PyThaiNLP: Thai Natural Language Processing in Python

Wannaphong Phatthiyaphaibun♢, Korakot Chaovavanich†, Charin Polpanumas†,
Arthit Suriyawongkul‡, Lalita Lowphansirikul♢, Pattarawat Chormai§¶,
Peerat Limkonchotiwat♢, Thanathip Suntorntip♣, Can Udomcharoenchaikit♢

♢VISTEC, †PyThaiNLP, ‡Trinity College Dublin,
§Technische Universität Berlin, ¶Max Planck School of Cognition,
♣Wisesight

wannaphong.p_s21@vistec.ac.th
https://github.com/pythainlp/PyThaiNLP

What's PyThaiNLP?

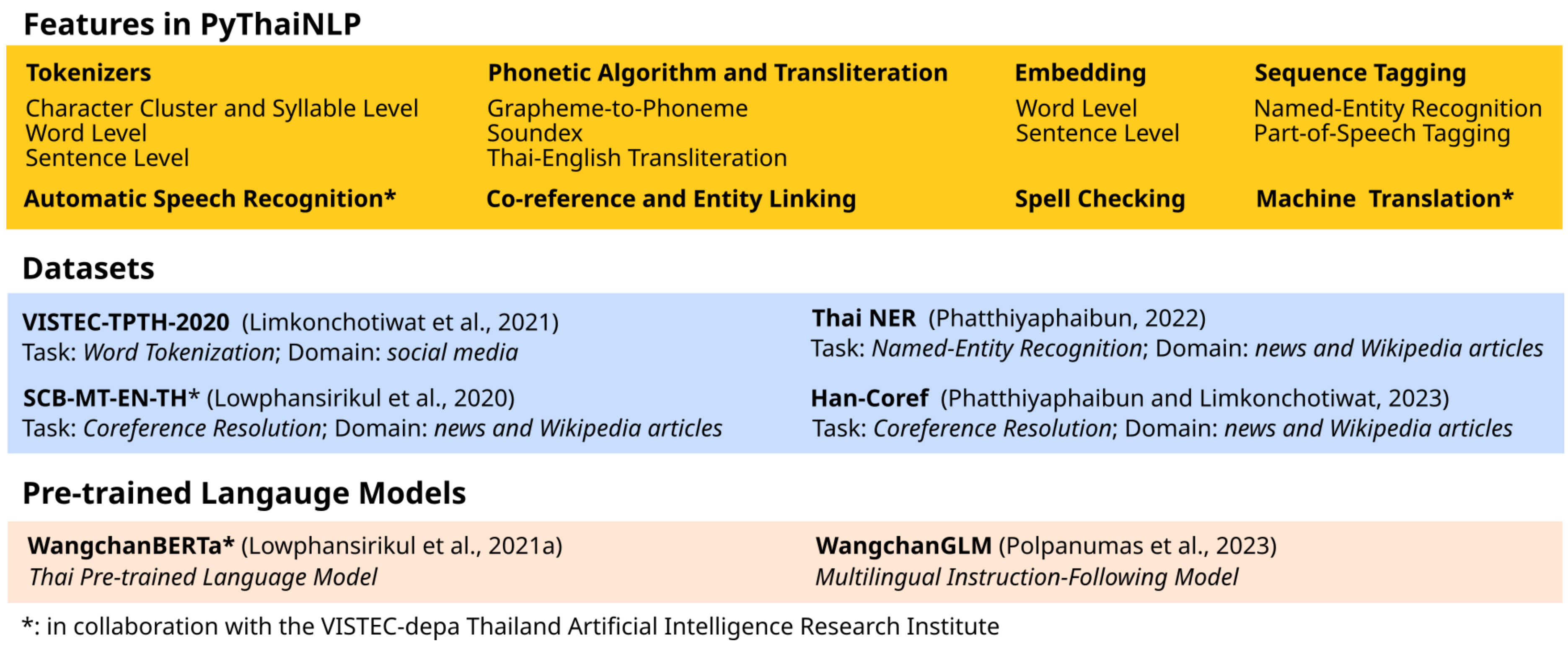
- PyThaiNLP is a free and open-source natural language processing (NLP) library for Thai language implemented in Python.
- We provide a wide range of software, models, and datasets for Thai language.
- Currently, there are more than 10 datasets and 10 NLP models in PyThaiNLP.
- PyThaiNLP was developed in 2016 by a high-school student, currently, we have 8 open-source developers, and we are still going on!
- We found that PyThaiNLP has a big impact on industrial and research communities

Open-source Thai NLP before PyThaiNLP

Examples of Thai NLP tools and datasets before PyThaiNLP (pre-2016):

- **Word tokenization**: ICU BreakIterator, KU Wordcut, SWATH, LexTo, OpenNLP, TLex, and wordcutpy
- **Part-of-speech (POS) tagging**: OpenNLP and RDRPOSTagger
- **Named-entity recognition (NER)**: Polyglot
- **Automatic speech recognition (ASR)**: Thai ARC corpus, NECTEC-ATR, and more but their licenses are not fully open.

PyThaiNLP and Its Ecosystem



Features in PyThaiNLP

Tokenizers Character Cluster and Syllable Level Word Level Sentence Level	**Phonetic Algorithm and Transliteration** Grapheme-to-Phoneme Soundex Thai-English Transliteration	**Embedding** Word Level Sentence Level	**Sequence Tagging** Named-Entity Recognition Part-of-Speech Tagging
Automatic Speech Recognition*	**Co-reference and Entity Linking**	**Spell Checking**	**Machine Translation***

Datasets

VISTEC-TPTH-2020 (Limkonchotiwat et al., 2021) Task: *Word Tokenization*; Domain: *social media*	**Thai NER** (Phatthiyaphaibun, 2022) Task: *Named-Entity Recognition*; Domain: *news and Wikipedia articles*
SCB-MT-EN-TH* (Lowphansirikul et al., 2020) Task: *Coreference Resolution*; Domain: *news and Wikipedia articles*	**Han-Coref** (Phatthiyaphaibun and Limkonchotiwat, 2023) Task: *Coreference Resolution*; Domain: *news and Wikipedia articles*

Pre-trained Langauge Models

WangchanBERTa* (Lowphansirikul et al., 2021a) *Thai Pre-trained Language Model*	**WangchanGLM** (Polpanumas et al., 2023) *Multilingual Instruction-Following Model*

*: in collaboration with the VISTEC-depa Thailand Artificial Intelligence Research Institute

PyThaiNLP in the Wild

- **Research Impact**
  - Researchers worldwide use PyThaiNLP to work with Thai language.
  - Example: XLM, universal dependency parsing, and XLM-R
- **Industry Impact**
  - PyThaiNLP is used in many real-world business use cases by firms of all sizes both domestic and international.
  - **SCB** finetuned WangchanBERTa for QA system.
  - **True Corporation** used PyThaiNLP both for digital media analysis and for recommendation engine on production.
  - **Central Retail Digital** used PyThaiNLP mainly to enhance search and recommendation system.
  - **AIA Thailand** used PyThaiNLP for analyzing their inbound and outbound call logs to improve customer retention
  - **VISAI** used PyThaiNLP for facilitate text processing for all their NLP-based products.

Community and Project Milestones

- **Foundation Years (2016-2019)**
  - PyThaiNLP was started by Wannaphong Phatthiyaphaibun, a high school student in 2016.
  - We created the “Thai Natural Language Processing” group to be a hub for Thai NLP researchers and practitioners to discuss the field.
  - In addition to Wannaphong, we have Korakot Chaovavanich, Charin Polpanumas and Arthit Suriyawongkul as main contributors to create foundational capabilities for Thai Language.
- **Gaining Resources for Large Language Models (2019-present)**
  - Our project began a collaboration with AIResearch.in.th to create and distribute open-source models and datasets.
  - These include English-Thai sentence-pair dataset, Thai Wav2Vec2 ASR model, WangchanBERTa: the RoBERTa-based monolingual language model, and WangChanGLM: the multilingual instruction-following model.
- **Community and Infrastructure for Software Quality**
  - On the infrastructure side, test automation and continuous integration (CI) help us systematically reinforce code style, detect code security vulnerabilities, maintain code coverage, and test the library in different computer configurations.

Conclusion and Future Works

- **Domain-specific datasets/models**: Some capabilities are not performing well on specific use cases; for instances, named-entity recognition in financial reports, medical terms translation, and legal documents question answering. We believe more domain-specific datasets and models will help close this gap.
- **Robust benchmark for Thai NLP tasks**: As NLP garner more attention, more models and datasets, both open- and closed-source, will become available. It will be imperative to have robust benchmarks to compre model performance and dataset quality.
- **Correctness and consistency**: Search key generation (such as Soundex), sorting, and tokenization have to be deterministic and strictly follow a specification, or an application may behave in an unexpected fashion. More test cases and verification might be needed for these features.
- **Efficient mechanism to load and manage datasets/models**: To reduce the size of the library and usage in a system with a restricted network connection.
- **Seamless integration with language-agnostic tools**: The ultimate goal is for developers to no longer need PyThaiNLP as Thai language is supported by standard NLP libraries such as spaCy and Hugging Face. We have begun this work with integrating our text processing functions and models to spaCy.