Ẹ KÚ [MASK]: INTEGRATING YORÙBÁ CULTURAL GREETINGS INTO MACHINE TRANSLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper investigates the performance of multilingually trained neural machine translation (NMT) systems in translating Yorùbá greetings which are a big part of Yorùbá language and culture into English. To evaluate these models, we present IkiniYorùbá a Yorùbá-English translation dataset containing some Yorùbá greetings, and sample use cases. We analyzed the performance of different multilingual NMT systems including Google and NLLB and show that these models struggle to accurately translate Yorùbá greetings into English. In addition, we trained a Yorùbá-English model by finetuning an existing NMT model on the training split of IkiniYorùbá and this achieved better performance when compared to the pre-trained multilingual NMT models although they were trained on a large volume of data.

1 INTRODUCTION

In recent years, multilingual neural machine translation (NMT) models have shown remarkable improvement in translating both high and low-resource languages and have become widely used in various applications (Kudugunta et al., 2019; Aharoni et al., 2019; NLLB Team et al., 2022; Bapna et al., 2022). Despite this progress, NMT models still struggle to accurately translate idiomatic expressions (Fadaee et al., 2018; Baziotis et al., 2022), cultural concepts such as proverbs (Alkhresheh & AlMaaytah, 2018; Adelani et al., 2021), and common greetings, particularly in African languages like Yorùbá which has a rich cultural heritage. The table below illustrates a Yorùbá sentence containing frequently used greeting phrases by the Yorùbá people, and the corresponding translations generated from three multilingual NMT systems, Google[1], Meta's NLLB[2], a model we trained.

Source: Ẹ kú ojúmọ́, ẹ sì kú dédé àsìkò yìí.
Target: Good morning and compliment for this period.
NLLB: You have died, and you have died to this hour.
Google: Die every day, and die at this time.
Our Model: Good morning and compliment for this time.

Table 1: Translation outputs of 3 different NMT models

An examination of Google and NLLB's outputs reveals that they both failed to produce accurate translations for the input sentence. One possible explanation for this is the lack of sufficient training data including these types of greetings, although they were trained on a large volume of multilingual data. Furthermore, Kú which is a common word in these kinds of greetings has two main interpretations that could mean either death or a compliment, depending on the context. Similarly, the syntactic frame of occurrence also determines the meaning of the verb (the type of complement and adjunct), and this is due to the ambiguous nature of Yorùbá verbs. Hence, it is possible that these models were trained on data with Kú having the meaning death.

To address this issue, this paper introduces a new dataset for Yorùbá-English translation dubbed IkiniYorùbá, a Yorùbá-English translation dataset of popular Yorùbá greetings. We evaluate the

[1] https://translate.google.com/
[2] https://ai.facebook.com/blog/nllb-200-high-quality-machine-translation/

performance of existing multilingual NMT systems on this dataset, and the results demonstrate that although current multilingual NMT systems are good at translating Yorùbá sentences into English, they struggle to accurately translate Yorùbá greetings, highlighting the need for further research in translating such cultural concepts on low-resource African languages.

2 Related Work

2.1 Yorùbá greetings

The Yorùbá people cherish greeting so much and they place enough significance on it, as a result, greeting is inseparable from the Yorùbá people, largely because it is their identity and the first thing a stranger will notice when they come in contact for the first time. After the abolishment of the slave trade at the beginning of the 19th century, the Yorùbá indigenes that were rescued by the British warship settled in Freetown, a place known as present-day Sierra Leone. In no time, people began to call them *a kú* which is a fragment attached to all forms of greetings in Yorùbá (Webster, 1966). This is not a surprise because while an English man will say *good morning*, *happy birthday*, *merry Christmas*, and so on, the Yorùbá people would say *ẹ káàrọ̀*, *ẹ kú ọjọ́ ìbí*, and *ẹ kú ọdún kérésìmesì*. The recurring of *ẹ kú* in their everyday conversation was responsible for the appellation. *Ẹ kú* has the same semantic importance with 'good-,', 'merry-' and 'happy-' in English greetings. Without the fragment *ẹ kú* in the communication frame of greeting, the cultural innate shared by interlocutors will be lost.

Another way of analysing this fragment of greeting is to examine the internal structure of the word using the syntagmatic and paradigmatic relationship. de Saussure (1983) introduced Syntagmatic relation and paradigmatic relation to distinguish two kinds of signifiers: one concerns positioning (syntagmatic relation) and the other focuses on substitution (paradigmatic relation). Syntagmatic relation is a type of sematic relations between words that co-occur in the same sentence or text (Asher & Simpson, 1994). Paradigmatic relation is a different type of semantic relation between words that can be substituted with another word in the same categories (Hjørland, 2015). Structurally, *ẹ kú* can be syntactically explained to have a subject-predicate relationship, rather than being a single lexeme or a prefix as claimed by most scholars. Using the paradigmatic relationship lens, *ẹ* can be replaced with any pronoun or nominal item (as described by interlocutors) with +human feature and still fit in perfectly. The +human feature becomes necessary because compliments are mainly for humans and *kú* requires a selection restriction to sieve out the non-human elements. Table 2 shows some of these constructions. It is equally worthy to note here that *ẹ kú* can also be used for supernatural beings or metaphysical beings which in this form sounds like a personification.

Greeting	Person	Compliment
O kú ìrìn	2nd person singular	Compliment for walking
Ẹ kú òdè	1st person plural	Compliment for attending a party
Wọ̀n kú ìjóòkó	3rd person plural	Compliment for sitting

Table 2: Translation output of some NMT systems

Kú on the other hand is a transitive predicate that requires a compliment. This compliment could either be a noun that signifies time like; *àárọ̀* (morning), a noun that denotes season like; *ọ̀ririn/òtútù* (cold), a noun that points to the celebration like; *kérésìmesì* (Christmas), and a nominalized verb that describes an event or action like; *ìjókòó* (sitting), and many more. The omission or loss of a compliment in a greeting construction will alter the interpretation of the expression which may also change the meaning of *kú* to death.

2.2 Machine translation for Yorùbá

The development of machine translation systems for low-resource languages such as Yorùbá has seen a significant amount of research efforts in recent years. One major area of focus has been on curating translation datasets for these languages, which are collected using either automatic or manual methods. Examples of automatically collected datasets that include Yorùbá are JW300 (Agić & Vulić, 2019), CCMatrix (Schwenk et al., 2021), and CCAligned (El-Kishky et al., 2020). On

the other hand, examples of manually translated datasets for Yoruba include MENYO-20k (Adelani et al., 2021), MAFAND-MT (Adelani et al., 2022), FLORES-101 (Goyal et al., 2022), and NTREX (Federmann et al., 2022). These datasets have been instrumental in the study, development, and improvement of machine translation systems for Yorùbá. For example, (Adelani et al., 2021) investigated how domain data quality and the use of diacritics, a crucial aspect of Yorùbá orthography, impact Yorùbá-English translations. Adebara et al. (2022) examined the effectiveness of Yorùbá-English machine translation in translating Bare nouns (BN), by comparing the results obtained from using statistical machine translation methods and neural approaches. Adelani et al. (2022) investigated how to use effectively leverage pretrained models for translation of African languages including Yorùbá. Despite the attempts to create datasets and develop translation systems for Yorùbá, to the best of our knowledge, only Adelani et al. (2021) has examined a cultural aspect of Yorùbá by evaluating their models on Yorùbá proverbs, which are a significant part of Yorùbá tradition. However, this research has not looked into how these models perform on another cultural aspect which is Yorùbá greetings. Therefore, in this work, we investigate the performance of Yorùbá-English translation models on Yorùbá greetings.

3 DATASET

IkiniYorùbá dataset: We introduce IkiniYorùbá, a Yorùbá-English translation dataset for Yorùbá greetings and their usage in various contexts, containing 960 parallel instances. The data curation process involves three key stages. Firstly, we gathered commonly used Yorùbá greetings that cover a variety of situations such as time, season, celebration, and more, as outlined in Section 2.1, resulting in a total of 160 Yorùbá greetings. Secondly, we created 5 different example sentences for each greeting, where the greetings are used in context, by expert users of the language, resulting in 800 use cases in total. Lastly, we had an expert translator translate the seed data and the use cases into English. Table 3 illustrates the sample sentences in the IkiniYorùbá dataset. It also presents the statistics of the data used. We split the data into train/dev/test splits with 100/20/40 seed greeting instances and each of the the splits having the 5 different use cases for each instance.

Other datasets: For our experiment, we used the movie transcripts subset of the MENYO-20k dataset, which is a human-translated English-Yorùbá dataset for movie transcripts. We selected this dataset because it is a conversational type of data.

Yorùbá	English
Seed Greetings	
Ẹ kú ìfẹ́	Thanks for the love
Ọkọ̀ á rẹ̀fó	Safe ride
Greetings with contexts	
Ẹ kú ìfẹ́, Ire là ó má bá ara wa ṣe.	Thanks for the love, may we continue to celebrate one another.
A ó ma fojú sónà láti rí yín, ọkọ̀ á rẹ̀fó	Looking forward to seeing you, safe ride.

	Number of Sentences		
Data	**Train. Set**	**Dev. Set**	**Test Set**
IkiniYoruba	100/500	20/100	40/200
Movie Transcript	–	–	775

Table 3: **Left:** Sample of the created corupus. **Right:** The split of the data

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Greetings play a crucial role in Yorùbá culture and are widely used in daily conversations by Yorùbá people. For every action, they have a customary way of greeting or complimenting those involved using the phrase "Ẹ ku". Despite efforts to create translation datasets and the development of Yorùbá-English translation systems, little attention has been given to the analysis of how well these systems perform with Yorùbá greetings, particularly when translating into English. In this work, we compare several existing translation systems and evaluate their performance on Yorùbá greetings. We demonstrate the effectiveness of these translation systems by testing them on movie transcripts, which are conversational in nature. Below, we outline our experiments.

Translation Models: In this study, we evaluate the performance of three multilingual NMT systems. These systems were pre-trained on various languages, and they are Google translation engine, the distilled version of Meta's NLLB (NLLB Team et al., 2022) with 600M parameters, and with a publicly available M2M-100 (Fan et al., 2020) with 418M parameters fine-tuned on the MENYO-20k dataset. We generated translations for the test sets using the Google Translate web application[3], while for Meta's M2M-100 and NLLB models, we used the HuggingFace transformers[4].

Data preprocessing and evaluation: To standardize the format of the two parallel datasets, we converted the Yorùbá texts in the dataset to Unicode Normalization Form Composition (NFC). And to automatically assess the performance of the models, we utilized BLEU (Papineni et al., 2002) score implemented in SacreBLEU (Post, 2018).

4.2 System Comparison and evaluation

Table 4 shows the results of evaluating the three different models on the two datasets: IkiniYorùbá test split and Movie Transcripts. The results showed that the models obtained impressive performance on the Movie Transcript data with high BLEU scores but poorly on the IkiniYorùbá data with significantly lower scores. This highlights their inability to translate Yorùbá cultural content such as greetings. The best-performing model, M2M-100, had a BLEU score of 34.70 on Movie Transcript data as it was trained on this same data by its authors. However, it had a score of 4.3 on greetings data. The second-best model, Google Translate, was -3.65 below the best model on Movie Transcript but better on greetings data with a score of 9.47, though still lower compared to its performance on Movie Transcript data.

	yo-en			
	BLEU		**Adequacy**	**CCP**
	Movie Transcript	**IkiniYorùbá**	**IkiniYorùbá**	
Google	31.05	9.47	2.46	0.13
NLLB	27.19	5.03	2.16	0.13
M2M-100	34.70	4.33	2.16	0.08
+ IkiniYorùbá	26.05	**29.67**	**3.01**	**0.43**
+ Movie Transcript	-	6.25	-	-
+ IkiniYorùbá + Movies Transcript	-	29.49	-	-

Table 4: Sample of the created corpus.

In addition, we finetuned the M2M-100 model on IkiniYorùbá, Movie Transcripts, and a combination of both data sources and evaluated the models on the IkiniYorùbá's test split. Our results show that finetuning the M2M-100 on Movie Transcripts improved the model's performance by 2 BLEU points compared to the original M2M-100. However, the best performance was achieved when the M2M-100 was finetuned on the IkiniYorùbá training split, with a BLEU score of 29.67. On the other hand, finetuning the M2M-100 on the combination of both datasets did not result in any improvement.

To understand the performance of individual models on the IkiniYorùbá test set, we conducted human evaluations of the translated outputs from Google, NLLB, M2M-100, and M2M-100 finetuned on IkiniYorùbá dataset. We ask two native Yorùbá speakers fluent in English to rate the 240 sentences for each system on two criteria: adequacy (on a Likert scale of 1 to 5) and cultural content preservation (binary scale of 0 or 1). Where adequacy describes how much of the meaning of the reference translation was preserved in the MT output, CCP indicates whether the greetings/compliments within the translation are preserved or not. The results showed that the NMT systems struggle at translating Yorùbá greetings accurately, and as confirmed by the automatic evaluation, M2M-100 finetuned on IkiniYorùbá outperformed all other models. Overall, we observed that human evaluation fairly agrees with automatic evaluation.

[3] https://translate.google.com/

[4] https://github.com/huggingface/transformers

5 CONCLUSION

In this study, we analyzed the performance of machine translation models in translating Yorùbá greetings into English. To achieve this objective, we introduced a novel dataset called IkiniYorùbá, which contains a collection of Yorùbá greetings and their respective sentence use cases. We evaluated three publicly available machine translation models on this dataset and found that, despite their ability to translate other Yorùbá texts, they failed to accurately translate Yorùbá greetings, which are a crucial aspect of Yorùbá culture. In future research, we aim to expand the IkiniYorùbá dataset by adding more profession-based greetings and exploring ways to enhance the performance of machine translation models with these data.

ACKNOWLEDGMENTS

Use unnumbered third level headings for the acknowledgments. All acknowledgments, including those to funding agencies, go at the end of the paper.

REFERENCES

Ife Adebara, Muhammad Abdul-Mageed, and Miikka Silfverberg. Linguistically-motivated Yorùbá-English machine translation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 5066–5075, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.449.

David Adelani, Dana Ruiter, Jesujoba Alabi, Damilola Adebonojo, Adesina Ayeni, Mofe Adeyemi, Ayodele Esther Awokoya, and Cristina España-Bonet. The effect of domain and diacritics in Yoruba–English neural machine translation. In *Proceedings of Machine Translation Summit XVIII: Research Track*, pp. 61–75, Virtual, August 2021. Association for Machine Translation in the Americas. URL https://aclanthology.org/2021.mtsummit-research.6.

David Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen Muhammad, Guyo Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ajibade, Tunde Ajayi, Yvonne Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. A few thousand translations go a long way! leveraging pre-trained models for African news translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3053–3070, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.223. URL https://aclanthology.org/2022.naacl-main.223.

Željko Agić and Ivan Vulić. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3204–3210, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1310. URL https://aclanthology.org/P19-1310.

Roee Aharoni, Melvin Johnson, and Orhan Firat. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3874–3884, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1388. URL https://aclanthology.org/N19-1388.

Mohammad H. Al-khresheh and Shahab A. AlMaaytah. English proverbs into arabic through machine translation. *International Journal of Applied Linguistics and English Literature*, 2018.

R.E. Asher and J.M.Y. Simpson. *The Encyclopedia of Language and Linguistics*. Number Bd. 9 in The Encyclopedia of Language and Linguistics. Pergamon Press, 1994. ISBN 9780080359434. URL https://books.google.de/books?id=5kdiAAAAMAAJ.

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Nikhil Baljekar, Xavier Garcia, Wolfgang Macherey, Theresa Breiner, Vera Saldinger Axelrod, Jason Riesa, Yuan Cao, Mia Chen, Klaus Macherey, Maxim Krikun, Pidong Wang, Alexander Gutkin, Apu Shah, Yanping Huang, Zhifeng Chen, Yonghui Wu, and Macduff Richard Hughes. Building machine translation systems for the next thousand languages. Technical report, Google Research, 2022.

Christos Baziotis, Prashant Mathur, and Eva Hasler. Automatic evaluation and analysis of idioms in neural machine translation, 2022. URL https://arxiv.org/abs/2210.04545.

Ferdinand de Saussure. Course in general linguistics. 1983. (trans. Roy Harris).

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5960–5969, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.480. URL https://aclanthology.org/2020.emnlp-main.480.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. Examining the tip of the iceberg: A data set for idiom translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1148.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. Beyond english-centric multilingual machine translation. *arXiv preprint*, 2020.

Christian Federmann, Tom Kocmi, and Ying Xin. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pp. 21–24, Online, nov 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.sumeval-1.4.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538, 2022. doi: 10.1162/tacl_a_00474. URL https://aclanthology.org/2022.tacl-1.30.

Birger Hjørland. Are relations in thesauri "context-free, definitional, and true in all possible worlds"? 66(7):1367–1373, jul 2015. ISSN 2330-1635. doi: 10.1002/asi.23253. URL https://doi.org/10.1002/asi.23253.

Sneha Kudugunta, Ankur Bapna, Isaac Caswell, and Orhan Firat. Investigating multilingual NMT representations at scale. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 1565–1575, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1167. URL https://aclanthology.org/D19-1167.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. No language left behind: Scaling human-centered machine translation. 2022.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040.

Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 186–191, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6319. URL https://aclanthology.org/W18-6319.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. CCMatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 6490–6500, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.507. URL https://aclanthology.org/2021.acl-long.507.

J. B. Webster. A preface to modern nigeria: The 'sierra leonians' in yoruba, 1830–1890. by jean herskovits kopytoff. *Africa*, 36(4):452–453, 1966. doi: 10.2307/1158056.

A Appendix

You may include other additional sections here.