Fine-tuning mSLAM for the SIGMORPHON 2022 Shared Task on Grapheme-to-Phoneme Conversion

Dan Garrette
Google Research
dhgarrette@google.com

Abstract

Grapheme-to-phoneme (G2P) conversion is a task that is inherently related to both written and spoken language. Therefore, our submission to the G2P shared task builds off of mSLAM (Bapna et al., 2022), a 600M parameter encoder model pretrained simultaneously on text from 101 languages and speech from 51 languages. For fine-tuning a G2P model, we combined mSLAM's text encoder, which uses characters as its input tokens, with an uninitialized single-layer RNN-T decoder (Graves, 2012) whose vocabulary is the set of all 381 phonemes appearing in the shared task data. We took an explicitly multilingual approach to modeling the G2P tasks, fine-tuning and evaluating a single model that covered all the languages in each task, and adding language codes as prefixes to the input strings as a means of specifying the language of each example.

Our models perform well in the shared task's "high" setting (in which they were trained on 1,000 words from each language), though they do poorly in the "low" task setting (training on only 100 words from each language). Our models also perform reasonably in the "mixed" setting (training on 100 words in the target language and 1000 words in a related language), hinting that mSLAM's multilingual pretraining may be enabling useful cross-lingual sharing.

References

Ankur Bapna, Colin Cherry, Yu Zhang, Ye Jia, Melvin Johnson, Yong Cheng, Simran Khanuja, Jason Riesa, and Alexis Conneau. 2022. mSLAM: Massively multilingual joint pre-training for speech and text. *arXiv preprint arXiv:2202.01374*.

Alex Graves. 2012. Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711*.

Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 250–250