Teaching Endangered World Languages with a Conversational Artificial Intelligence Application

CS4991 Capstone Report, 2023
Charles Beall
Computer Science
The University of Virginia
School of Engineering and Applied Science
Charlottesville, Virginia USA
ceb7zn@virginia.edu

ABSTRACT

Languages are an integral part of global cultures, but many languages are losing native speakers and are on track to die out as they are not being taught to younger generations. To teach endangered languages with dwindling numbers of native speakers, I propose that communities turn to natural language processing (NLP) and generating chatbots that are trained to speak endangered languages. Implementing this solution would require the use of NLP algorithms optimized to work with small amounts of language data. The tool would need to be trained to produce speech, listen, and process text to interact with users. The desired outcome of the development of a product used for this specific purpose would enable people in communities of endangered languages to pass their language on through traditional language acquisition to future generations. Future work would involve long-term studies to validate the usefulness of these tools.

1. INTRODUCTION

Imagine a world where nobody spoke English anymore, and what that would mean for all media, literature, and cultural identity associated with the English language. This is the reality for many communities around the globe whose languages are endangered. Endangered languages are those that are at risk of remaining without any speakers soon (Woodbury, n.d.). Woodbury notes that, even though languages that are endangered are those that are not generally seen as important to the global order, these languages make up the majority of those that are spoken today and are at risk of falling out of use by the end of the century. Languages can become endangered and die for reasons such as language evolution, genocide, social or legal pressure (Woodbury, n.d.). Language is linked to culture, so the loss of languages in turn leads to the loss of a large part of communities' cultural identity, traditions, and history.

Even though many languages are dying or have become extinct, efforts can be made to revitalize or revive them. One example of a successful language revival is Hebrew, which at one point was a dead language, but is now spoken by millions of people. On the other hand, the Irish language has been on a slow path to revitalization despite all efforts (UNESCO, n.d.). Even so, many languages are not so fortunate as to have government- and community-led revitalization efforts. There are varying methods to revitalize a language, but for success the focus must be on the creation of new speakers (Rehg & Campbell, 2018). To teach new speakers of a language, there must be resources available for educational use.

2. RELATED WORKS

While language revitalization can be and has been done without the use of modern technology, software tools can bolster community efforts. One language learning platform that many are already familiar with is Duolingo. Over the past decade Duolingo has worked to add support for many endangered languages in addition to its more popular ones (Griffiths, 2019). Adding Irish to Duolingo has been valuable because it provides an easy to access resource for Irish language learners across the globe. Fortunately, adding languages to Duolingo does not take as much effort as it did for the first languages, meaning that the possibility of expanding to endangered languages is a feasible way to provide value to the company and its users (Griffiths, 2019).

Another program that can be used as a language learning tool is ChatGPT. This program can process and generate text in multiple languages and hold text conversations (Simon, 2023). ChatGPT can simulate different situations, so users can focus on the specific vocabulary or grammar that they want to practice. Simon also describes how it is also possible to use ChatGPT for reading and listening comprehension practice, or to generate study plans and exercises. It has even already been harnessed to such ends by the app Quizlet for language learning (Bayer, 2023). With all these capabilities, ChatGPT can make for a valuable language learning resource.

A different technology that acts more like a conversation partner is a voice bot, such as Amazon's Alexa. (Microsoft, 2023). They can understand the speech of users in addition to responding with audio output (UNESCO, 2023). This idea, in tandem with the ideas behind Duolingo and ChatGPT, is key to the construction of a product designed as conversational AI to facilitate endangered language instruction.

Other important related works critical to the success of this project are in the field of training language models on small datasets. Since this project would focus on endangered languages, it would be necessary to employ methods such as transfer learning (Orynycz, 2022) and pretraining models on monolingual data before training on low-resource languages (Zheng et al., 2021), in addition to gathering training data from resources such as radio archives when there are limited amounts of literature or media (Doumbouya et al., 2021).

3. PROPOSED SYSTEM DESIGN

The system I propose would need to be built from the ground up as a web and mobile application by a team of engineers. The project would take structural inspiration from the previously mentioned applications, each of which have qualities that lend themselves to this system's design. This system would be made up of a user interface constructed by frontend code, and further logic that controls processes invisible to the user. To scale, processes would be broken down into microservices, and cloud containers would be used. Additionally, machine learning models would have to be trained and evaluated on datasets of language content. Together, each of these components would make up the entire application, as discussed below.

3.1 User Interface Design

The user interface would be rather simple to be leveraged by users from different language backgrounds with varying levels of technological literacy. At the front and center of the main page would be the face of a human avatar, meant to represent the bot that the user is conversing with. The avatar's mouth and face would move to mimic talking and create expressions. Directly below would be a large rectangle containing all interaction methods between the user and the bot. At the top of this rectangle, there would be text representation of the bot's speech, with the option to toggle the text on and off, as well as an option to toggle the audio response on and off. Initially, it would display a welcome message. At the bottom of this rectangle would be an input text bar with a send button to the right, and a button to start and end audio recordings for the user's speech to the left. Just above this section would be some suggested actions for the user to take, such as "Talk", "Lesson", or "Exercises". Taking up a small portion of the left-hand side of the main page would be a menu including details about the user's profile, and a list of previous conversations and lessons. There would be the option to toggle the menu open and closed. These specifications would allow the user to interact with the bot via text or audio and customize the page to fit their preferences.

3.2 System Architecture

The overall system would require a variety of services calling for collaboration among the engineers working on each of its parts. System needs would include a web server to host and display application content and use Websockets for client-server communication to handle user requests. The bulk of the development would occur in a web framework such as Django, which would mesh with a frontend framework such as React. The API would be constructed to handle each of the potential interactions that the user can have with the page, implementing a microservice architecture. The user interface would be constructed using React and HTML, CSS, and TypeScript for structure, style, and control. The web framework would connect to the system's database for API calls that require data retrieval. The system's web server and database systems would be hosted in a container system such as Docker.

3.3 Machine Learning

The proposed system would need to be able to leverage machine learning and NLP to train a Natural Language Understanding unit that can communicate with and understand users. For this to work, the machine learning model would need to be trained on a large dataset made of language education resources and other multimedia in target languages with tools such as scikit-learn and NLTK. To further improve the model, a team would need to evaluate and provide feedback to its actions. Since endangered languages have limited resources available, we could utilize data augmentation to increase training resources in addition to the methods for training on limited amounts of language data as described above in the Related Works section.

3.4 Challenges

The limited number of resources for endangered languages is likely to be one of the biggest challenges for this project. Another potential challenge would be that the interactions of the bot with the users would be based upon the multimedia that it is provided in training, which could be biased towards certain ideas, regional dialects, and target audiences. If these challenges persist, users could see inaccurate use of the target language by the program. Fine-tuning the program to use accurate language may also pose a challenge, because it would be increasingly difficult to find people to give it feedback as the number of speakers for a language decreases.

4. ANTICIPATED RESULTS

Once development is complete, the application would need to be able to run without any security or reliability issues and would need to be able to support users on the site. To accomplish this, the bot would be able to respond to users appropriately and correctly in any of the supported languages by sustaining audio and text conversations, creating lessons, and creating language exercises for users. As a result, this portable application could be used on computers or mobile devices, and by language learners in both educational and informal settings. By using this product as a teacher and a conversational partner, users would be able to learn their target language to an extent that they feel confident to use it in their local community, or in online communities.

5. CONCLUSION

Language and culture go together, and as languages fall out of use over time the knowledge of many cultures will also disappear. Developing this project would not be the end-all solution to this problem, but it would be a key tool in allowing communities to combat it. The customizable conversational design and vast linguistic knowledge of this program would offer a comfortable and effective way for users to learn their target language, while access over the internet would allow people from across the globe to use the service. The focus on endangered languages would have to overcome the issue of limited resources, but NLP techniques to work around this challenge would be used to train the language model. As a result, this project has the potential to empower communities that have yet to see the benefits that language technology can bring to the table.

6. FUTURE WORK

Since this project only exists as an idea thus far, it would need to be built from scratch. One of the first steps would be assembling a team of engineers and scientists to go through the entire software development process and train the AI models being used. Another necessary task is spending a lot of time and money collecting the necessary data for low-resource languages so that the models can be trained in the first place. Further research will need to be done to improve the quality of the language models based on small amounts of data relative to other languages so that users can count on the program's accuracy, expanding on the current progress of the related works in the field. The resource-intensity that the undertaking of this project would require would likely need outside funding, and studies to evaluate its effectiveness.

REFERENCES

Bayer, L. (2023, March 1). *Introducing Q-chat, the world's first AI tutor built with OpenAI's ChatGPT*. Quizlet. https://quizlet.com/blog/meet-q-chat

Doumbouya, M., Einstein, L., & Piech, C. (2021, May). Using radio archives for low-resource speech recognition: towards an intelligent virtual assistant for illiterate users. In *Proceedings of the AAAI Conference on Artificial Intelligence*(Vol. 35, No. 17, pp. 14757-14765).

Griffiths, J. (2019, October 4). *The internet threatened to speed up the death of endangered languages. could it save them instead? | CNN business*. CNN. https://www.cnn.com/2019/10/04/tech/duolingo-endangered-languages-intl-hnk/index.html

Learning a New Language with ChatGPT. Microsoft. (2023, May 5). https://www.microsoft.com/en-us/microsoft-365-life-hacks/writing/using-chatgpt-for-foreign-language-learning

Orynycz, P. (2022, May). Say It Right: AI Neural Machine Translation Empowers New Speakers to Revitalize Lemko. In *International Conference on Human-Computer Interaction* (pp. 567-580). Cham: Springer International Publishing.

Rehg, K. L., & Campbell, L. (2018). Abstract. In *The Oxford Handbook of Endangered Languages*. essay, Oxford University Press.

Simon, E. (2023, July 22). *6 ways to use CHATGPT to learn a foreign language: ICLS: International Center for Language Studies: Washington D.C.* ICLS. https://www.icls.edu/6-ways-to-use-chatgpt-to-learn-a-foreign-language/

UNESCO WAL. (n.d.). https://en.wal.unesco.org/

Voice Bots and conversational AI. Voice bots and Conversational AI | Microsoft Power Virtual Agents. (2023). https://powervirtualagents.microsoft.com/en-us/voicebots-conversational-ai/

Woodbury, A. C. (n.d.). *What is an endangered language? - Linguistic Society of America*. Linguistic Society of America. https://www.linguisticsociety.org/sites/default/files/Endangered_Languages_0.pdf

Zheng, F., Reid, M., Marrese-Taylor, E., & Matsuo, Y. (2021, June). Low-resource machine translation using cross-lingual language model pretraining. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas* (pp. 234-240)