A Large-Scale Data Collection from Internet for Thai Language and Speech Processing

Chompakorn Chaksangchaichot, Korakot Chaovavanich

VISTEC-DEPA Artificial Intelligence Research Thailand
{chompakornc_pro, korakotc_pro}@vistec.ac.th

Abstract

The acoustic model, lexicon, and language model are fundamental building blocks for any traditional speech recognition model. Despite having a plentiful resource for training language models in Thai, the resource for training a good lexicon and acoustic model is still scarce. Many open-source grapheme-to-phoneme (G2P) datasets used to train a lexicon have different annotation guidelines, making it hard to combine many datasets into one to build a good model. In this work, we aim to create an open-source dataset for both lexicon and the acoustic model by combining various G2P datasets with different annotation guidelines and crawling license-free videos from available websites.

Unified Thai G2P

There are many G2P datasets that are publicly available. For example, NECTEC's LOTUS corpus (Kasuriya et al., 2003), Thainotes (Thai Notes, n.d.), Wikitionary (Wikitionary, n.d.), and so forth. However, these datasets share different annotation standards, which makes combining impossible. For example, in NECTEC's LOTUS Corpus (Kasuriya et al., 2003), we label phoneme for "ออ" as "@@" while Thainotes (Thai Notes, n.d.) use "ɔɔ." Thus, our work aims to standardize these different annotation guidelines by looking at mutual samples on different dataset and perform simple mapping rules.

License-free Speech Dataset: Thai MOOC

There are very few open-source datasets for Thai speech recognition. For instance, NECTEC's LOTUS corpus (Kasuriya et al., 2003) and Gowajee Smart home corpus (Chuangsuwanich et al., 2018). Despite these two datasets combining, they both merely reach 10 hours of speech-transcription pairs. Therefore, it is impossible to make a good speech recognition model out of these two datasets. We aim to create an open-source speech recognition dataset by scraping videos and transcription from the internet. We choose Thai-MOOC websites (Thai MOOC, n.d.) which contain online courses with video transcription where all of its videos are Creative Common License.

Current states and Future Works

The current state of G2P dataset is in progress of comparing different G2P standards from different datasets. We have downloaded most G2P datasets and are currently working with mapping rules for different phonemes.

For the speech dataset, we used Scrapy (Kouzis-Loukas, 2016) to crawl a total of 305 available online courses from Thai MOOC (Thai MOOC, n.d.). Still, there are several issues about the subtitles to be addressed: bad subtitle alignment, too many special tokens, and [music] token subtitles. Currently, we are in the process of fixing these issues before releasing the dataset to the public.

Conclusion

Our work presents an in-progress dataset creation for building a lexicon and speech recognition model. These datasets are obtained by gathering available resources on the internet and will be publicly available to the community. We hope that our work will benefit Thai language processing community by releasing these dataset publicly.

Reference

Gowajee: Chuangsuwanich, E., Suchato, A., Karunratanakul, K., & Naowarat, B. (2018). Gowajee Corpus. Chulalongkorn University, Faculty of Engineering, Computer Engineering Department. https://github.com/ekapolc/gowajee_corpus

Kasuriya, S., Sornlertlamvanich, V., Cotsomrong, P., Kanokphara, S., & Thatphithakkul, N. (2003, June). Thai speech corpus for Thai speech recognition. In Proceedings of Oriental COCOSDA (pp. 54-61).

Kouzis-Loukas, D. (2016). Learning Scrapy. Packt Publishing Ltd.

Thai MOOC. (n.d.). Retrieved from https://lms.thaimooc.org/

Thai Notes (n.d.). Retrieved from http://thai-notes.com/

Wiktionary. (n.d.). Retrieved from https://www.wiktionary.org/