Grapheme-To-Phoneme Models for Cross-Lingual Transfer in South African Languages

Johannes Abraham Louw
Council for Scientific and Industrial Research (CSIR)

Abstract

The digital divide disproportionately affects speakers of low-resourced languages. This work introduces a novel approach to address this disparity by leveraging publicly available data and scraping information from Wiktionary to train models capable of generating word pronunciations. The envisioned models have dual utility: firstly, they empower the creation of speech technologies tailored to serve under-resourced languages, and secondly, they facilitate the generation of new pronunciations, thereby contributing to the expansion of entries on Wiktionary.

Introduction

The scarcity of pronunciation sources for low-resource languages poses a significant challenge in the development of speech technologies tailored to bridge the digital divide [1]. However, innovative approaches, such as leveraging multilingual grapheme-to-phoneme (G2P) models, offer a promising solution to address this gap [2]. By training on diverse datasets from sources like Wiktionary, the models can generalize patterns that extend to low-resource languages in a cross-lingual transfer learning manner [3]. The developed multilingual G2P models can be employed to transcribe new pronunciation sources for low-resource languages in Wiktionary, recorded sources such as Lingua Libre[1], as well as provide a valuable resource for Wikispeech and their aim to develop TTS voices.

From a scientific perspective, this study aims to compare G2P models trained on Wiktionary-scraped data, similar to the approach outlined in [4], against models trained on existing pronunciation sources as documented in [5]. The work will primarily be done in the 11 official spoken South African languages and all of the collected data and models will be made publicly available, but the outcomes should be applicable to any low-resourced language. Commencing in June 2024, this research is planned to conclude by December 2024.

Related work

The study that closely aligns with our proposed research was conducted in [4], where Wiktionary data was scraped to develop pronunciation dictionaries and train generic

[1] https://lingualibre.org/

G2P models. Our proposed research differs from this prior work in two key aspects. Firstly, we intend to train *multilingual* G2P models, specifically aiming to induce cross-lingual transfer. This implies supplementing data from low-resourced languages with information from resource-rich languages, with the goal of enhancing the performance of G2P models for low-resourced languages. This is similar to the work done in [2], but our attention will at this point in time be limited to the official South African spoken languages. Secondly, we will conduct a comparative analysis between established pronunciation dictionaries [5] and those derived from Wiktionary, seeking to establish criteria for determining the requisite volume of data essential for training baseline G2P models that are adequate for speech technology applications.

Methods

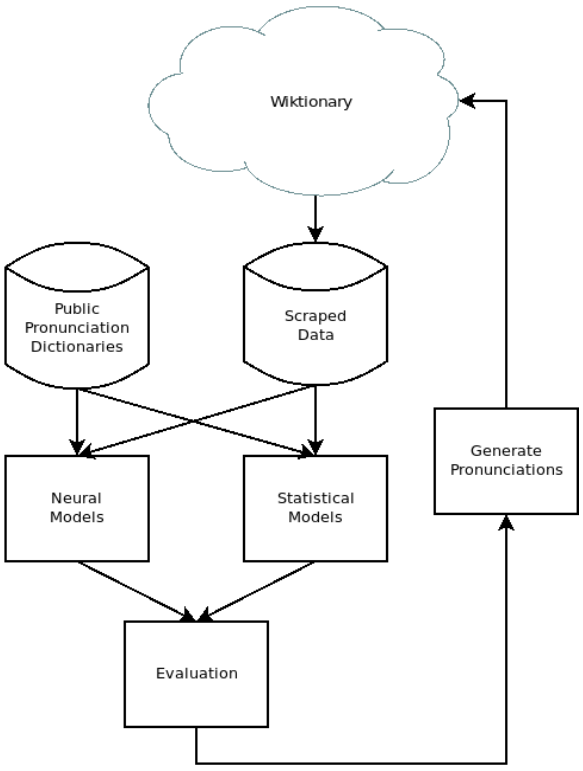

Figure 1: Proposed method

The project will commence with an extensive literature review focusing on the efficacy of statistical and neural models in G2P learning through cross-lingual transfer. Subsequent to this, a data collection process will ensue, encompassing conventional pronunciation dictionaries, such as those referenced in [5], and harvested data from Wiktionary, analogous to the approach in [4]. Data cleaning and appropriate segmentation for training purposes will be conducted.

The models identified in the literature will be translated into executable code, and the identified data sets will be utilized for training purposes. A comprehensive evaluation will be executed, and the findings will be disseminated through presentation at a suitable conference. All code, scripts, and data sources will be released as open-source, fostering transparency and collaboration within the research community. Figure 1 gives an overview of the proposed method.

Expected output

The outputs for the project include a research paper and open-source contributions:

1. A research paper will be written, with the main aim to a) compare the differences in error rates between G2P models trained from established datasets and those derived from Wiktionary; b) compare G2P models based neural networks and those based on traditional statistical approaches; and c) analysis of error rates in trained G2P models.
2. Open-source all G2P models, data and scripts.
3. Open-sourced scripts for creating new Wiktionary pronunciation entries.

Risks

The availability and retention of skilled personnel, especially in specialized fields such as multilingual G2P modeling, can be a challenge. Turnover or unexpected departures may impact project continuity and should be planned for and mitigated. The project being funded in dollars while expenses are incurred in South African rand exposes the project to currency exchange rate fluctuations. Unfavorable changes may impact the budget and financial planning.

Community impact plan

Although this study will be rooted in the scientific exploration of the ability of neural G2P models to perform cross-lingual transfer in multilingual environments, the primary impact will be the ability of said models to expand the IPA pronunciation definitions of the entries in low-resourced South African languages in Wiktionary. The resultant entries can serve as valuable resources for the development of speech technologies facilitating communication across language barriers as well as repositories of knowledge.

Evaluation

The outcomes of the project will be evaluated using standard objective measures used in evaluating G2P models, phone-error-rate (PER) and word-error-rate (WER) (both minimum edit distance [6] measures). The proposed multilingual neural models will also be evaluated in their ability to perform cross-lingual transfer learning.

Budget

The exchange rate used was R18.50/$, which is subject to change and has been identified as one of the risks.

Budget		
Task	**Manpower**	**Running**
Work package 1		
Literature study		
Documentation		
Paper writing		
	$11,351.35	$2,702.70
Work package 2		
Resource collection		
Data processing		
	$7,837.84	
Work package 3		
Baseline models (statistical)		
Neural models		
Architecture experiments		
Training		
Evaluation		
	$13,513.51	
Work package 4		
Source code publication and curating		
Wiktionary generating code		
	$7,837.84	
Sub Total	$40,540.54	$2,702.70
Institutional overhead	$4,864.86	
Project Total	**$48,108.11**	

Prior contributions

The author has been actively publishing in the speech technology field since 2004 (https://scholar.google.com/citations?user=chGDsc4AAAAJ&hl=en) and the research group at the CSIR of which the author is a member has been actively working on human language technology since 2003. The author has made contributions to various open-source speech technology packages [7,8,9,10,11].

References

[1] J.A. Louw and A. Moodley, "Rhonda: The architecture of a multilingual speech-to-speech translation pipeline." in Proc International Conference on Intelligent and Innovative Computing Applications (ICONIC), Plaine Magnien, Mauritius , Dec. 2018, pp. 1-7.

[2] J. Zhu, C. Zhang, and D. Jurgens, "ByT5 model for massively multilingual grapheme-to-phoneme conversion". in Proc. INTERSPEECH 2022 – 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, Sep. 2022, pp. 446-450.

[3] D. Wang, and T.F. Zheng, "Transfer learning for speech and language processing." in Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), Hong Kong, Dec. 2015, pp. 1225-1237.

[4] J.L. Lee, L.F. Ashby, M.E. Garza, Y. Lee-Sikka, S. Miller, A. Wong, A.D. McCarthy and K. Gorman, "Massively multilingual pronunciation modeling with WikiPron." in Proc. The 12th Language Resources and Evaluation Conference, Marseille, France, May 2020, pp. 4223-4228.

[5] K.V. Calteaux, F. De Wet, C. Moors, D.R. van Niekerk, B. McAlister, A.S. Grover, T. Reid, M. Davel, E. Barnard and C. van Heerden, "Lwazi II Final Report: Increasing the impact of speech technologies in South Africa". Technical Report, 2013.

[6] G. Navarro, "A guided tour to approximate string matching." ACM computing surveys (CSUR) 33, no. 1, pp. 31-88, 2001

[7] https://github.com/mmorise/World

[8] https://github.com/CSTR-Edinburgh/merlin

[9] https://github.com/TensorSpeech/TensorFlowTTS

[10] https://github.com/waywardgeek/sonic

[11] https://github.com/festvox/festival/blob/master/COPYING