Experimental Protocol

XCS224U - Natural Language Processing, Fall 2022

Mario Peng Lee

1 Hypotheses

Phonetics, despite being a core component in the field of Linguistics, remains relatively unexplored as compared to semantics and syntax in language models. The main focus of my research is to explore whether a transformer language model can yield interesting results on phonetic tasks. Traditional transformers or TTS (text-to-speech) models are not designed to respond in tangible phonetic output, therefore ill-equipped for tasks such that require a phonetic response as analysis.

To resolve this question, I plan to design and train a PLM (phonological language model). This model will read in normal graphemic English text and output in IPA (international phonetic alphabet) phonemes. I seek to answer whether phonetic analysis leads to a deeper exploration of the linguistic understanding of language models.

More specifically, my main hypotheses are:

- A PLM will not perform worse than a conventional language model.
- A PLM will yield human-aligned results on phonetic tasks such as the wug test (Gleason), inspired by previous research in Neural Morphological Inflection Models (Liu and Hulden)
- A PLM will predict the correct pronunciation of borrowed or made up words, and even match human predictions.
- A PLM will be able to give accurate english-to-IPA transcriptions of unprecedented words.
- There will be a specific model size or configuration that will start yielding correct results, and I will be able to pinpoint the exact moment when the model starts understanding certain phonological rules.

2 Data

In terms of evaluation, for the model's G2P (Grapheme to Phoneme) performance, I will be using the SIGMORPHON Shared Task G2P, a dataset that maps English words to their respective IPA transcription(Pimentel et al.). This is to ensure the PLM was successfully implemented and is able to give accurate IPA transcriptions of unseen English words. Then, to evaluate the language modeling, the OpenWebText (Gokaskan et al.) data with each word undergoing an IPA conversion through the OpenDict IPA dictionary or the SIGMORPHON shared task dataset will be used. Finally, for wug-test-like evaluation, the SIGMORPHON Shared Task in Morphological Inflection Generation evaluation dataset will be used, which compares the model output to human production probabilities.

3 Metrics

In order to assess the English to IPA performance, I will use word error rate (WER), which is the percentage of words for which the hypothesized transcription sequence does not match the gold transcription. Since this is a straightforward task the evaluation metric will be accordingly simple.

To evaluate how natural sounding the phonetic transcription of borrowed words or made up words are, I will conduct a survey on native English speakers from my Stanford XCS224U and UCLA LING185A classes. I will compare the results between humans and the model, using the human dataset as the gold transcription to get a WER on this task.

Finally, to assess wug-test performance, accuracy against human-probability data, a macro F1 scores.

4 Models

For my baseline models, I will utilize 3 different checkpoints of T5 (Raffel et al.), the text-to-text transfer transformer model, provided by huggingface. It is an encoder-decoder multi-task multilingual pretrained model.

- T5 small: 60 million parameters
- T5 base: 220 million parameters
- T5 large: 770 million parameters

Afterwhich, I have planned to train a neural model from scratch using only IPA symbols. The specifics of this model are note detailed out yet.

5 General Reasoning

By finetuning these baseline T5 pretrained models, we might notice some emergent behavior regarding IPA understanding, if this is the case, we can extrapollate the findings to develop questions such as what would happen if a model was solely trained in IPA, or how does the size of a language model correlate to its phonological understanding. This task will not be easy and my experiment might not yield significant results, however, I believe the concept of a phonological language model is still worth exploring through this method.

6 Summary of Progress So Far

Progress so far can be seen in a Google Colab[1].

Data

The OpenDict for English to IPA data was parsed into test and evaluation sets, I also managed to start translating openwebtext into IPA using a simplistic linear parser over the OpenDict dataset, however results were not great since openwebtext contains many non alphanumeric symbols, therefore I will either have to clean them up using an algorithm or simply discard the idea of developing an IPA labelled corpus due to the time constraints of this course.

Training

I've set up the functions and environment for training T5 using the huggingface library. Most of the time was spent ensuring the code did not crash and could run properly. As of lately, T5 small was trained with the OpenDict dataset, yielding a minimum loss of 0.42 so far (currently training epoch 2). This result should improve once I start tuning hyperparameters and make more modifications to the dataset and training algorithm.

Next Steps

My next objective is to train the different sizes of T5 and evaluate them on the metrics. Depending on the results, one of the models will be taken as the main evaluator for the wug test and other hypotheses.

Obstacles & Reflections

It seems like computation will be my most significant bottleneck in this project. Running T5 small over a section of the data requires at least 6 hours of local computation. Alternatives will have to be

[1] https://colab.research.google.com/drive/1YwSLszEWBhstnHXIVCW44KOiPPSxBpr-?usp=sharing

sought. Perhaps trying to train T5 is unrealistic given the capabilities within my reach for this project's duration.

7 References

1. Aaron Gokaslan, Vanya Cohen, Ellie Pavlick, Stefanie Tellex 2019. "Open Web Text". Online: http://Skylion007.github.io/OpenWebTextCorpus.
2. Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," October. https://doi.org/10.48550/arXiv.1910.10683.
3. Jean Berko Gleason. 2014. "Wug Test and the Elicited Production Paradigm." In *Encyclopedia of Language Development*, 687–88. Thousand Oaks: SAGE Publications, Inc. https://doi.org/10.4135/9781483346441.
4. Ling Liu and Mans Hulden. 2021. "Can a Transformer Pass the Wug Test? Tuning Copying Bias in Neural Morphological Inflection Models." arXiv. https://doi.org/10.48550/arXiv.2104.06483.
5. Tiago Pimentel, Maria Ryskina, Sabrina J. Mielke, Shijie Wu, Eleanor Chodroff, Brian Leonard, Garrett Nicolai, et al. 2021. "SIGMORPHON 2021 Shared Task on Morphological Reinflection: Generalization Across Languages." In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, 229–59. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.sigmorphon-1.25.