Thai POS Tagger Project

**Our potential contributions**

* We aim to create an accurate UD-style Part-of-speech tagger for Thai, and the model will be deployed to PyThaiNLP.
* We prove that BPE/syllable features can help with Thai part-of-speech tagging in both discrete and continuous models.

**Previous approaches**

We need to try all of the methods that were previously applied to English and Thai.

* MaxEnt model
* CRF model
* Structured Perceptron
* Bi-LSTM CRF

Features for part-of-speech taggers are usually some word shape features, which obviously depend on the languages.

**Evaluation**

* ORCHID data converted to UD
* Evaluation should also be done on unknown words (words not found in the training set)

**Annotated Bibliography**

ORCHID corpus (a bit revised from 1997 paper)

<https://www.researchgate.net/publication/2768729_Thai_Part-Of-Speech_Tagged_Corpus_Technical_Report>

Very relevant recent work

<https://www.aclweb.org/anthology/C16-1031.pdf>

Dataset

<https://gist.github.com/wannaphongcom/a8224d8fb19eb32c3c86335e180bcc9e>

More information on Thai PUD

<https://universaldependencies.org/treebanks/th_pud/index.html>

UD Annotation guidelines by Aj.Wirote

<http://www.arts.chula.ac.th/~ling/contents/File/UD%20Annotation%20for%20Thai.pdf>

A comparative study on different techniques for Thai part-of-speech tagging (2013)

<https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6559527>

State-of-the-art POS tagging for English

<https://aclweb.org/aclwiki/POS_Tagging_(State_of_the_art)>

Joint Segmentation and POS Tagging

<https://www.researchgate.net/publication/318167037_Bidirectional_Deep_Learning_of_Context_Representation_for_Joint_Word_Segmentation_and_POS_Tagging>