A Survey of Machine Translation Tasks on Nigerian Languages

Ebelechukwu Nwafor, Ph.D[1]. , Anietie Andy Ph.D[2].

[1]Department of Computing Sciences, Villanova University, Villanova, PA, 19805

[2]Penn Medicine, Center for Digital Health, University of Pennsylvania, Philadelphia, PA, 19104.

Abstract

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 research, 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.

Overview of Nigerian Languages

Nigeria consists of one of the most culturally diverse group of languages in Africa with over 500 indigenous languages spoken. There are three major languages (Igbo, Yoruba and Hausa) which are spoken by the major tribes in Nigeria. Most of the languages spoken are from the Niger-Congo origin with English as the official language. Out of all of the languages spoken in Nigeria, there are over 16 endangered Nigerian languages which are mostly spoken by indigenous groups from the northern part of Nigeria. The table below 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

Select Nigerian languages consisting of language families, estimated total number of speakers and geographic regions [2].

Machine Translation Techniques

- Rule-based Approach
  - This approach is based on understanding the linguistic properties of the source and target languages using dictionaries and expert knowledge to define grammar rules.
- 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.
- 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. 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 utilize transformer based models with encoder-decoder architecture.

State of the Art

Approach	Author	Paper	Description
Rule-based	S F Ayegba, O E Osuagwu, N D Okechukwu	Machine translation of noun phrases from English to Igala using the rule-based approach.	This approach utilizes noun phrases for 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
	Akinwale O. I., Adetunmbi A. O., Obe O. O., Adesuyi A. T.	Web-Based English to Yoruba Machine Translation	This approach utilizes a set of twenty rules which were specified using context free grammar.
	Safiriyu Eludiora, Benjamin Ajibade	Design and implementation of English to Yoruba verb phrase machine translation system.	The authors 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.
Statistics-based	Safiriyu Eludiora, Benjamin Ajibade	Automatic restoration of diacritics for Igbo language	This approach develops a model using the Igbo Bible corpus to detect and restore missing didactics in texts at word level toknization. Their approach on didatic replacement consists of using Hidden Markov Model in which the input text is viewed as a stochastic process.
	Ikechukwu E Onyenwe, Mark Hepple, Uchechukwu Chinedu, Ignatius Ezeani	Toward an effective Igbo part-of-speech tagger	This approach 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.
Neural-based	Iroro Orife	Towards neural machine translation for Edoid languages	The authors developed a neural machine translation model for translating Edoid languages to English utilizing transformer models with encoder decoder and multi-head self attention. The training was conducted using tokenization processes such as Byte-pair encoding (BPE) and word-level tokenization .
	Orevaoghene Ahia, Kelechi Ogueji	Towards supervised and unsupervised neural machine translation baselines for Nigerian pidgin	The authors 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 while experimenting with word-level and Byte-Pair encoding subword tokenization.
	Toan Nguyen, David Chiang	Nguyen, T. and Chiang, D. (2018). Improving lexical choice in neural machine translation	The authors 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.
	Kelechi Ogueji	Small Data? No Problem! Exploring the Viability of Pretrained Multilingual Language Models for Low-resourced Languages	The authors developed AfriBERTa, an approach which involves training multilingual models on low-resource language. 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.

Limitations

Limited Open Dataset: 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 to provide translation from native speakers.

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. 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. More emphasis needs to be placed on evaluating the fairness of machine learning (ML) and artificial intelligence (AI) algorithms 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

Community Partnership: 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.

Conclusion

In this research, 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.

References

1. 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.
2. 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.
3. Onyenwe, I. E., Hepple, M., Chinedu, U., and Ezeani, I.(2019). Toward an effective igbo part-of-speech tagger. 18(4).
4. Orife, I. (2020). Towards neural machine translation for edoid languages
5. Simard, M. (1998). Automatic insertion of accents in french text
6. Ahia, O. and Ogueji, K. (2020). Towards supervised and unsupervised neural machine translation baselines for Nigerian pidgin.