LaoPLM: Pre-trained Language Models for Lao

Nankai Lin, Yingwen Fu, Ziyu Yang, Chuwei Chen and Shengyi Jiang

Guangdong University of Foreign Studies, Guangzhou 510006, China

Guangzhou Key Laboratory of Multilingual Intelligent Processing, Guangzhou, China

Background

The use of pre-trained language models (PLMs) represented by BERT in natural language processing (NLP) has achieved great success in multiple areas. PLMs do not rely on any manually annotated training data but help to produce significant performance gains for various NLP tasks, making them recently become extremely popular.

Problem for Lao's PLM

There are currently no monolingual PLMs for Lao, which has brought certain restrictions to the development of Lao language technology.

Many monolingual and multilingual models are only pre-trained on Wikipedia corpus. At the same time, the size of Lao Wiki data is relatively small, which brings a serious impact on the performance of the pre-trained models.

Multilingual pre-trained models struggle to explain their applicability in acquiring language-invariant knowledge for downstream tasks of various languages. However, due to the different pre-training corpus sizes for different languages, the multilingual pre-trained model tend to be biased towards high-resource languages, such as English.

As Lao is a language that has no explicit delimiters between words, directly applying Byte-Pair encoding (BPE) methods to the Lao pre-training data may bring a performance drop to the pre-trained models.

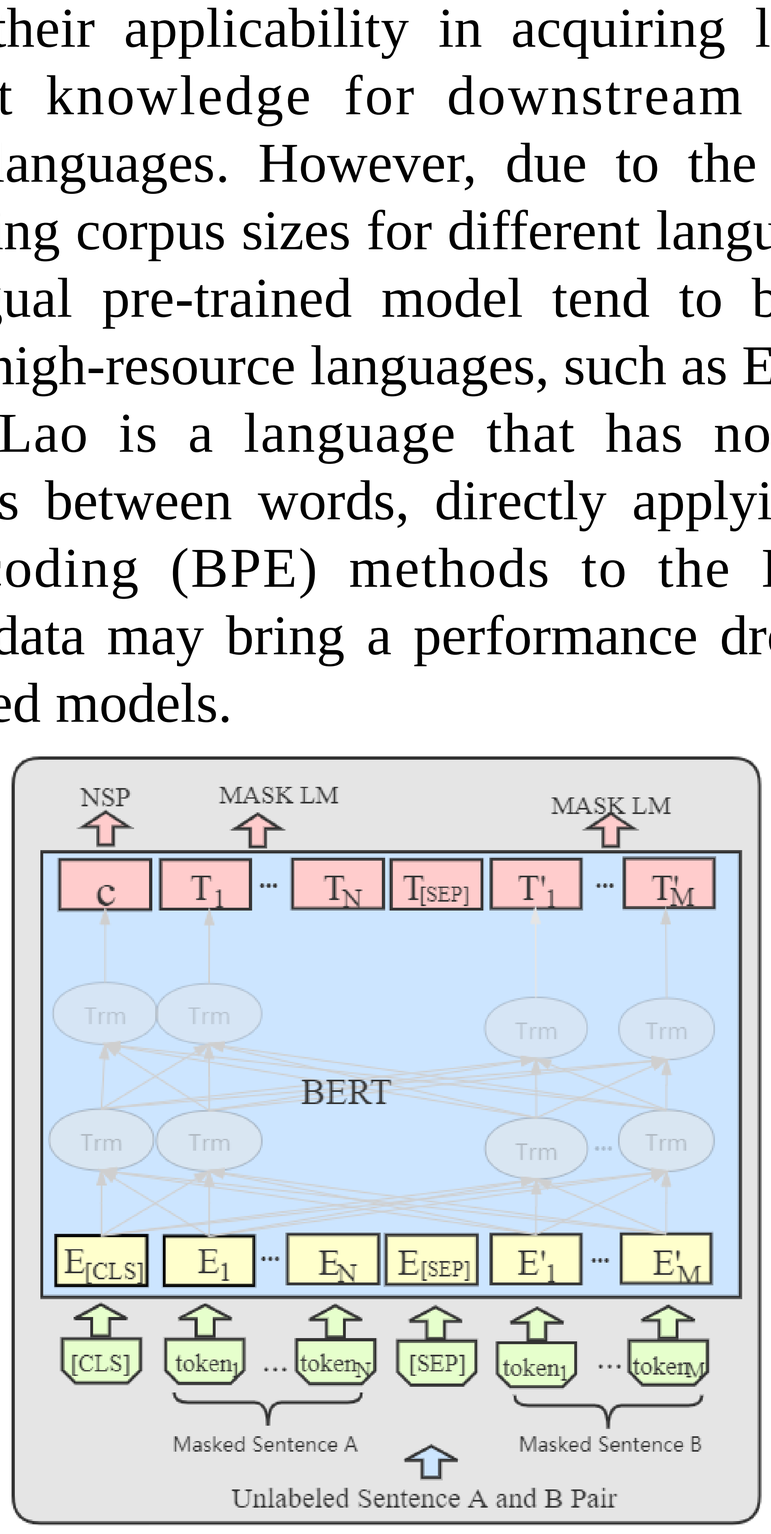

Figure 1: BERT Model

Model

We pre-train two kinds of transformer-based models, namely BERT (figure 1) and ELECTRA.

Data

Table 1. Statistics of the POS tagging dataset.

Data	Num. of Sentence	Num. of Token
Train	6400	94464
Dev	1600	23686
Test	3000	44849
Total	10000	162999

Table 2. Statistics of our dataset for Lao news categorization.

Category	Num. of articles	Num. of articles in the training set	Num. of articles in the validation set	Num. of articles in the test set
Politics	754	526	76	152
Economy	494	344	50	100
Society	947	662	95	190
Military	103	70	11	22
Environment	80	56	8	16
Culture	119	83	12	24
Technology	102	70	11	21
Others	369	258	37	74

Table 3. Statistics of the pre-training corpus.

Source	Num. of Lines	Size of File
Oscar	143888	113m
CC100	2570964	625m
All	2714852	738m

Result

Table 4. Performance of POS tagging.

Model	Accuracy
AMFF	90.32%
BERT-Small	92.37%
BERT-Base	87.18%
ELECTRA-Small	88.47%
ELECTRA-Base	89.78%
XLM-RoBERTa-Base	88.40%

Table 5. Performance of News Classification.

Model	Strategy	F1-Score	Accuracy
BERT-Small	-	66.03%	71.95%
	Upsampling	66.61%	72.45%
	EasyEnsemble	66.47%	71.11%
BERT-Base	-	67.87%	72.95%
	Upsampling	68.33%	72.95%
	EasyEnsemble	66.37%	71.79%
ELECTRA-Small	-	64.65%	71.62%
	Upsampling	67.55%	71.29%
	EasyEnsemble	65.57%	70.62%
ELECTRA-Base	-	53.03%	62.94%
	Upsampling	58.26%	66.44%
	EasyEnsemble	52.83%	62.27%