Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 6480–6486
Marseille, 20-25 June 2022

A Survey of Machine Translation Tasks on Nigerian Languages

Ebelechukwu Nwafor[1], **Anietie Andy**[2]
[1]Department of Computing Sciences, [2] Penn Medicine.
[1]Villanova University, Villanova, PA, 19085, [2]University of Pennsylvania, Philadephia, PA, 19104.
Email: ebelechukwu.nwafor@villanova.edu, anietie.andy@pennmedicine.upenn.edu

Abstract

Machine translation is an active area of research that has received a significant amount of attention over the past decade. With the advent of deep learning models, the translation of several languages has been performed with high accuracy and precision. In spite of the development in machine translation techniques, there is very limited work focused on translating low-resource African languages, particularly Nigerian languages. Nigeria is one of the most populous countries in Africa with diverse language and ethnic groups. In this paper, we survey the current state of the art of machine translation research on Nigerian languages with a major emphasis on neural machine translation techniques. We outline the limitations of research in machine translation on Nigerian languages and propose future directions in increasing research and participation.

Keywords: Machine translation, low-resource languages

1. Introduction

Machine translation is an important and very beneficial computing task that provides an automated mechanism for communicating across languages. In recent years, there has been an increase in available tools and techniques for machine translation such as statistical (Koehn et al., 2007b) and neural machine translation (Bahdanau et al., 2016; Luong et al., 2015) techniques. Recently, neural machine translation has become the standard means for language translation tasks with advanced techniques such as encoder-decoder models (Vaswani et al., 2017) which have proven to produce remarkable state of the art results.

With the advancements in neural machine translation, there has been a lot of work in the literature which focuses on machine translation task of high-resource languages (Wu et al., 2016; Koehn and others, 2005) such as from English to French (Meyer et al., 2013; Brown et al., 1988; He et al., 2016) or English to German (Cheng et al., 2018). However, there are very few literature on machine translation of African languages (Nekoto et al., 2020b; Abbott and Martinus, 2018) and even fewer works centered on Nigerian languages. This might be attributed to socio-economic issues such as the high cost of training language models and the effort involved in curating high-quality datasets. Nigeria is one of the most populous country in Africa (Akinyemi and Isiugo-abanihe, 2014) with a population of over 150 million inhabitants and over 500 languages [1]. According to the United Nations (noa,), Nigeria is the 7th largest world population, and it is estimated that the population of Nigeria will surpass that of the United States and become the third largest nation before 2050. In addition to being one of the most populous country in Africa with immense human resources, Nigeria is poised for business growth with an increase in emerging markets and an interest from foreign investors, particularly in the financial technology ecosystem. One can see that it is imperative to provide a diverse set of machine translation tools that provides high-quality translation of Nigerian languages. This helps provide a low-cost means to breaking language barriers while connecting speakers of indigenous languages with investors and tourists from other parts of the world. In addition, providing a means of language translation further preserves endangered languages as it brings people closer to properly understanding how to speak these languages. One of the impediments to the furtherance of research in machine translation of low-resource languages is the lack of high-quality parallel and monolingual datasets with most of the available open-source datasets consisting of sentence pairs without human validation. In addition, there is a lack of diversity in sentence pairs as such sentences are mostly focused on particular topic areas such as religious texts. It is imperative that we provide high-quality open-source datasets and benchmarks which enables researchers to effectively implement and evaluate machine translation tasks. To this end, this paper makes the following contributions:

- We survey the state of the art in machine translation tasks on Nigerian languages with major emphasis on neural machine translation techniques.
- We outline the limitations of current research in machine translation of Nigerian languages and propose future directions to increasing research and participation.

The remaining portion of the paper is outlined as follows. Section 2 outlines background information on Nigerian languages, and types of machine translation techniques. Section 3 provides related work on machine translation techniques of Nigerian languages. We categorize the related works based on various approaches as outlined in the background section. Section 4 provides information on limitations and future directions in the application to machine translation techniques on Nigerian languages. Finally, we conclude in section 5.

2. Background Information

2.1. Overview of Nigerian Languages

Nigeria consists of one of the most culturally diverse group of languages (Blench, 2012) in Africa with over 500 indigenous languages spoken. There are three major languages

[1] https://www.cia.gov/the-world-factbook/countries/nigeria/

(Igbo, Yoruba and Hausa) which are spoken by the major tribes in Nigeria. Most of the languages spoken are from the Niger-Congo origin (Williamson and Blench, 2000) with English as the official language. Out of all of the languages spoken in Nigeria, there are over 16 endangered Nigerian languages (Blench, 2012) which are mostly spoken by indigenous groups from the northern part of Nigeria. Table 2.1 represents some of the major languages spoken, the language origin, and the estimated total number of speakers.

Language	Family	Speakers	Region
Igbo	Niger-Congo	27M	East
Yoruba	Niger-Congo	42M	West
Hausa	Afro-Asiatic	63M	North
Nigerian Pidgin	English Creole	75M	All

Table 1: Select Nigerian languages consisting of language families, estimated total number of speakers and geographic regions (Ogueji et al., 2021).

2.2. Machine Translation Techniques

Machine translation has been extensively studied for decades (Bahdanau et al., 2016; Luong et al., 2015; Koehn et al., 2007b) with neural machine translation providing the most recent state-of-the art results. There are three types of machine translation techniques that have been explored in the literature – Rule-based machine translation (RBMT), Statistical machine translation (SMT), and Neural machine translation (NMT). A high level overview of these techniques are outlined as follows:

Rule-based Approach: This is one of the oldest form of machine translation technique used. This approach is based on understanding the linguistic properties of the source and target languages using dictionaries and expert knowledge to define grammar rules. This process involves morphology analysis, syntax, and lexical semantics. Linguistic analysis is performed on the source language to identify morphology, parts of speech, phrases, named entity, and word disambiguation. Each word is replaced in the target language using a dictionary which represents mappings between source and target words. In order to preserve sentence semantics across translated languages, most RBMT approach utilizes a combination of finite state machines to develop their knowledge graphs (Forcada et al., 2011; Scott and Barreiro, 2009).

(Forcada et al., 2011) utilizes finite-state transducers for lexical processing, Hidden Markov models for part-of-speech tagging, and multi-stage finite-state chunking for structural transfer. (Eisele et al., 2008) utilizes a modified phrase table with entries from translating various data with rule-based systems. One of the main advantage of this approach is that it does not require as much parallel sentence pairs as with most NMT approaches. Also, translation errors can be corrected by updating the dictionary. This allows for flexibility in updating language constructs. Consequently, one major drawback of this approach is that the translation quality is mostly defined by the strength of the dictionary which requires frequent updates from domain experts. RBMT also tends to produce translations that are more repetitive and less fluid which can be attributed to its mechanical approach of using rules for translation.

Statistical-based Approach: This approach involves the use of statistical techniques such as probability distribution models to provide a means for machine translation between source and target languages. This is achieved by assigning a probability score to word or phrase contained in every target sentence where words or phrases with the highest probability contains the best translation for the target sentence (Koehn et al., 2007b; Brown et al., 1993). SMT can be applied at a word or phrase level and consists of a translation and language model. The translation model is defined as the probability that the source sentence is the translated version of the target word. The language model tries to describe how representative the target sentence is to the natural spoken language. It assigns probabilities to sentence similar to the sentence ordering. One approach utilized in developing the probability distributions is the use of Bayes theorem (Zens et al., 2002) and Hidden Markov Model (Deng and Byrne, 2008; Alkhouli et al., 2016). (Koehn et al., 2007a) developed Moses, an open-source machine translation toolkit which utilizes linguistic information that captures semantics in mapping text phrases and a confusion network decoding for translating ambiguous text inputs.

One advantage of SMT approach over RBMT is the improved translation quality. It allows for translation that captures not just linguistic morphology but the use of a probability distribution which improves with semantic quality.

Neural-based Approach: This approach is referred to as the state of the art in machine translation as it is widely used and has shown to provide results with higher accuracy as compared to the other approaches (Bahdanau et al., 2016; Luong et al., 2015; Cho et al., 2014). Neural machine translation involves the use of deep learning techniques to provide a means of inferring high level semantics from language translations. A popular neural machine translation approach (Vaswani et al., 2017) utilize transformer based models with encoder-decoder architecture. These models consists of stacks of multiple hidden layers with multi-head attention mechanisms and have been shown (Vaswani et al., 2017) to outperform traditional neural architecture such as Recurrent Neural Networks for machine translation task.

Current implementation for language models consists of multilingual language model embeddings (Pires et al., 2019; Lample and Conneau, 2019) where one language model is trained on multiple languages. This allows for zero-shot transfer learning where cross language representation is learned without the need for a parallel language corpus across all language pairs. This has been shown to produce better results than monolingual model training (Conneau et al., 2020) especially for low-resource languages. Supervised neural approach relies heavily on a large corpus of quality translated sentence pairs; as such this poses a limitation to the quality of language translation. There are some approaches that work well with limited datasets (Mikolov et al., 2013; Artetxe et al., 2018) and can provide a means of translating from one language to another based on translations derived from a similar language (Johnson et al., 2017; Zoph et al., 2016).

3. State of the Art

There has been a limited number of work centered on machine translation of Nigerian languages. Most of the cutting edge research on machine translation utilizes neural machine translation approaches (Stahlberg, 2020). However, most of the machine translation work on Nigerian languages focuses on rule-based approach using context free grammars while a few focuses on neural machine translation techniques such as transformer-based models. We outline the work that has been conducted over the years and categorize each work based on the different approach utilized.

3.1. Rule-based Approach

(Ayegba et al., 2014) utilizes a rule-based approach for machine translation of English to Igala language. This approach utilizes noun phrases from English language while performing a series of processes such as parts of speech tagging, morphological analysis which analyzes words based on its root or base form, and comparing noun phrases to components contained in a bilingual dictionary. Their approach was tested on 120 randomly selected English noun phrases and achieves a Bilingual Evaluation Understudy (BLEU) accuracy of 90.9%. (Akinwale et al., 2015) proposed a web-based English to Yoruba translation model utilizing a similar approach as (Ayegba et al., 2014). The translator component utilizes a set of twenty rules which were specified using context free grammar. This approach achieves an accuracy of 90.5%. (Eludiora and Ajibade, 2021) proposed a rule-based model for English to Yoruba translation of Yoruba verbs based on tone changing. It is their intuition that some Yoruba verbs change tone in the bilingual dictionary from low-tone to mid-tone which sometimes changes the meaning of the sentence. Their approach is implemented using 20 tone changing verbs. They evaluate the efficacy of their approach by performing language expert evaluation which entails comparing the output derived from their model with the output generated from Google translation. According to the authors, this approach is very time-consuming but very extensive. In addition, they evaluate their approach using human evaluators. In a total of 70 respondents, 69% of the respondents agree that their system correctly translates verb-phrases while 29% of the respondents agrees that Google translation works efficiently.

3.2. Statistical-based Approach

(Ezeani et al., 2016) developed a model using the Igbo Bible corpus to detect and restore missing didactics in texts at word level toknization. Their approach on didatic replacement utilizes work conducted by (Simard, 1998) which consists of using Hidden Markov Model in which the input text is viewed as a stochastic process. (Onyenwe et al., 2019) develops a parts of speech (POS) tagger for Igbo language. Their approach utilizes a host of post tagging approach including Hidden Markov Model. They achieve an accuracy of of 93.17% to 98.11% on the overall words, and 7.13% to 83.95% on unknown words.

3.3. Neural-based Approach

(Orife, 2020) developed a neural machine translation model for translating Edoid languages to English. Edoid languages are primarily spoken by the southern Nigeria (Edo and Delta states) consisting of Edo, Esan, Urhobo, and Isoko languages. They utilize transformer models with encoder decoder and multi-head self attention. To evaluate the effectiveness of their approach, they trained their model using JW300 dataset (Agić and Vulić, 2019) consisting of over 100 African languages The training was conducted using tokenization processeses such as Byte-pair encoding (BPE) and word-level tokenization . The results shows that Urhobo and Isoko consists of larger training dataset performed best with higher BLEU scores. BPE tokenization provided a 37% boost for the development and test dataset of Edo and Esan languages and a 32% boost for Urhobo language. However, BPE produced worse results when compared to word-level tokenization for Isoko languages. (Ahia and Ogueji, 2020) developed supervised and unsupervised neural machine translation models to serve as a baseline for future works to come in the translation of Nigerian pidgin. For their approach, they utilized a transformer architecture proposed by (Vaswani et al., 2017) while experimenting with word-level and Byte-Pair encoding subword tokenization. The supervised approach produced a BLEU score of 17.73 while the unsupervised model produced a BLEU score of 5.18 for English to Pidgin Translation.

(Nguyen and Chiang, 2018) developed a model that improves the mistranslation of rare words. This approach is based on a modified version of attention based encoder-decoder models. Their approach hones on the premise that the output layer which consists of the inner product of the context vector and all possible word embeddings improperly rewards frequently occurring words. In their approach, instead of using the dot product, the norm vectors are set to a constant value. In addition, they include new terms which provides direct connection from source sentence and this makes the model properly memorize rare word translations. They evaluate their approach on 8 language pairs which includes Hausa to English language pair. (Hedderich et al., 2020) demonstrates that a transfer learning approach through multilingual transformer models (mBERT and XLM-RoBERTa) can be utilized for tasks such as name entity recognition and topic classification on low-resource languages. The approach involves fine-tuning the target language dataset on high-resource language models. Their approach is evaluated on three African languages Hausa, isiXhosa and Yoruba out of which two of the languages (Hausa, and Yoruba) are Nigerian languages. They produce results comparable to the state-of-the-art with as little as 10 or 100 labelled sentences. They achieve at least an improvement of 10 points in the F-1 score for a shared label of named entity recognition. Their result shows promise and is consistent with their hypothesis which also validates work shown in prior research. Their approach however does not produce good results for topic classification. This might be as a result of mismatch in the label set.

(Ogueji et al., 2021) developed AfriBERTa, an approach which involves training multilingual models on low-resource language. According to the authors, it is a general assumption that low-resource multilingual language models benefit from being trained in combination with high-resource languages. low-resource multilingual models do not need to be trained in combination with high-resource languages and does not require as much dataset used for training high-resource languages. The authors accomplish multilingual model training on low-resource languages with a dataset consisting of 11 African languages of which Igbo, Yoruba, Hausa, and Nigerian Pidgin are Nigerian languages. They also show that the state of the art accuracy can be achieved with training on less than 1GB of data. Furthermore, they apply their pre-trained transformer model on downstream tasks such as name entity recognition and text classification task. Their model outperforms the state of the art multilingual models such as mBERT and XLM-R.

3.4. Data Acquisition

One of the major impediments to corpus-based machine translation of low-resource languages is the quality and quantity of the dataset utilized for model training. However, there has been limited work in generating datasets for machine translation tasks of Nigerian languages. Some of the prominent dataset utilized for this task are outlined below.

(Gutkin et al., 2020) developed an open-source dataset of Yoruba speech which consists of over four hours of recordings from 36 male and female volunteers with transcription and disfluency annotation. (Adelani et al., 2021) developed a publicly available parallel corpus known as MENYO-20K which consists of a parallel corpus of texts in English-Yoruba language with over 20,000 sentences obtained from news articles, TED talks, movie and radio transcripts, science and technology texts and short articles from the web which were annotated by professional translators with proficiency in Yoruba language. (Butryna et al., 2020) developed a crowd-sourced speech corpus for low-resource languages which consists of languages in South and Southeast Asia, Africa (South Africa, and Nigeria), Europe and South America. The only Nigerian language supported was Nigerian Pidgin. They achieve this task by partnering with local communities and universities in the region. (Agić and Vulić, 2019) introduces JW300, a parallel corpus of over 300 languages containing around over 100,000 sentences per language pair. The corpus consists of a total of 1,335,376 articles with over 109 million sentences and 1.48 billion tokens. They achieve this by crawling publications from jw.org. OPUS (Tiedemann, 2012), is one of the largest open source parallel corpora repository of translated text. It consists of over 90 languages with a total of 3,800 language pairs comprising of over 40 billion tokens in 2.7 billion parallel units. (Goyal et al., 2021) introduces Flores-101, an open-source benchmark for evaluating low-resource multilingual machine translation task. This dataset consists of 3,000 sentences extracted from Wikipeadia. In addition, the sentences have been converted into 101 languages which includes three major languages in Nigeria (Igbo, Yoruba, and Hausa). (Ezeani et al., 2020) developed a publicly available standard evaluation benchmark dataset for Igbo to English machine translation. This includes over 10,000 high quality English to Igbo sentence pairs which were derived mostly from news (BBC Igbo [2] and PUNCH newspaper [3]) domains.

4. Discussion and Future Directions

Machine translation technology has improved tremendously over the years with neural machine translation producing remarkable results; however there are still limitations surrounding the development of tools for automatic machine translation of low-resource languages such as Nigerian languages. We outline some of the key limitations facing the widespread adoption of machine translation of Nigerian languages.

4.1. Limited Open Source Datasets

There is a strong need to create more high-quality dataset that can be used for neural machine translation. Most of the parallel corpora available consists of less than 100,000 translated sentence pairs. One approach to generating high-quality parallel corpora in addition to utilizing linguistic experts with domain knowledge is to leverage crowd-sourcing platforms like Amazon Mechanical Turk [4] to provide translation from native speakers. It has been shown in previous literature (Bloodgood and Callison-Burch, 2014) that crowd-sourcing platforms can provide translation at an expert level with a reduced cost. One drawback to the use of crowd-sourcing platforms is the difficulty in evaluating the competency of the reviewers (Allahbakhsh et al., 2013). There are metrics to circumvent this issue such as evaluating the reviewer translation (Nowak and Rüger, 2010). Another drawback to using crowd-sourcing for machine translation is that it often does not establish real connections between the linguist and the language community which is an essential component for fostering an efficient translation ecosystem and also for understanding the needs of the community (Bird, 2020). For more information on all of the stakeholders contained in the language translation process, we refer the reader to view the work by (Nekoto et al., 2020a). In the absence of high quality large training datasets, one can employ the use of unsupervised learning approaches (Artetxe et al., 2018), zero-shot learning (Johnson et al., 2017) and various data augmentation and transfer-learning approaches (Zoph et al., 2016; Nguyen and Chiang, 2017) which requires minimal training datasets.

4.2. Fairness in Language Models

A number of language models are developed without considering the variety of the training dataset and as such might not effectively transfer to low-resource languages (Wu and Dredze, 2020). Ensuring that our language models are able to cater to a diverse set of machine translation tasks while producing appropriate results is as crucial as the machine translation task (Nekoto et al., 2020b). More emphasis needs to be placed on evaluating the fairness of machine learning (ML) and artificial intelligence (AI) algorithms

[2] https://www.bbc.com/igbo

[3] https://punchng.com/

[4] https://www.mturk.com/

with a focus on learning algorithms used to develop these machine translation models while taking into consideration the effects of the diversity of its training dataset. It is important to note that AI fairness has become a focal point and an active area of research amongst the machine learning community (Mehrabi et al., 2019). It is important that fairness in incorporated into the entire machine translation process as a lack of fairness can possibly lead to socio-economic inequalities and also language misrepresentation based on gender (Vanmassenhove et al., 2018) ethnic groups (Nekoto et al., 2020a).

4.3. Community Partnerships

In order to ensure an effective machine translation ecosystem, there must be a cohesive synergy between all stakeholders involved in the process–from community members, native speakers, to linguistic experts. All of these participants have to play an active role in the development of translation tasks. Individuals with expert domain knowledge are crucial to the machine translation process. In order to get the best performance of our models, we require high quality translated datasets and this can be accomplished by including domain experts early in the process. By building partnerships with linguistic experts across universities in low-resource language regions, we can help foster translation efforts on a large scale while providing high quality translation resources to the machine translation community. This partnership is not only important to the researchers, it also helps foster language education across the community. The community members are on either sides of the source and/or target language. Community members help provide the necessary resources such as native speakers that enables the process of translation. In addition, the technology developed by the machine translation efforts provides valuable resources that is utilized by the community. Not incorporating community members in the process of machine translation might be detrimental to the development and in some instance might be considered unethical.

5. Conclusion

In this paper, we survey the current state of machine translation tasks on Nigerian languages. We outline limitations and provide future directions on increasing research participation. While machine translation tasks on Nigerian languages is still in its infancy, there exists promising work in this field. In the future, we hope that more emphasis and mechanisms will be put in place to acquire high quality datasets and in addition generate diverse models which cater to the development of both low and high resource languages.

6. References

Abbott, J. Z. and Martinus, L. (2018). Towards neural machine translation for african languages.

Adelani, D. I., Ruiter, D., Alabi, J. O., Adebonojo, D., Ayeni, A., Adeyemi, M., Awokoya, A., and España-Bonet, C. (2021). Menyo-20k: A multi-domain english-yorùbá corpus for machine translation and domain adaptation.

Agić, Ž. and Vulić, I. (2019). JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy, July. Association for Computational Linguistics.

Ahia, O. and Ogueji, K. (2020). Towards supervised and unsupervised neural machine translation baselines for nigerian pidgin.

Akinwale, O., Adetunmbi, A. O., Obe, O., and Adesuyi, A. (2015). Web-based english to yoruba machine translation. *International Journal of Language and Linguistics*, 3:154.

Akinyemi, A. and Isiugo-abanihe, U. (2014). Demographic dynamics and development in nigeria : Issues and perspectives.

Alkhouli, T., Bretschner, G., Peter, J.-T., Hethnawi, M., Guta, A., and Ney, H. (2016). Alignment-based neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*, pages 54–65.

Allahbakhsh, M., Benatallah, B., Ignjatovic, A., Motahari-Nezhad, H. R., Bertino, E., and Dustdar, S. (2013). Quality control in crowdsourcing systems: Issues and directions. *IEEE Internet Computing*, 17(2):76–81.

Artetxe, M., Labaka, G., Agirre, E., and Cho, K. (2018). Unsupervised neural machine translation.

Ayegba, S. F., Osuagwu, O. E., and Okechukwu, N. D. (2014). Machine translation of noun phrases from english to igala using the rule-based approach. *West African Journal of Industrial and Academic Research*, 11:18–28.

Bahdanau, D., Cho, K., and Bengio, Y. (2016). Neural machine translation by jointly learning to align and translate.

Bird, S. (2020). Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Blench, R. (2012). An atlas of nigerian languages.

Bloodgood, M. and Callison-Burch, C. (2014). Using mechanical turk to build machine translation evaluation sets.

Brown, P., Cocke, J., Della Pietra, S. A., Della Pietra, V. J., Jelinek, F., Mercer, R. L., and Roossin, P. (1988). A statistical approach to french/english translation. In *Proceedings of the Second Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Comput. Linguist.*, 19(2):263–311, June.

Butryna, A., Chu, S.-H. C., Demirsahin, I., Gutkin, A., Ha, L., He, F., Jansche, M., Johny, C., Katanova, A., Kjartansson, O., Li, C., Merkulova, T., Oo, Y. M., Pipatsrisawat, K., Rivera, C., Sarin, S., de Silva, P., Sodimana, K., Sproat, R., Wattanavekin, T., and Wibawa, J. A. E. (2020). Google crowdsourced speech corpora and related open-source resources for low-resource languages and dialects: An overview.

Cheng, Y., Tu, Z., Meng, F., Zhai, J., and Liu, Y. (2018). Towards robust neural machine translation. *arXiv preprint arXiv:1805.06130*.

Cho, K., van Merrienboer, B., Bahdanau, D., and Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale.

Deng, Y. and Byrne, W. (2008). Hmm word and phrase alignment for statistical machine translation. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(3):494–507.

Eisele, A., Federmann, C., Saint-Amand, H., Jellinghaus, M., Herrmann, T., and Chen, Y. (2008). Using Moses to integrate multiple rule-based machine translation engines into a hybrid system. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 179–182, Columbus, Ohio, June. Association for Computational Linguistics.

Eludiora, S. and Ajibade, B. (2021). Design and implementation of english to yoruba verb phrase machine translation system.

Ezeani, I., Hepple, M., and Onyenwe, I. (2016). Automatic restoration of diacritics for igbo language. In Petr Sojka, et al., editors, *Text, Speech, and Dialogue*, pages 198–205, Cham. Springer International Publishing.

Ezeani, I., Rayson, P., Onyenwe, I., Uchechukwu, C., and Hepple, M. (2020). Igbo-english machine translation: An evaluation benchmark.

Forcada, M. L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Sánchez-Martínez, F., Ramírez-Sánchez, G., and Tyers, F. M. (2011). Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, 25(2):127–144.

Goyal, N., Gao, C., Chaudhary, V., Chen, P.-J., Wenzek, G., Ju, D., Krishnan, S., Ranzato, M., Guzman, F., and Fan, A. (2021). The flores-101 evaluation benchmark for low-resource and multilingual machine translation.

Gutkin, A., Demirsahin, I., Kjartansson, O., Rivera, C. E., and Túbòsún, K. (2020). Developing an open-source corpus of yoruba speech.

He, D., Xia, Y., Qin, T., Wang, L., Yu, N., Liu, T.-Y., and Ma, W.-Y. (2016). Dual learning for machine translation. *Advances in neural information processing systems*, 29:820–828.

Hedderich, M. A., Adelani, D., Zhu, D., Alabi, J., Markus, U., and Klakow, D. (2020). Transfer learning and distant supervision for multilingual transformer models: A study on african languages.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., et al. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Koehn, P. et al. (2005). Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007a). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007b). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180.

Lample, G. and Conneau, A. (2019). Cross-lingual language model pretraining.

Luong, M.-T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2019). A survey on bias and fairness in machine learning.

Meyer, T., Grisot, C., and Popescu-Belis, A. (2013). Detecting narrativity to improve english to french translation of simple past verbs. In *Proceedings of the Workshop on Discourse in Machine Translation*, pages 33–42.

Mikolov, T., Le, Q. V., and Sutskever, I. (2013). Exploiting similarities among languages for machine translation.

Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T., Fagbohungbe, T., Akinola, S. O., Muhammad, S., Kabongo Kabenamualu, S., Osei, S., Sackey, F., Niyongabo, R. A., Macharm, R., Ogayo, P., Ahia, O., Berhe, M. M., Adeyemi, M., Mokgesi-Selinga, M., Okegbemi, L., Martinus, L., Tajudeen, K., Degila, K., Ogueji, K., Siminyu, K., Kreutzer, J., Webster, J., Ali, J. T., Abbott, J., Orife, I., Ezeani, I., Dangana, I. A., Kamper, H., Elsahar, H., Duru, G., Kioko, G., Espoir, M., van Biljon, E., Whitenack, D., Onyefuluchi, C., Emezue, C. C., Dossou, B. F. P., Sibanda, B., Bassey, B., Olabiyi, A., Ramkilowan, A., Öktem, A., Akinfaderin, A., and Bashir, A. (2020a). Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online, November. Association for Computational Linguistics.

Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T., Kolawole, T., Fagbohungbe, T., Akinola, S. O., Muhammad, S. H., Kabongo, S., Osei, S., Freshia, S., Niyongabo, R. A., Macharm, R., Ogayo, P., Ahia, O., Meressa, M., Adeyemi, M., Mokgesi-Selinga, M., Okegbemi, L., Martinus, L. J., Tajudeen, K., Degila, K., Ogueji, K., Siminyu, K., Kreutzer, J., Webster, J., Ali, J. T., Abbott, J., Orife, I., Ezeani, I., Dangana, I. A., Kamper, H., Elsahar, H., Duru, G., Kioko, G., Murhabazi, E., van Biljon, E., Whitenack, D., Onyefuluchi, C., Emezue, C., Dossou, B., Sibanda, B., Bassey, B. I., Olabiyi, A., Ramkilowan, A., Öktem, A., Akinfaderin, A., and Bashir, A. (2020b). Participatory research for low-resourced machine translation: A case study in african languages.

Nguyen, T. Q. and Chiang, D. (2017). Transfer learning across low-resource, related languages for neural machine translation.

Nguyen, T. and Chiang, D. (2018). Improving lexical choice in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 334–343, New Orleans, Louisiana, June. Association for Computational Linguistics.

). World population projected to reach 9.8 billion in 2050, and 11.2 billion in 2100 | UN DESA | United Nations Department of Economic and Social Affairs.

Nowak, S. and Rüger, S. (2010). How reliable are annotations via crowdsourcing: A study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the International Conference on Multimedia Information Retrieval*, MIR '10, page 557–566, New York, NY, USA. Association for Computing Machinery.

Ogueji, K., Zhu, Y., and Lin, J. (2021). Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Onyenwe, I. E., Hepple, M., Chinedu, U., and Ezeani, I. (2019). Toward an effective igbo part-of-speech tagger. 18(4).

Orife, I. (2020). Towards neural machine translation for edoid languages.

Pires, T., Schlinger, E., and Garrette, D. (2019). How multilingual is multilingual bert?

Scott, B. and Barreiro, A. (2009). OpenLogos MT and the SAL representation language. In *Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation*, pages 19–26, Alacant, Spain, November 2-3.

Simard, M. (1998). Automatic insertion of accents in french text.

Stahlberg, F. (2020). Neural machine translation: A review and survey.

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Vanmassenhove, E., Hardmeier, C., and Way, A. (2018). Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008, Brussels, Belgium, October-November. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Williamson, K. and Blench, R. (2000). Niger-congo. *African languages: An introduction*, 1:42.

Wu, S. and Dredze, M. (2020). Are all languages created equal in multilingual bert?

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Zens, R., Och, F. J., and Ney, H. (2002). Phrase-based statistical machine translation. In *Annual Conference on Artificial Intelligence*, pages 18–32. Springer.

Zoph, B., Yuret, D., May, J., and Knight, K. (2016). Transfer learning for low-resource neural machine translation.