Improving Large-scale Language Models and Resources for Filipino

Jan Christian Blaise Cruz and Charibeth Cheng

The 13th Language Resources and Evaluation Conference (LREC 2022)

Goals

- Create a larger, more topically diverse **dataset** for pretraining and general purpose use in Filipino.

- Create **new large language models** using this dataset to improve performance.

TLUnified

We construct a large-scale pretraining dataset for Filipino which we call **TLUnified**. We source the datasets included in this compilation from the following:

- Bilingual Text Data – Bitexts used for Filipino translation. See our paper for full list!

- OSCAR – We use the Filipino part of the Open Super-Large Crawled Aggregated Corpus (Suarez et al., 2019).

- NewsPH – We use a large-scale crawled News dataset for Filipino (Cruz et al., 2021).

TLUnified – Preprocessing

Since a large percentage of the combined dataset is crawled, we expected out-of-the-box quality to be low. We apply the following filters and cleaning steps:

Non-latin Filter – We remove sentences where >= 15% of the total characters are non-latin.

Length Filter – We only keep sentences which have a number of space-split tokens N where 4 <= N <= 150.

Punctuation Filter – We use a filtering script to remove sentences that have “too many punctuations”

Average Word Filter – If a sentence has tokens that are significantly longer than the other tokens in the sentence, we remove the sentence entirely. We first take the sum of the character lengths of each token, then divide it by the number of tokens to get a ratio r. Only sentences with ratio 3 <= r <= 18 are kept in the corpus.

HTML Filter – All sentences with HTML and URL-related tokens (e.g. “.com” or “http://”) are removed.

We then create a BPE tokenizer (Sennrich et al., 2015) using TLUnified, limited to 32,000 BPE subwords. We train with a character coverage of 1.0 and keep capitalization intact. This tokenizer is what we use for all models that use TLUnified.

RoBERTa-TL

We then pretrain large language models using our TLUnified dataset, which we call **RoBERTa-Tagalog**. Following the original (Liu et al., 2019), we produce two variants: a **110M parameter version (base)** and a **330M parameter version (large)**. Both models use the same dataset (TLUnified) and the same BPE Tokenizer (32k BPE Subwords).

Our hyperparameters are otherwise standard for our use case: we construct batches of **8192 tokens** and train with the **Adafactor** (Shazeer and Stern, 2018) optimizer, setting Beta2 to 0.98 and weight decay to 0.01. The base model is trained for 100k steps at 6e-4, while the large model is trained for 300k steps at 4e-4. Both models use a linear warmup learning rate schedule that warms-up until 25k steps have passed, then decaying to zero.

Benchmarking

To test the efficacy of our models, we test them on three Filipino benchmarks:

- Hatespeech Classification (Cabasag et al., 2019) – Binary classification task from a dataset of hatespeech collected during the 2016 elections in the Philippines.

- Dengue Topic Classification (Livelo and Cheng, 2018) – Multiclass classification task (5-way) for tweets collected about dengue.

- News NLI (Cruz et al., 2021) – Entailment classification tasks from news articles.

See our paper for details on hyperparameter choices.

Model	Hatespeech		Dengue		NewsPH-NLI Med.	
	Val. Acc	Test Acc.	Val. Acc	Test Acc.	Val. Acc	Test Acc.
BERT Base Cased	0.7479	0.7417	0.7720	0.7580	0.8838	0.8874
ELECTRA Base Cased	0.7491	0.7250	0.7400	0.6920	0.9094	0.9106
RoBERTa Base	0.7866	0.7807	0.8180	0.8020	0.9492	0.9501
RoBERTa Large	0.7897	0.7824	0.8281	0.8110	0.9499	0.9510

RoBERTa-TL **outperforms both our previous models (BERT and ELECTRA) in all tasks**, even when comparing using the identical model sizes (Base, 110M parameters)

Interestingly, the Large variant **only marginally outperforms** the Base variant. We hypothesize that, while the dataset is improved, we still do not have enough data to make the most out of the model capacity.