A neural machine translation system for Kreol Repiblik Moris and English

Sameerchand Pudaruth[1], Sheeba Armoogum[1], Nirmal Kumar Betchoo[2], Aneerav Sukhoo[3], Vandanah Gooria[4], Abdallah Peerally[1], Mohammad Zafar Khodabocus[1]

[1]Department of Information and Communication Technology, Faculty of Information, Communication and Digital Technologies, University of Mauritius, Moka, Mauritius
[2]Department of Management, Faculty of Business and Management, Université des Mascareignes, Beau Bassin-Rose Hill, Mauritius
[3]Central Information Systems Division, Ministry of Information Technology, Communication and Innovation, Port Louis, Mauritius
[4]Academic Affairs Division, Open University of Mauritius, Reduit, Mauritius

Article Info

Article history:

Received May 8, 2024
Revised Jun 22, 2024
Accepted Jun 28, 2024

Keywords:

Deep learning
Kreol Repiblik Moris
Machine translation
Mauritius
Transformer model

ABSTRACT

Although Google Translate is a widely used machine translation service that supports 133 languages, it does not incorporate support for the Kreol Repiblik Moris (KRM) language. Addressing this limitation, the current research focuses on enhancing the accuracy and fluency of machine translation between KRM and English through natural language processing and deep neural machine translation techniques. In this study, a machine translation system using a transformer model trained with a dataset of 50,000 parallel corpora has been developed. The model was evaluated using manual translations and the bilingual evaluation understudy (BLEU) score. A score of 31.46 for translating from KRM to English and 28.15 for translating from English to KRM was achieved. To our knowledge, these are the highest BLEU scores for translation between these two languages. This is due to utilising the largest dataset and extensive atomic words from the KRM dictionary. This successful interdisciplinary funded project led to the setting up of a free online translation service and a smartphone app for Mauritian citizens and tourists.

Corresponding Author:

Sameerchand Pudaruth
Department of Information and Communication Technology
Faculty of Information, Communication and Digital Technologies, University of Mauritius
Moka, Mauritius
Email: s.pudaruth@uom.ac.mu

1. INTRODUCTION

As per Statistics Mauritius [1], Creole is spoken by 90% of the Mauritian population. The Creole language is formally known as Kreol Morisien (KM) or as Kreol Repiblik Moris (KRM). In 2012, the government of Mauritius included KM as one of the subjects in primary education, and in 2017, 4,000 students took this language for their primary school achievement certificate (PSAC). A significant proportion of the Mauritian population is not proficient in English which leads to difficulties in understanding English used in public communication and the media. This limitation also hinders their interaction with tourists and other foreigners. Despite the availability of numerous online translation services, they have yet to address translating between KRM and English effectively.

Advocacy for using KRM as a national language has been advanced by Dev Virahsawmy, an author, poet, and political figure [2]. Virahsawmy has authored numerous texts and poems in Kreol Morisien. The development of a harmonised writing system called Grafi-larmoni aimed at establishing a standardised form for writing the language. Furthermore, Carpooran introduced an updated Kreol Morisien dictionary in 2011, and subsequent versions have included new words [3]. These efforts towards standardisation and linguistic development in KRM have contributed to its recognition and inclusion in education.

KRM or KM has a sentence structure mainly influenced by its historical formation and evolution from a pidgin language during the French colonisation period to an established one. Kreol Morisien's structure is shaped by its parent languages, mainly French and the languages of the African slaves and indentured labourers from India who contributed to it is formation [4]. However, many English words have also found their way into the language. The sentence structure of Kreol Morisien closely resembles English but with some distinctive variations. However, unlike English, Kreol Morisien does not have plural forms for words and frequently positions adjectives behind the object. The translation process often entails omitting state-of-being verbs. For instance, "He is good at dancing" would be translated as "Li bon dan danse", where "he" becomes "Li", and "good at dancing" becomes "bon dan danse". The verb "is" is omitted in the translation. By understanding the distinct sentence structure of Kreol Morisien, educators can effectively teach and include the language in education. The following are the characteristics of the sentence structure: subject-word-object (SVO) order, pre-verbal markers for tense and aspect, negation and question formation.

Like English and French, Kreol Morisien usually uses the subject-verb-object word order. The subject is placed first, followed by the verb and then the object that receives the action. This sentence format facilitates clear communication and effortless understanding in Kreol Morisien, thus facilitating effective instruction and integration into the education system. In addition to the SVO word order, Kreol Morisien exhibits distinctive sentence structure characteristics.

Pre-verbal markers in Kreol Morisien express tense and aspect within sentences. These tags precede the verb and offer vital details about the timing of the action or its duration. Tense, mood, and aspect markers in Kreol Morisien indicate an action's time, perspective, and duration. For instance: "Mo pe manze." (I am eating)-"pe" signifies a continuous aspect. "Li ti pe ale." (He was leaving.)-"ti" is an expression of the past tense. "Zot pou vini"-"pou" denotes the future tense. Kreol Morisien commonly uses the marker "pa" before the verb to express negation. For instance, "Li pa pe danse." translates to "He/She is not dancing."

The formation of questions in Kreol Morisien differs between English and French. It often involves subject-verb inversion or the use of question words. Yes/no questions can frequently be formed by intonation alone, raising the pitch of the voice towards the end of the sentence. Wh-questions use question words like "kifer" (why), "ki" (what), and "kot" (where), which are placed at the beginning of the sentence. Kreol Morisien's sentence structure is straightforward and practical, allowing speakers to communicate intricate concepts within a versatile and minimalist grammatical system. The research demonstrates that Kreol Morisien displays a unique sentence structure distinguished by it is subject-verb-object word order, pre-verbal markers for tense, aspect and negation, and distinct methods for forming questions.

Recent advancements in machine translation (MT) technology have been substantial, but its ability to translate Kreol Morisien is constrained by the language's distinct sentence structure, grammar and lack of resources. Thus, it is vital to enhance the development of machine translation systems that can effectively manage the unique characteristics and subtleties present in Kreol Morisien to attain improved precision in translations. MT is a branch of computational linguistics that examines the utilisation of software for converting text or speech from one language to another. Literal machine translation involves replacing words in one language with equivalent words in another. However, more is needed to generate an accurate translation. It is essential to recognise entire phrases and find their most appropriate equivalents in the target language. Not all terms in a particular language have direct counterparts in another language, and numerous words carry multiple interpretations. While the principles of machine translation may seem straightforward, the underlying science and technologies involved are highly intricate. Since the beginning of the 2010s, a new form of artificial intelligence technology known as deep neural networks has provided speech recognition technologies and machine translation systems with the capability to achieve a satisfactory level of quality [5]. This research fully supports more fluid communication across borders and different cultures.

The rule-based machine translation (RBMT) model encompassed transfer-based, inter-lingual, and dictionary-based machine translation approaches. This form of interpretation was primarily utilised in developing lexicons and linguistic software. RBMT involves more information about the linguistics of the source and target languages. The fundamental strategy connects the input sentence's format with the output sentence's structure. The RBMT approach faced limitations such as inadequate high-quality dictionaries, manual setup of linguistic information, and challenges with rule interactions in complex systems such as ambiguity and idiomatic expressions [6].

Statistical machine translation (SMT) is the automated transformation of sentences from one human language to another using statistics and probability. The original language is termed the "source," while the secondary language is labelled the "target". This procedure can be conceptualised as a probabilistic process. Williams *et al.* [7] suggest that various SMT variants exist, depending on their approach to translation modelling. These approaches include string-to-string mapping, tree-to-strings, and tree-to-tree models. However, they all share the familiar principle that translation is automated and involves models derived from parallel corpora (source-target pairs) and monolingual corpora (target sentence examples). SMT faces several significant obstacles: i) generating and training parallel data is expensive and time-consuming, ii) it necessitates large volumes of parallel data comprising at least 2 million words, iii) anticipating and rectifying specific translation errors can be challenging, and iv) SMT is not well-suited for language pairs with differing word orders.

Neural machine translation (NMT) represents a significant shift from traditional approaches to machine translation, as it utilises continuous representations rather than discrete symbolic representations employed in SMT. Conversely, NMT employs a single extensive neural network to represent the entire translation process, eliminating the necessity for abundant feature engineering [8]. NMT employs deep neural networks to map the source and target languages directly, significantly advancing translation performance. It has now become the dominant paradigm of machine translation [9]. Ragni and Vieira [10] express particular concerns regarding NMT. First, it is important to delve more deeply into the NMT editing process and its specific aspects from the perspective of translators. Second, there should be a stronger emphasis on highlighting the usefulness of NMT as a tool for professionals. Third, there has been a tendency to narrowly conceptualise translation productivity based solely on processing time or throughput measures. Lastly, it has become apparent that investigations related to NMT involving end-users are still relatively rare.

The initial success of Morisia 1.0 [11], a multi-disciplinary project funded by the higher education commission (HEC) of Mauritius, led to the development of Morisia 2.0. The primary focus initially was on establishing a platform that would incorporate specific parallel corpora. The current study builds upon prior research by significantly expanding the corpus to enhance user translation efficiency. In this context, there is a clear requirement for extensive collections of parallel texts, or parallel corpora, to support various linguistic studies that rely on sentence-level alignments within these corpora. When translating, the translator has the flexibility to split, combine, remove, add or rearrange sentences. Alignment is a significant undertaking. This paper addresses the efficient translation between KRM and English and vice-versa, aiming to ensure user satisfaction and successful message interpretation through a high-quality translation system.

The research project aims to develop an automated translation system that seamlessly translates between English and KRM using deep neural networks. The project includes the development of a web portal translator with support from a mobile app. It seeks to achieve several objectives, including building a parallel corpus of 50,000 sentences to improve translation accuracy between KM/KRM and English, assessing the system’s performance using metrics such as the bilingual evaluation understudy (BLEU) score, designing a web portal app with accessibility features tailored for individuals with disabilities (e.g., larger font size and colour differentiation), as well as incorporating voice-to-text and text-to-voice functionality in the mobile app specifically for English language translations. The ultimate goal of this project is to provide a user-friendly and efficient translation system that bridges the language gap between English and Kreol Morisien to facilitate effective communication and promote cultural exchange between the users of these two languages.

After presenting the topic and highlighting the challenges associated with the development of Morisia 2.0, section 2 examines the relevant literature on machine translation and emphasises the importance of utilising metrics, such as BLEU, to evaluate translation quality. Section 3 delves into the technical aspects by discussing the methodology for machine translation in this project, followed by a detailed exploration of experimental design. Section 4 outlines the results and evaluation of the system, while section 5 presents the concluding part of this research.

2. LITERATURE REVIEW

This section aims to comprehensively understand the existing research and establish a foundation for developing an automated translation system for English and Kreol Morisien. It provides an in-depth examination of the BLEU score, metric for evaluation of translation with explicit ordering (METEOR) Score, and the NST method, along with a comparison. The researchers have also explored prior research on machine translation between English and Kreol Morisien to identify deficiencies and potential areas for enhancement.

Nath *et al.* [12] define machine translation as a computational process for translating a given set of words from one human-readable language into another. Machine translation models can fall into three categories: rule-based, statistical-based, or neural network-based. The emergence of the NMT model was driven by the constraints encountered with rule-based and statistical-based machine translation models. Neural network-based machine translation models have demonstrated potential for various human languages, featuring an extensive vocabulary acquired from a substantial dataset [12].

In the beginning, machine translation relied heavily on RBMT to formulate grammatical rules for both the source and target languages [13]. SMT was thus developed to address this issue [14]. A statistical model was developed by analysing a matched collection of sentences in the source and target languages (training set), which was then utilised to generate a translation. NMT is the most recent system, which considers the entire sentence and can identify connections between phrases even when located further apart. This leads to enhanced grammatical precision in comparison to SMT.

2.1. BLEU scores

Phan-Vu *et al.* [15] noted a transition in machine translation towards an end-to-end strategy utilising deep neural networks. Significant advancements have been made in the state of the art for widely spoken language pairs like English-French or English-Chinese. Their study focused on enhancing English-Vietnamese translations: i) constructing the most extensive open Vietnamese-English corpus and ii) conducting comprehensive trials using state-of-the-art neural models to attain the highest BLEU scores. Krüger [16] explored the cognitive linguistic viewpoint to examine how human translation could be modelled in terms of context and meaning. This research demonstrated how NMT interprets linguistic meaning and to what degree it incorporates contextual information into this process.

Batsukh [17] examined the advantages and disadvantages of contemporary neural machine translation. While NMT provided a more straightforward approach to modelling, leading to effective word and sentence structure implementation, it sometimes led to distorted sentence structures and boundaries when translating untrained data. Zhang *et al.* [18] suggested that despite the impressive performance of NMT, it needs help to accurately capture the alignment between the inputs and the outputs during the translation process. This lack of alignment gives rise to three challenges: interpreting the translation process, imposing lexical constraints, and applying structural constraints. These issues complicate the development of new NMT architectures and restrict their practical applications.

Emna *et al.* [19] studied the importance of comprehending and processing unclearly articulated speech. They developed a NMT system for translating the tunisian dialect (TD) to modern standard Arabic (MSA). This type of NMT task faced challenges due to limited training data available for low-resource languages such as TD. By building a parallel corpus of TD-MSA and effectively utilising it, they formulated a setup for a neural translation model that achieved an impressive BLEU score of 67.56%.

2.2. The purpose of the BLEU algorithm

Adlaon and Marcos [20] sought to create a parallel corpus as an essential tool in machine learning-based translation by employing a recurrent neural network (RNN) within the OpenNMT framework. The quality of the translation was assessed using the BLEU score. A subword unit translation was conducted to rectify inconsistencies in the original dataset, leading to an improved BLEU score of 22.87 compared to the initial 20.01. Villanueva *et al.* [21] investigated the intricacies of converting the traditional Philippine language to English. They proposed a mobile-oriented translation system equipped with object detection to aid travellers. The system utilised NMT to convert Filipino to Cebuano language and vice versa, drawing input from the user’s keyboard and extracting text strings from identified objects in images. These studies highlight the challenges faced in neural machine translation, such as capturing alignment between input and output, dealing with limited training data for low-resource languages, and improving the quality of translation through techniques like subword unit translation and object detection. The model for Filipino-Cebuano achieved a BLEU score of 31.1, while the Cebuano-Filipino model scored 31.6. The BLEU score measures the precision of matching word sequences between a “candidate” machine translation and one or more “reference” human translations [22]. The algorithm was created to compare sentences within a corpus, calculating n-gram matches at the sentence level and then aggregating them into an overall score for the corpus. Using the BLEU algorithm in these studies demonstrates its effectiveness in evaluating the quality and accuracy of machine translations.

2.3. Empirical review of translation quality with BLEU scores

An empirical evaluation compared the translation quality of different available systems for end users in Thailand to gain insight into the overall quality of translations used [23]. The difficulty of translating Thai to English is evident from the high error rate of 47.2% and a low BLEU score of 21. However, despite the high translation error rate, users correctly answered approximately 60% of questions in reading comprehension tests using output from machine translation systems. Papineni *et al.* [22] previously claimed that BLEU could speed up machine translation by enabling researchers to focus on practical modelling ideas quickly. The results of BLEU were closely associated with human evaluations as it averaged out individual sentence judgment errors across a test corpus rather than striving for precise human judgment for each sentence, demonstrating that quantity can lead to quality. Using BLEU scores to evaluate machine translations has proven effective and reliable. Therefore, it is appropriate to consider the BLEU score as a viable metric for assessing the quality and accuracy of machine translations.

2.4. Comparison between BLEU and METEOR

Banerjee and Lavie [24] suggested that the static brevity penalty in BLEU did not sufficiently address the recall issue. They proposed that a direct assessment of grammaticality (or word order) could more effectively capture the significance of grammaticality as a component in the machine translation metric, leading to improved alignment with human evaluations of translation quality. Consequently, sentence or segment-level BLEU scores may not be considered relevant. METEOR was designed to rectify the shortcomings of BLEU by assessing a translation by calculating a score derived from explicit word-to-word correspondences between the translation and a reference version [24]. Agarwal and Lavie [25] endorsed the METEOR measure, which involves an initial phase of precisely mapping words from two texts, followed by a second phase where mapped n-grams are divided into subsets. The sequence then selects the most significant subset as the resulting alignment set. Each n-gram from the candidate text can be paired with its closest corresponding n-gram from the reference text [24].

2.5. NIST method of evaluation

Since 2002, National Institute of Standards and Technology (NIST) has been leading open evaluations like OpenMT, which serve as a platform for experimenting with evaluation methods applicable to sponsored MT technology assessments. NIST's Metrics for machine translation challenge offers an opportunity to explore and advocate for new techniques that enhance the measurement sciences in MT evaluations. NIST coordinated and executed the defense advanced research projects agency (DARPA) broad operational language translation (BOLT) assessments of speech-to-text and text-to-text MT technology and the end-to-end MT systems that facilitate real-time spoken communication between speakers of different languages [26]. The NIST method of evaluation, exemplified through initiatives such as OpenMT and the metrics for machine translation challenge, has played a pivotal role in advancing the measurement sciences in machine translation evaluations.

The BOLT program's goal of bridging the language barrier between English-speaking individuals and non-English-speaking populations underscores the importance of effective communication and efficient information retrieval through machine translation technology. With a focus on enabling multi-turn communication in both text and speech, the program aimed to facilitate seamless interactions between individuals across different languages. This involved enabling English speakers to comprehend a wide range of foreign-language sources, such as chat conversations and informal messaging; equipping them with the capacity to locate specific information in these sources using natural-language queries swiftly; and facilitating multi-turn communication in both text and speech with non-English speakers [26].

The existing body of research provides valuable insights into the challenges and opportunities of developing an automated translation system for English and Kreol Morisien. As indicated by previous studies, building automated translation systems for diverse languages like English and Kreol Morisien presents both hurdles and potential benefits. Consistent enhancement of translation precision and assessment standards is crucial to improving the overall calibre of machine translations. Additionally, automatic evaluation systems, such as the BLEU metric and its modified versions, have proven beneficial in evaluating machine translation quality and ensuring that it meets the necessary standards for effective communication in various language pairs and domains.

3. METHODOLOGY

The methodology used in this study is crucial for ensuring the validity and reliability of the research findings. This section presents the systematic approach to developing the machine translation system for converting Kreol Morisien to English and vice versa. The methodology employed in this study involved a combination of quantitative and qualitative research methods.

3.1. Dataset and data collection

The study utilises a dataset of 50,000 English sentences paired with their translated counterparts in KRM. The dataset includes all the words from the third edition of Diksioner Morisien. The dataset features a large variety of short sentences (around 4 words), medium-sized sentences (between 5 to 10 words), and long sentences (above 10 words). These sentences primarily reflect everyday use for general purposes. The dataset was collected through various sources, including online resources, books, and conversations with native speakers of KRM.

3.2. Data pre-processing

The dataset was prepared for analysis by implementing data pre-processing methods. These methods involved eliminating punctuation, converting all text to lowercase, and dividing the sentences into separate words. The data pre-processing stage was necessary to ensure a cleaner and more manageable dataset for analysis. The steps for data pre-processing details can be divided as follows.

3.2.1. Data tokenisation

The data pre-processing stage in this study involves loading the dataset of English sentences paired with their translated counterparts in the system. This study employed data tokenisation as a data pre-processing method to convert the English and KRM sentences into a list of individual words or tokens. Several tokenisation methods are available in the system, including SubwordTextEncoder for breaking down text data into subword units and ByteTextEncoder for dividing text data into tokens at the byte level.

3.2.2. Vocabulary generation

Vocabulary generation is a crucial step in data pre-processing. It involves creating a comprehensive list of all unique words in the dataset. This list maps words to numerical representations (word embeddings) for further analysis. The vocabulary generation step in data pre-processing ensures that every unique word in the dataset is accounted for and assigned a numerical representation, allowing for further study and modelling. This procedure offers various methods for creating vocabularies. It also identifies N-grams within sentences to effectively characterise the essence and significance conveyed by the words.

3.2.3. Feature scaling

Feature scaling is adjusting the scale of features to a consistent range to prevent certain features from disproportionately impacting the model's performance. Various methods are available for feature scaling, including Standardize, which enables normalisation of the input data. Normalization play a significant role in ensuring the stability and efficiency of training in neural machine translation models.

3.2.4. Encoding categorical variables

This data pre-processing step involves converting categorical variables into numerical representations that the algorithms can understand. Categorical variables must be converted to numerical values before they can be utilised in machine learning models. There are different ways in which this can done. This include methods such as label encoding, one-hot encoding, binary encoding and frequency encoding.

3.3. Tokenization

Tokenisation is a fundamental step in data preprocessing. It involves converting the input text into a series of tokens (words or subwords) suitable for input into the deep neural machine translation (DNMT) model. Several tokenisation approaches have been suggested for NMT and DNMT, such as word-level, character-level, and subword-level tokenisation [27]. Tokenising words at the level of individual word units is a simple method, but it may need help with out-of-vocabulary (OOV) terms and languages with complex morphology. Tokenising at the character level can effectively process out-of-vocabulary words and is resilient to variations in spelling.

However, this method produces extended sequences that may pose challenges for the model during processing. To address the limitations of word-level and character-level tokenisation, subword-level tokenisation has gained popularity. Subword-level tokenisation breaks down words into smaller units or subwords, which allows for better handling of OOV terms and languages with complex morphology. Tokenisation at the subword level using byte pair encoding (BPE) has become widely adopted for NMT and DNMT tasks because it offers a middle ground between word- and character-level tokenisation. BPE divides the text into subword units that are shared across different languages, thereby reducing the size of the vocabulary and better managing out-of-vocabulary words [27].

3.4. Training using the transformer model

The Transformer model is a type of deep-learning model which can be used for machine translation [28]. The core component of this model is the self-attention mechanism, which allows the model to evaluate the value of each word in the sentence both at encoding time and at decoding time. A transformer model processes all words in a sentence at the same time. The encoder takes the input sequence in the source language and processes it using multiple self-attention layers and feed-forward neural networks. Each layer in the encoder refines the representation of the input sequence. The decoder takes the output of the encoder and generates the output sequence in the target language [28]. Similar to the encoder, the decoder consists of multiple layers of self-attention and feed-forward neural networks. However, the decoder also incorporates an additional attention mechanism called encoder-decoder attention, allowing it to focus on relevant parts of the input sequence during decoding [28].

The standard transformer model from the Tensor2Tensor library was used in this study. Tensor2Tensor offers a comprehensive framework for training, evaluating, and deploying machine translation models [29]. Tensor2Tensor also supports distributed training across multiple GPUs and/or TPUs. This allows larger datasets and more complex models to be accessed. A dataset of 48,000 parallel sentences was used for training. A set of 1,000 sentences was used for testing, and another set of 1,000 sentences was used for validation. All the sentences were prepared by the research team. The first 25,810 parallel sentences came from the Morisia 1.0 project [11]. The training was run for 100k steps on a Windows 10 laptop with 16 GB RAM and a 256 GB SSD. The BLEU score, available in the Tensor2Tensor library, was used to evaluate the quality of the translations. Tensor2Tensor also provides utilities for loading and running trained models in production environments. The best model was saved and uploaded on DigitalOcean [30], which provides a cloud-based infrastructure for hosting scalable virtual machines (droplets). A web interface to access this translation service was then developed on kreolrepiblikmoris.net. The mobile app also provides the same service for translating from KRM to English and vice versa.

4. RESULTS AND DISCUSSION

In this part, the system's findings and critical assessment are presented. The implications of these findings for future research are also discussed. Based on these results, recommendations for practitioners are also provided. Figure 1 depicts the translation tool on the KreolRepiblikMoris.net website, which enables users to translate from KRM to English and vice-versa. Additionally, as users input text for translation in the provided textbox, the tool offers potential word suggestions in KRM. Figure 2 illustrates a similar translation approach from English to KRM.

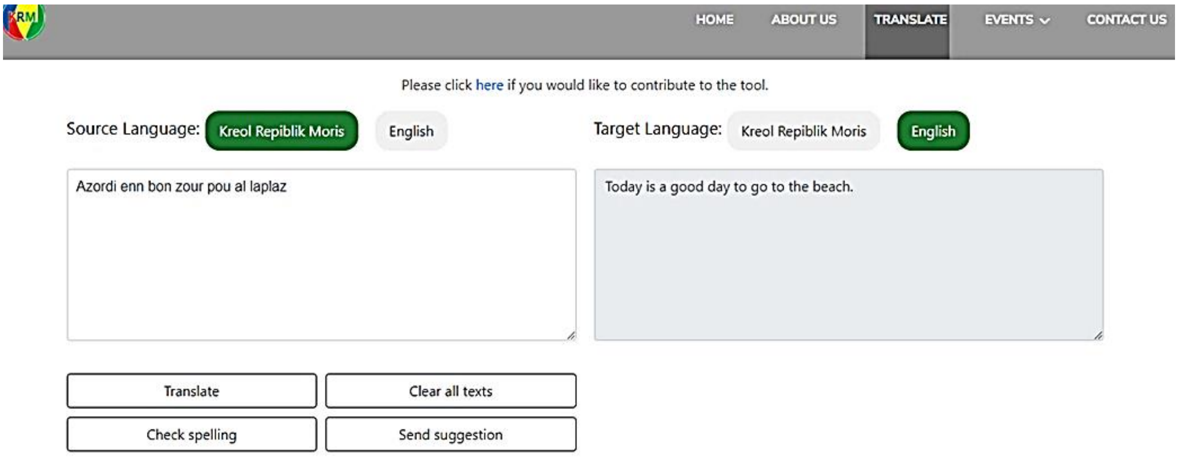

Figure 1. Translation tool to translate from KRM to English

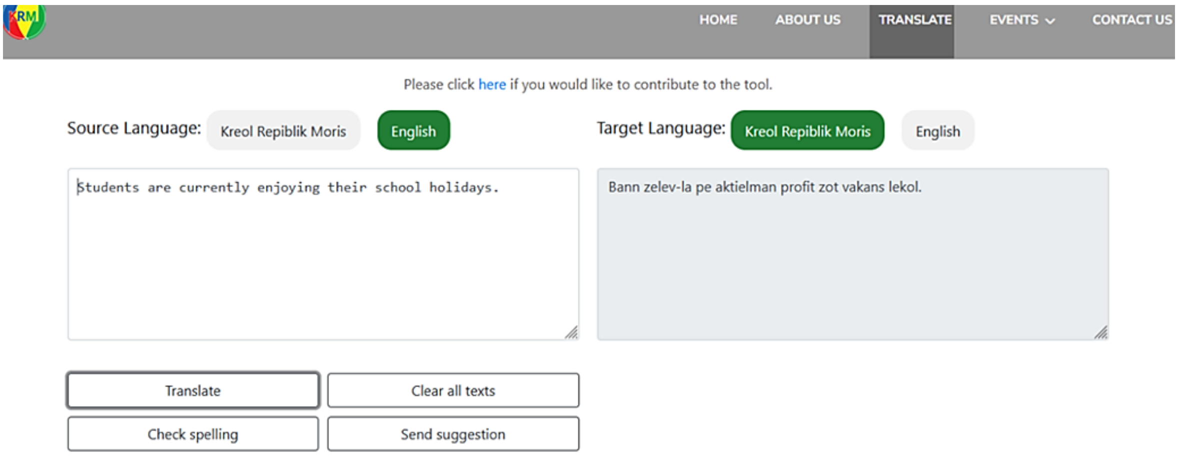

Figure 2. Translation tool to translate from English to KRM

The user must choose the source language and enter the correct word, phrase, or sentence before clicking the Translate button to receive the translated text in the target language. The source phrase in KRM, "Azordi enn bon zour pou al laplaz", is converted into the corresponding English sentence, "Today is a good day to go to the beach", as shown in Figure 1. While undergoing translation, a message 'Translation in progress, please wait a moment' appears on the screen to notify users about the ongoing translation and advises them to wait for the result. A basic spelling checker functionality is also available in the portal. Figure 3 shows a scenario where the words 'siklonn' and 'souvant' are not appropriately written but the tool can provide the correct spelling for these two words.

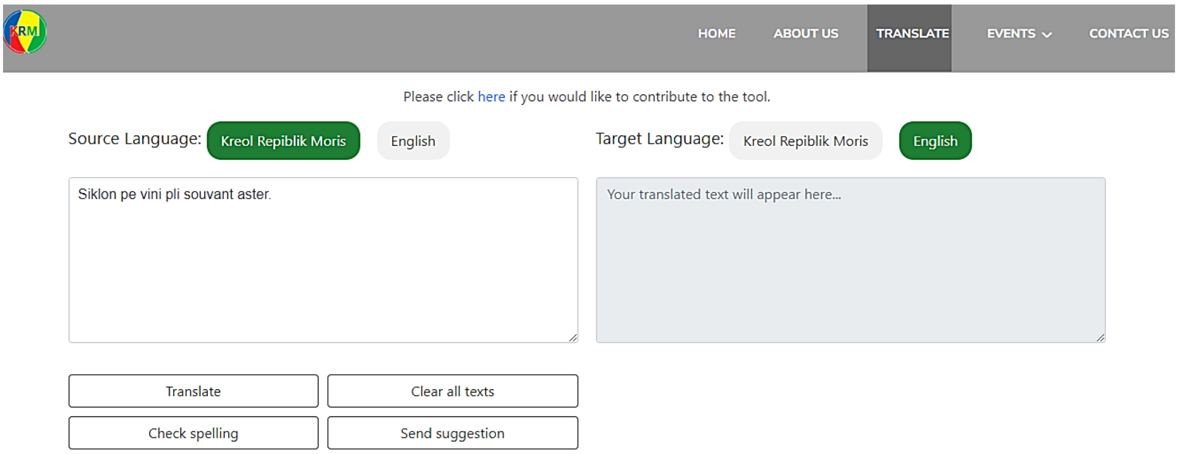

Figure 3. Translation tool showing spelling checker functionality

The translation service available on the website is also accessible as a convenient and user-friendly Android mobile application, making it easier for users to access the language translation service on the go. Figure 4 depicts the mobile app interface design explicitly created for translating between KRM and various other languages. The mobile application allows users to select their desired source and target languages, input their text for translation, and receive the translated result. Figure 5 shows some sample translations recorded in the translation history section. The translation system has undergone extensive testing with different KRM sentences and has been compared with the previous translation system, Morisia 1.0. The testing results revealed that the new translation tool on the KreolRepiblikMoris.net website outperforms Morisia 1.0.

Table 1 presents the conversion of sentences from KRM to English. The table also compares the earlier translation system Morisia 1.0 and the updated system Morisia 2.0. The term 'akredite' was absent in the original Morisia 1.0 dataset and could not be accurately translated. As a result, it has been included in the new Morisia 2.0 dataset, leading to improved sentence translations. Similarly, although the word 'dekatlon' was part of the dataset, it could not be translated correctly. A single occurrence of a word within a dataset may not be enough for the model to learn how to use it properly. The term 'fiziyad' has been interpreted as 'troublemaker'. While this does not perfectly match the original sentence in English, it presents a more reliable translation than Morisia 1.0. The translation accuracy also improves significantly for the fourth and fifth sentences. In summary, there is a marked improvement in translation accuracy from Morisia 1.0 to Morisia 2.0 when converting KRM to English.

Table 2 presents the translation of sentences from English to KRM. The comparison between the former system, Morisia 1.0, and the present system, Morisia 2.0, is illustrated in Table 2. In the first sentence, the term 'gentle' was absent in the Morisia 1.0 dataset and could not be accurately translated. Nevertheless, it has been included in the Morisia 2.0 dataset, resulting in an improved translation for this particular sentence. Sentences 2, 3, and 4 indicate that there have been enhancements in the quality of translated sentences when moving from Morisia 1.0 to Morisia 2.0. The term 'restaurants' in the fifth sentence was missing from the Morisia 1.0 dataset and was not translated accurately. Its addition to the Morisia 2.0 dataset has improved the output quality. Overall, there is a noticeable improvement in translation quality from Morisia 1.0 to Morisia 2.0 when translating from English to KRM. Thus, increasing the size of the dataset by about 24,000 new sentences positively impacts the translation quality. However, the BLEU score increased only from 30.30 (Morisia 1.0) to 31.46 (Morisia 2.0) for translation from KRM to English. For translation from English to KRM, the BLEU score increased from 26.34 (Morisia 1.0) to 28.15 (Morisia 2.0).

Figure 4. Translation App

Figure 5. Translation history

Table 1. Translation from KRM to English

#	Kreol Repiblik Moris	English translation in Morisia 1.0	English translation in Morisia 2.0
1	Pa tou bann liniversite ki akredite.	Not all universities who are prected.	Not all the universities that are accredited.
2	Li finn sorti premie dan dekatlon.	She came out first in the first stage.	She came out first in the meditation.
3	Finn ena ankor enn fiziyad dan lamerik hier.	There was once more of fild in America.	There was one more troublemaker in America.
4	Tom inn kokin plin larzan depi Mary.	Tom has fooled money from Mary.	Tom could have stolen the money from Mary.
5	To bizin evit fer bann erer koumsa.	You must avoid making such a mistake.	You should avoid making mistakes like that.

Table 2. Translation from English to KRM

#	English	KRM translation in Morisia 1.0	KRM translation in Morisia 2.0
1	He was as gentle a man as ever lived.	Li ti kouma enn misie ki zame viv.	Li ti osi dou ki enn dimounn kapav existe.
2	She made the same mistake again.	Li finn fer mem erer.	Li finn refer mem erer.
3	I listened to the music of birds.	Mo ti ekout lamizik so bann swazo.	Mo ti ekout lamizik bann zwazo.
4	She'll be up around by this afternoon.	Nou bizin fer pre pou sa lapremidi-la.	Li pou leve dan lapremidi.
5	The city has an abundance of fine restaurants.	Lavil ena enn abondans korek.	Lavil ena enn abondans bann bon restoran.

4. CONCLUSION

The Kreol Morisien or KRM language has become increasingly popular, creating a demand for an accessible technology platform facilitating its learning. To address this need, an online platform was developed to facilitate the translation of KRM to English and vice versa. This system is capable of translating both individual words and complete sentences. Morisia 2.0 has been developed as an Android application and will be accessible on PlayStore. The translation quality from KRM to English and vice versa is comparable based on the BLEU score evaluation. Morisia 2.0 achieved a BLEU score of 31.46 for translating KRM into English. The earlier evaluation for Morisia 1.0 yielded a score of 30.30. Morisia 2.0 achieved a BLEU score of 28.15 for translating English into KRM, representing an improvement from the previous score of 26.34 in Morisia 1.0, indicating that the expanded dataset has notably enhanced translation quality. Among other comparable

Kreol language translation platforms, Morisia 2.0 has demonstrated it is capability to deliver higher-quality translations.

ACKNOWLEDGEMENTS

This paper is based on a work supported by the Higher Education Commission (HEC) of Mauritius under the award number INT-2022-03.

REFERENCES

[1] "Economic and social indicators," *Statistics Mauritius*. Accessed: Feb. 21, 2024. [Online]. Available: https://statsmauritius.govmu.org/Documents/Statistics/ESI/2022/EI1687/2022 Population Census_Main Results_18112022.pdf.

[2] "Interview: Dernie dialog avek Dev Virahsawmy (1)," *Le Mauricien*. Accessed: Feb. 21, 2024. [Online]. Available: https://www.lemauricien.com/week-end/interview-dernie-dialog-avek-dev-virahsawmy-1/607465/

[3] Z. Boodeea and S. Pudaruth, "Kreol Morisien to English and English to Kreol Morisien translation system using attention and transformer model," *International Journal of Computing and Digital Systems*, vol. 9, no. 6, pp. 1143–1153, Nov. 2020, doi: 10.12785/ijcds/0906012.

[4] S. Paoli and H. Davidson, "Pragmatic markers and verba dicendi: An investigation of Mauritian Creole," *Journal of Pragmatics*, vol. 214, pp. 107–126, Sep. 2023, doi: 10.1016/j.pragma.2023.06.011.

[5] Y. Wu and Y. Qin, "Machine translation of English speech: Comparison of multiple algorithms," *Journal of Intelligent Systems*, vol. 31, no. 1, pp. 159–167, Jan. 2022, doi: 10.1515/jisys-2022-0005.

[6] J.-X. Huang, K.-S. Lee, and Y.-K. Kim, "Hybrid translation with classification: revisiting rule-based and neural machine translation," *Electronics*, vol. 9, no. 2, Jan. 2020, doi: 10.3390/electronics9020201.

[7] P. Williams, R. Sennrich, M. Post, and P. Koehn, "Syntax-based statistical machine translation," in *Synthesis Lectures on Human Language Technologies*, Cham: Springer International Publishing, 2016, doi: 10.1007/978-3-031-02164-0.

[8] Z. Tan *et al.*, "Neural machine translation: A review of methods, resources, and tools," *AI Open*, vol. 1, pp. 5–21, 2020.

[9] J. Zhang and C. Zong, "Neural machine translation: Challenges, progress and future," *Science China Technological Sciences*, vol. 63, no. 10, pp. 2028–2050, Oct. 2020, doi: 10.1007/s11431-020-1632-x.

[10] V. Ragni and L. N. Vieira, "What has changed with neural machine translation? A critical review of human factors," *Perspectives*, vol. 30, no. 1, pp. 137–158, Jan. 2022, doi: 10.1080/0907676X.2021.1889005.

[11] S. Pudaruth *et al.*, "Morisia: a neural machine translation system to translate between Kreol Morisien and English," *inTRAlinea*, vol. 23, 2021.

[12] B. Nath, C. Kumbhar, and B. T. Khoa, "A study on approaches to neural machine translation," *Journal of Logistics, Informatics and Service Science*, vol. 9, no. 1, pp. 271–283, Aug. 2022, doi: 10.33168/LISS.2022.0319.

[13] L. Benkova and L. Benko, "Neural machine translation as a novel approach to machine translation," in *Divai 2020: 13Th International Scientific Conference on Distance Learning in Applied Informatics*, pp. 499–508, 2020.

[14] H. Phan and A. Jannesari, "Statistical machine translation outperforms neural machine translation in software engineering: why and how," in *Proceedings of the 1st ACM SIGSOFT International Workshop on Representation Learning for Software Engineering and Program Languages*, New York, NY, USA: ACM, Nov. 2020, pp. 3–12, doi: 10.1145/3416506.3423576.

[15] H.-H. Phan-Vu, V. T. Tran, V. N. Nguyen, H. V. Dang, and P. T. Do, "Neural machine translation between Vietnamese and English: an empirical study," *Journal of Computer Science and Cybernetics*, vol. 35, no. 2, pp. 147–166, Jun. 2019, doi: 10.15625/1813-9663/35/2/13233.

[16] R. Krüger, "Explicitation in neural machine translation," *Across Languages and Cultures*, vol. 21, no. 2, pp. 195–216, Dec. 2020, doi: 10.1556/084.2020.00012.

[17] B.-E. Batsukh, "Sentence structure and boundary for deep neural machine translation alignment model," in *Proceedings of the Future Technologies Conference*, 2023, pp. 508–520, doi: 10.1007/978-3-031-18344-7_36.

[18] J. Zhang, H. Luan, M. Sun, F. Zhai, J. Xu, and Y. Liu, "Neural machine translation with explicit phrase alignment," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1001–1010, 2021, doi: 10.1109/TASLP.2021.3057831.

[19] A. Emna, S. Kchaou, and R. Boujelban, "Neural machine translation of low resource languages: application to transcriptions of Tunisian Dialect," in *International Conference on Intelligent Systems and Pattern Recognition*, 2022, pp. 234–247, doi: 10.1007/978-3-031-08277-1_20.

[20] K. M. M. Adlaon and N. Marcos, "Building the language resource for a Cebuano-Filipino neural machine translation system," in *Proceedings of the 2019 3rd International Conference on Natural Language Processing and Information Retrieval*, New York, USA: ACM, Jun. 2019, pp. 127–132, doi: 10.1145/3342827.3342833.

[21] A. Villanueva *et al.*, "Mobile-based translation system for Cebuano language with object detection for travel assistance using neural machine translation," in *2019 International Conference on Information and Communications Technology (ICOIACT)*, IEEE, Jul. 2019, pp. 523–528, doi: 10.1109/ICOIACT46704.2019.8938565.

[22] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics-ACL '02*, Morristown, NJ, USA: Association for Computational Linguistics, 2001, doi: 10.3115/1073083.1073135.

[23] S. Lyons, "Quality of Thai to English machine translation," in *Knowledge Management and Acquisition for Intelligent Systems: 14th Pacific Rim Knowledge Acquisition Workshop, PKAW 2016, Phuket, Thailand, August 22-23, 2016, Proceedings 14*, 2016, pp. 261–270, doi: 10.1007/978-3-319-42706-5_20.

[24] S. Banerjee and A. Lavie, "METEOR: An automatic metric for mt evaluation with improved correlation with human judgments," in *Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Proceedings of the Workshop ACL 2005*, 2005, pp. 65–72.

[25] A. Agarwal and A. Lavie, "Meteor, m-bleu and m-ter: Evaluation metrics for high-correlation with human rankings of machine translation output," in *3rd Workshop on Statistical Machine Translation, WMT 2008 at the Annual Meeting of the Association for Computational Linguistics, ACL 2008*, 2008, pp. 115–118.

[26] "Broad operational language translation (BOLT)," *National Institute of Standards and Technology*. Accessed: Mar. 12, 2024. [Online]. Available: https://www.nist.gov/itl/iad/mig/broad-operational-language-translation-bolt.

[27] R. Sennrich, B. Haddow, and A. Birch, “Neural machine translation of rare words with subword units,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2016, pp. 1715–1725, doi: 10.18653/v1/P16-1162.
[28] A. Vaswani *et al.*, “Attention is all you need,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach*, 2017.
[29] A. Vaswani *et al.*, “Tensor2tensor for neural machine translation,” in *AMTA 2018-13th Conference of the Association for Machine Translation in the Americas, Proceedings*, 2018, pp. 193–199.
[30] “Learn here. Dream here. Develop here.” *DigitalOcean*. Accessed: Jun. 16, 2024. [Online]. Available: https://www.digitalocean.com/.

BIOGRAPHIES OF AUTHORS

Sameerchand Pudaruth is an Associate Professor at the University of Mauritius. He has a Ph.D. in Artificial Intelligence. He also holds an LLB from the University of London. He is currently at the ICT Department in the Faculty of Information, Communication and Digital Technologies. He is a senior member of ACM, a senior member of IEEE and a founding member of the IEEE Mauritius Section. He is also a member of the British Computer Society (BCS). His research interests are artificial intelligence, machine learning, data science, machine translation, computer vision, robotics, mobile applications, web technologies, educational technologies, and information technology law. He has written more than 90+ papers for national and international journals and conferences. He has worked on many funded projects from the Higher Education Commission (HEC), Mauritius Research and Innovation Council (MRIC), Human Resources Development Council (HRDC) and the University of Mauritius. He has been in the organising committee of many successful international conferences. He has also written a book entitled Python in One Week. He can be contacted at email: s.pudaruth@uom.ac.mu.

Sheeba Armoogum is an Associate Professor at the University of Mauritius. She has a Ph.D. in Cybersecurity and is notable for inventing a cybersecurity patent on the Intrusion Detection and Prevention System, classified under H04L, an IP classification under WIPO with a 20-year validity. Her research spans numerous IT fields-including Cybersecurity, Cyber Forensics, CyberPsychology and artificial intelligence and she contributes extensively to academia and community through papers and speaking roles. Internationally engaged, she holds positions as a Research Fellow, International Doctoral Examiner, and Executive Committee member for IT and cyber-related organizations, where she focuses on proliferating cybersecurity knowledge and combatting cybercrimes globally. She’s recognized for her contributions and has received multiple awards, including the GlobalWIIN 2023 special recognition. She is dedicated to promoting cybersecurity awareness and combating cybercrimes globally, striving to create a safe digital environment and empower individuals and organizations against cyber threats. She can be contacted at email: s.armoogum@uom.ac.mu.

Nirmal Kumar Betchoo was a former Dean of Faculty of Business and Management at Mauritius’ Université des Mascareignes. He has been in education since 1986 and taught at the tertiary level for nearly 30 years. He specialises in the social sciences, human resource management, business, and strategy. He published fifteen books, over 60 peer-reviewed international research articles, and took part in around 50 local and international conferences. He, moreover, writes for the national press on culture and economy since 2002 with over 125 papers including occasional exclusive reports for leading local papers. ‘He recently graduated with a Post graduate diploma Universite Caen as Ingenieur de formation DU INES 2020. He can be contacted at email: nbetchoo@udm.ac.mu.

Aneerav Sukhoo is the Deputy Director of the Central Information Systems Division of the Ministry of Information Technology, Communication and Innovation of the Republic of Mauritius. He has held responsibilities as Systems Analyst, Project Manager, Technical Manager, Deputy Director and Director of institutions spearheading the computerisation programme in Government for the last 30 years. He holds a Ph.D. in Computer Science from the University of Sout Africa (UNISA) and conducted postdoctoral research at the Indian Institute of Technology (IIT), Bombay. He was Professor and Dean of IT at the Amity Institute of Higher Education, Mauritius on a full-time basis from 2019 to 2020. He has also provided lectures at various universities and supervised several doctoral students. He can be contacted at email: aneeravsukhoo@yahoo.com.

Vandanah Gooria is a programme manager and lecturer in Marketing and Management at the Open University of Mauritius. She has 13 years of experience in administration, then has over 10 years of professional and academic experience encompassing university teaching, market research and surveys, development and authoring course materials, book chapters and publishing research papers. She has a specific interest in serving vulnerable groups and involved in social activities for more than 5 years. Her areas of interest are mainly Special education needs, information technology and open distance learning, marketing, management, Open Educational Resources (OER). She can be contacted at email: v.gooria@open.ac.mu.

Abdallah Peerally is an IT Lecturer at the Polytechnics Mauritius, deliver courses in Cybersecurity, Emerging Technologies, Artificial Intelligence, and Programming. Noteworthy is his pivotal role in establishing Mauritius' inaugural light rail system. As a founding member, he actively contributed to designing and implementation of processes across IT & Operations, steering the project from phase 1 to completion. Simultaneously, he engages in research ventures at the University of Mauritius, relating to fields such as Artificial Intelligence, Game Development, Data Analytics, Machine Translation, Computer Vision, Robotics; Automation, Mobile Applications, Web Technologies, RF technology, and signal processing. Additionally, he pursued postgraduate studies, earning a Masters in Software Project Management and a Masters in Business Administration. He can be contacted at email: abdallahpeerally@gmail.com.

Mohammad Zafar Khodabocus has earned a bachelor's degree in Software Engineering from the University of Mauritius. He has worked as Research Assistant for the project entitled, "Creole to English and English to Creole machine translation using natural language processing techniques and deep learning neural networks", at the University of Mauritius from 2018 to 2020. The project was funded by the Tertiary Education Commission (TEC). He has acquired skills in the following fields: internet of things (IoT), machine learning (ML), artificial intelligence (AI), deep learning (DL), machine translation (MT), internet technologies and game development. He is currently working as a Software Engineer in the ICT industry. He can be contacted at email: zafar.khodabocus@gmail.com.