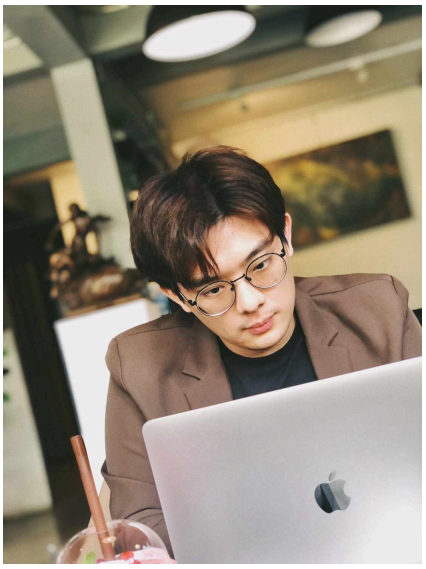
Language Understanding $\lambda(\mu)$ Laboratory

LU Talks

Myanmar Prosodic Structure Analysis and MY-ToBI Annotation System Development in Speech Corpus: Implications for Myanmar-Thai Machine Translation in Medical Domain



MR. TANINTORN LIMPISIRI
Mahidol University, Research Institute of Language and Culture of Asia
(Majored in Linguistics)

Biography

Tanintorn Limpisiri went from being an archaeology graduate student to earning a doctorate degree in Linguistics. He worked as a researcher in the field of anthropology and history, conducting field work, interviews, archival research, studying annals, etc. Then he encountered problems from communicating with foreign workers. Therefore, he shifted away from his original field of study to linguistics. Phonetics and phonology became his most interested areas of study and research. He chose to research the Myanmar or Burmese language because he wanted to support the immigrant community from Myanmar in Thailand. He found that natural language processing could be helpful, but it was also difficult to apply from a linguistic perspective.

Abstract

The study findings reveal that in the prosodic structure at sentence level, the intonation pattern begins with a high pitch marked by %h, %hH, %H, or %Hh boundary tone, followed by a declination throughout the sentence marked by H L. Emphasis in sentences follows the patterns !H*L and H*L, especially in broad-focus statements and verb-focus statements. Additionally, two intonation patterns are found at the end of sentences: falling or lowering offset intonation (L%) and rising intonation (Lh% or LH%), sometimes marked as h% or H%. The rising patterns Lh%, LH%, h%, and H% share common characteristics, featuring a lower pitch at the end of the sentence before the last syllable. This rising intonation is realized, especially in final particles. However, Lh or LH rising intonation is not realized in the middle of phrases.

Proposing the Myanmar prosodic annotation system (MY-ToBI), the symbols are assigned to align with the intonation patterns in the language. MY-ToBI includes 13 symbols, including H, h, L, *, !, %H, %h, L%, H%, h%, LH%, and Lh%. These symbols are combined into a set of 13 tags, serving as a tagset to annotate prosody in the Myanmar-Thai parallel corpus, which comprises 87,525 sentence pairs, as well as in the Myanmar speech corpus, totaling 55 hours. This corpus serves as a dataset for training a transformer-based machine translation. The evaluation result of machine translation reveals that fine-tuning the model with a ToBI tag increases the BLEU score by 7.47 points, the METEOR score by 0.07 points, and the chrF score by 3.3 points compared to the baseline model. The ToBI-tag idea augmented in the transformer-based model originates from the POS-tag.

May 18, 2024 at 10:00 AM (Thailand Time)
Google Meet Link: meet.google.com/gpo-riai-axq

Chairperson: Dr. Ye Kyaw Thu, Lab. Leader

Contact Info: yktnlp@gmail.com