was adjusted adaptively with Noam decay, and the training was stable.

Care was made to choose hyperparameters well, balancing accuracy and efficiency. In particular, the model had 512-dimensional embeddings, six encoder/decoder layers, eight attention heads, a batch size of 4096 tokens, a dropout rate of 0.2, and 4000 warmup steps. Early stopping with a patience of 5 was implemented to halt training if validation performance plateaued, preventing overfitting and unnecessary computations. Although the training was set for 20,000 steps, it was stopped at 12,000 steps when validation performance stabilized, ensuring optimal model performance. Checkpoints were saved every 1,000 steps.

The test dataset was separated from training and validation. Generalization to real-world translations was assessed on unseen test data by evaluating the model. It had a robust evaluation via an independent test set and an 80-20 training-validation split. It achieved high accuracy in the final model and generalized to new translations outside the training data.

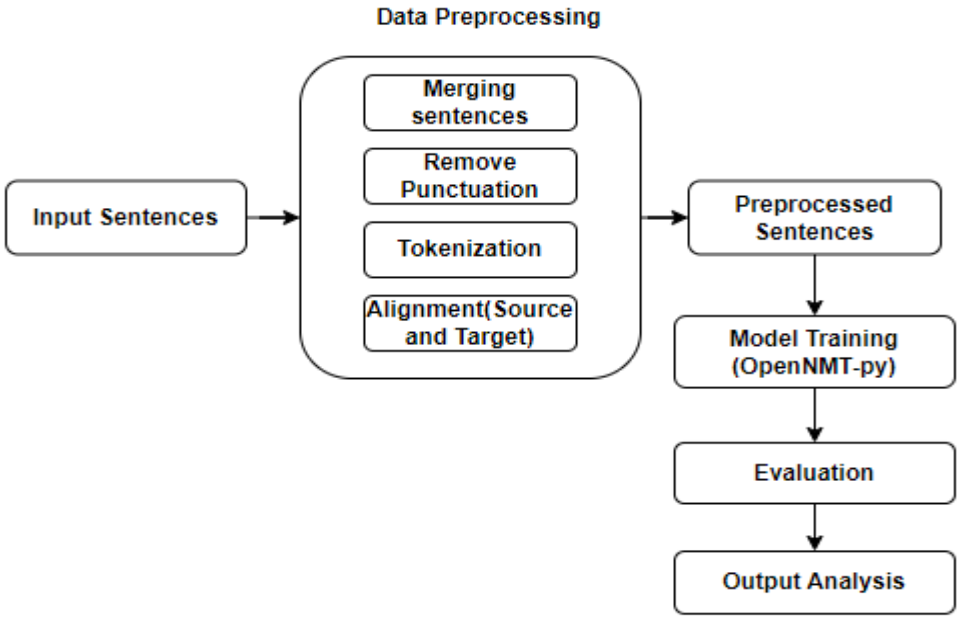

Figure 2: Steps followed in Methodology

4. Technology Description

To deploy the NMT, OpenNMT-py was utilized, a free tool for Neural Machine Translation and Neural sequence modeling, released in December 2016. OpenNMT-py was developed with PyTorch, and is a tool for those who want to do translation, summarization, or morphological analysis without restrictions [19]. It was created specifically for sequence-to-sequence models and is very well suited for machine translation tasks. The text was

Authorized licensed use limited to: Dedan Kimathi University of Technology. Downloaded on July 23,2025 at 09:40:42 UTC from IEEE Xplore. Restrictions apply.

tokenized by SentencePiece, a popular tool for subword segmentation, which improved the handling of unknown words. The training was performed in a Kaggle GPU -P100. This environment allowed us to optimize both speed and memory efficiency.

5. Developments

Several optimizations were made while developing the Dholuo-Swahili NMT model, which aimed at improving translation quality and training stability. The most important technique was dropout regularization, set at 0.2, to prevent overfitting. The model was trained with 20% of neurons randomly deactivated, encouraging the model to generalize better and not just memorize the training data. This is particularly important when the data set is relatively small, as models trained on small data are prone to overfitting. Dropout was applied at several layers of the encoder and decoder of the Transformer, which means that both input encoding and output generation will benefit from regularization.

Gradient clipping was employed with a maximum norm of 5.0 to stabilize training. This prevented sudden weight update spikes that could make training unstable. Transformer models, which are mainly based on multi-head attention, are susceptible to the exploding gradient issue. This is where very large gradient values diverge the training. Clipping makes the updates within a certain range and hence smooths out convergence. Noam decay learning rate scheduling helped adjust the learning rate throughout the training. In the beginning, a large learning rate helps faster learning. However, as training continues, the learning rate gradually decreases, so the translation accuracy can be fine-tuned. This ensured the model did not converge too fast into a suboptimal solution, and the weight updates became more refined over time.

6. Results

The OpenNMT-based Dholuo-to-Swahili neural machine translation (NMT) model was trained and evaluated using a parallel dataset of 21,932 training sentence pairs, 5,483 validation sentence pairs, and 5,848 test sentence pairs. The decision to adopt a unidirectional translation model (Dholuo-to-Swahili) was primarily driven by the need to increase accessibility by translating from a low-resource language into a more widely spoken language. While a bidirectional model could have been considered, the limited dataset size and computational constraints made a single-source-to-target approach more feasible.

The BLEU score, METEOR, prediction perplexity (PPL), and prediction score were tracked across multiple training steps to evaluate the model's performance. Training steps refer to the number of times the model updates its parameters based on mini-batches of data. Each step improves the model’s ability to translate by minimizing the error in predictions. The model was trained up to 12,000 steps, with evaluations conducted at regular intervals.

Figure 3: Blue Score Over Training steps

Authorized licensed use limited to: Dedan Kimathi University of Technology. Downloaded on July 23,2025 at 09:40:42 UTC from IEEE Xplore. Restrictions apply.

The BLEU score was selected as the primary evaluation metric due to its widespread adoption in machine translation to measure the similarity between model-generated translations and human references. BLEU measures n-gram overlap between model-generated translations and reference translations. As shown in Figure 3, the BLEU score results indicate a steady improvement in translation quality as training progresses, peaking at 16.29 at 8,000 training steps before experiencing slight fluctuations, eventually settling at 15.78 at 12,000 steps. This suggests that while the model learns effectively in the early stages, additional training does not significantly enhance performance and may cause minor degradation.

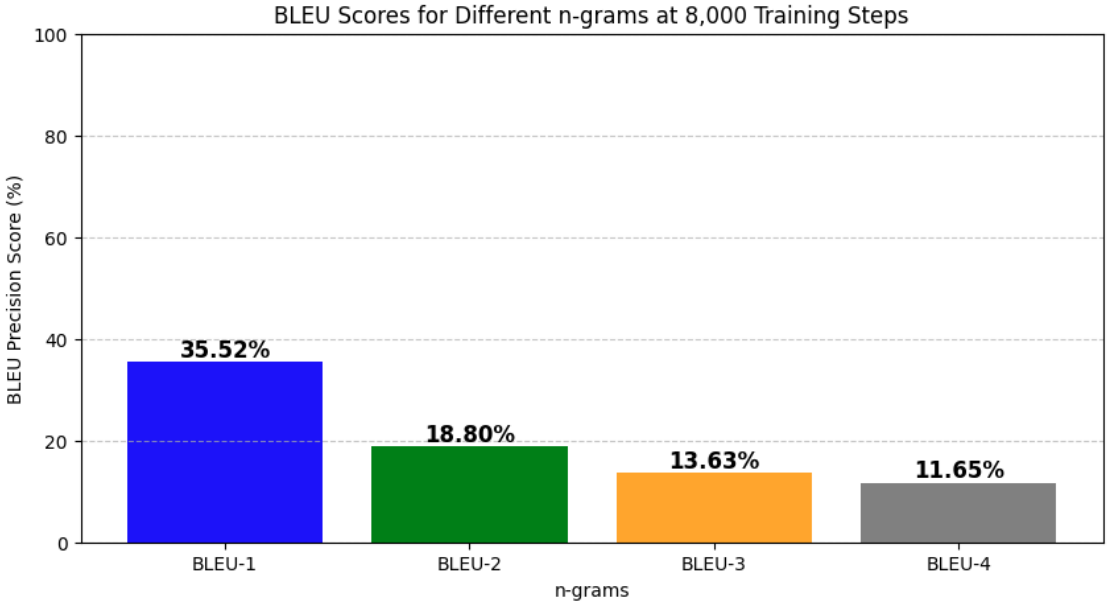

Figure 4: BLUE Score for Different n-grams at 8,000 Training Steps

Figure 4 presents the n-gram BLEU scores at 8,000 training steps, offering a detailed analysis of the model's linguistic capabilities. BLEU-1 (35.52%) indicates strong unigram precision, showing that the model correctly predicts many individual words. However, the sharp decline in BLEU-2 (18.80%), BLEU-3 (13.63%), and BLEU-4 (11.65%) suggests difficulty in maintaining coherence over longer phrase structures. While a BLEU score of 16.29 reflects moderate translation accuracy, it highlights that the model struggles with fluency and precise word choices, producing understandable translations but lacking in grammatical correctness and natural flow.

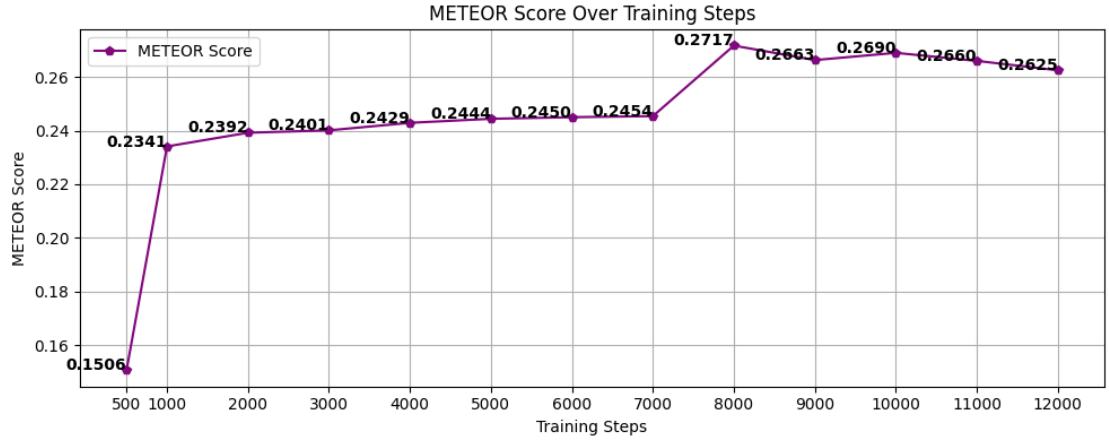

Figure 5: METEOR Score Over Training Steps

METEOR evaluates machine translation quality by considering synonym matches, stemming, and word order, making it more reflective of human judgment than BLEU. The METEOR score (as shown in Figure 5) shows a steady improvement from 0.1506 at 500

Authorized licensed use limited to: Dedan Kimathi University of Technology. Downloaded on July 23,2025 at 09:40:42 UTC from IEEE Xplore. Restrictions apply.

steps to a peak of 0.2717 at 8000 steps. This indicates that the model progressively enhances its translation fluency and recall. However, after reaching this peak, the score slightly declines and stabilizes around 0.2625 at 12000 steps, suggesting diminishing returns in learning.

The increase in METEOR suggests that while the model becomes more accurate, it also captures semantic meaning better. However, the slight drop after 8000 steps suggests the model might be memorizing rather than generalizing.

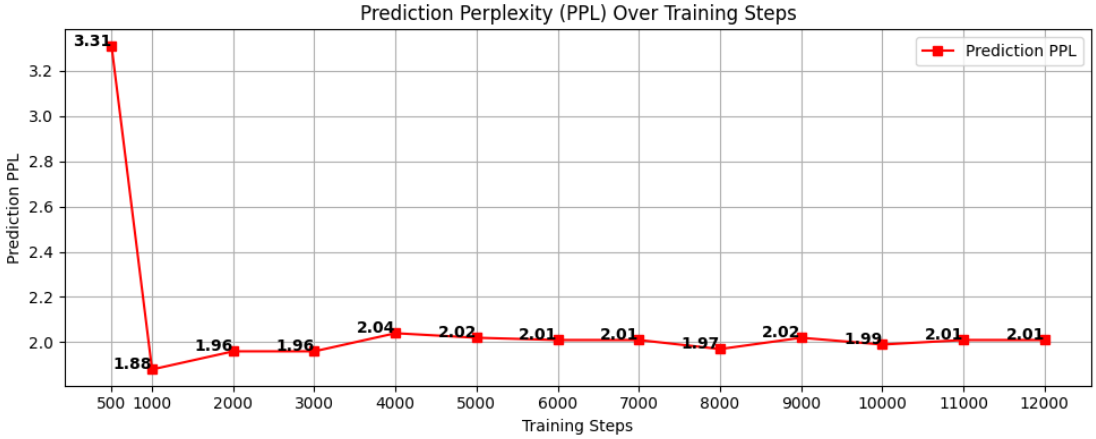

Figure 6: Prediction Perplexity(PPL) Over Training Steps

Perplexity (PPL) measures how uncertain the model is when predicting the next word in a translation. A lower perplexity score indicates a higher confidence in predictions. The best perplexity score achieved was 1.88 at 1,000 training steps, after which it fluctuated slightly before stabilizing around 2.01 at later steps (as shown in Figure 6). A perplexity close to 1.0 is expected in an ideal scenario, meaning the model makes near-perfect predictions. Perplexity values above 10 indicate significant uncertainty, while values around 2.0, such as those achieved in this study, suggest that the model is relatively confident in its output.

However, perplexity varied throughout training due to multiple factors. Initially, as the model learned fundamental translation patterns, PPL dropped significantly, reflecting improved word prediction. However, as training progressed, the model encountered more complex sentence structures and unseen patterns, leading to slight increases in PPL at certain checkpoints. This variation suggests that while the model became more confident in structured translations, it struggled with linguistic diversity and rare vocabulary, preventing PPL from decreasing consistently.

Low perplexity does not always correlate with high translation quality, as the model may still generate fluent but incorrect translations. This explains the gap between the low PPL and the moderate BLEU score. The model might have learned syntactic fluency but struggled with semantic accuracy, producing grammatically coherent, less meaningful, or contextually inaccurate translations. This issue is common in low-resource language pairs, where limited training data restricts the model's ability to generalize beyond seen examples, leading to a confidence gap between perplexity and actual translation accuracy.

The prediction score is another key metric to evaluate the model's performance. It measures the log-likelihood of the predicted translations, where higher (less negative) scores indicate better predictions. In this study, as illustrated in Figure 5, the best prediction score achieved was -0.63 at 1,000 training steps, with scores fluctuating slightly to around -0.70 at later steps. Ideally, a prediction score closer to zero is desired, which implies that the model assigns higher probabilities to correct translations.

Authorized licensed use limited to: Dedan Kimathi University of Technology. Downloaded on July 23,2025 at 09:40:42 UTC from IEEE Xplore. Restrictions apply.

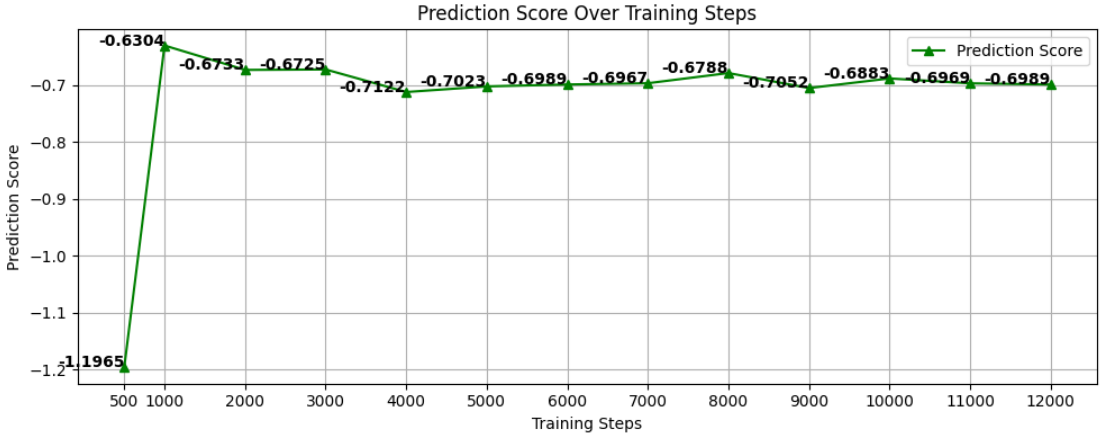

Figure 7: Prediction Score Over Training Steps

The variation in prediction scores over training steps can be attributed to the model's learning process. Initially, prediction scores improved as the model learned to generate reasonable translations, becoming less negative. However, as training continued, the model encountered more complex linguistic structures and unseen patterns, leading to slight fluctuations in prediction scores.

7. Business Benefits

Building a Dholuo-Swahili Neural Machine Translation (NMT) benefits business and commerce, government services, and access to digital content. It helps businesses overcome language barriers and thus expand their market reach to engage with customers better, offer localized content, and make cost-effective communication. The model can be leveraged by e-commerce, banking, and telecommunications sectors to deliver services in both languages, decrease the translation staff, and improve customer experience. Furthermore, the translation system can also be used by government agencies to enhance the accessibility of public services to Dholuo speakers, particularly in legal, healthcare, and administrative processes.

After commerce and governance, digital inclusion and education will have great potential. Such businesses can leverage this model in websites, mobile apps, and chatbots to improve user experience and stimulate digital services' adoption. The model can advocate for translating learning materials and research papers in education to make knowledge available for Dholuo speakers. Cross-border trade within East Africa will also increase if Dholuo-speaking traders begin to work more effectively in Swahili-dominated markets. This NMT model could reduce the translation cost, streamline the operations, and geographically open the market to realize the possibility of the businesses operating more efficiently, promoting economic integration.

8. Conclusions

This study trained a Dholuo-Swahili NMT model using OpenNMT-py, revealing the strengths and limitations of low-resource translation. Evaluations included BLEU, METEOR, PPL, and prediction scores. The best BLEU was 16.29 at 8,000 steps, showing mid-range quality, while METEOR was 0.27, reflecting improved recall and synonym matching. The lowest PPL of 1.88 at 1,000 steps showed early confidence, but prediction scores (-1.20 to -0.63) showed inconsistent fluency and alignment. Despite the low PPL, moderate BLEU and METEOR scores showed fluency and lexical choice struggles, illustrating that confidence doesn't guarantee natural translations.

Key limitations included the small dataset, limiting generalizability, and morphological differences between Dholuo and Swahili. Important lessons learned include the crucial role

Authorized licensed use limited to: Dedan Kimathi University of Technology. Downloaded on July 23,2025 at 09:40:42 UTC from IEEE Xplore. Restrictions apply.

of data availability in determining quality, difficulties with long sentences, and the mismatch of a low PPL to the true translation quality. Recommendations are to expand high-quality parallel data, to use multilingual models like LLaMA 3.2, to integrate BLEU, METEOR, and human evaluations, and to combine rule-based, statistical, and neural models for grammar.

In future work, we shall consider expanding the dataset, focusing on Transformer models, fine-tuning multilingual models such as LLaMA 3.2, and experimenting with BiLSTM for improved fluency and accuracy. These steps would help address the current limitations and improve translation quality.

Acknowledgments

Declaration of Use of Content generated by Artificial Intelligence (AI) (including but not limited to Generative-AI) in the paper

The authors confirm that there has been no use of content generated by Artificial Intelligence (AI) (including but not limited to text, figures, images, and code) in the paper entitled "Neural Machine Translation in Low-Resourced Languages: Case of Dholuo – Swahili Translation".

References

[1] Wang Q, Li B, Xiao T, Zhu J, Li C, Wong DF, et al. Learning Deep Transformer Models for Machine Translation [Internet]. arXiv; 2019 [cited 2024 Jun 18]. Available from: http://arxiv.org/abs/1906.01787

[2] Kituku B, Muchemi L, Nganga W. A Review on Machine Translation Approaches. Indones J Electr Eng Comput Sci. 2016 Jan 1;1(1):182.

[3] Oladosu J, Esan A, Adeyanju I, Adegoke B, Olaniyan O, Omodunbi B. Approaches to Machine Translation: A Review. FUOYE J Eng Technol [Internet]. 2016 Sep 30 [cited 2024 Jun 18];1(1). Available from: https://journal.engineering.fuoye.edu.ng/index.php/engineer/article/view/26

[4] Benson K. An Approach to Bootstrapping the Development of Multilingual Rule-based Grammars for Under-Resourced Languages Using Cross-linguistic Similarities: A Case Study of a Sub-set of Kenyan Bantu Languages.

[5] 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) [Internet]. Bahir Dar, Ethiopia: IEEE; 2021 [cited 2024 Jun 18]. p. 83–8. Available from: https://ieeexplore.ieee.org/document/9672222/

[6] Pauw GD, Wagacha PW. Bootstrapping morphological analysis of gĩkũyũ using unsupervised maximum entropy learning. In: Interspeech 2007 [Internet]. ISCA; 2007 [cited 2024 Oct 30]. p. 1517–20. Available from: https://www.isca-archive.org/interspeech_2007/pauw07_interspeech.html

[7] N. Kituku B, George M, Wagacha P. Kamba Part of Speech Tagger Using Memory Based Approach. Int J Nat Lang Comput. 2015 Apr 30;4(2):43–53.

[8] Kizza JM, Muhirwe J, Aisbett J, Getao K, Mbarika VW, Patel D, Rodrigues AJ. Strengthening the Role of ICT in Development. Special Topics in Computing and ICT Research. 2007.

[9] Duong L. Natural language processing for resource-poor languages. University of Melbourne. 2017 Oct.

[10] Krauwer S. The basic language resource kit (BLARK) as the first milestone for the language resources roadmap. InProceedings of SPECOM 2003 Oct 27 (Vol. 2003, No. 8, p. 15).

[11] Yigezu MG, Woldeyohannis MM, Tonja AL. Multilingual neural machine translation for low resourced languages: Ometo-english. In2021 International Conference on Information and Communication Technology for Development for Africa (ICT4DA) 2021 Nov 22 (pp. 89-94). IEEE.

[12] Lankford S, Afli H, Way A. adaptNMT: an open-source, language-agnostic development environment for neural machine translation. Language Resources and Evaluation. 2023 Dec;57(4):1671-96.

[13] Bengi M, Amalia A, Budiman MA. A Transformer-Based Approach with OpenNMT for Indonesian to Gayo Language Translation. In2024 2nd International Symposium on Information Technology and Digital Innovation (ISITDI) 2024 Jul 24 (pp. 149-153). IEEE.

[14] Li Y. Comparison and analysis of two methods for improving the accuracy of OpenNMT in literary and vernacular Chinese translation. Applied and Computational Engineering. 2024 Mar 28; 53:238-45.

[15] Luvaga D, Enaka DT, Issa MY. The Role of Kiswahili in Promoting Intercultural Communication and Strengthening National Ethos Amongst the Co-Cultures of the Rift Valley in Kenya. 2021;

Authorized licensed use limited to: Dedan Kimathi University of Technology. Downloaded on July 23,2025 at 09:40:42 UTC from IEEE Xplore. Restrictions apply.

[16] Ngoni VN. *English–Bukusu Automatic Machine Translation for Digital Services Inclusion in E-governance* (Doctoral dissertation, university of nairobi).
[17] Guzmán F, Chen PJ, Ott M, Pino J, Lample G, Koehn P, et al. The FLoRes Evaluation Datasets for Low-Resource Machine Translation: Nepali-English and Sinhala-English [Internet]. arXiv; 2019 [cited 2024 Jun 18]. Available from: http://arxiv.org/abs/1902.01382
[18] Klein G, Kim Y, Deng Y, Nguyen V, Senellart J, Rush AM. OpenNMT: Neural Machine Translation Toolkit [Internet]. arXiv; 2018 [cited 2024 Jun 18]. Available from: http://arxiv.org/abs/1805.11462
[19] Wanjawa B, Wanzare L, Indede F, McOnyango O, Ombui E, Muchemi L. Kencorpus: A Kenyan Language Corpus of Swahili, Dholuo and Luhya for Natural Language Processing Tasks. J Lang Technol Comput Linguist. 2023 Jun 21;36(2):1–27.
[20] Vidal-Silva C, Barriga NA, Ortega-Cordero F, Gonzalez-Lopez J, Jimenez-Quintana C, Pezoa-Fuentes C, et al. Developing Computing Competencies Without Restrictions. IEEE Access. 2022; 10:106568–80.
[21] Nekoto W, Marivate V, Matsila T, Fasubaa T, Kolawole T, Fagbohungbe T, et al. Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages [Internet]. arXiv; 2020 [cited 2024 Jun 18]. Available from: http://arxiv.org/abs/2010.02353
[22] Mbogho A, Kipkebut A, Wanzare L, Awuor Q, Oloo V, Lugano R. waleghwa/low-resource-language-data: Parallel Corpora for Kiswahili and Kidaw'ida, Kalenjin and Dholuo. Zenodo; 2024.

Authorized licensed use limited to: Dedan Kimathi University of Technology. Downloaded on July 23,2025 at 09:40:42 UTC from IEEE Xplore. Restrictions apply.