WangchanBERTa: Pretraining Transformer-based Thai Language Models

Lalita Lowphansirikul∗†, Charin Polpanumas∗ †, Nawat Jantrakulchai †, and Sarana Nutanong †

†PyThaiNLP
charin.polpanumas@datatouille.org
†School of Information Science and Technology, Vidyasirimedhi Institution of Science and Technology
{lalitalpro, nawatjpro, snutanon}@vistec.ac.th

January 26, 2021

ABSTRACT

Transformer-based language models, more specifically BERT-based architectures have achieved state-of-the-art performance in many downstream tasks. However, for a relatively low-resource language such as Thai, the choices of models are limited to training a BERT-based model based on a much smaller dataset or finetuning multi-lingual models, both of which yield suboptimal downstream performance. Moreover, large-scale multi-lingual pretraining does not take into account language-specific features for Thai. To overcome these limitations, we pretrain a language model based on RoBERTa-base architecture on a large, deduplicated, cleaned training set (78GB in total size), curated from diverse domains of social media posts, news articles and other publicly available datasets. We apply text processing rules that are specific to Thai most importantly preserving spaces, which are important chunk and sentence boundaries in Thai before subword tokenization. We also experiment with word-level, syllable-level and SentencePiece tokenization with a smaller dataset to explore the effects on tokenization on downstream performance. Our model wangchanberta-base-att-spm-uncased trained on the 78.5GB dataset outperforms strong baselines (NBSVM, CRF and ULMFit) and multi-lingual models (XLMR and mBERT) on both sequence classification and token classification tasks in human-annotated, mono-lingual contexts.

Keywords Language Modeling · BERT · RoBERTa · Pretraining · Transformer · Thai Language

1 Introduction

Transformer-based language models, more specifically BERT-based architectures [Devlin et al., 2018b], [Liu et al., 2019], [Lan et al., 2019], [Clark et al., 2020], and [He et al., 2020], have achieved state-of-the-art performance in downstream tasks such as sequence classification, token classification, question answering, natural language inference and word sense disambiguation [Wang et al., 2018, Wang et al., 2019]. However, for a relatively low-resource language such as Thai, the choices of models are limited to training a BERT-based model based on a much smaller dataset such as BERT-th [ThAIKeras, 2018] trained on Thai Wikipedia Dump, or finetuning multi-lingual models such as XLMR [Conneau et al., 2019] (100 languages) and mBERT [Devlin et al., 2018b] (104

∗Equal contribution. Listed in random order.
languages). Training on a small dataset has a detrimental effect on downstream performance. BERT-th underperforms RNN-based ULMFit [Polpanumas and Phatthiyaphaibun, 2021] trained on Thai Wikipedia Dump on sequence classification task Wongnai Reviews [Wongnai.com, 2018]. For multi-lingual training, we can see from comparison between multi-lingual and mono-lingual models such as [Martin et al., 2019] that multi-lingual models underperform mono-lingual models. Moreover, large-scale multi-lingual pretraining does not take into account language-specific features for Thai. To overcome these limitations, we pretrain a language model based on RoBERTa-base architecture on a large, deduplicated, cleaned training set (78GB in total size), curated from diverse domains of social media posts, news articles and other publicly available datasets. We apply text processing rules that are specific to Thai most importantly preserving spaces, which are important chunk and sentence boundaries in Thai before subword tokenization.

In this report, we describe a language model based on RoBERTa-base architecture and SentencePiece [Kudo and Richardson, 2018] subword tokenizer on 78GB cleaned and deduplicated data from publicly available social media posts, news articles, and other open datasets. We also pretrain four other language models using different tokenizers, namely SentencePiece [Kudo and Richardson, 2018], dictionary-based tokenizer (PyThaiNLP’s newmm [Phatthiyaphaibun et al., 2020]), syllable tokenizer (ssg [Chormai et al., 2020]), and SEFR tokenizer [Limkonchotiwat et al., 2020], on Thai Wikipedia Dump to explore how tokens affect downstream performance.

To assess the effectiveness of our language model, we conducted an extensive set of experimental studies on the following downstream tasks: sequence classification (multi-class and multi-label) and token classification. Our model wangchanberta-base-att-spm-uncased outperforms strong baseline models (NBSVM [Wang and Manning, 2012] and CRF [Okazaki, 2007]), ULMFit [Howard and Ruder, 2018] (thai2fit [Polpanumas and Phatthiyaphaibun, 2021]) and multi-lingual transformer-based models (XLMR [Conneau et al., 2019] and mBERT [Devlin et al., 2018a]) on both sequence and token classification tasks.

The remaining sections of this report are organized as follows. In Section 2, we describe the methodology in pretraining the language models including raw data, preprocessing, train-validation-test split preparation and training the models. In Section 3, we introduce the downstream tasks we use to test the performance of our language models. In Section 4, we demonstrate the results of our language modeling and finetuning for downstream tasks. In Section 5, we discuss the results and next steps for this work.

The pretrained language models and finetuned models2 are publicly available at Huggingface’s Model Hub. The source code used for the experiments can be found at our GitHub repository3.

## 2 Methodology

We train one language model on the Assorted Thai Texts dataset including all available raw datasets and four language models on the Wikipedia-only dataset, each with a different tokenizer.

### 2.1 Raw Data

The raw data are obtained from (statistics after preprocessing):

| Dataset name       | Data size            | Description                                                                                                                                 |
|--------------------|----------------------|-------------------------------------------------------------------------------------------------------------------------------------------|
| wisesight-large    | 51.44GB (314M lines) | a large dataset of social media posts provided by the social listening platform Wisesight4 for this study. The dataset contains posts Twitter, Facebook, Pantip, Instagram, YouTube and other websites sampled from 2019. |

---

2https://huggingface.co/airesearch
3https://github.com/vistec-AI/thai2transformers
4https://wisesight.com/
| Dataset                  | Size          | Description                                                                                                                                                                                                 |
|-------------------------|---------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| pantip-large            | 22.35GB (95M lines) | A collection of posts and answers of Thailand’s largest online bulletin board Pantip.com from 2015 to 2019 provided by audience analytics platform Chaos Theory.                                                  |
| Thairath-222k          | 1.48GB (5M lines)     | A collection of articles published on newspaper website Thairath.com up to December 2019.                                                                                                                     |
| prachathai-67k         | 903.1MB (2.7M lines) | A collection of articles published on newspaper website Prachathai.com from August 24, 2004 to November 15, 2018.                                                                                             |
| Thai Wikipedia Dump    | 515MB (843k lines)    | The Wikipedia articles extracted using Giuseppe Attardi’s WikiExtractor in September 2020. All HTML tags, bullet points, and tables are removed.                                                            |
| OpenSubtitles          | 468.8MB (5M lines)    | A collection of movie subtitles translated by crowdsourcing from OpenSubtitles.org. We use only the portions containing Thai texts.                                                                                 |
| ThaiPBS-111k           | 372.3MB (858k lines)| A collection of articles published on newspaper website ThaiPBS.or.th up to December 2019.                                                                                                               |
| Thai National Corpus   | 366MB (797k lines)    | A 14-million-word corpus of Thai texts containing 75% non-fiction and 25% fiction works. Media source breakdown is 60% books, 25% magazines, and the rest from other publications and writings. Most of the texts are curated from 1998 to 2007. |
| scb-mt-en-th-2020      | 290.4MB (947k lines)| A parallel corpus of English-Thai sentence pairs curated news, Wikipedia articles, SMS messages, task-based dialogs, web-crawled data, government documents, and machine-generated text. |
| JW300                   | 182.8MB (727k lines)| A parallel corpus of religion texts from jw.org that includes Thai texts.                                                                                                                                     |
| wongnai-corpus         | 64MB (101k lines)      | A collection of restaurant reviews and ratings (1 to 5 stars) published on Wongnai.com.                                                                                                                     |
| QED                     | 42MB (407k lines)      | A collection of transcripts for educational videos and lectures collaboratively created on the AMARA web-based platform.                                                                                     |
| bibleuedin             | 2.18MB (62k lines)     | A multilingual corpus of the Bible created by Christos Christodoulopoulos and Mark Steedman.                                                                                                                |
| wisesight-sentiment    | 5.3MB (22k lines)       | A collection of Twitter posts about consumer products and services from 2016 to early 2019 labeled positive, negative, neutral and question.                                                            |
| tanzil                  | 2.4MB (6k lines)        | A collection of Quran translations compiled by the Tanzil project.                                                                                                                                          |
| tatoeba                 | 1MB (2k lines)          | A collection of translated sentences from the crowdsourced multilingual dataset Tatoeba.                                                                                                                   |

5 https://www.facebook.com/ChaosTheoryCompany/
6 https://github.com/nakhunchumpolsathien/TR-TPBS
7 https://github.com/PyThaiNLP/prachathai-67k
8 https://dumps.wikimedia.org/backup-index.html
9 https://github.com/attardi/wikiextractor/
10 https://github.com/nakhunchumpolsathien/TR-TPBS
11 https://github.com/wongnai/wongnai-corpus
2.2 Preprocessing

We apply preprocessing rules to the raw datasets before using them as our training sets. This effectively demands the preprocessing rules to be applied before finetuning for both domain-specific language modeling and other downstream tasks.

**Text Processing** A large portion of our training data (*wisesight-large* and *pantip-large*) comes from social media, which usually have a lot of unusual spellings and repetitions. For such noisy data, [Raffel et al., 2020] reports that pretraining on a cleaned corpus *C4* yields better performance in downstream tasks. Therefore, we opted to perform the following processing rules, in order:

- Replace HTML forms of characters with the actual characters such as `nbsp;` with a space and `<br />` with a line break [Howard and Ruder, 2018].
- Remove empty brackets ([], {}, and []) than sometimes come up as a result of text extraction such as from Wikipedia.
- Replace line breaks with spaces.
- Replace more than one spaces with a single space.
- Remove more than 3 repetitive characters such as ดีมากกก to ดีมาก [Howard and Ruder, 2018].
- Word-level tokenization using [Phatthiyaphaibun et al., 2020]’s `newmm` dictionary-based maximal matching tokenizer.
- Replace repetitive words; this is done post-tokenization unlike [Howard and Ruder, 2018] since there is no delimitation by space in Thai as in English.
- Replace spaces with `<>`. The SentencePiece tokenizer combines the spaces with other tokens. Since spaces serve as punctuation in Thai such as sentence boundaries similar to periods in English, combining it with other tokens will omit an important feature for tasks such as word tokenization and sentence breaking. Therefore, we opt to explicitly mark spaces with `<>.

For *Wikipedia-only dataset*, we only replace non-breaking spaces with spaces, remove an empty parenthesis that occur right after the title of the first paragraph, and replace spaces with `<>.

**Sentence Breaking** Each row of all datasets are originally delimited by line breaks. Due to memory constraints, in order to train the language model, we need to limit our maximum sequence length to 416 subword tokens (tokenized by SentencePiece [Kudo and Richardson, 2018] unigram model) or roughly 300 word tokens (tokenized by dictionary-based maximal matching [Phatthiyaphaibun et al., 2020]). In order to do so, we use the sentence breaking model CRFCut ([Lowphansirikul et al., 2020]). CRFCut is a conditional random fields (CRF) model trained on English-to-Thai translated texts of [Sornlertlamvanich et al., 1997] (23,125 sentences), TED transcripts (136,463 sentences; [Lowphansirikul et al., 2020]) and generated product reviews (217,482 sentences; [Lowphansirikul et al., 2020]). It uses English sentence boundary as sentence boundary labels for translated Thai texts. CRFCut has sentence-boundary F1 score of 0.69 on [Sornlertlamvanich et al., 1997], 0.71 on TED Transcripts, and 0.96 on generated product reviews. We keep only sentences that are 5 to 300 words long to not exceed 416-subword maximum sequence length and also not have a sequence too short for language modeling.

**Tokenizers** For the model trained on *Assorted Thai Texts dataset*, in the same manner as [Martin et al., 2019], we use SentencePiece [Kudo and Richardson, 2018] unigram language model [Kudo, 2018] to tokenize sentences of training data into subwords. The tokenizer has a vocabulary size of 25,000 subwords, trained on 15M sentences. To construct the training set for the tokenizer, we first take 2.5M randomly sampled sentences from *pantip-large*, 3.5M randomly sampled sentences from *wisesight-large* and all sentences of the remaining datasets, resulting in 20,961,306 total sentences. Out of those, we randomly sampled 15M sentences to train the tokenizer.
For the models trained on Wikipedia-only dataset, we use four different tokenizers to examine their effects on language modeling and downstream tasks. We use the same training set of 944,782 sentences sampled from Thai Wikipedia Dump

- **SentencePiece tokenizer**: we train the SentencePiece [Kudo and Richardson, 2018] unigram language model [Kudo, 2018] using 944,782 sentences from Thai Wikipedia Dump, resulting in a tokenizer with vocab size of 24,000 subwords.

- **Word-level tokenizer**: the word-level, dictionary-based tokenizer newmm [Phatthiyaphaibun et al., 2020] is used to create a tokenizer with vocab size of 97,982 words.

- **Syllable-level tokenizer**: CRF-based syllable tokenizer ssg [Chormai et al., 2020] is used to create a tokenizer with vocab size of 59,235 syllables.

- **SEFR tokenizer**: Stacked Ensemble Filter and Refine tokenizer (engine=¨best¨) [Limkonchotiwat et al., 2020] based on probabilities from CNN-based deepcut [Kittinaradorn et al., 2019] with a vocab size of 92,177 words.

### 2.3 Train-Validation-Test Splits

**Assorted Thai Texts Dataset** After preprocessing and deduplication, we have a training set of 381,034,638 unique, mostly Thai sentences with sequence length of 5 to 300 words (78.5GB). The training set has a total of 16,957,775,412 words as tokenized by dictionary-based maximal matching [Phatthiyaphaibun et al., 2020], 8,680,485,067 subwords as tokenized by SentencePiece [Kudo and Richardson, 2018] tokenizer, and 53,035,823,287 characters. We also randomly sampled 99,181 sentences (19.28MB) as validation set and 42,238,656 sentences (8GB) as test set. Both are preprocessed in the same manner as the training set.

**Wikipedia-only Dataset** From Thai Wikipedia Dump, we extract in a uniformly random manner 944,782 sentences for training set, 24,863 sentences for validation set and 24,862 sentences for test set.

### 2.4 Language Modeling

**Architecture** We use the transformer [Vaswani et al., 2017] architecture of BERT (Base) (12 layers, 768 hidden dimensions, 12 attention heads) [Devlin et al., 2018]. Our setup is very similar to [Martin et al., 2019] replacing BERT’s WordPiece tokenizer with a SentencePiece tokenizer, with the exception of preprocessing rules applied before subword tokenization.

**Pretraining Objective** We train the model with masked language modeling. To circumvent the word boundary issues in Thai, we opted to perform this at the subword level instead of whole-word level, even though the latter is reported to have better performance in English [Joshi et al., 2020]. In the same manner as BERT [Devlin et al., 2018b] and RoBERTa [Liu et al., 2019], for each sequence, we sampled 15% of the tokens and replace them with <mask> token. Out of the 15%, 80% is replaced with a <mask> token, 10% is left unchanged and 10% is replaced with a random token. The objective is to predict the tokens replaced with <mask> using cross entropy loss.

**Pretraining** We pretrain RoBERTaBASE on both the Assorted Thai Texts dataset and Wikipedia-only dataset. The size of Wikipedia-only dataset is about 0.57 GB which is comparatively low compared to the Assorted Thai Texts dataset. Therefore, we manually tune the hyperparameters used for RoBERTaBASE pretraining for each training set in order to control the loss stability. The hyperparameters of the RoBERTaBASE architecture and model pretraining are listed in Table 2.
| Hyperparameters | RoBERTa\textsubscript{BASE} (Wikipedia-only Dataset) | RoBERTa\textsubscript{BASE} (Assorted Thai Texts Dataset) |
|-----------------|---------------------------------|---------------------------------|
| Number of Layers | 12                              | 12                              |
| Hidden size     | 768                             | 768                             |
| FFN hidden size | 3,072                           | 3,072                           |
| Attention heads | 12                              | 12                              |
| Dropout         | 0.1                             | 0.1                             |
| Attention dropout | 0.1                            | 0.1                             |
| Max sequence length | 512                          | 416                             |
| Effective batch size | 8,192                     | 4,092                           |
| Warmup steps    | 1,250                           | 24,000                          |
| Peak learning rate | 7e-4                         | 3e-4                            |
| Learning rate decay | Linear                       | Linear                          |
| Max steps       | 31,250                          | 500,000                         |
| Weight decay    | 0.01                            | 0.01                            |
| Adam $\epsilon$ | 1e-6                            | 1e-6                            |
| Adam $\beta_1$  | 0.9                             | 0.9                             |
| Adam $\beta_2$  | 0.98                            | 0.999                           |
| FP16 training   | True                            | True                            |

Table 2: Hyperparameters of RoBERTa\textsubscript{BASE} used when pretrain on Assorted Thai Texts dataset and Wikipedia-only dataset.

WangchanBERTa We name our pretrained language models according to their architectures, tokenizers and the datasets on which they are trained on. The models can be found on HuggingFace.

| Architecture             | Dataset            | Tokenizer   |
|--------------------------|--------------------|-------------|
| wangchanberta-base-wiki-spm | RoBERTa-base       | SentencePiece |
| wangchanberta-base-wiki-newmm  | RoBERTa-base       | word (newmm) |
| wangchanberta-base-wiki-ssg    | RoBERTa-base       | syllable (ssg) |
| wangchanberta-base-wiki-sefr   | RoBERTa-base       | SEFR |
| wangchanberta-base-att-spm-uncased | RoBERTa-base   | Assorted Thai Texts | SentencePiece |

Table 3: WangchanBERTa model names

3 Downstream Tasks

We evaluate the downstream performance of our pretrained Thai RoBERTa\textsubscript{BASE} models on existing Thai sequence-classification and token-classification benchmark datasets.

12 https://huggingface.co/models
3.1 Datasets

We use train-validation-test split as provided by each dataset as hosted on Huggingface Datasets. When not all splits are available, namely for Wongnai Reviews and ThaiNER, we sample respective splits in a uniformly random manner. The descriptive statistics of each datasets are as follows:

| Datasets             | Label       | Style                   | Tasks                              | Labels | Train     | Eval      | Test     |
|----------------------|-------------|-------------------------|------------------------------------|--------|-----------|-----------|----------|
| wisesight_sentiment  | category    | informal; social media  | multi-class sequence classification| 4      | 21628     | 2404      | 2671     |
| wongnai_reviews      | rating      | informal; restaurant    | multi-class sequence classification| 5      | 36000     | 4000      | 6203     |
| generated_reviews_enth | review_star | informal; product       | multi-class sequence classification| 5      | 141369    | 15708     | 17453    |
| prachathai67k        | tags        | formal; news            | multi-label sequence classification| 12     | 54379     | 6721      | 6789     |
| thainer              | ner_tags    | formal; news and other  | token classification              | 13**   | 5079      | 635       | 621***   |
| lst20                | pos_tags    | formal; news and other  | token classification              | 16     | 67104     | 6094      | 5733     |
| lst20                | ner_tags    | formal; news and other  | token classification              | 10     | 67104     | 6094      | 5733     |

*Uniform randomly split with seed = 2020

**We replace B-ไม/uni0E48-lowยืนยัน and I-ไม/uni0E48-lowยืนยัน which are extremely rare tags from ThaiNER with O

***We removed examples on test set which did not fit in mBERT max token length to have a fair comparison among all models

Table 4: Datasets for downstream tasks

3.1.1 Sequence Classification

Wisesight Sentiment [Suriyawongkul et al., 2019] is a multi-class text classification dataset (sentiment analysis). The data are social media messages in Thailand collected from 2016 to early 2019. Each message is annotated as positive, neutral, negative, or question.

Wongnai Reviews [Wongnai.com, 2018] is a multi-class text classification dataset (rating classification). The data are restaurant reviews and their respective ratings from 1 (worst) to 5 (best) stars.

Generated Reviews EN-TH [Lowphansirikul et al., 2020] is a dataset that originally consists of product reviews generated by CTRL [Keskar et al., 2019] in English. It is translated to Thai as part of the scb-mt-en-th-2020 machine translation dataset. Translation is performed both by human annotators and models. We use only the translated Thai texts as a feature to predict review stars from 1 (worst) to 5 (best).

Prachathai-67k is a multi-label text classification dataset (topic classification) based on news articles of Prachathai.com from August 24, 2004 to November 15, 2018 packaged by [Phatthiyaphaibun et al., 2020]. We perform topic classification of the headline of each article, which can contain none to all of the following labels: politics, human rights, quality of life, international, social, environment, economics, culture, labor, national security, ICT, and education.

3.1.2 Token Classification

ThaiNER v1.3 [Phatthiyaphaibun, 2019] is a 6,456-sentence named entity recognition (NER) dataset created by expanding an unnamed, 2,258-sentence dataset by [Tirasaroj and Aroonmanakun, 2012]. The NER tags are annotated by humans in IOB format.

LST20 [Boonkwan et al., 2020] is a dataset with 5 layers of linguistic annotations: word boundaries, POS tagging, NER, clause boundaries, and sentence boundaries. NER tags are in IOBE format. We use the dataset for POS tagging and NER tasks.

https://huggingface.co/datasets
3.2 Benchmarking Models

We provide benchmarks using traditional models (NBSVM for sequence classification and CRF for token classification), RNN-based models (ULMFit; only for sequence classification) and transformer-based models.

**NBSVM** [Wang and Manning, 2012] We adopt the NBSVM implementation by Jeremy Howard as our strong baselines for sequence classification both multi-class and multi-label. The notable differences are substituting binarized ngram features with tf-idf features (uni- and bi-grams; minimum document frequency of 3, maximum document frequency of 90%). We also apply the same cleaning rules as the language model, with the differences being adding repeated character tokens `<rep>` and repeated word tokens `<wrep>` instead of space tokens `<>.

We perform hyperparameter tuning for penalty types (L1 and L2) and inverse of regularization strength (C=[1.0, 2.0, 3.0, 4.0]) and choose the models with the highest F1 scores (micro-averaged for multi-class and macro-averaged for multi-label classification). See Table 8. For multi-label classification, we search for the best set of thresholds (ranging between 0.01 – 0.99 with the step size of 0.01) that maximize macro-average F1-score on validation set.

**ULMFit (thai2fit)** is an implementation of ULMFit language model finetuning for text classification [Howard and Ruder, 2018]. [Polpanumas and Phatlhiyaphaibun, 2021] pretrained a language model with vocab size of 60,005 words (tokenized by PyThaiNLP's `newmm`) on Thai Wikipedia Dump. We finetune the language model on the training set of each dataset for 5 epochs. Then that, we finetune for the sequence classification tasks using gradual unfreezing from the last one, two and three parameter groups with discriminative learning rates, for one epoch each. After that, we finetune all the weights of the model for 5 epochs. The checkpoints with the highest accuracy scores (validation losses for multi-label classification) are chosen to perform on the test sets. See Table 9. Lastly, we search for the best set of thresholds (ranging between 0.01 – 0.99 with the step size of 0.01) that maximize macro-average F1-score on validation set.

**Conditional Random Fields (CRF)** [Lafferty et al., 2001] We use the CRFSuite implementation [Okazaki, 2007] of conditional random fields as a strong baseline for POS and NER tagging tasks. We generate the features by extracting unigrams, bigrams and trigrams features within a sliding window of three timesteps, before and after the current token (beginning and ending of sentences are padded with `xxpad` tokens). We finetune L1 and L2 penalty combinations using 10,000 randomly sampled sentences for LST20 and the entire training set for ThaiNER. With hyperparameters with the best F1 score (micro-averaged) on the validation set, we train on the entire training sets and report performances on the test sets. See Table 10. We run each CRF model for 500 iterations.

**Transformer-based models** We use the same finetuning scheme for all transformer-based models, namely XLM-RoBERTa-base [Conneau et al., 2019], BERT-base-multilingual-cased [Devlin et al., 2018a], wangchanberta-base-wiki-tokenizer (`spm`, `newmm`, `ssg`, `sefr`), and wangchanberta-base-att-spm-uncased. For the sequence classification task, we preprocess each dataset with the rules described in 2.2. We then finetune each pretrained language model on downstream tasks for 3 epochs. The criteria to select the best epoch is the validation micro-average F1-score for multi-class classification and macro-average F1-score for multi-label classification. The batch size is set to 16. The The learning rate is warmed up over the first 10% of steps to the value of 3e-5 and linearly decayed to zero. We finetune models with FP16 mixed precision training. All models are optimized with Adam [Kingma and Ba, 2014] ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e^{-8}$, $L_2$ weight decay = 0.01) with corrected bias. For multi-label classification head, we search for the best set of thresholds (ranging between 0.01 – 0.99 with the step size of 0.01) that maximize macro-average F1-score on validation set.

For the token classification tasks, we finetune each pretrained language models for 6 epochs. The criteria to select the best epoch is the validation loss. The batch size is set to 32. The learning rate is warmed up over the first 10% of steps to the value of 3e-5 and linearly decayed to zero. We finetune models with FP16 mixed precision training. All models are optimized with Adam with the parameters as same as the sequence classification task.

14https://www.kaggle.com/jhoward/nb-svm-strong-linear-baseline
4 Results

4.1 Language Modeling

The following table shows the performance RoBERTa\textsubscript{BASE} trained on \textit{Wikipedia-only dataset}. There are four variations of tokenization including subword-level with SentencePiece \cite{Kudo2018}, word-level with PyThaiNLP \cite{Phatthiyaphaibun2020} dictionary-based tokenizer \texttt{newmm}, subword-level with a CRF-based Thai syllable tokenizer \texttt{ssg} \cite{Chormai2020}, and stacked-ensemble, word-level tokenizer \texttt{sefr} \cite{Limkonchotiwat2020}.

| Model Name | Vocab Size | Number of Training Examples | Best Checkpoint |
|------------|------------|-----------------------------|-----------------|
|             |            |                             | Validation loss | Steps        |
| Pretraining on Wikipedia-only dataset: |            |                             |                 |              |
| wangchanberta-base-wiki-spm | 24,000     | 116,715                     | 1.5127          | 7,000        |
| wangchanberta-base-wiki-newmm | 97,982     | 119,074                     | 1.4990          | 5,000        |
| wangchanberta-base-wiki-ssg | 59,235     | 167,279                     | 0.8068          | 8,000        |
| wangchanberta-base-wiki-sefr | 92,177     | 125,177                     | 1.2995          | 4,500        |

Pretraining on Assorted Thai Texts dataset (Currently, the model has not reached the max steps):

| Model Name                  | Vocab Size | Number of Training Examples | Best Checkpoint            |
|-----------------------------|------------|-----------------------------|-----------------------------|
| wangchanberta-base-att-spm-uncased | 25,000     | 382 M                       | 2.551                       | 360,000 (latest checkpoint) |

Table 5: The vocab size, number of training examples, and best checkpoint of the RoBERTa\textsubscript{BASE} models trained on Thai Wikipedia corpus for each type of input tokens and Assorted Thai Texts dataset.

For the RoBERTa\textsubscript{BASE} trained on \textit{Assorted Thai Texts dataset}, we only trained with subword token built with SentencePiece \cite{Kudo2018} due to the limited computational resources.

4.2 Downstream Tasks

We choose models to perform on the test set based on their performance on the validation sets. For multi-class sequence classification and token classification, we optimize our models for the highest micro-averaged F1 score. For multi-label sequence classification, we optimize for the highest macro-averaged F1 score, as it is less affected by class imbalance. Moreover, for multi-label sequence classification, we also find the best probability threshold for each label based on the validation set. We report the performance of these optimized models on the test sets.

For sequence classification tasks, our model trained on the \textit{Assorted Thai Texts dataset} outperforms both strong baselines and other transformer-based architecture on all downstream tasks except Generated Reviews (EN-TH). This may be attributed to the fact that the dataset is translated from generated texts in English, thus multi-lingual pretraining of XLMR gives it the advantage.\footnote{This dataset is not available.}

For token classification tasks, our model trained on the \textit{Assorted Thai Texts dataset} achieves the highest micro-averaged F1 score in all tasks except POS tagging in \textit{ThaiNER} dataset. This could be attributed to the fact that the POS tags in \textit{ThaiNER} are machine-generated and thus more suited for the baseline model CRF. See Table\footnote{This table is not available.}.


5 Discussions and Future Works

Consistent with previous works on language modeling, we found that training on large datasets such as our Assorted Thai Texts dataset yield better downstream performance. The only case when a multi-lingual model (XLMR) outperforms our largest mono-lingual model is when the training data include multi-lingual elements namely the English-to-Thai translated texts of Generated Reviews EN-TH. From our experiments on the Wikipedia-only dataset, we did not find any notable difference in downstream performance for sequence classification or token classification tasks.
Another area we will explore in the future is the inherent biases on our relatively large language models. Previous works including [Sheng et al., 2019] [Nadeem et al., 2020] [Nangia et al., 2020] have detected social biases within large language models trained in English. Our next step in this direction is to create similar bias-measuring datasets in Thai contexts to detect the biases in our language models.

We pretrain our language models on publicly available datasets. Two main concerns that have been raised about similar models are copyrights and privacy. All datasets used to train our models are based on publicly available data. Publicly available social media data are packaged and provided to use by Wisesight (wisesight-large) and Chaos Theory (pantip-large). Unless specified otherwise in the distribution of datasets, all rights belong to the content creators. We provide the weights of our pretrained language models under CC-BY-SA 4.0. Our models are trained as feature extractors for downstream tasks, and not generative tasks. Reproduction of training data can happen [Carlini et al., 2020] albeit at much lower chance than language models trained specifically for generative tasks.

6 Acknowledgements

We thank Wisesight16, Chaos Theory17 and Pantip.com for providing what has become, to the best of our knowledge, the largest and most diverse high-quality training data in Thai for language modeling.

References

[Abdelali et al., 2014] Abdelali, A., Guzman, F., Sajjad, H., and Vogel, S. (2014). The AMARA corpus: Building parallel language resources for the educational domain. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14), pages 1856–1862, Reykjavik, Iceland. European Language Resources Association (ELRA).

[Aroonmanakun et al., 2009] Aroonmanakun, W., Tansiri, K., and Nittayanuparp, P. (2009). Thai national corpus: a progress report. In Proceedings of the 7th Workshop on Asian Language Resources (ALR7), pages 153–160.

[Boonkwan et al., 2020] Boonkwan, P., Luantangrisuk, V., Phaholphinyo, S., Kriengket, K., Leenoi, D., Phrombut, C., Boriboon, M., Kosawat, K., and Supnithi, T. (2020). The annotation guideline of lst20 corpus. arXiv preprint arXiv:2008.05055.

[Carlini et al., 2020] Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., et al. (2020). Extracting training data from large language models. arXiv preprint arXiv:2012.07805.

[Chormai et al., 2020] Chormai, P., Prasertsom, P., Cheevaprawatdomrong, J., and Rutherford, A. (2020). Syllable-based neural thai word segmentation. In Proceedings of the 28th International Conference on Computational Linguistics, pages 4619–4637.

[Clark et al., 2020] Clark, K., Luong, M.-T., Le, Q. V., and Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:2003.10555.

[Conneau et al., 2019] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116.

[Devlin et al., 2018a] Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018a). BERT: pre-training of deep bidirectional transformers for language understanding. CoRR, abs/1810.04805.

[Devlin et al., 2018b] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018b). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

15https://wisesight.com
16https://www.facebook.com/ChaosTheoryCompany/
[He et al., 2020] He, P., Liu, X., Gao, J., and Chen, W. (2020). Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

[Howard and Ruder, 2018] Howard, J. and Ruder, S. (2018). Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

[Joshi et al., 2020] Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., and Levy, O. (2020). Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

[Keskar et al., 2019] Keskar, N. S., McCann, B., Varshney, L. R., Xiong, C., and Socher, R. (2019). CTRL: A conditional transformer language model for controllable generation. *CoRR*, abs/1909.05858.

[Kingma and Ba, 2014] Kingma, D. and Ba, J. (2014). Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

[Kittinaradorn et al., 2019] Kittinaradorn, R., Achakulvisut, T., Chaovavanich, K., Srithaworn, K., Chormai, P., Kaewkasi, C., Ruangrong, T., and Oparad, K. (2019). DeepCut: A Thai word tokenization library using Deep Neural Network.

[Kudo, 2018] Kudo, T. (2018). Subword regularization: Improving neural network translation models with multiple subword candidates. *arXiv preprint arXiv:1804.10959*.

[Kudo and Richardson, 2018] Kudo, T. and Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

[Lafferty et al., 2001] Lafferty, J., McCallum, A., and Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

[Lan et al., 2019] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

[Limkonchotiwat et al., 2020] Limkonchotiwat, P., Phatthiyaphaibun, W., Sarwar, R., Chuangsawanich, E., and Nutanong, S. (2020). Domain adaptation of thai word segmentation models using stacked ensemble. Association for Computational Linguistics.

[Lison and Tiedemann, 2016] Lison, P. and Tiedemann, J. (2016). OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

[Liu et al., 2019] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arxiv 2019. *arXiv preprint arXiv:1907.11692*.

[Lowphansirikul et al., 2020] Lowphansirikul, L., Polpanumas, C., Rutherford, A. T., and Nutanong, S. (2020). scbmt-en-th-2020: A large english-thai parallel corpus. *arXiv preprint arXiv:2007.03541*.

[Martin et al., 2019] Martin, L., Muller, B., Suárez, P. J. O., Dupont, Y., Romary, L., de la Clergerie, É. V., Seddah, D., and Sagot, B. (2019). Camembert: a tasty french language model. *arXiv preprint arXiv:1911.03894*.

[Nadeem et al., 2020] Nadeem, M., Bethke, A., and Reddy, S. (2020). Stereoset: Measuring stereotypical bias in pretrained language models.

[Nangia et al., 2020] Nangia, N., Vania, C., Bhalerao, R., and Bowman, S. R. (2020). Crows-pairs: A challenge dataset for measuring social biases in masked language models.

[Okazaki, 2007] Okazaki, N. (2007). Crfsuite: a fast implementation of conditional random fields, 2007.

[Phatthiyaphaibun, 2019] Phatthiyaphaibun, W. (2019). wannaphongcom/thai-ner: Thainer 1.3.
[Phatthiyaphaibun et al., 2020] Phatthiyaphaibun, W., Chaovavanich, K., Polpanumas, C., Suriyawongkul, A., Lowphansirikul, L., and Chormai, P. (2020). Pythainlp/pythainlp: Pythainlp 2.1.4.

[Polpanumas and Phatthiyaphaibun, 2021] Polpanumas, C. and Phatthiyaphaibun, W. (2021). thai2fit: Thai language implementation of umlfit.

[Raffel et al., 2020] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

[Sheng et al., 2019] Sheng, E., Chang, K.-W., Natarajan, P., and Peng, N. (2019). The woman worked as a babysitter: On biases in language generation. *arXiv preprint arXiv:1909.01326*.

[Sornlertlamvanich et al., 1997] Sornlertlamvanich, V., Charoenporn, T., and Isahara, H. (1997). Orchid: Thai part-of-speech tagged corpus. *National Electronics and Computer Technology Center Technical Report*, pages 5–19.

[Suriyawongkul et al., 2019] Suriyawongkul, A., Chuangsawanich, E., Chormai, P., and Polpanumas, C. (2019). Pythainlp/wisesight-sentiment: First release.

[ThAIKeras, 2018] ThAIKeras (2018). Thaikeras bert. [https://github.com/ThAIKeras/bert](https://github.com/ThAIKeras/bert).

[Tiedemann, 2012] Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In Chair), N. C. C., Choucri, K., Declerck, T., Dogan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC’12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

[Tirasaroj and Aroonmanakun, 2012] Tirasaroj, N. and Aroonmanakun, W. (2012). Thai ner using crf model based on surface features. pages 176–180. SNLP-AOS 2011.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

[Wang et al., 2019] Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2019). Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in neural information processing systems*, pages 3266–3280.

[Wang et al., 2018] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2018). Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

[Wang and Manning, 2012] Wang, S. I. and Manning, C. D. (2012). Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 90–94.

[Wongai.com, 2018] Wongai.com (2018). Wongai-corpus. [https://github.com/wongnai/wongnai-corpus](https://github.com/wongnai/wongnai-corpus).
### Table 8: NBSVM Hyperparameter Tuning Results

| datasets[features:labels] | penalty | C   | f1      |
|---------------------------|---------|-----|---------|
| wisesight_sentiment[texts:category] |   |     |         |
| l2 3                       | 0.720466 |
| l2 2                       | 0.718386 |
| l2 4                       | 0.715474 |
| l2 1                       | 0.707571 |
| l1 1                       | 0.706323 |
| l1 4                       | 0.705075 |
| l2 2                       | 0.57125  |
| l2 4                       | 0.5705   |
| l2 3                       | 0.56675  |
| wongnai_reviews[review_body:star_rating] |   |     |         |
| l2 2                       | 0.5635   |
| l2 3                       | 0.5605   |
| l2 4                       | 0.55425  |
| l2 1                       | 0.55325  |
| l1 1                       | 0.54     |
| generated_reviews_enth[translation[th]:review_star] |   |     |         |
| l2 4                       | 0.593265 |
| l2 3                       | 0.591609 |
| l2 1                       | 0.590718 |
| prachathai67k[title:multilabel] |   |     |         |
| l2 4                       | 0.601663 |
| l2 1                       | 0.59017  |
| l2 2                       | 0.59017  |
| datasets [features:labels] | finetuning | unfreezing | epoch | train_loss | valid_loss | accuracy |
|---------------------------|------------|------------|-------|------------|------------|----------|
| wisegait [sentiment] [texts:category] | LM all | 0 | 4.459779 | 4.256248 | 0.330593 | |
|  | LM all | 1 | 4.261896 | 4.081441 | 0.348543 | |
|  | LM all | 2 | 4.042319 | 3.979969 | 0.358797 | |
|  | LM all | 3 | 3.854878 | 3.935666 | 0.364824 | |
|  | LM all | 4 | 3.754257 | 3.932823 | 0.365743 | |
|  | CLS last 1 | 0 | 0.90485 | 0.800313 | 0.653078 | |
|  | CLS last 2 | 0 | 0.834293 | 0.762427 | 0.675957 | |
|  | CLS last 3 | 0 | 0.783797 | 0.724123 | 0.68594 | |
|  | CLS all | 0 | 0.729717 | 0.73506 | 0.673877 | |
|  | CLS all | 1 | 0.744124 | 0.707241 | 0.702579 | |
|  | CLS all | 2 | 0.721162 | 0.694311 | 0.714642 | |
|  | CLS all | 3 | 0.719528 | 0.698624 | 0.710899 | |
|  | CLS all | 4 | 0.659977 | 0.691418 | 0.711314 | |
| wongnai [reviews] [review body:star rating] | LM all | 0 | 3.844957 | 3.675409 | 0.358546 | |
|  | LM all | 1 | 3.640318 | 3.511868 | 0.375098 | |
|  | LM all | 2 | 3.521275 | 3.422731 | 0.383874 | |
|  | LM all | 3 | 3.423584 | 3.377162 | 0.388852 | |
|  | LM all | 4 | 3.333337 | 3.370303 | 0.389565 | |
|  | CLS last 1 | 0 | 1.039886 | 0.981237 | 0.54275 | |
|  | CLS last 2 | 0 | 0.952058 | 0.913627 | 0.58675 | |
|  | CLS last 3 | 0 | 0.917949 | 0.884318 | 0.595 | |
|  | CLS all | 0 | 0.881919 | 0.882625 | 0.595 | |
|  | CLS all | 1 | 0.879615 | 0.883927 | 0.59825 | |
|  | CLS all | 2 | 0.865561 | 0.889925 | 0.58675 | |
|  | CLS all | 3 | 0.831835 | 0.894447 | 0.602 | |
|  | CLS all | 4 | 0.808713 | 0.895076 | 0.59925 | |
| generated [review] [en][translation] [th]: review [star] | LM all | 0 | 3.562119 | 3.389167 | 0.347284 | |
|  | LM all | 1 | 3.425128 | 3.265404 | 0.362213 | |
|  | LM all | 2 | 3.312375 | 3.198505 | 0.370227 | |
|  | LM all | 3 | 3.235396 | 3.164119 | 0.374517 | |
|  | LM all | 4 | 3.184817 | 3.157655 | 0.37286 | |
|  | CLS last 1 | 0 | 1.097455 | 0.98512 | 0.586516 | |
|  | CLS last 2 | 0 | 0.976767 | 0.902084 | 0.62204 | |
|  | CLS last 3 | 0 | 0.925023 | 0.874969 | 0.631653 | |
|  | CLS all | 0 | 0.892837 | 0.870975 | 0.637 | |
|  | CLS all | 1 | 0.884311 | 0.859921 | 0.636555 | |
|  | CLS all | 2 | 0.852318 | 0.856317 | 0.638464 | |
|  | CLS all | 3 | 0.840453 | 0.85957 | 0.64012 | |
|  | CLS all | 4 | 0.827038 | 0.859206 | 0.636674 | |
| prachathai [title: multilabel] | LM all | 0 | 4.371334 | 4.142264 | 0.347872 | |
|  | LM all | 1 | 4.150784 | 3.989359 | 0.359503 | |
|  | LM all | 2 | 3.950324 | 3.895626 | 0.370871 | |
|  | LM all | 3 | 3.784429 | 3.858943 | 0.37453 | |
|  | LM all | 4 | 3.709645 | 3.854859 | 0.374904 | |
|  | CLS last 1 | 0 | 0.263054 | 0.240299 | NA | |
|  | CLS last 2 | 0 | 0.246976 | 0.22738 | NA | |
|  | CLS last 3 | 0 | 0.234152 | 0.217878 | NA | |
|  | CLS all | 0 | 0.224548 | 0.216462 | NA | |
|  | CLS all | 1 | 0.219356 | 0.211842 | NA | |
|  | CLS all | 2 | 0.213312 | 0.2097 | NA | |
|  | CLS all | 3 | 0.206874 | 0.208715 | NA | |
|  | CLS all | 4 | 0.203129 | 0.208698 | NA | |

finetuning LM: language model, CLS: classification; unfreezing all: all layers, last X: last X layer groups

Table 9: ULMFit (thai2fit) Hyperparameter Tuning Results
| datasets[features:labels] | l1 penalty | l2 penalty | f1_micro  | f1_macro  |
|--------------------------|------------|------------|-----------|-----------|
|                          | 0.5        | 0          | 0.717296  | 0.627721  |
|                          | 1          | 0          | 0.716445  | 0.622451  |
|                          | 0          | 0          | 0.688666  | 0.615289  |
|                          | 0          | 0.5        | 0.703625  | 0.602803  |
| lst20[tokens:ner_tags]   | 0.5        | 0.5        | 0.699979  | 0.590872  |
|                          | 0          | 1          | 0.694041  | 0.586303  |
|                          | 1          | 0.5        | 0.692479  | 0.585022  |
|                          | 0.5        | 1          | 0.686875  | 0.580285  |
|                          | 1          | 1          | 0.679075  | 0.574405  |
|                          | 1          | 0          | 0.952271  | 0.803645  |
|                          | 0.5        | 0          | 0.951856  | 0.80205   |
|                          | 0          | 0.5        | 0.950801  | 0.801114  |
|                          | 0.5        | 0.5        | 0.950444  | 0.798502  |
| lst20[tokens:pos_tags]   | 1          | 0.5        | 0.949195  | 0.797815  |
|                          | 0.5        | 1          | 0.94809   | 0.796299  |
|                          | 0          | 1          | 0.948829  | 0.796092  |
|                          | 1          | 1          | 0.946645  | 0.795668  |
|                          | 0          | 0          | 0.934692  | 0.790179  |
|                          | 0.5        | 0          | 0.810651  | 0.76763   |
|                          | 1          | 0          | 0.799159  | 0.770863  |
|                          | 0.5        | 0.5        | 0.776165  | 0.749212  |
|                          | 0          | 0.5        | 0.79007   | 0.745939  |
| thainer[tokens:ner_tags] | 1          | 0.5        | 0.762203  | 0.741308  |
|                          | 0          | 0          | 0.766292  | 0.739058  |
|                          | 0          | 1          | 0.770964  | 0.732445  |
|                          | 0.5        | 1          | 0.755218  | 0.727801  |
|                          | 1          | 1          | 0.742729  | 0.722583  |

Table 10: CRF Hyperparameter Tuning Results
| Tag     | Precision | Recall | F1-score | Support |
|---------|-----------|--------|----------|---------|
| DATE    | 0.8758    | 0.8221 | 0.8481   | 163     |
| EMAIL   | 1.0000    | 1.0000 | 1.0000   | 1       |
| LAW     | 0.9000    | 0.6000 | 0.7200   | 15      |
| LEN     | 0.9412    | 0.8000 | 0.8649   | 20      |
| LOCATION| 0.7747    | 0.6770 | 0.7226   | 452     |
| MONEY   | 1.0000    | 0.9138 | 0.9550   | 58      |
| ORGANIZATION | 0.8550 | 0.7400 | 0.7934   | 550     |
| PERCENT | 0.9375    | 0.9375 | 0.9375   | 16      |
| PERSON  | 0.8816    | 0.7941 | 0.8356   | 272     |
| PHONE   | 0.7500    | 0.6000 | 0.6667   | 10      |
| TIME    | 0.8154    | 0.6235 | 0.7067   | 85      |
| URL     | 1.0000    | 0.8571 | 0.9231   | 7       |
| ZIP     | 1.0000    | 0.5000 | 0.6667   | 2       |
| Micro avg | 0.8458 | 0.7408 | 0.7898   | 1651    |
| Macro avg | 0.9024 | 0.7589 | 0.8185   | 1651    |
| Weighted avg | 0.8450 | 0.7408 | 0.7889   | 1651    |

Table 11: CRF – per-class precision, recall and F1-score on test set of ThaiNER

| Tag | Precision | Recall | F1-score | Support |
|-----|-----------|--------|----------|---------|
| NN  | 0.9699    | 0.9780 | 0.9740   | 58568   |
| VV  | 0.9567    | 0.9670 | 0.9618   | 42586   |
| PU  | 0.9999    | 0.9998 | 0.9999   | 37973   |
| CC  | 0.9485    | 0.9671 | 0.9577   | 17613   |
| PS  | 0.9413    | 0.9421 | 0.9417   | 10886   |
| AX  | 0.9514    | 0.9427 | 0.9470   | 7556    |
| AV  | 0.8881    | 0.7889 | 0.8356   | 6722    |
| FX  | 0.9955    | 0.9928 | 0.9941   | 6918    |
| NU  | 0.9684    | 0.9559 | 0.9621   | 6256    |
| AJ  | 0.8974    | 0.8506 | 0.8734   | 4403    |
| CL  | 0.8781    | 0.8422 | 0.8598   | 3739    |
| PR  | 0.8479    | 0.8523 | 0.8501   | 2139    |
| NG  | 1.0000    | 0.9953 | 0.9976   | 1694    |
| PA  | 0.8122    | 0.8918 | 0.8501   | 194     |
| XX  | 0.0000    | 0.0000 | 0.0000   | 27      |
| IJ  | 0.0000    | 0.0000 | 0.0000   | 4       |
| Accuracy | 0.9628 |        |          | 207278  |
| Macro avg | 0.8160 | 0.8104 | 0.8128   | 207278  |
| Weighted avg | 0.9624 | 0.9628 | 0.9624   | 207278  |

Table 12: CRF – per-class precision, recall and F1-score on test set of LST20
### Table 13: wangchanberta-thwiki-spm – per-class precision, recall and F1-score on test set of ThaiNER

| Tag      | Precision | Recall  | F1-score | Support |
|----------|-----------|---------|----------|---------|
| DATE     | 0.2297    | 0.3988  | 0.2915   | 163     |
| EMAIL    | 1.0000    | 1.0000  | 1.0000   | 1       |
| LAW      | 0.2593    | 0.4667  | 0.3333   | 15      |
| LEN      | 0.1053    | 0.2000  | 0.1379   | 20      |
| LOCATION | 0.7635    | 0.8429  | 0.8013   | 452     |
| MONEY    | 0.1038    | 0.1897  | 0.1341   | 58      |
| ORGANIZATION | 0.7496 | 0.8491  | 0.7962   | 550     |
| PERCENT  | 0.3077    | 0.5000  | 0.3810   | 16      |
| PERSON   | 0.2414    | 0.4118  | 0.3043   | 272     |
| PHONE    | 0.6923    | 0.9000  | 0.7826   | 10      |
| TIME     | 0.1824    | 0.3176  | 0.2318   | 85      |
| URL      | 1.0000    | 1.0000  | 1.0000   | 7       |
| ZIP      | 1.0000    | 1.0000  | 1.0000   | 2       |

Micro avg 0.4922 0.6669 0.5664 1651  
Macro avg 0.5104 0.6213 0.5534 1651  
Weighted avg 0.5511 0.6669 0.5994 1651

### Table 14: wangchanberta-thwiki-spm – per-class precision, recall and F1-score on test set of LST20

| Tag | Precision | Recall  | F1-score | Support |
|-----|-----------|---------|----------|---------|
| AJ  | 0.8847    | 0.8378  | 0.8606   | 4403    |
| AV  | 0.8881    | 0.7885  | 0.8353   | 6722    |
| AX  | 0.9399    | 0.9423  | 0.9411   | 7556    |
| CC  | 0.9552    | 0.9601  | 0.9576   | 17613   |
| CL  | 0.8311    | 0.8804  | 0.8551   | 3739    |
| FX  | 0.9938    | 0.9929  | 0.9933   | 6918    |
| IJ  | 0.5000    | 0.5000  | 0.5000   | 4       |
| NG  | 0.9988    | 0.9953  | 0.9970   | 1694    |
| NN  | 0.9780    | 0.9706  | 0.9743   | 58568   |
| NU  | 0.9565    | 0.9735  | 0.9649   | 6256    |
| PA  | 0.7521    | 0.9072  | 0.8224   | 194     |
| PR  | 0.8082    | 0.8668  | 0.8365   | 2139    |
| PS  | 0.9348    | 0.9433  | 0.9391   | 10886   |
| PU  | 0.9998    | 0.9974  | 0.9986   | 37973   |
| VV  | 0.9532    | 0.9715  | 0.9623   | 42586   |
| XX  | 0.0000    | 0.0000  | 0.0000   | 27      |

Accuracy 0.9618 207278  
Macro avg 0.7004 0.7287 0.7132 18352  
Weighted avg 0.7480 0.7988 0.7718 18352

| Tag  | Precision | Recall  | F1-score | Support |
|------|-----------|---------|----------|---------|
| _BRN | 0.2692    | 0.2979  | 0.2828   | 47      |
| _DES | 0.8658    | 0.8776  | 0.8716   | 1176    |
| _DTM | 0.6494    | 0.7055  | 0.6763   | 1331    |
| _LOC | 0.6523    | 0.7395  | 0.6931   | 2349    |
| _MEA | 0.6657    | 0.7505  | 0.7056   | 3166    |
| _NUM | 0.6673    | 0.5857  | 0.6238   | 1243    |
| _ORG | 0.6978    | 0.7705  | 0.7323   | 4261    |
| _PER | 0.9186    | 0.9523  | 0.9352   | 3272    |
| _TRM | 0.6780    | 0.6250  | 0.6504   | 128     |
| _TTL | 0.9403    | 0.9826  | 0.9610   | 1379    |

Accuracy 0.9618 207278  
Macro avg 0.7453 0.7988 0.7712 18352  
Weighted avg 0.7480 0.7988 0.7718 18352
### ThaiNER (NER)

| Tag     | Precision | Recall | F1-score | Support |
|---------|-----------|--------|----------|---------|
| DATE    | 0.2456    | 0.4233 | 0.3108   | 163     |
| EMAIL   | 0.0000    | 0.0000 | 0.0000   | 1       |
| LAW     | 0.2800    | 0.4667 | 0.3500   | 15      |
| LEN     | 0.1053    | 0.2000 | 0.1379   | 20      |
| LOCATION| 0.7933    | 0.8407 | 0.8163   | 452     |
| MONEY   | 0.1028    | 0.1897 | 0.1333   | 58      |
| ORGANIZATION | 0.7764 | 0.8836 | 0.8265 | 550  |
| PERCENT | 0.3200    | 0.5000 | 0.3902   | 16      |
| PERSON  | 0.2567    | 0.4228 | 0.3194   | 272     |
| PHONE   | 0.6429    | 0.9000 | 0.7500   | 10      |
| TIME    | 0.1985    | 0.3176 | 0.2443   | 85      |
| URL     | 1.0000    | 0.8571 | 0.9231   | 7       |
| ZIP     | 1.0000    | 1.0000 | 1.0000   | 2       |

Micro avg 0.5135 0.6808 0.5854 1651
Macro avg 0.4401 0.5386 0.4771 1651
Weighted avg 0.5725 0.6808 0.6177 1651

Table 15: wangchanberta-thwiki-newmm – per-class precision, recall and F1-score on test set of ThaiNER

### LST20 (POS)

| Tag | Precision | Recall | F1-score | Support |
|-----|-----------|--------|----------|---------|
| AJ  | 0.8882    | 0.8358 | 0.8612   | 4403    |
| AV  | 0.8888    | 0.7873 | 0.8350   | 6722    |
| AX  | 0.9281    | 0.9453 | 0.9367   | 7556    |
| CC  | 0.9526    | 0.9566 | 0.9546   | 17613   |
| CL  | 0.8339    | 0.8850 | 0.8587   | 3739    |
| FX  | 0.9941    | 0.9929 | 0.9935   | 6918    |
| IJ  | 0.5000    | 0.2500 | 0.3333   | 4       |
| NG  | 0.9994    | 0.9953 | 0.9973   | 1694    |
| NN  | 0.9773    | 0.9717 | 0.9745   | 58568   |
| NU  | 0.9528    | 0.9714 | 0.9620   | 6256    |
| PA  | 0.8037    | 0.9072 | 0.8523   | 194     |
| PR  | 0.8102    | 0.8719 | 0.8399   | 2139    |
| PS  | 0.9338    | 0.9407 | 0.9372   | 10886   |
| PU  | 0.9998    | 0.9972 | 0.9985   | 37973   |
| VV  | 0.9549    | 0.9701 | 0.9625   | 42586   |
| XX  | 0.0000    | 0.0000 | 0.0000   | 27      |

Accuracy 0.9614 207278
Macro avg 0.8386 0.8299 0.8311 207278
Weighted avg 0.9613 0.9614 0.9612 207278

Table 16: wangchanberta-thwiki-newmm – per-class precision, recall and F1-score on test set of LST20
| Tag    | Precision | Recall  | F1-score | Support |
|--------|-----------|---------|----------|---------|
| DATE   | 0.8114    | 0.8712  | 0.8402   | 163     |
| EMAIL  | 0.0000    | 0.0000  | 0.0000   | 1       |
| LAW    | 0.7333    | 0.7333  | 0.7333   | 15      |
| LEN    | 0.8261    | 0.9500  | 0.8837   | 20      |
| LOCATION | 0.7792 | 0.8274  | 0.8026   | 452     |
| MONEY  | 0.8833    | 0.9138  | 0.8983   | 58      |
| ORGANIZATION | 0.7800 | 0.8636  | 0.8197   | 550     |
| PERCENT | 0.9375 | 0.9375  | 0.9375   | 16      |
| PERSON | 0.8869    | 0.9228  | 0.9045   | 272     |
| PHONE  | 0.7143    | 1.0000  | 0.8333   | 10      |
| TIME   | 0.7527    | 0.8235  | 0.7865   | 85      |
| URL    | 0.8571    | 0.8571  | 0.8571   | 7       |
| ZIP    | 1.0000    | 0.5000  | 0.6667   | 2       |

Micro avg = 0.8026, Recall = 0.8643, F1-score = 0.8323, Support = 1651
Macro avg = 0.7663, Recall = 0.7846, F1-score = 0.7664, Support = 1651
Weighted avg = 0.8041, Recall = 0.8643, F1-score = 0.8327, Support = 1651

Table 17: wangchanberta-thwiki-ssg – per-class precision, recall and F1-score on test set of ThaiNER

| Tag    | Precision | Recall  | F1-score | Support |
|--------|-----------|---------|----------|---------|
| AJ     | 0.8901    | 0.8428  | 0.8658   | 4403    |
| AV     | 0.8899    | 0.7876  | 0.8356   | 6722    |
| AX     | 0.9382    | 0.9419  | 0.9400   | 7556    |
| CC     | 0.9495    | 0.9593  | 0.9544   | 17613   |
| CL     | 0.8367    | 0.8783  | 0.8570   | 3739    |
| FX     | 0.9936    | 0.9857  | 0.9896   | 6918    |
| JJ     | 0.5000    | 0.5000  | 0.5000   | 4       |
| NG     | 0.9976    | 0.9935  | 0.9956   | 1694    |
| NN     | 0.9753    | 0.9699  | 0.9726   | 58568   |
| NU     | 0.9550    | 0.9711  | 0.9630   | 6256    |
| PA     | 0.7719    | 0.9072  | 0.8341   | 194     |
| PR     | 0.8116    | 0.8560  | 0.8332   | 2139    |
| PS     | 0.9333    | 0.9395  | 0.9364   | 10886   |
| PU     | 0.9998    | 0.9976  | 0.9987   | 37973   |
| VV     | 0.9528    | 0.9700  | 0.9613   | 42586   |
| XX     | 0.0000    | 0.0000  | 0.0000   | 27      |

Accuracy: 0.9606, Support: 207278
Macro avg: 0.7352, Recall: 0.7964, F1-score: 0.7645, Support: 18352
Weighted avg: 0.7384, Recall: 0.7964, F1-score: 0.7658, Support: 18352

Table 18: wangchanberta-thwiki-ssg – per-class precision, recall and F1-score on test set of LST20
| Tag     | Precision | Recall | F1-score | Support |
|---------|-----------|--------|----------|---------|
| DATE    | 0.8480    | 0.8896 | 0.8683   | 163     |
| EMAIL   | 0.0000    | 0.0000 | 0.0000   | 1       |
| LAW     | 0.5625    | 0.6000 | 0.5806   | 15      |
| LEN     | 0.8182    | 0.9000 | 0.8571   | 20      |
| LOCATION| 0.8038    | 0.8518 | 0.8271   | 452     |
| MONEY   | 0.9483    | 0.9483 | 0.9483   | 58      |
| ORGANIZATION | 0.8242 | 0.8782 | 0.8504 | 550 |
| PERCENT | 0.9375    | 0.9375 | 0.9375   | 16      |
| PERSON  | 0.8961    | 0.9191 | 0.9074   | 272     |
| PHONE   | 0.8750    | 0.7000 | 0.7778   | 10      |
| TIME    | 0.6970    | 0.8118 | 0.7500   | 85      |
| URL     | 0.7500    | 0.8571 | 0.8000   | 7       |
| ZIP     | 1.0000    | 1.0000 | 1.0000   | 2       |
| Micro avg | 0.8275 | 0.8746 | 0.8504 | 1651 |
| Macro avg | 0.7662 | 0.7918 | 0.7773 | 1651 |
| Weighted avg | 0.8290 | 0.8746 | 0.8509 | 1651 |

Table 19: wangchanberta-thwiki-sefr – per-class precision, recall and F1-score on test set of ThaiNER

| Tag     | Precision | Recall | F1-score | Support |
|---------|-----------|--------|----------|---------|
| AJ      | 0.9143    | 0.8188 | 0.8639   | 4403    |
| AV      | 0.8847    | 0.8079 | 0.8446   | 6722    |
| AX      | 0.9688    | 0.9277 | 0.9478   | 7556    |
| CC      | 0.9497    | 0.9678 | 0.9586   | 17613   |
| CL      | 0.8479    | 0.8869 | 0.8669   | 3739    |
| FX      | 0.9954    | 0.9915 | 0.9934   | 6918    |
| IJ      | 1.0000    | 0.5000 | 0.6667   | 4       |
| NG      | 0.9994    | 0.9941 | 0.9967   | 1694    |
| NN      | 0.9763    | 0.9747 | 0.9755   | 58568   |
| NU      | 0.9569    | 0.9731 | 0.9650   | 6256    |
| PA      | 0.7344    | 0.9124 | 0.8138   | 194     |
| PR      | 0.8346    | 0.8518 | 0.8431   | 2139    |
| PS      | 0.9295    | 0.9475 | 0.9384   | 10886   |
| PU      | 0.9999    | 0.9987 | 0.9993   | 37973   |
| VV      | 0.9566    | 0.9713 | 0.9639   | 42586   |
| XX      | 0.0000    | 0.0000 | 0.0000   | 27      |
| Accuracy | 0.9636 | 0.9636 | 0.9636 | 207278 |
| Macro avg | 0.8718 | 0.8453 | 0.8524 | 207278 |
| Weighted avg | 0.9634 | 0.9636 | 0.9633 | 207278 |

Table 20: wangchanberta-thwiki-sefr – per-class precision, recall and F1-score on test set of LST20
### ThaiNER (NER)

| Tag    | Precision | Recall | F1-score | Support |
|--------|-----------|--------|----------|---------|
| DATE   | 0.8198    | 0.8650 | 0.8418   | 163     |
| EMAIL  | 0.0000    | 0.0000 | 0.0000   | 1       |
| LAW    | 0.5263    | 0.6667 | 0.5882   | 15      |
| LEN    | 0.8261    | 0.9500 | 0.8837   | 20      |
| LOCATION | 0.8143   | 0.8827 | 0.8471   | 452     |
| MONEY  | 0.7895    | 0.7759 | 0.7826   | 58      |
| ORGANIZATION | 0.8777 | 0.9000 | 0.8887 | 550     |
| PERCENT | 0.9375 | 0.9375 | 0.9375 | 16     |
| PERSON | 0.8897    | 0.9485 | 0.9181   | 272     |
| PHONE  | 1.0000    | 1.0000 | 1.0000   | 10      |
| TIME   | 0.7500    | 0.7765 | 0.7630   | 85      |
| URL    | 0.8571    | 0.8571 | 0.8571   | 7       |
| ZIP    | 1.0000    | 1.0000 | 1.0000   | 2       |

Micro avg 0.8430 0.8879 0.8649 1651
Macro avg 0.7760 0.8123 0.7929 1651
Weighted avg 0.8439 0.8879 0.8652 1651

Table 21: wanchanberta-base-att-spm-uncased – per-class precision, recall and F1-score on test set of ThaiNER

### LST20 (POS)  
### LST20 (NER)

| Tag | Precision | Recall | F1-score | Support |
|-----|-----------|--------|----------|---------|
| AJ  | 0.9027    | 0.8685 | 0.8853   | 4403    |
| AV  | 0.9163    | 0.7849 | 0.8455   | 6722    |
| AX  | 0.9494    | 0.9464 | 0.9479   | 7556    |
| CC  | 0.9561    | 0.9611 | 0.9585   | 17613   |
| CL  | 0.8430    | 0.8930 | 0.8673   | 3739    |
| FX  | 0.9949    | 0.9928 | 0.9938   | 6918    |
| IJ  | 0.4286    | 0.7500 | 0.5455   | 4       |
| NG  | 1.0000    | 0.9970 | 0.9985   | 1694    |
| NN  | 0.9819    | 0.9744 | 0.9781   | 58568   |
| NU  | 0.9685    | 0.9823 | 0.9753   | 6256    |
| PA  | 0.7662    | 0.9124 | 0.8329   | 194     |
| PR  | 0.8240    | 0.9018 | 0.8612   | 2139    |
| PS  | 0.9383    | 0.9486 | 0.9434   | 10886   |
| PU  | 0.9999    | 0.9999 | 0.9999   | 37973   |
| VV  | 0.9566    | 0.9765 | 0.9665   | 42586   |
| XX  | 1.0000    | 0.0370 | 0.0714   | 27      |

Accuracy 0.9662 207278
Macro avg 0.7651 0.7957 0.7801 18352
Macro avg 0.7254 0.7230 0.7225 18352
Macro avg 0.7664 0.7957 0.7805 18352

Table 22: wanchanberta-base-att-spm-uncased – per-class precision, recall and F1-score on test set of LST20
| Tag     | Precision | Recall | F1-score | Support |
|---------|-----------|--------|----------|---------|
| DATE    | 0.8258    | 0.9018 | 0.8622   | 163     |
| EMAIL   | 0.0000    | 0.0000 | 0.0000   | 1       |
| LAW     | 0.6471    | 0.7333 | 0.6875   | 15      |
| LEN     | 0.7391    | 0.8500 | 0.7907   | 20      |
| LOCATION| 0.7571    | 0.8208 | 0.7877   | 452     |
| MONEY   | 0.9298    | 0.9138 | 0.9217   | 58      |
| ORGANIZATION | 0.7905 | 0.8436 | 0.8162 | 550     |
| PERCENT | 0.6000    | 0.7500 | 0.6667   | 16      |
| PERSON  | 0.8551    | 0.8897 | 0.8721   | 272     |
| PHONE   | 0.9091    | 1.0000 | 0.9524   | 10      |
| TIME    | 0.6019    | 0.7294 | 0.6596   | 85      |
| URL     | 0.8750    | 1.0000 | 0.9333   | 7       |
| ZIP     | 1.0000    | 0.5000 | 0.6667   | 2       |

Micro avg 0.7857 0.8462 0.8148 1651
Macro avg 0.7331 0.7640 0.7397 1651
Weighted avg 0.7878 0.8462 0.8155 1651

Table 23: mBERT – per-class precision, recall and F1-score on test set of ThaiNER

| Tag     | Precision | Recall | F1-score | Support |
|---------|-----------|--------|----------|---------|
| AJ      | 0.9112    | 0.8626 | 0.8862   | 4403    |
| AV      | 0.9177    | 0.7792 | 0.8428   | 6722    |
| AX      | 0.9525    | 0.9453 | 0.9489   | 7556    |
| CC      | 0.9598    | 0.9654 | 0.9626   | 17613   |
| CL      | 0.8140    | 0.9128 | 0.8606   | 3739    |
| FX      | 0.9958    | 0.9928 | 0.9943   | 6918    |
| JJ      | 1.0000    | 0.5000 | 0.6667   | 4       |
| NG      | 1.0000    | 0.9953 | 0.9976   | 1694    |
| NN      | 0.9791    | 0.9700 | 0.9745   | 58568   |
| NU      | 0.9655    | 0.9783 | 0.9718   | 6256    |
| PA      | 0.8178    | 0.9021 | 0.8578   | 194     |
| PR      | 0.8023    | 0.9392 | 0.8654   | 2139    |
| PS      | 0.9357    | 0.9612 | 0.9483   | 10886   |
| PU      | 0.9997    | 0.9979 | 0.9988   | 37973   |
| VV      | 0.9547    | 0.9693 | 0.9620   | 42586   |
| XX      | 0.0000    | 0.0000 | 0.0000   | 27      |

Accuracy 0.9644 207278
Macro avg 0.8754 0.8545 0.8586 207278
Weighted avg 0.9648 0.9644 0.9643 207278

Table 24: mBERT – per-class precision, recall and F1-score on test set of LST20
### ThaiNER (NER)

| Tag   | Precision | Recall  | F1-score | Support |
|-------|-----------|---------|----------|---------|
| DATE  | 0.8047    | 0.8344  | 0.8193   | 163     |
| EMAIL | 0.0000    | 0.0000  | 0.0000   | 1       |
| LAW   | 0.4375    | 0.4667  | 0.4516   | 15      |
| LEN   | 0.7895    | 0.7500  | 0.7692   | 20      |
| LOCATION | 0.7761 | 0.8053  | 0.7904   | 452     |
| MONEY | 0.9310    | 0.9310  | 0.9310   | 58      |
| ORGANIZATION | 0.7977 | 0.8964  | 0.8442   | 550     |
| PERCENT | 0.7222 | 0.8125  | 0.7647   | 16      |
| PERSON | 0.9078    | 0.9412  | 0.9242   | 272     |
| PHONE | 0.9091    | 1.0000  | 0.9524   | 10      |
| TIME  | 0.7561    | 0.7294  | 0.7425   | 85      |
| URL   | 0.6667    | 0.8571  | 0.7500   | 7       |
| ZIP   | 0.0000    | 0.0000  | 0.0000   | 2       |

Micro avg 0.8087 0.8577 0.8325 1651  
Macro avg 0.6537 0.6942 0.6723 1651  
Weighted avg 0.8077 0.8577 0.8315 1651

Table 25: XLMR – per-class precision, recall and F1-score on test set of ThaiNER

### LST20 (POS)

| Tag | Precision | Recall | F1-score | Support |
|-----|-----------|--------|----------|---------|
| AJ  | 0.8750    | 0.8858 | 0.8804   | 4403    |
| AV  | 0.8831    | 0.8026 | 0.8409   | 6722    |
| AX  | 0.9686    | 0.9275 | 0.9476   | 7556    |
| CC  | 0.9514    | 0.9742 | 0.9626   | 17613   |
| CL  | 0.8603    | 0.8874 | 0.8736   | 3739    |
| FX  | 0.9949    | 0.9929 | 0.9939   | 6918    |
| IJ  | 0.5000    | 0.5000 | 0.5000   | 4       |
| NG  | 0.9994    | 0.9970 | 0.9982   | 1694    |
| NN  | 0.9836    | 0.9705 | 0.9770   | 58568   |
| NU  | 0.9750    | 0.9771 | 0.9760   | 6256    |
| PA  | 0.8812    | 0.9175 | 0.8990   | 194     |
| PR  | 0.8753    | 0.8069 | 0.8397   | 2139    |
| PS  | 0.9420    | 0.9522 | 0.9471   | 10886   |
| PU  | 0.9998    | 1.0000 | 0.9999   | 37973   |
| VV  | 0.9512    | 0.9778 | 0.9643   | 42586   |
| XX  | 0.0000    | 0.0000 | 0.0000   | 27      |

Accuracy 0.9657 207278  
Macro avg 0.8526 0.8481 0.8500 207278  
Weighted avg 0.9656 0.9657 0.9655 207278

Table 26: XLMR – per-class precision, recall and F1-score on test set of LST20