ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24OCT410

# Advancements in Natural Language Understanding-Driven Machine Translation: Focus on English and the Low Resource Dialectal Lusoga

Azizi Wasike[*1]; Ismail Kamukama[2]; Yusuf Abass Aleshinloye[3]; Adeleke Raheem Ajiboye[4]; Jamir Ssebadduka[5]
[1,2,3,4,5]Department of Computer Science
Kampala International University
Kampala, Uganda

**Abstract:- This review explores recent advancements in Natural Language Understanding-driven Machine Translation (NLU-MT) with a focus on English and the low-resource dialectal Lusoga. A Low-resource language, such as Lusoga, faces significant challenges in Machine Translation (MT) due to the scarcity of high-quality parallel corpora, the complex morphology inherent in Bantu languages, and the dialectal variations within Lusoga itself, particularly between Lutenga and Lupakoyo. This paper examines the role of NLU-based MT systems in overcoming these challenges by shifting from word-for-word mapping to meaning-based translations, enabling better handling of these dialectal differences. We highlight the success of leveraging linguistic similarities between Lusoga and related languages, such as Luganda, to improve translation performance through multilingual transfer learning techniques. Key advancements include the use of transformer-based architectures such as Multilingual Bidirectional and Auto-Regressive Transformer (mBART) and Multilingual Text-To-Text Transfer Transformer (mT5), specifically selected for their effectiveness in NLU-driven contexts, which have shown promise in enhancing translation accuracy for African low-resource languages. However, the review also identifies ongoing obstacles, including historical low demand and the lack of well-developed corpora, which hinder scalability. The paper concludes by emphasizing the potential of hybrid approaches that combine community-driven corpus-building initiatives with improved model architectures to drive further progress in low-resource MT. Ultimately, NLU-MT is positioned as a crucial tool not only for bridging communication gaps but also for preserving linguistic diversity and cultural heritage.**

***Keywords:-*** *Natural Language Understanding; Machine Translation; Low-Resource Languages; Lusoga, Dialectal Variations; Transfer Learning; Community-driven Corpus Building; mBART; mT5;mBERT.*

## I. INTRODUCTION

### *A. Background*

Machine Translation (MT) for low-resource languages, like Lusoga, a dialectal Bantu language spoken in Uganda, has become vital for preserving cultural identity and facilitating communication across linguistic boundaries [1], [2]. However, the lack of large parallel corpora and computational resources makes developing accurate translation models challenging for Lusoga and similar languages [3], [4].

Bantu languages, including Lusoga, present unique challenges due to their rich morphology, agglutination, and syntactic complexity. Traditional MT methods, which depend on large bilingual datasets, often struggle with these complexities, leading to subpar translations [5].

Natural Language Understanding (NLU) offers a solution by emphasizing semantic meaning and linguistic context. NLU-driven MT (NLU-MT) systems focus on meaning-based translation rather than word-for-word mapping, allowing them to handle dialectal variations, such as Lutenga and Lupakoyo, more effectively. By leveraging deep contextual representations, NLU-based systems can generalize well, even with limited data, enhancing fluency, adequacy, and overall translation quality in low-resource languages like Lusoga [6], [7].

### *B. Objective of the review*

This paper reviews NLU-MT model architectures for English-Lusoga translation, focusing on key advancements that improve MT performance in low-resource settings. It also highlights current limitations and suggests future research areas to enhance MT for Lusoga and similar dialects.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24OCT410

## II. NATURAL LANGUAGE UNDERSTANDING AND MACHINE TRANSLATION

### A. Overview of NLU-MT

NLU-MT is a machine translation system that integrates Natural Language Understanding (NLU), emphasizing semantic comprehension, contextual awareness, and disambiguation. Unlike Statistical Machine Translation (SMT), which relies on large aligned corpora and probabilistic models, or Neural Machine Translation (NMT), which predicts word sequences through neural networks, NLU-MT deeply analyses the meaning behind words and phrases. This semantic understanding enables it to better grasp context, resolve ambiguities, and handle linguistic nuances, making it particularly effective for complex, morphologically rich languages like Lusoga. NLU-driven MT focuses on meaning rather than statistical patterns, allowing for more accurate translations in low-resource languages, where large datasets are unavailable and cultural subtleties must be carefully considered [6], [7].

### B. Role of NLU in Low-Resource MT

NLU enhances translation quality for low-resource languages by focusing on meaning rather than word- or phrase-based methods. In cases with limited parallel corpora, traditional models struggle due to their reliance on large datasets. NLU addresses this by incorporating semantic understanding, allowing models to better capture context and meaning. Lusoga, a Bantu language with rich morphology and dialectal variations, challenges literal translation models [5], [8]. NLU-driven systems can disambiguate polysemous words and handle context-sensitive phrases, adapting well to complex grammar. This approach improves translation accuracy for Lusoga by capturing its unique structure and cultural nuances, without needing extensive datasets [9], [10].

## III. KEY MODEL ARCHITECTURES FOR NLU-MT

Lusoga, with its two dialects—Lutenga and Lupakoyo—can benefit from multilingual machine translation, which handles multiple language pairs with a single model [4]. This approach is scalable, easy to manage, and promotes knowledge transfer between related languages through shared representations. It also improves translations for low-resource languages and enables zero-shot translation between pairs not seen during training [11], [12].

Recent advancements in NMT have been driven by several key models that have greatly enhanced translation quality and efficiency. Recent advancements in Natural Language Processing (NLP) have underscored the importance of pre-trained language models, which leverage large-scale corpora for initial training before being fine-tuned for specific tasks, such as machine translation in low-resource languages. At the heart of this progress is the Transformer architecture, introduced by [13], which is employed in three primary configurations: The Transformer encoder, used in models like Bidirectional Encoder Representations from Transformers (BERT) and its variants (Robustly Optimized BERT Approach (RoBERTa), Span-based BERT (SpanBERT)); the Transformer decoder, found in generative models such as Generative Pre-trained Transformer (GPT) and its successors (GPT-2, GPT-3); and the Transformer encoder-decoder, exemplified by models like Text-To-Text Transfer Transformer (T5) and Bidirectional and Auto-Regressive Transformer (BART). Each of these architectures offers distinct advantages for various NLP tasks, driving significant advancements in the field [12], [14]. Despite these gains, [12] highlights a critical challenge in using deep learning methods for NMT: their heavy reliance on vast amounts of data and computational resources, making it difficult to apply these models effectively to low-resource languages.

The BERT architecture, developed by [15] enhanced contextual understanding in NMT. BERT's bidirectional text representation enables the model to grasp context from both preceding and following text, improving the accuracy and fluency of translations. Although originally designed for tasks such as question answering, named entity recognition and text classification, BERT has been effectively adapted for translation tasks, showcasing its versatility and deep contextual understanding.

RoBERTa, an improved version of BERT, enhances NLP task performance through optimized pre-training and uses the Transformer encoder to learn bidirectional context. It excels in tasks like text classification and sentiment analysis, thanks to its deep linguistic modelling. However, RoBERTa is not ideal for translation, especially for low-resource languages like Lusoga, as it lacks the sequence-to-sequence architecture needed for effective translation. While strong in language understanding, its longer execution time and focus on comprehension over generation limit its use for translating between languages like English and Lusoga [16].

SpanBERT, an extension of BERT, improves predictions by focusing on spans of text, capturing relationships between multiple tokens. It uses the Transformer encoder to learn context and is particularly strong in tasks like coreference resolution and question answering. However, SpanBERT is not designed for translation, especially for low-resource languages like Lusoga, as it lacks the sequence-to-sequence capabilities needed for effective machine translation. While it excels in span-based tasks and text classification, it is less suited for translation between languages [17].

Reference [18] highlight that while multilingual models like mBERT, trained on both high and low-resource languages, aim to improve translation for low-resource languages, their performance often falls short. The increased diversity of languages in the model can degrade translation quality for low-resource languages. They suggest that training transformer models on smaller, related language datasets is more effective than using large, unrelated datasets. The “small data” approach, as seen in AfriBERTa [4], focuses on pretraining with limited data in low-resource languages. Unlike mBERT, which is primarily for tasks like text classification, mBART is specifically designed for translation tasks, making it more effective in this area.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24OCT410

Reference [4] addressed key challenges in developing language models for low-resource languages, proposing innovative solutions that emphasize the importance of focusing on language similarity rather than relying on high-resource languages. They demonstrated that pretraining on similar low-resource languages can yield better results, challenging the common assumption in the NLP community that combining them with high-resource languages is always beneficial. The study also tackled the limited data availability by successfully training the AfriBERTa model with less than one gigabyte of text data from African languages, proving that competitive multilingual models can be pretrained from scratch using only low-resource languages. Additionally, [4] explored the relationship between vocabulary size and model performance, finding that medium-sized vocabularies often outperform larger ones, a critical insight for optimizing model training in low-resource settings. Finally, the researchers addressed ethical concerns, focusing on reducing societal bias by developing language technology for underserved languages and using smaller datasets to facilitate cleaner and more socially responsible model training.

The GPT series, including GPT-2, GPT-3, and GPT-4, uses a transformer architecture to generate text by predicting the next word based on context. While effective for languages with abundant data, GPT struggles with low-resource languages like African Bantu due to limited training data, leading to less accurate translations and biases [19]. GPT-2 introduced strong text generation capabilities but had limitations with complex translation. GPT-3, with 175 billion parameters, improved coherence and context understanding but still lacks effectiveness for under-resourced languages. GPT-4 further enhances contextual accuracy but remains challenged by low-resource data availability. In contrast, mBART, optimized for sequence-to-sequence tasks, combines a bidirectional encoder with an autoregressive decoder [12]. It excels in both understanding and generating text, particularly when fine-tuned on low-resource languages [20]. Studies, like [21], show significant improvements in translation quality for languages such as Nepali, Sinhala, and Gujarati, making mBART a more effective solution than GPT for low-resource translation.

Reference [12] developed a multilingual corpus for five Ugandan native languages to train and evaluate mBART models, addressing several challenges in machine translation for low-resource languages. A key challenge was the lack of standardized writing systems, which they tackled by involving local translators familiar with linguistic nuances to create a more consistent dataset. To address the limited availability of professional translators, they recruited and trained local translators, aiming to minimize biases and ensure contextually accurate translations. Recognizing the complications of dialectal variations, [12] included a wide range of dialectal expressions to reflect local usage accurately. The scarcity of reliable translation systems, due to insufficient training data, was mitigated by developing a parallel text corpus, Sunbird African language Technology (SALT), specifically for these languages. Additionally, to avoid copyright issues and capture informal tones, they generated prompts from diverse sources like social media and local news. Finally, they employed a community-based approach to collect a comprehensive dataset that meets the broader linguistic needs of the local population, overcoming the narrow focus of existing public datasets.

The Text-To-Text Transfer Transformer (T5) model, introduced by [22] frames all NLP tasks as text-to-text problems, simplifying training and enhancing adaptability, which results in strong performance across tasks, including translation. Several studies have explored Afri-centric T5 models like AfriMT5, AfriByT5, AfriTeVa, and AfriTeVa V2 [23]. Notably, AfriTeVa V2, fine-tuned on the filtered Afri-centric WURA dataset, showed significant improvements in translation quality. Reference [23] identified several key challenges in improving T5-based models for Afri-centric multilingual machine translation. They noted that datasets from African news websites often fail to capture the continent's linguistic and cultural diversity, leading to models that inadequately reflect African languages. To address this, they recommend rigorous auditing of document sources, including translation to ensure high-quality data. Furthermore, [23] highlights the bias in existing data sources, like the mC4 dataset, which can skew model accuracy. To combat this, they propose enhanced web crawling strategies to create a more diverse and balanced dataset. Additionally, they emphasize the need to address quality issues in corpora for low-resource African languages, which often rely on limited sources like religious texts. To overcome these challenges, they suggest developing new multilingual datasets, such as the WURA corpus for 16 African languages, ensuring clean and relevant data for training.

In the study "Building a Parallel Corpus and Training Translation Models Between Luganda and English," [24] employed the mT5 model, a pre-trained variant of the transformer architecture recognized for its proficiency in multilingual translation tasks. This model was chosen for its ability to effectively handle sequential data, a critical requirement in neural machine translation. During training, a hyperparameter search was carried out to optimize the model’s performance, adjusting parameters to achieve better results. The mT5 model utilized a 10,000-word vocabulary and a 6-layer encoder-decoder architecture, specifically designed to capture complex linguistic relationships. Notably, [25] observed that between 82% and 86% of words in Lusoga are similar or identical in form and meaning to those in Luganda. Therefore, this mT5 model, pretrained on Luganda, can be fine-tuned on a Lusoga dialectal corpus, leveraging transfer learning to address the low-resource nature of Lusoga [12]. Reference [24] addressed key challenges in developing T5 machine translation models for English and Luganda, including the lack of high-quality parallel corpora and small datasets from previous research. They built a bilingual corpus with 41,070 pairwise sentences by combining open datasets, significantly enhancing the data available for model training. To overcome the high computational demands of modern NMT models, the researchers used a pre-trained mT5 model variant, optimized for efficiency in multilingual settings, allowing them to achieve strong translation results despite limited resources.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24OCT410

Reference [26] developed an English-Swahili corpus for news domain NMT using several methods. They primarily employed the Transformer architecture, known for its efficiency with long-range dependencies, which achieved the highest Bilingual Evaluation Understudy (BLEU) scores for Swahili to English translations. They also used Recurrent Neural Networks (RNNs) with attention mechanisms, which enhanced translation quality for English to Swahili tasks by focusing on specific input parts. Reference [26] optimized hyperparameters through a grid search for both architectures to maximize performance. Additionally, they integrated monolingual data using back-translation techniques to boost translation quality and incorporated linguistic information by encoding tags in the training corpus to improve grammatical accuracy. Several challenges were faced by [26] in their study, which they tackled with various strategies. The primary issue was the scarcity of parallel corpora for English and Swahili, leading them to crawl additional data from the Internet to expand their corpus and improve model training. Linguistic differences, such as Swahili's noun classes and agreement rules, further complicated translation; the researchers addressed this by incorporating linguistic information into their models. In addition, [26] managed word order differences, especially in noun phrases, by training models to handle specific reordering challenges. To address the absence of articles and gender marking in Swahili, the models were designed to maintain grammatical accuracy in English translations. Additionally, to resolve inconsistencies in manual evaluation, the researchers refined their methods to ensure more reliable assessments of translation quality.

While developing the low-resource NMT model for Wolaytta-English, [27] employed a transformer-based NMT approach to address challenges in low-resource languages. They tackled the issue of limited training data by using source-side monolingual datasets to supplement scarce parallel data, enhancing model performance through self-learning techniques. To overcome domain-specific limitations, they used both authentic and synthetic datasets for the Wolaytta-English translation task, creating a more adaptable model. Additionally, they balanced authentic and synthetic data in their models to avoid overfitting and improve translation quality, achieving significant gains in BLEU scores for both translation directions. Reference [28] proposed several techniques to improve NMT for low-resource languages. They recommended back-translation to generate additional training data by translating from the target language back into the source language, which significantly enhances translation performance. They also highlighted transfer learning, where models trained on high-resource languages are adapted to low-resource ones, leveraging linguistic similarities, as seen with languages like Bavarian. To tackle data scarcity, [28] suggested data augmentation through synthetic data and monolingual data, which helps the model generalize better. Additionally, incorporating data from closely related languages was shown to improve translation accuracy by exploiting linguistic similarities. Finally, they stressed the need for standardizing training data to reduce noise and ensure high-quality examples, which is crucial for achieving accurate and reliable translations.

Architectures such as mBART and mT5 are effective for multilingual tasks, but their ability to handle Lusoga's rich morphology depends on having enough training data, which is often lacking for low-resource languages like Lusoga [29]. This underrepresentation limits the models' ability to fully learn Lusoga's inflectional and agglutinative patterns, leading to errors like incorrect noun-verb agreement or mistranslation of inflected forms. Additionally, Lusoga has dialects such as Lutenga and Lupakoyo, which differ in vocabulary and grammar, making translation even more challenging. Transfer learning offers a solution by first training these models on related languages like Luganda, which shares linguistic similarities with Lusoga, before fine-tuning on Lusoga-specific or dialectal data. This approach improves the model's ability to handle Lusoga's unique structure and dialectal variations [30], [31].

## IV. CORPUS DEVELOPMENT FOR LUSOGA

### *A. Lusoga Language*

Lusoga, a Bantu language used by the Basoga in Eastern Uganda, has three dialects: Lutenga, Lulamoogi, and Lusiki. The latter two are often grouped together under the name Lupakoyo. The differences between Lutenga and Lupakoyo extend beyond mere pronunciation; they are also lexical, phonological, and morphological, which can make it difficult for speakers of one dialect to understand speakers of the other [2]. The author highlights that, despite notable linguistic differences, the deep attachment speakers feel towards their particular dialects is significant. Past attempts to create both single-dialect and multi-dialect orthographies for Lusoga have faced skepticism, particularly as the National Curriculum Development Centre (NCDC) has advocated for standardizing the language with Lutenga as the chosen form. Lusoga is a low-resource language with minimal linguistic resources and corpora, as it predominantly exists in oral form and remains mostly undocumented [25], [32].

Interest in Lusoga surged after its official recognition in 2005, leading to Lusoga Orthography efforts by [33], Lusoga sound system descriptions, creation of the first monolingual offline and electronic Lusoga dictionary, and the 1.7m Lusoga corpus [25], [34]–[37]. Despite these efforts and its over two million speakers, Lusoga remains severely undocumented and under-resourced [25], [38]. According to [33], Lusoga utilizes 51 letters and letter combinations, ranging from A(a) to Z(z). In the Lutenga dialect, these letters are called "nhukuta," while in the Lupakoyo dialect, they are referred to as "nyukuta." Numbers, on the other hand, are known as "nhuguta" or "nyuguta". The following list displays the various letters: A (a), B (b), BW (bw), By (by), C (c), D (d), Dh (dh), E (e), F (f), G (g), Gh (gh), ŋ (ŋ), Gw (gw), H (h), I (i), J (j), Gy (gy), K (k), Kw (kw), Ky (ky), L (l), M (m), Mb (mb), Mp (mp), Mw (mw), N (n), Nd (nd), Ndw (ndw), Nf (nf), Nh (nh), Nhw (nhw), Nw (nw), Nk (nk), Nkw (nkw), O (o), P (p), Th (th), R (r), S(s), Shy (shy), Sy (sy),T (t), Tw (tw), Ty (ty), U (u), V (v), W (w), Y (y), Z (z), Zw (zw), and Zy (zy).

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24OCT410

The author also notes that Lusoga uses 25 vowels called endhatuza. These include; A (a), E (e), I (i), O (o), U (u), aa, ae, ai, ao, au, ae, ee, ei, eo, eu, ia, ie, ii, io, iu, oa, oe, oi, oo and ou. The language's consonants include; W (w), Y (y), C (c), H (h), ŋ (ŋ), B (b), P (p), V (v), F (f), M (m), D (d), DH (dh), T (t), L (l), R (r), N (n), Z (z), S (s), J (j), G (g), and K (k). The letter J (j) and C (c) are commonly used in borrowed words such as Chai (tea), Cooka (chalk) and Jiija (compound grass). Furthermore, Lusoga is a noun centric language and organizes its nouns into a structured system of 19 distinct classes, each distinguished by specific prefixes that indicate both singular and plural forms [33], [39].

*B. Overview of Corpus Creation*

NMT models depend on extensive bilingual and monolingual corpora to generate accurate translations. Reference [26] emphasize the importance of bilingual corpora for NMT training, noting that synthetic corpora derived from back-translating monolingual data are commonly used. For their English-Kiswahili modelling project, the GoURMET project supplied English and Kiswahili bilingual and monolingual corpora, and the SAWA corpus added more parallel texts. The GoURMET corpora was crawled using Bitextor a free opensource tool that automatically extracts parallel corpora from multilingual websites for training and evaluating NMT models. Additional data came from the OPUS website, which offers a large, multilingual dataset from various sources, improving translation accuracy.

Several publicly available datasets have been used for NLP tasks. These datasets include Common Crawl, a large-scale web corpus with raw web page data [40]; The Pile, a diverse collection of English text from sources like books, GitHub, and academic papers, designed specifically for language modelling [41]; Wikipedia dumps, often used for NLP tasks due to their well-structured, high-quality text on various topics [42]; and OpenSubtitles, which offers English subtitles from movies and TV shows, providing a wide range of conversational text [43].

Reference [40] notes that web-crawled datasets like Common Crawl often contain noise, such as duplicate texts, non-linguistic content (e.g., HTML and script tags), and incorrect language use, including incomplete sentences, slang, and errors. They emphasize the need to clean these datasets before using them for model training, as done in the creation of the English Colossal Clean Crawled Corpus (C4). Moreover, these corpora often underrepresent low-resource Bantu languages [23]. The authors emphasize that high-quality datasets with robust representation of African languages are essential for improving translation model performance, as exemplified by their African-focused WURA multilingual dataset.

Masakhane, a pan-African NLP network, is exploring participatory methods for sustainable language data collection, including for languages like Setswana and Sepedi. Reference [12] developed the SALT dataset, a parallel corpus for English and five low-resource Ugandan languages—Acholi, Ateso, Luganda, Lugbara, and Runyankole—by collaborating with local communities. The dataset includes 40,000 Luganda sentences and 25,000 sentences for each of the other languages, all professionally translated on relevant topics. This effort reduced the bias found in datasets like MT560, which are mostly based on religious texts.

Reference [26] developed an English-Swahili corpus for news domain NMT using the GlobalVoices parallel corpus. They selected 4,000 sentences from GlobalVoices-v2015 and GlobalVoices-v2017q3, dividing them into 2,000 sentences for the development corpus and 2,000 for the test corpus. To ensure test data quality, they removed sentences overlapping with any monolingual corpora and pre-processed the remaining data by tokenization, true-casing, and applying Byte Pair Encoding (BPE). For manual evaluation, they added five Swahili sentences written by a translator to the test set, resulting in 210 sentences.

Reference [24] developed a bilingual parallel corpus for Luganda and English by merging multiple open datasets. One dataset from Zenodo (2022) contained 1,042 English-Luganda parallel sentences, while another from 2021 included 15,022 sentences. A third dataset, released in 2021 by Sunbird AI and Makerere AI lab, featured 25,006 phrases translated into five local languages. The English texts, sourced from social media, radio transcripts, online newspapers, blogs, and farmer surveys, are available in JSON format on GitHub [12].

When using Source-Side Monolingual data to improve low-resource NMT, [27] developed their corpus by combining authentic parallel datasets, monolingual datasets, and synthetic datasets to enhance translation quality. They trained the baseline model using the Wolaytta-English parallel dataset and, for the self-trained model, incorporated authentic parallel sentences with Wolaytta monolingual and English synthetic data. Data pre-processing involved removing duplicate sentences, converting text to lowercase, eliminating special characters except the significant apostrophe in Wolaytta, and tokenizing sentences into sub-word tokens using BPE. Additionally, they made the training scripts and datasets publicly available, supporting further research in Wolaytta-English machine translation.

Reference [44] developed low-resource machine translation models for South African languages, focusing on isiZulu and Sepedi. They used datasets from the National Centre for Human Language Technology (NCHLT), which provided monolingual corpora for all 11 official South African languages, and included news articles from the isiZulu Isolezwe newspaper, one of the largest publicly available African language corpora.

Reference [44] collected data from all 11 languages but focused primarily on isiZulu and Sepedi for detailed modelling. The pre-processing steps included filtering out irrelevant text and normalizing the data to reflect natural language usage. The datasets, containing between 1 and 3 million tokens, were divided into training, validation, and test sets with an 80%/10%/10% split. In contrast, [24] opted for a different allocation, designating 94% of the dataset for training, and 3% each for testing and evaluation.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24OCT410

Reference [12] also recommend extending training datasets using data augmentation techniques. After training their models on the original dataset, they merged it with the FLORES-101 and MT560 datasets and retrained their models resulting in improved translation quality. They used back-translation to expand their dataset by translating locally relevant texts from news sites back into the source language and employed paraphrasing as another augmentation strategy. The authors also addressed challenges with out-of-vocabulary (OOV) terms, especially named entities, which models often attempt to translate but should remain unchanged. To tackle this, they created a dataset of named entities with identical source and target texts from WNUT17 and WikiGold databases, teaching the model to preserve OOV terms and thus significantly improving translation quality.

Reference [45] developed a multilingual corpus for five Ugandan languages—Ateso, Luganda, Lugbara, and Runyankole—using a community-based approach. They gathered data from sources like religious texts, magazines, and the Bible to ensure a diverse dataset. By merging data from three research Centre corpora, they created a comprehensive dataset to enhance translation model training. They also expanded existing bilingual datasets, creating a larger Luganda-English dataset with 41,070 sentence pairs from open-source corpora. Reference [46] developed a small bilingual corpus of English and Lumasaaba using Bible verses, which was then used to train and evaluate MT models. Reference [47] used Luo and English language bilingual experts to translate from English texts but observed that the resulting translations were too formal. Reference [48] identified gender bias as the most common issue in MT systems.

To demonstrate a solution to this for the case of English-Luganda MT, [48] created an English-Luganda MT model using a dataset of 1000 gender-sensitive sentences that was then made this dataset publicly available for others to use. It should be noted however that much of the early machine translation for African native languages relied heavily on Biblical texts to create bilingual datasets which may not adequately reflect the local context of these languages [11].

### *C. Data Augmentation techniques*

In low-resource settings like English-Lusoga, data augmentation techniques such as paraphrasing and backtranslation are crucial for generating additional sentence pairs for MT models. Paraphrasing rewords sentences while preserving their meaning, enriching the data with diverse structures. However, it may introduce inaccuracies if the meaning shifts. Backtranslation expands the dataset by translating from the target language (Lusoga) back to the source (English), but its effectiveness depends on the quality of the initial model [49].

Advanced techniques further improve NMT models. Adversarial Data Augmentation generates virtual sentences to enhance robustness, though it can be computationally demanding. Doubly Adversarial Inputs test models with difficult sentences but risk overfitting if overused [50]. Noise Integration, introduced by Michel and Neubig, trains models using noisy data from sources like Reddit, improving adaptability but potentially lowering translation quality if too much noise is included [51].

Incorporating monolingual data boosts fluency, reduces overfitting, and aids domain adaptation, though biases may emerge if the data is misaligned with the target domain [52]. Overall, these techniques are essential for improving low-resource NMT models like English-Lusoga, compensating for the lack of large parallel corpora.

### *D. Utilizing Native Lutenga and Lupakoyo Speakers for Corpus Creation*

In Translation task-based data collection with bilingual English and Lusoga experts is crucial for creating quality translations in the absence of large corpora. These experts, often linguists, provide key insights into dialectal variations and syntax, ensuring translations preserve both linguistic and cultural authenticity. Their input helps navigate Lusoga's dialects, like Lutenga and Lupakoyo, ensuring translations are grammatically accurate and contextually appropriate. This process, where experts translate English sentences into Lusoga, captures idiomatic expressions and nuances that automated systems might miss. Their involvement minimizes translation errors and improves the performance of NLU-MT models, enhancing translation quality in low-resource settings [12], [53].

## V. EVALUATION OF NLU-DRIVEN MT MODELS

Evaluating MT models is critical for developing effective translation systems, particularly for low-resource languages. Common metrics like BLEU, chrF, and TER are widely used, each offering distinct strengths and limitations. BLEU, which measures n-gram overlap between machine and reference translations, is simple and fast but struggles with flexible word order and rich morphology, making it less suitable for languages like Lusoga [54]. chrF, which evaluates character-level matches, is better suited for morphologically rich languages, capturing nuances BLEU may miss, though it can be less effective for assessing broader syntactic and semantic qualities [27]. TER, which measures the number of edits needed to align translations with reference texts, mirrors human judgment well but can penalize minor changes unnecessarily [28]. In low-resource settings, qualitative analysis by bilingual speakers is also valuable, offering insights into translation nuances that quantitative metrics may overlook, though it is time-consuming, expensive and requires expert knowledge. For example, [24] used BLEU to evaluate an English-Luganda MT model, achieving scores of 21.28 for Luganda-to-English and 17.47 for English-to-Luganda, highlighting BLEU's utility in these contexts but acknowledging its limitations. Similarly, other studies, such as those by [27] and [26], combine BLEU, Character F-score (chrF), and Transfer Error Rate (TER) to gain a more comprehensive understanding of MT model performance across various linguistic challenges, emphasizing the need for multiple metrics in low-resource language evaluation.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24OCT410

## VI. CHALLENGES AND OPPORTUNITIES IN NLU-BASED MT FOR LUSOGA

### *A. Resource Constraints*

Gathering sufficient training data for Lusoga presents significant challenges, primarily due to its status as a low-resource language and the dialectal variations within the language itself, such as Lutenga and Lupakoyo. The scarcity of parallel corpora—texts available in both English and Lusoga—further complicates efforts to build robust MT systems [35]. Unlike high-resource languages, where large datasets are readily available for training, the Lusoga language lacks comprehensive digitized content and formal linguistic documentation. This shortage restricts the ability to develop high-quality translation models, as machine learning systems rely heavily on large datasets to improve accuracy and generalization [48].

Dialectal differences between Lutenga and Lupakoyo add another layer of complexity [2]. Since these dialects are spoken in different regions of the Busoga subregion, ensuring that the data is representative of both requires careful sampling and inclusion of varied linguistic sources. Without accounting for dialectal variation, MT systems risk generating translations that are regionally biased or linguistically inconsistent, thereby limiting their effectiveness across the wider Lusoga-speaking population [55].

Ethical considerations are also crucial when developing language resources for underrepresented languages like Lusoga. Data collected should be representative and thus not marginalize the dialects that constitute the language. Collecting language data from native speakers and linguistic experts must respect cultural sensitivities [56], and researchers must ensure that the process is non-exploitative. Consent and community involvement are essential to prevent the misuse or commodification of linguistic heritage [57]. Furthermore, the resulting datasets should ideally serve the community by contributing to educational resources or preserving linguistic diversity. Making these datasets freely accessible helps ensure that the language remains a vibrant and integral part of the community's cultural heritage [58]. Balancing these ethical concerns with the technical need for extensive, high-quality data remains one of the key challenges in building MT systems for low-resource languages like Lusoga.

### *B. Transfer Learning and Zero-Shot Translation*

In low-resource settings, transfer learning and zero-shot translation offer effective methods to enhance MT for languages like Lusoga, which lacks extensive data.

- *Transfer Learning*

Several strategies enhance translation performance in low-resource contexts. One effective approach is parent-child model initialization, where a high-resource parent model initializes a low-resource child model, resulting in better performance compared to random initialization [59]. This method effectively leverages existing knowledge but may be limited if the languages are too dissimilar, leading to inadequate model adaptation. Language relatedness is also critical; shared vocabularies between related languages can significantly aid the transfer process. Techniques such as subword segmentation, like BPE, improve vocabulary overlap, particularly for distantly related languages, although they may not fully address grammatical differences. Transliteration serves closely related languages with different scripts, enhancing transferability, but it may struggle with phonetic nuances. Additionally, syntactic reordering techniques improve translation quality by aligning language structures, though this requires accurate linguistic rules for each language pair [60], [61].

Multi-stage transfer learning introduces a pivot language, allowing the model to learn from multiple sources, thus further strengthening low-resource translations. However, this approach can be computationally intensive and may require extensive parallel data. Finally, using pre-trained models from high-resource language pairs, such as Turkish-English, can benefit low-resource pairs like Kazakh-English, offering a comprehensive approach to enhancing translation accuracy and efficiency. However, the effectiveness of this method can be limited by the specificity of the high-resource data to the target languages [62], [63]. Together, these strategies provide a multifaceted approach to improving machine translation in challenging linguistic contexts.

- *Zero-Short Translation*

Zero-shot translation (ZST) allows models to translate between languages without explicit parallel data for those languages, relying on their ability to generalize across language pairs by identifying shared linguistic features. For instance, a model trained on an English-Luganda dataset can effectively translate between English and Lusoga, despite the absence of direct English-Lusoga training examples [31], [64]. Key techniques that enhance ZST include language-agnostic representations, which capture universal linguistic features, allowing for effective translation across diverse languages; however, they may struggle with distant language families [65]. Cross-lingual Consistency Regularization (CrossConST) further refines this process by enforcing prediction consistency across languages, leveraging Kullback-Leibler (KL) regularization to minimize discrepancies, although its effectiveness can diminish in low-resource contexts with limited data [66], [67]. Agreement-based training ensures coherence in multilingual predictions, but performance may wane with highly divergent syntactic structures [68]. Techniques like auxiliary training objectives and model architecture modifications facilitate the learning of more universal representations, providing robustness but potentially requiring substantial computational resources [69]. Language tag strategies enhance model understanding of distinct languages, improving translation accuracy; however, their efficacy may decline with less clear language distinctions [70]. While these approaches have demonstrated significant advancements in ZST, challenges remain for certain languages or domains, particularly those that are morphologically complex or distant, emphasizing the need for additional fine-tuning or supplemental data. Despite these limitations, ZST offers practical solutions for addressing data scarcity and improving translation quality for low-resource languages.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24OCT410

### C. The Transformative Impact of Machine Translation

Machine translation plays a transformative role in bridging language barriers across key sectors, significantly contributing to education, healthcare, business, language preservation, cultural continuity, and intergenerational knowledge transfer [71]. In education, Machine translation facilitates the conversion of learning materials into indigenous languages, promoting inclusivity and reducing language barriers. This enhancement improves comprehension and academic performance, particularly in rural areas where limited English proficiency can hinder student progress [72]. In the healthcare sector, MT ensures clear communication between medical professionals and patients from diverse linguistic backgrounds, which enhances healthcare delivery and outcomes [73]. In business, MT fosters cross-lingual communication, enabling companies to enter new markets and engage more effectively with local customers, thereby driving economic growth and increasing market inclusivity, especially for speakers of underrepresented languages [74]. Moreover, limited English proficiency often restricts access to job opportunities and business interactions, reinforcing social inequality. By promoting language learning, MT expands access to these opportunities and helps reduce disparities [75].

MT also plays a crucial role in preserving and revitalizing African native languages, many of which face significant threats [76]. Additionally, MT supports cultural preservation by documenting and sharing stories, proverbs, and songs, thereby maintaining the cultural identities of diverse linguistic communities [77]. Importantly, MT facilitates intergenerational knowledge transfer by bridging the language gap between older custodians of traditional knowledge and younger generations, ensuring that valuable cultural and historical knowledge is preserved and passed down [78]. Thus, in Africa, MT not only enhances communication but also plays a critical role in preserving cultural heritage and promoting linguistic diversity.

## VII. FUTURE DIRECTIONS

### A. Improving Corpus Quality

Improving the quality of English-Lusoga corpora, with particular attention to dialects like Lulamoogi and Lusiki (which together form the Lupakoyo dialect), can be achieved through a blend of crowdsourcing, community collaboration, and linguistic research [8], [79]. Crowdsourcing platforms, whether through online platforms or translation community challenges, can engage native speakers to contribute translations and speech data. These initiatives, supported by incentives, would help gather a wide range of dialect-specific examples. Complementing this, workshops and fieldwork in Lusoga-speaking regions would allow researchers to collect natural spoken data, folklore, and oral history, while fostering local involvement in the preservation and expansion of dialectal corpora [80].

### B. Expanding to Other Bantu Languages

Uganda is home to many native languages, most of which are not yet integrated into machine translation systems. To bridge this gap, efforts should focus on building robust language corpora and exploring multilingual MT techniques [30]. For instance, linguistic similarities among Eastern Uganda Bantu languages like Lusoga, Lunyole, and Lugwere can be leveraged to develop multilingual MT models, enabling effective communication between speakers of these languages and English. However, the lack of well-developed corpora remains a major limitation, highlighting the need for dedicated corpus-building initiatives. Without comprehensive datasets, the full potential of MT to support these underrepresented languages cannot be realized. The English-Lusoga NLU-MT model could be fine-tuned using corpora from related languages, extending its translation capabilities to include those languages as well.

## VIII. CONCLUSION

This review underscores significant advancements in Natural Language Understanding-driven Machine Translation for low-resource Bantu languages like Lusoga. Leveraging linguistic similarities between languages such as Lusoga and Luganda has proven effective in improving MT performance. Multilingual transfer learning model architectures, particularly mBART and mT5, have enhanced translation accuracy for African languages, even with limited data [30]. However, the historically low demand and scarcity of high-quality dialectal parallel corpora for languages like Lusoga remain significant obstacles, limiting scalability for both the languages and their dialects [81]. Despite these challenges, expanding community-driven corpus-building efforts and refining model architectures hold great potential for advancing low-resource machine translation, which in turn could bridge language barriers across sectors like education, healthcare, business, and cultural preservation.

## REFERENCES

[1]. A. B. Olani, A. B. Olani, T. B. Muleta, and D. H. Rikitu, “Impacts of language barriers on healthcare access and quality among Afaan Oromoo - speaking patients in Addis Ababa ,” BMC Health Serv. Res., pp. 1–12, 2023, doi: 10.1186/s12913-023-09036-z.

[2]. C. W. Gulere, “Standardised Language-Based Orthographies,” in LITERACY: A BRIDGE TO EQUITY, 2019.

[3]. J. Dong, “Transfer Learning-Based Neural Machine Translation for Low-Resource Languages,” ACM Trans. Asian Low-Resource Lang. Inf. Process., 2023, doi: 10.1145/3618111.

[4]. K. Ogueji, Y. Zhu, and J. Lin, “Small Data ? No Problem ! Exploring the Viability of Pretrained Multilingual Language Models for Low-Resource Languages,” in Proceedings ofthe 1st Workshop on Multilingual Representation Learning, 2021, pp. 116–126.

[5]. A. Nzeyimana, “Low-resource neural machine translation with morphological modeling,” Find. Assoc. Comput. Linguist. NAACL 2024 - Find., pp. 182–195, 2024.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24OCT410

[6]. A. Hernández, R. M. Ortega-mendoza, E. Villatoro-tello, C. J. Camacho-bello, and O. Pérez-cortés, "Natural Language Understanding for Navigation of Service Robots in Low-Resource Domains and Languages : Scenarios in Spanish and Nahuatl," Mathematics, vol. 12, no. 8, 2024, doi: https://doi.org/10.3390/math12081136.
[7]. S. Ghosh, "Natural Language Processing: Basics, Challenges, and Clustering Applications," in A Handbook of Computational Linguistics: Artificial Intelligence in Natural Language Processing, 2024. doi: http://dx.doi.org/10.2174/9789815238488124020006.
[8]. M. M. I. Alam, S. Ahmadi, and A. Anastasopoulos, "CODET: A Benchmark for Contrastive Dialectal Evaluation of Machine Translation," EACL 2024 - 18th Conf. Eur. Chapter Assoc. Comput. Linguist. Find. EACL 2024, pp. 1790–1859, 2024.
[9]. P. Prasada, M. Vishwanatha, and P. Rao, "Reinforcement of low-resource language translation with neural machine translation and backtranslation synergies," Int. J. Artif. Intell., vol. 13, no. 3, pp. 3478–3488, 2024, doi: 10.11591/ijai.v13.i3.pp3478-3488.
[10]. T. Sumanth, "Deep Learning for Natural Language Processing," Int. J. Adv. Res. Eng. Technol., vol. 190, no. 5, pp. 523–533, 2021, doi: 10.1007/978-981-16-0882-7_45.
[11]. C. C. Emezue and B. F. P. Dossou, "MMTAfrica: Multilingual Machine Translation for African Languages," WMT 2021 - 6th Conf. Mach. Transl. Proc., pp. 398–411, 2021.
[12]. B. Akera et al., "Machine translation for african languages: commutity creation of datasets and models in Uganda," 3rd Work. African Nat. Lang. Process. 2022., no. 61733011, pp. 1–13, 2022, [Online]. Available: https://openreview.net/forum?id=BK-z5qzEU-9
[13]. A. Vaswani, "Attention Is All You Need," in Conference on Neural Information Processing Systems (NIPS 2017), 2017.
[14]. Y. Zhao, J. Zhang, and C. Zong, "Transformer: A General Framework from Machine Translation to Others," Mach. Intell. Res., vol. 20, no. 4, pp. 514–538, 2023, doi: 10.1007/s11633-022-1393-5.
[15]. M. C. Kenton, L. Kristina, and J. Devlin, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2019.
[16]. M. Rahman, A. I. Shiplu, Y. Watanobe, and A. Alam, "RoBERTa-BiLSTM : A Context-Aware Hybrid Model for Sentiment Analysis," 2021. [Online]. Available: https://arxiv.org/pdf/2406.00367
[17]. M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, and O. Levy, "SpanBERT: Improving Pre-training by Representing and Predicting Spans," 2020.
[18]. D. Jurafsky and C. D. Manning, "Mini But Mighty : Efficient Multilingual Pretraining with Linguistically-Informed Data Selection," in Findings of the Association for Computational Linguistics: EACL 2023, 2023, pp. 1251–1266.
[19]. N. R. Robinson, P. Ogayo, D. R. Mortensen, and G. Neubig, "ChatGPT MT : Competitive for High- ( but not Low- ) Resource Languages," 2023. [Online]. Available: https://aclanthology.org/2023.wmt-1.40.pdf
[20]. T. B. Brown et al., "Language Models are Few-Shot Learners," 2022.
[21]. Y. Tang et al., "Multilingual Translation with Extensible Multilingual Pretraining and Finetuning," 2020.
[22]. C. Raffel et al., "Exploring the limits of transfer learning with a unified text-to-text transformer," J. Mach. Learn. Res., vol. 21, pp. 1–67, 2020.
[23]. A. Oladipo et al., "Better Quality Pretraining Data and T5 Models for African Languages," 2023. [Online]. Available: https://aclanthology.org/2023.emnlp-main.11.pdf
[24]. R. Kimera, D. N. Rim, and H. Choi, "Building a Parallel Corpus and Training Translation Models Between Luganda and English," J. KIISE, vol. 49, no. 11, pp. 1009–1016, 2022, doi: 10.5626/jok.2022.49.11.1009.
[25]. M. Nabirye, G. M. De Schryver, and J. Verhoeven, "Lusoga (Lutenga)," J. Int. Phon. Assoc., vol. 46, no. 2, pp. 219–228, 2016, doi: 10.1017/S0025100315000249.
[26]. F. Śanchez-Martinez et al., "An English-Swahili parallel corpus and its use for neural machine translation in the news domain," Proc. 22nd Annu. Conf. Eur. Assoc. Mach. Transl. EAMT 2020, pp. 299–308, 2020.
[27]. A. L. Tonja, O. Kolesnikova, A. Gelbukh, and G. Sidorov, "Low-Resource Neural Machine Translation Improvement Using Source-Side Monolingual Data," Appl. Sci., vol. 13, no. 2, 2023, doi: 10.3390/app13021201.
[28]. H. Wan-hua and U. Kruschwitz, "Investigating Neural Machine Translation for Low-Resource Languages : Using Bavarian as a Case Study," 2024. [Online]. Available: https://aclanthology.org/2024.sigul-1.20.pdf
[29]. T. Ngo, P. Nguyen, V. V. Nguyen, T. Ha, and L. Nguyen, "An Efficient Method for Generating Synthetic Data for Low-Resource Machine Translation An Efficient Method for Generating Synthetic Data for Low-Resource Machine Translation," Appl. Artif. Intell., vol. 36, no. 1, 2022, doi: 10.1080/08839514.2022.2101755.
[30]. E. S. A. Lee et al., "Pre-Trained Multilingual Sequence-to-Sequence Models: A Hope for Low-Resource Language Translation?," in Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2022, pp. 58–67. doi: 10.18653/v1/2022.findings-acl.6.
[31]. E. Nyoni and B. A. Bassett, "Low-Resource Neural Machine Translation for Southern African Languages," arXiv:2104.00366. Accessed: Apr. 09, 2023. [Online]. Available: http://arxiv.org/abs/2104.00366
[32]. F. N. Al-Wesabi, H. J. Alshahrani, A. E. Osman, and E. S. Abd Elhameed, "Low-Resource Language Processing Using Improved Deep Learning with Hunter–Prey Optimization Algorithm," Mathematics, vol. 11, no. 21, 2023, doi: 10.3390/math11214493.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24OCT410

[33]. C. Gulere, An Introduction to Lusoga Orthography, 2nd ed. Mpolyabigere RC – RICED Center Ltd. Plot, 2012. [Online]. Available: https://shorturl.at/3lJl8
[34]. M. Nabirye, "Compiling the first monolingual Lusoga dictionary," Lexikos, vol. 19, pp. 177–196, 2009, doi: 10.4314/lex.v19i1.49125.
[35]. M. Nabirye and G. M. De Schryver, "Digitizing the Monolingual lusoga dictionary: Challenges and prospects," Lexikos, vol. 23, pp. 297–322, 2013, doi: 10.5788/23-1-1217.
[36]. G. M. De Schryver and M. Nabirye, "Corpus-driven Bantu Lexicography Part 2: Lemmatisation and rulers for Lusoga," Lexikos, vol. 28, pp. 79–111, 2018, doi: 10.5788/28-1-1458.
[37]. G. M. De Schryver and M. Nabirye, "Corpus-driven Bantu Lexicography Part 3: Mapping meaning onto use in Lusoga," Lexikos, vol. 28, pp. 112–151, 2018, doi: 10.5788/28-1-1459.
[38]. M. R. Marlo, M. Nabirye, and G. M. de Schryver, "Reduplication in Lusoga," Africana Linguist., vol. 28, pp. 147–197, 2022.
[39]. G. M. De Schryver and M. Nabirye, "A quantitative analysis of the morphology, morphophonology and semantic import of the Lusoga noun," Africana Linguist., vol. 16, pp. 97–153, 2010, doi: 10.3406/aflin.2010.989.
[40]. J. Dodge et al., "Documenting Large Webtext Corpora : A Case Study on the Colossal Clean Crawled Corpus," 2020.
[41]. L. Gao, S. Biderman, and S. Black, "The Pile: An 800GB Dataset of Diverse Text for Language Modeling," 2020.
[42]. N. Aspert, V. Miz, B. Ricaud, P. Vandergheynst, and I. R. Mar, "A graph-structured dataset for Wikipedia research," 2013.
[43]. B. Thompson, "subs2vec : Word embeddings from subtitles in 55 languages," Behav. Res. Methods, vol. 53, pp. 629–655, 2021.
[44]. S. Mesham, L. Hayward, J. Shapiro, and J. Buys, "Low-Resource Language Modelling of South African Languages," 2021. [Online]. Available: http://arxiv.org/abs/2104.00772
[45]. N.-N. Joyce et al., "Applied AI Letters - 2024 - Nakatumba-Nabende - Building Text and Speech Benchmark Datasets and Models for Low-Resourced.pdf," Appl. AI Lett., p. 18, 2024.
[46]. P. Nabende, "Towards Data-Driven Machine Translation for Lumasaaba BT - Digital Science," in The 2018 International Conference on Digital Science, T. Antipova and A. Rocha, Eds., Cham: Springer International Publishing, 2019, pp. 3–11.
[47]. J. Omona and N. Groce, "Translation and research outcomes of the Bridging the Gap project: A case of the Luo language, spoken in northern Uganda," Transl. Stud., vol. 14, no. 3, pp. 282–297, 2021, doi: 10.1080/14781700.2021.1888784.
[48]. E. P. Wairagala, J. Mukiibi, J. F. Tusubira, C. Babirye, and J. Nakatumba-Nabende, "Gender Bias Evaluation in Luganda-English Machine Translation," Zenodo. [Online]. Available: https://zenodo.org/records/5864560
[49]. M. Mager, E. Mager, K. Kann, and N. T. Vu, "Ethical Considerations for Machine Translation of Indigenous Languages: Giving a Voice to the Speakers," in Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2023, pp. 4871–4897. doi: 10.18653/v1/2023.acl-long.268.
[50]. Z. Zhou, J. Chen, N. Wang, L. Li, and D. Wang, "Adversarial Data Augmentation for Robust Speaker Verification," in Proceedings of the 2023 9th International Conference on Communication and Information Processing, in ICCIP '23. New York, NY, USA: Association for Computing Machinery, 2024, pp. 226–230. doi: 10.1145/3638884.3638917.
[51]. E. M. Provost, "Best Practices for Noise-Based Augmentation to Improve the Performance of Deployable Speech-Based Emotion Recognition Systems," 2023.
[52]. L. Pandey et al., "Towards scalable efficient on-device ASR with transfer learning," 2024.
[53]. S. Cahyawijaya et al., "NusaWrites : Constructing High-Quality Corpora for Underrepresented and Extremely Low-Resource Languages," vol. 1, pp. 921–945, 2023.
[54]. T. Glushkova, C. Zerva, and A. F. T. Martins, "BLEU Meets COMET: Combining Lexical and Neural Metrics Towards Robust Machine Translation Evaluation," in Proceedings of the 24th Annual Conference of the European Association for Machine Translation, EAMT 2023, 2023, pp. 47–58.
[55]. S. M. Lakew, A. Erofeeva, M. Federico, F. B. Kessler, and M. M. T. Srl, "Neural Machine Translation into Language Varieties," in Proceedings ofthe Third Conference on Machine Translation (WMT), 2018, pp. 156–164.
[56]. P. Helm, G. Bella, G. Koch, and F. Giunchiglia, "Diversity and language technology : how language modeling bias causes epistemic injustice," Ethics Inf. Technol., vol. 26, no. 1, pp. 1–15, 2024, doi: 10.1007/s10676-023-09742-6.
[57]. T. Kunz and T. Gummer, "Understanding Respondents' Attitudes Toward Web Paradata Use," Soc. Sci. Comput. Rev., vol. 38, no. 6, pp. 739–753, Feb. 2019, doi: 10.1177/0894439319826904.
[58]. M. Nurminen and M. Koponen, "Machine translation and fair access to information," Transl. Spaces, vol. 9, no. 1, pp. 150–169, 2020, doi: 10.1075/ts.00025.nur.
[59]. T. Kocmi and O. Bojar, "Trivial Transfer Learning for Low-Resource Neural Machine Translation," in Proceedings ofthe Third Conference on Machine Translation (WMT), 2018, pp. 244–252.
[60]. S. M. Lakew, A. Erofeeva, M. Negri, M. Federico, and M. Turchi, "Transfer Learning in Multilingual Neural Machine Translation with Dynamic Vocabulary Transfer Learning in Multilingual Neural Machine Translation with Dynamic Vocabulary," in International Workshop on Spoken Language Translation, 2018.
[61]. C.-K. Wu, C.-C. Shih, Y.-C. Wang, and R. T.-H. Tsai, "Improving low-resource machine transliteration by using 3-way transfer learning," Comput. Speech Lang., vol. 72, p. 101283, 2022, doi: https://doi.org/10.1016/j.csl.2021.101283.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24OCT410

[62]. R. Dabre and A. Fujita, “Exploiting Multilingualism through Multistage Fine-Tuning for Low-Resource Neural Machine Translation,” in Conference on Empirical Methods in Natural Language Processing, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:208163390
[63]. X. Han et al., “Pre-trained models : Past , present and future,” AI Open, vol. 2, no. August 2021, pp. 225–250, 2023, doi: 10.1016/j.aiopen.2021.08.002.
[64]. P. Gao, L. Zhang, Z. He, H. Wu, and H. Wang, “Improving Zero-shot Multilingual Neural Machine Translation by Leveraging Cross-lingual Consistency Regularization,” Find. ofthe Assoc. Comput. Linguist., no. 10, pp. 12103–12119, 2023.
[65]. X. Chen and C. Zhang, “Language-agnostic Zero-Shot Machine Translation with Language-specific Modeling,” in 2024 International Joint Conference on Neural Networks (IJCNN), 2024, pp. 1–8. doi: 10.1109/IJCNN60899.2024.10649983.
[66]. B. Zheng et al., “Consistency Regularization for Cross-Lingual Fine-Tuning,” 2020.
[67]. N. Vieillard et al., “Leverage the Average : an Analysis of KL Regularization in Reinforcement Learning,” in 34th Conference on Neural Information Processing Systems (NeurIPS 2020), 2020.
[68]. S. Gu and Y. Feng, “Improving Zero-Shot Multilingual Translation with Universal Representations and Cross-Mappings,” in Findings ofthe Association for Computational Linguistics, 2022, pp. 6492–6504.
[69]. D. Liu, J. Niehues, J. Cross, and F. Guzm, “Improving Zero-Shot Translation by Disentangling Positional Information,” in Proceedings ofthe 59th Annual Meeting ofthe Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021, pp. 1259–1273.
[70]. Z. Sun, Y. Liu, F. Meng, J. Xu, Y. Chen, and J. Zhou, “LCS: A Language Converter Strategy for Zero-Shot Neural Machine Translation,” 2021.
[71]. R. Taye et al., “Language As A Barrier In Health Care Communication-A Comparative Study On Rural And Urban hospitals,” J. Pharm. Negat. Results ¦, vol. 14, no. February, p. 2023, 2023, doi: 10.47750/pnr.2023.14.S02.271.
[72]. M. Phiri, C. C. Thelma, and N. H. Mwanapabu, “The Effect of Using Local Languages as A Medium of Instruction on Academic Performance of Learners : A Case of Selected Primary Schools in Solwezi District of North- Western Province , Zambia,” Int. J. Nov. Res. Humanit. Soc. Sci., vol. 11, no. 3, 2024, doi: 10.5281/zenodo.11178057.
[73]. F. Thonon et al., “Electronic tools to bridge the language gap in health care for people who have migrated: Systematic review,” J. Med. Internet Res., vol. 23, no. 5, pp. 1–14, 2021, doi: 10.2196/25131.
[74]. H. Gao, “Research on Automatic Business English Text Translation Technology Based on Intelligent Computing,” Appl. Math. Nonlinear Sci., vol. 9, no. 1, pp. 1–15, 2024.
[75]. S. Mumtaz, S. P. Chandio, and D. A. K. Malokani, “The correlation between English language proficiency and perceived career opportunities. Empirical Analysis,” Remit. Rev., vol. 8, no. 4, pp. 4818–4827, 2023, doi: 10.33182/rr.v8i4.310.
[76]. R. Mlambo and M. Matfunjwa, “The use of technology to preserve indigenous languages of South Africa,” J. Lit. Crit. Comp. Linguist. Lit. Stud., no. Etim, pp. 1–8, 2024.
[77]. I. Jibreel, “Online Machine Translation Efficiency in Translating Fixed Expressions Between English and Arabic ( Proverbs as a Case-in-Point ),” Theory Pract. Lang. Stud., vol. 13, no. 5, pp. 1148–1158, 2023.
[78]. N. Rupčić, “Intergenerational Learning and Knowledge Transfer BT - Managing Learning Enterprises: Challenges, Controversies and Opportunities,” N. Rupčić, Ed., Cham: Springer Nature Switzerland, 2024, pp. 201–211. doi: 10.1007/978-3-031-57704-8_13.
[79]. S. Nisioi, A. S. Uban, and L. P. Dinu, “Identifying Source-Language Dialects in Translation,” Mathematics, vol. 10, no. 9, 2022, doi: 10.3390/math10091431.
[80]. S. Luger, M. Leventhal, C. M. Homan, M. Zampieri, and M. Zampieri, “Towards a Crowdsourcing Platform for Low Resource Languages – A Semi-Supervised Approach,” in Conference on Human Computation and Crowdsourcing (HCOMP), 2020, pp. 1–3.
[81]. W. Nekoto et al., “Participatory Research for Low-resourced Machine Translation : A Case Study in African Languages,” in Findings of the Association for Computational Linguistics: EMNLP 2020, 2020, pp. 2144–2160.