Natural Language Processing for Low-Resource Languages: Training and Integrating Speech Recognition Software for Low-Resource Wikimedia Languages

Ebube Chuba
Nkọwa okwu

Egbe Eugene
Wikimedia

Ijemma Onwuzulike
Nkọwa okwu

Abstract

This research proposes to build an automatic speech recognition (ASR) model from scratch for the low-resource languages within the Volta-Niger[1] language family which includes languages such as Igbo, Yoruba, Edo, Ikwerre, and at least 23 other languages. This model will be primarily trained on Igbo text-audio data collected from Igbo Wikipedia articles. The final model will showcase ~75% transcription accuracy for Igbo and ~30% without extra training data for other Volta-Niger languages.

Introduction

Low-resource African languages have extremely active speaker populations. The majority of the Nigerian population speaks at least two languages, one being a native language. However, writing and typing in native languages have proven to be tedious due to a lack of writing standards surrounding grammar and punctuation. Wikimedians User Group community members for low-resource languages face this challenge daily: inconsistent spelling, incorrect diacritic applications, and confusing grammatical structures. In recent years, speech has been shown to be one of the most intuitive forms of human-computer interaction[2]. Building a speech recognition system for an active language family will not only improve the existing language cataloging workflows but also push for advancements in natural language processing for African languages.

Date: June 1, 2024 - June 30, 2025

Related work

The team at Nkọwa okwu has built a model that is 51% accurate (WER 49) Igbo speech recognition model with approximately 20 hours of training and testing data. The model is currently available under private beta testing[3]. This model was fine-tuned based on OpenAI's Whisper model[4] which was trained on English data that greatly differs in vocabulary and tokens from the Igbo language. Retraining with no reliance on English data will likely give better results. All data was collected on Nkọwa okwu's Igbo data entry platform, the Igbo API Editor Platform.

Methods

We are working in partnership with the Igbo Wikimedians User Group and will contract

audio recorders from this community to record, review, and annotate audio for the 250,000+ Igbo sentences available on Wikipedia. Audio recorders will use their mobile devices to record, upload, and annotate their audio.

As data is getting collected, primary researchers, linguists, and project collaborators will be annotating, organizing, and analyzing the data into knowledge domains (i.e. education, politics, medicine, transportation, etc.) and speaker demographics (i.e. gender, dialect, age, etc.) to have a clear break down of the type of collected data.

After organizing the dataset, our team will perform benchmark tests to better understand which specific knowledge domains, types of speakers, and other Volta-Niger languages the model understands the best.

Expected output

- Release the Igbo audio-text dataset for 250,000+ Igbo sentences for future researchers to use in their projects.
- Deploy an API that can be accessed by developers and researchers.
- Publish a demo website to allow anyone to test and provide feedback for the model.
- Work closely with Igbo Wikimedians User Groups to understand how to directly integrate the model into existing Wikimedians data collection tools to improve their workflows.
- Create training sessions with Igbo Wikimedians User Groups to teach community members how to use the technology.

Risks

Potential risks include:

- Audio recorders not best understanding the rules and expectations for recording audio for each sentence.
- Speaker demographic imbalances in categories such as gender, dialect, age, etc.

Community impact plan

We will work with the Igbo Wikimedians User Group leaders to introduce the technology in their communities' contribution workflows. Our team will also work with community leaders to create and lead training sessions on how to use the model which will empower the entire Wikimedians User Group community to contribute more to Wikipedia by using speech. By introducing this technology to this community, contributors will be able to produce more and easy-to-edit low-resource language content, increasing the language's online presence and discoverability.

Evaluation

The main metric of success for this project is Word Error Rate (WER) to determine the model's accuracy. The WER is the ratio of errors in the transcription produced by the model to the total words spoken. WERs will be collected and analyzed based on different speakers, discussion topics, and language.

In the event that evaluation is done via translating between languages in the Volta-Niger family, we may use Bleu Score to accurately determine the quality of these translations.

Budget

The budget for this research project is $50,000 USD. Primary researchers (Chuba, Onwuzulike)

except Eugene will not be paid under this budget. A select number of Igbo Wikimedians User Group members will be contracted as audio recorders to collect language data. Volta-Niger linguists will be contracted to oversee data quality and establish data collection work process standards and expectations. Recording equipment and software costs will be covered under the budget. Finally, up to 10% of the budget will be reserved for emergency costs.

Prior contributions

Ijemma Onwuzulike founded the non-profit Nkọwa okwu with the focus on building open-source low-resource African language technology to minimize the barrier to interacting with computers. Her organization has built the Igbo API[5] which is an Igbo-English dictionary API that exposes 75,000 Igbo words and sentences. Each word and sentence is accompanied with dialectal variations, definitions, audio recordings, and 10 more metadata features. Additionally, her team built IgboSpeech, which is a 51% accurate Igbo automatic speech recognition model. Finally, her team built the Igbo API Editor Platform[6] which is an open-source data collection platform that has more than 500 active volunteers contributing to the data.

Egbe Eugene is a Wikimedian technical contributor and community builder who has been growing a community of developers for Wikimedia especially in the African region since 2016, User:Eugene233. He has worked on the Wikipedia project Scribe Project[7]: a project for collecting references used in citing while writing articles in low-resource languages which led to the Scribe[8] gadget. He has collaborated as a contractor with the Goethe Institute and is currently supporting the African German phrasebook project[9] which has led to the building of the AGPB web and mobile app[10] with contributions to over 3000 contributions on Wikidata and Wikimedia Commons in over 30 African languages[11]. He has also served as a technical consultant for the Wiki Mentor African Project [12].

Ebube Chuba is a Lead AI Platform Engineer at Virtualitics, an AI startup. Previously, he's also worked as a researcher at IBM Research Almaden where he developed several patents [13] and contributed to the academic literature in the fields of federated learning [14] and bias in ML systems [15]. He was also a primary contributor to ART [16], an open source repository hosting the state of the art methods in attacks and defenses to AI systems. Finally, he was the cofounder and technical director of CHIIN [17], a nonprofit whose goal is to send critical medical information to healthcare providers through SMS in regions lacking WiFi.

References

You can choose which format you use for showing the references. References will not be counted against your word limit.

[1] Wikipedia contributors, "Volta–Niger languages," *Wikipedia*, Dec. 29, 2022. https://en.wikipedia.org/wiki/Volta%E2%80%93Niger_languages

[2] A. A. Alnuaim *et al.*, "Human-Computer interaction for recognizing speech emotions using multilayer perceptron classifier," *Journal of Healthcare Engineering*, vol. 2022, pp. 1–12, Mar. 2022, doi: 10.1155/2022/6005446.

[3] "IgboSpeech API Demo." https://speech.igboapi.com/

[4] "Introducing whisper." https://openai.com/research/whisper

[5] Nkọwa okwu, "Igbo API." https://igboapi.com

[6] “Igbo API Editor Platform.” https://editor.igboapi.com/

[7] “Grants:Project/Frimelle and Hadyelsahar/Scribe: Supporting under-resourced Wikipedia editors in creating new articles - Meta.” https://meta.wikimedia.org/wiki/Grants:Project/Frimelle_and_Hadyelsahar/Scribe:_Supporting_Under-resourced_Wikipedia_Editors_in_Creating_New_Articles

[8] “Scribe - Meta.” https://meta.wikimedia.org/wiki/Scribe

[9] “Wikidata:WikiProject African-German phrases for Sub-Sahara Africa - Wikidata.” https://www.wikidata.org/wiki/Wikidata:WikiProject_African-German_Phrases_for_Sub-Sahara_Africa

[10] “African German Phrase Book - apps on Google Play.” https://play.google.com/store/apps/details?id=com.agpb&pcampaignid=web_share

[11] “Category:African German Phrasebook Project - Wikimedia Commons.” https://commons.wikimedia.org/wiki/Category:African_German_Phrasebook_Project

[12] “Wikidata:Wiki Mentor Africa - Wikidata.” https://wikidata.org/wiki/Wikidata:Wiki_Mentor_Africa

[13] “Google Patents.” https://patents.google.com/?inventor=%22ebube+chuba%22&oq=inventor:+%22ebube+chuba%22

[14] H. Ludwig, “IBM Federated Learning: an Enterprise Framework White Paper V0.1,” *arXiv.org*, Jul. 22, 2020. https://arxiv.org/abs/2007.10987

[15] A. Abay, “Mitigating bias in federated learning,” *arXiv.org*, Dec. 04, 2020. https://arxiv.org/abs/2012.02447

[16] Trusted-Ai, “GitHub - Trusted-AI/adversarial-robustness-toolbox: Adversarial Robustness Toolbox (ART) - Python Library for Machine Learning Security - Evasion, Poisoning, Extraction, Inference - Red and Blue Teams,” *GitHub*. https://github.com/Trusted-AI/adversarial-robustness-toolbox

[17] “CHIIN.” https://chiin.org/