Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 6506–6512
Marseille, 20-25 June 2022

LaoPLM: Pre-trained Language Models for Lao

Nankai Lin[1], Yingwen Fu[1], Ziyu Yang[1], Chuwei Chen[1] and Shengyi Jiang[1,2]
[1] Guangdong University of Foreign Studies, Guangzhou, China
[2] Guangzhou Key Laboratory of Multilingual Intelligent Processing, Guangzhou, China
{20191010004, 20201010002, 20201002958, 20202005086}@gdufs.edu.cn
jiangshengyi@163.com

Abstract

Trained on large corpora, pre-trained language models (PLMs) can capture different levels of concepts in context and hence generate universal language representations. They are benefitcial for multiple downstream natural language processing (NLP) tasks. Although PTMs have been widely used in most NLP applications, especially for high-resource languages such as English, it is under-represented in Lao NLP research. Previous works on Lao have been hampered by the lack of annotated datasets and the sparsity of language resources. In this work, we construct a text classification dataset to alleviate the resource-scarce situation of the Lao language. In addition, we present the first transformer-based PTMs for Lao with four versions: *BERT-Small*[1], *BERT-Base*[2], *ELECTRA-Small*[3], and *ELECTRA-Base*[4]. Furthermore, we evaluate them on two downstream tasks: part-of-speech (POS) tagging and text classification. Experiments demonstrate the effectiveness of our Lao models. We release our models and datasets to the community, hoping to facilitate the future development of Lao NLP applications.

Keywords: Pre-trained Language Model, Lao, Text Classification, Part-of-speech Tagging

1. Introduction

The use of pre-trained language models (PLMs) represented by BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019) 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. BERT-based PLMs could be categorized into two classes: (1) Monolingual models are language-specific models trained in monolingual datasets (Cui et al., 2019; de Vries et. al., 2019; Vu Xuan et al., 2019; Martin et al., 2020; Nguyen and Tuan Nguyen, 2020). However, the success of monolingual models has been primarily limited to high-resource languages represented by English. (2) Multilingual models (Devlin et al., 2019; Conneau and Lample, 2019; Conneau et al., 2020; Xue et al., 2021) are trained in datasets from multiple languages and simultaneously support downstream tasks for multiple languages.

When it comes to Lao language modeling, there are some concerns to the best of our knowledge:

- 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. It is worth noting that Wikipedia data is not representative of general language use. At the same time, the Lao Wiki data size is relatively small, which may negatively impact the performance of the pre-trained models. PLMs can be significantly improved by using more pre-training data (Liu et al., 2019).
- 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. In addition, as different languages have different sequence structures, multilingual pre-trained models are more suitable for application in cross-language research than in monolingual research. As Lao is a language with no explicit delimiters between tokens, directly applying Byte-Pair encoding (BPE) methods (as previously common BERT-based models) to the Lao pre-training data may bring a performance drop to the pre-trained models. It is necessary to pre-train monolingual models for Lao to improve the performance of Lao downstream tasks.

To alleviate the concerns above, in this paper, we use Oscar corpus (Suárez et al., 2019) and CC-100 corpus (Conneau et al., 2020) to train the first monolingual BERT-based PLMs for Lao with four versions: *BERT-Small*, *BERT-Base*, *ELECTRA-Small*, and *ELECTRA-Base*. Instead of directly adopting the BPE method, we utilize sentence-piece segmentation on Lao pre-training data to tackle the problem of no explicit delimiters between tokens. The pre-trained models are then evaluated on two NLP tasks: (1) a sequence labeling task of part-of-speech (POS) tagging and (2) a text classification task of news classification. The POS tagging dataset is an open-source dataset from Yunshan Cup 2020. The news classification dataset is self-constructed to alleviate the scarce classification resource in Lao. The dataset consists of 2968 news articles with 8 news classes.

In summary, our contributions are as follows:

- We present the first four BERT-based PLMs for Lao based on a large-size corpus.

[1] https://huggingface.co/GKLMIP/bert-laos-small-uncased
[2] https://huggingface.co/GKLMIP/bert-laos-base-uncased
[3] https://huggingface.co/GKLMIP/electra-laos-small-uncased
[4] https://huggingface.co/GKLMIP/electra-laos-base-uncased

Nankai Lin and Yingwen Fu are the co-first authors. They have worked together and contributed equally to the paper.

- A large-scale and high-quality Lao news classification dataset is constructed to alleviate the current situation of insufficient language resources.
- Our models achieve competitive performances on news classification and POS tagging tasks, showing the superiority of large-scale BERT-based monolingual language models for Lao.
- We publicly release all pre-trained models and datasets in an open repository[5].

2. Related Work

2.1 Lao Text Classification

Text classification is a fundamental supervised task in NLP that aims to assign one of the pre-defined categories for an input sequence. Lao is represented as a low-resource language and there is little classification research for Lao text. Most of the current classification methods for Lao are based on machine learning (ML): Vilavong and Huy (2015) utilize two advanced ML techniques, namely radial basis function (RBF) network, and support vector machines (SVM), to classify Lao documents. Chen et al. (2020) propose a KNN-based classification method for the Lao news text classification.

2.2 Lao Part-of-speech Tagging

Part-of-speech (POS) tagging is defined as a sequence labeling task that assigns the correct POS tag for each token in the input sequence based on its morphological and syntactic behaviors. Yang et al. (2016) present a semi-supervised approach for the Lao POS tagging task to alleviate the problem of little labeled resources. Wang et al. (2019) propose an approach combining neural token prediction with a hidden Markov-based (Rabiner and Juang, 1986) semi-supervised method to label Lao POS. Wang et al. (2019) study the structural characteristics of Lao tokens and propose a multi-task (Caruana, 1993) attention-based (Bahdanau et al., 2015) Lao POS tagging model with a combination of POS tagging loss with the main consonant auxiliary loss. Tang et al. (2021) propose a method for the Lao POS tagging task which integrates fine-grained token features to build an Attention-Bi-LSTM-CRF model.

2.3 Lao Pre-trained Language Model

There is no open-source monolingual pre-trained model for Lao. Open-source multilingual pre-trained models represented by mBERT (Devlin et al., 2019), XLM (Conneau and Lample, 2019), XLM-RoBERTa (Conneau et al., 2020), and mT5 (Xue et al., 2021) are trained in a large-scale multilingual dataset aiming to learn language-independent knowledge and then support various downstream NLP tasks for multiple languages. Among them, XLM-RoBERTa and mT5 support the Lao language while mBERT excludes the Lao language. However, because of the huge language discrepancy between Lao and other languages, multilingual pre-trained models do not perform well in Lao downstream tasks.

3. Model

We pre-train two kinds of transformer-based models, namely BERT (Devlin et al., 2019) and ELECTRA (Clark et al., 2020).

3.1 BERT

BERT is a transformer-based (Vaswani et al., 2017) language model that is designed to pre-train on a large unsupervised dataset to learn deep bidirectional representations. It can be fine-tuned to multiple benchmarks and achieves competitive results. It consists of two subtasks, namely Mask Language Model (MLM) and Next Sentence Prediction (NSP): (1) MLM is designed to mask some tokens in the input sequence and then predict the masked tokens according to the context; (2) NSP refers to predicting whether the sentence pair is continuous.

Our pre-trained LaoBERT has two versions, LaoBERT-Base and LaoBERT-Small. They follow the same architecture of BERT-Base (12 layers, 768 hidden units, 12 attention heads) and BERT-Small (4 layers, 512 hidden units, 8 attention heads), respectively. The two models accept a maximum sequence length of 512.

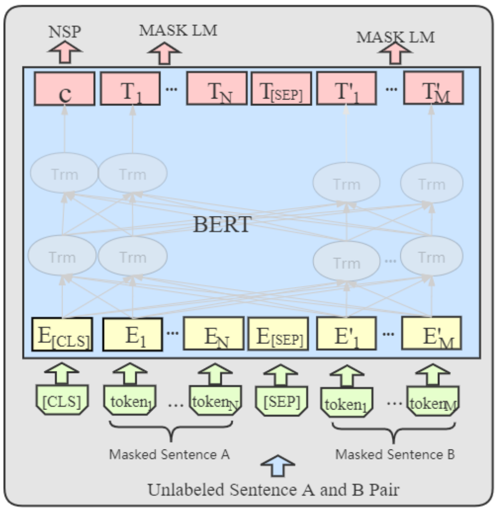

Figure 1: BERT model

3.2 ELECTRA

ELECTRA (Clark et al., 2020) is another transformer-based pre-training framework that leverages the combination of generator and discriminator based on the idea of GAN.

As shown in Figure 2, the structure of ELECTRA can be divided into two parts. The generator is similar to BERT which is pre-trained with MLM objective. The discriminator is trained with a novel binary classification task called replaced token detection (RTD). Rather than masking the input tokens randomly, RTD tries to construct a corrupted sequence by replacing some tokens in the original input with plausible alternatives sampled from the generator. And then the discriminator takes the corrupted sequence as input and identifies whether each token has been replaced by the generator or not. The loss function is summed up with MLM and RTD loss. ELECTRA poses an advantage over the widely used BERT in its ability to use pretraining data more efficiently, as BERT only uses 15% tokens of the training data for the MLM task per epoch which may lead to data inefficiency. As for fine-tuning, the generator is discarded and the discriminator is adopted as the pre-trained ELECTRA model.

[5] https://huggingface.co/GKLMIP

We produce two ELECTRA models respectively in the base size (12 layers, 768 hidden units, 12 attention heads), and the small size (4 layers, 512 hidden units, 8 attention heads). The two models accept a maximum sequence length of 512.

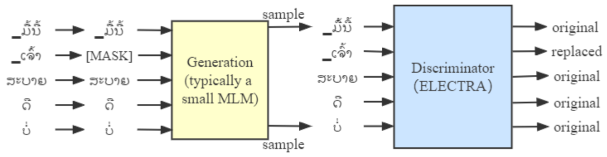

Figure 2: ELECTRA model. The meaning of the sentence example is “*How are you today*”.

4. Downstream Tasks

We evaluate our pre-trained models on two downstream NLP tasks: POS tagging and text classification. In the following sections, we will briefly introduce each task, along with the evaluation datasets and procedures.

4.1 POS tagging

The dataset utilized for POS tagging evaluation comes from Yunshan Cup 2020 Lao POS tagging track[6] (Fu et al., 2022). The dataset consists of 10000 sentences (162999 tokens totally) with 26 POS labels. We reassign the dataset into **(6400, 1600, 3000) sentences for (training, test, validation).** Some statistics about this dataset are shown in Table 1 and Table 2.

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

Table 1: Statistics of the POS tagging dataset

4.2 News Classification

We obtain 2968 Lao news articles from the China Radio International website[7]. According to the news system of Wang et al. (2021), we annotate the news as one of the following classes: ***politics***, ***economy***, ***society***, ***military***, ***environment***, ***culture***, ***technology***, and ***others***. We invite Laotian experts and scholars to label each sample. Each sample is annotated by two annotators. For each sample, if different annotated results are produced, a third person would further annotate the sample. The dataset is divided into three parts, **with the training/validation/test split of 70%/10%/20%**. It should be pointed out that since different classes have significantly different numbers of articles, the split is conducted on the class level instead of the dataset level to preserve the percentage of samples for each class. The detailed statistics of the Lao news classification dataset are presented in Table 3.

Tag	Proportion (%)	Explanation
IAC	0.7472	Indefinite determiner
COJ	5.2497	Conjunction
ONM	0.0251	Ordinal number
PRE	5.6386	Completed
PRS	2.8202	Preposition
V	19.6682	Verb
DBQ	0.3294	Pre-quantifier
IBQ	0.0190	Indefinite qualifier
FIX	0.5889	Preposition
N	30.7756	Common noun
ADJ	5.0184	Adjective
DMN	1.0374	Demonstrative
IAQ	0.0797	Indefinite qualifier
CLF	1.8202	Quantifier
PRA	2.7423	Pre-auxiliary verb
DAN	0.3007	Post-noun determiner
NEG	1.1441	Negative Words
NTR	0.7815	Interrogative pronouns
REL	1.1693	Relative pronouns
PVA	0.8423	Post auxiliary verb
TTL	0.3288	Title noun
DAQ	0.0226	Post-quantifier
PRN	10.1264	Proper nouns
ADV	3.6153	Adverb
PUNCT	4.8613	Punctuation
CNM	0.5411	Cardinal

Table 2: Tagset of the POS tagging dataset

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 our dataset for Lao news classification

[6] https://github.com/GKLMIP/Yunshan-Cup-2020

[7] http://laos.cri.cn/

5. Experiment

5.1 Pre-training

We collect texts from different sources to train our models. On the one hand, we utilize all the Lao data from the OSCAR corpus [8] (Suárez et al., 2019), a humongous multilingual corpus whose texts all come from the Common Crawl corpus[9]. Suárez et al. (2019) propose an architecture to perform language classification and apply the model to the Common Crawl corpus. At last, they obtain the language-classified and ready-to-use OSCAR, with 166 different languages available so far. In addition, articles on CC-100 (Conneau et al., 2020) are also used as part of our corpus for pre-training. This corpus is constructed for training XLM-R. It consists of monolingual data for 100+ languages and also includes data for Romanized languages. The corpus statistics for pre-training are shown in table 4.

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

Table 4: Statistics of the pre-training corpus

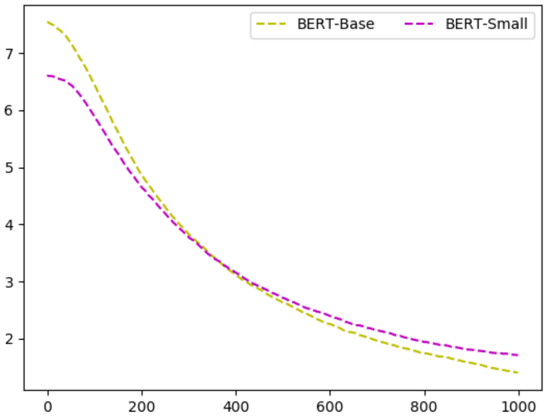

Figure 3: Pre-training losses for BERT over different training steps

The batch size for pre-training is set as 8. All the models are trained on the pre-training data for 1,000,000 steps. The learning rates of BERT-Small and ELECTRA-Base are all warmed up over the first 5,000 steps to a peak value of 1e-4, and then decay linearly. The learning rate of BERT-Base is 5e-5, and the learning rate of ELECTRA-Small is 2e-4. The weights are initialized randomly from a normal distribution with a mean of 0.0 and a standard deviation of 0.02. Instead of directly adopting the BPE method, we utilize sentence-piece segmentation on Lao pre-training data to tackle the problem of no explicit delimiters between tokens. We directly adopt the sentence-piece segmentation mode[10] trained by Heinzerling and Strube (2019), which has a vocabulary size of 25,000. Figure 3 and Figure 4 illustrate the pre-training loss for each model. It could be observed that given the same training time, the deeper and wider models could greatly help to achieve lower training loss than the shallower models.

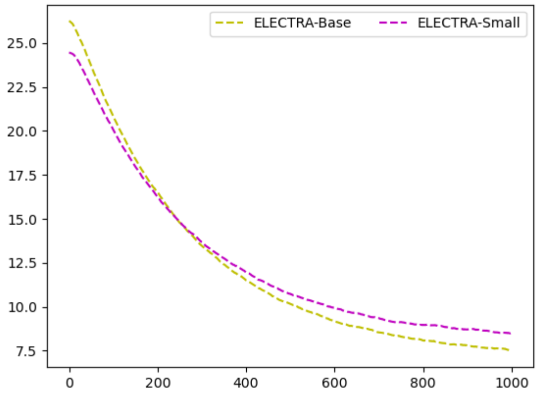

Figure 4: Pre-training losses for ELECTRA over different training steps

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: POS tagging performance

5.2 News Classification

Model	F1-Score	Accuracy
BERT-Small	66.03%	71.95%
BERT-Base	**67.87%**	**72.95%**
ELECTRA-Small	64.65%	71.62%
ELECTRA-Base	53.03%	62.94%
XLM-RoBERTa-Base	64.00%	71.12%

Table 6: News classification performance

5.3 News Classification

For two small models, we fine-tune for 5 epochs with a learning rate of 1e-4. For the BERT-Base model and XLM-RoBERTa-Base model, we fine-tune for 5 epochs with a learning rate of 5e-5. For the ELECTRA-Base model, we fine-tune for 10 epochs with a learning rate of 2e-5 because it needs a longer training time and a smaller learning rate to converge. As shown in Table 6, on the whole, BERT models outperform ELECTRA models on the Lao text classification task, and the base-size models perform better than the small-size models. Among them, BERT-Base achieves the best results with an F1 score of 66.03% and an accuracy of 71.95%. In addition, we find that compared

[8] https://oscar-corpus.com/
[9] https://commoncrawl.org/
[10] https://github.com/bheinzerling/bpemb

Category	BERT(Small)	BERT(Base)	ELECTRA(Small)	ELECTRA(Base)
Politics	80.27%	81.21%	**81.29%**	75.42%
Economy	77.88%	**81.52%**	78.43%	72.32%
Society	75.90%	**77.63%**	75.96%	68.62%
Military	61.90%	**69.77%**	57.78%	42.86%
Environment	68.75%	**77.78%**	64.52%	38.10%
Culture	**65.12%**	59.57%	61.22%	47.62%
Technology	51.43%	**52.63%**	51.28%	41.38%
Others	**46.98%**	42.86%	46.75%	37.96%

Table 7: F1 scores of each news class

to the other three models, the performance of the ELECTRA-Base model is relatively poor. Therefore, we conduct an error analysis by checking the model performances with the help of a confusion matrix. We consider the macro F1 scores on each news category (Table 7) and the top 5 mistakes on the test set (Table 8). Some in-depth analyses are drawn as follows:

- Firstly, the three classes with more samples (politics, economy, and society) perform better in each model, with all F1 scores above 0.7. In comparison, the other classes with fewer samples tend to have poor performances. For ELECTRA-Base, the information of the classes with fewer samples cannot be learned well, which explains why the F1 score of the ELECTRA-Base model is only 53.03%.
- Secondly, we find that all models tend to confuse between some classes, especially society and other classes.
- Thirdly, we realize that the models might suffer from class imbalance problems, as they perform relatively poorly in the class with the fewest articles.

To deal with the class imbalance problem, we employ two simple yet effective sampling strategies, EasyEnsemble (Liu et al., 2009) and Upsampling (Rajagede and Hastuti, 2021), These two strategies sample subsets from the majority classes, train learners on each of them, and ensemble all these weak learners for a final model. In the EasyEnsemble experiment, we generate a total of 5 subsets, each of which satisfies the same class distribution. For each subset, the sample ratios are 1.0 for military, environment, culture, and technology classes, 0.3 for politics and society classes, and 0.5 for others class. In the Upsampling experiments, for each subset, the sample times are 7 for military, environment, culture, and technology classes, 1 for politics and society classes, and 2 for others class. The results are shown in Table 9. As we can see, the two strategies above can significantly improve the model performances. BERT-Base model performs best on the Upsampling framework, with the F1 score of 68.33%. While using the Upsampling strategy, the F1-score of the ELECTRA-Base model is improved by 5.13%, which verifies that the purely ELECTRA-Base model is less effective due to the influence of imbalanced data.

We further consider the macro F1 scores on each news class of the BERT-Base model with the best performance and the ELECTRA-Base model with the most significant improvement. As can be observed from Table 10 and Table 11, experimental results show that in the ELECTRA-Base model, two strategies greatly improve the classification performance in small-sample classes (military, environment, and culture), while for the BERT-Base model, only the Upsampling strategy can improve the model to a certain extent. The Upsampling strategy brings significant improvement to classes with fewer samples, while the performance of the classes with more samples decreases slightly. It can be seen that the ELECTRA-Base model depends more on training data size for this task. With the Upsampling strategy, ELECTRA-Base could learn the class information more adequately.

Model	Ref	Hyp.	Freq.
BERT-Small	Others	Society	22
	Society	Others	17
	Politics	Society	13
	Politics	Economy	9
	Society	Politics	9
BERT-Base	Society	Others	21
	Others	Society	18
	Politics	Economy	10
	Politics	Others	9
	Society	Politics	9
ELECTRA-Small	Others	Society	20
	Society	Others	20
	Society	Politics	11
	Politics	Society	8
	Society	Economy	8
ELECTRA-Base	Society	Others	53
	Politics	Economy	20
	Others	Society	16
	Politics	Others	14
	Economy	Others	12

Table 8: Top 5 mistakes on the test set

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%

Table 9: Performance of two strategies

Category	BERT-Base	BERT-Base with Upsampling	BERT-Base with EasyEnsemble
Politics	**81.21%**	79.61%	78.81%
Economy	**81.52%**	81.31%	77.88%
Society	**77.63%**	76.88%	77.55%
Military	**69.77%**	65.22%	68.42%
Environment	**77.78%**	76.92%	71.79%
Culture	59.57%	**66.67%**	65.31%
Technology	52.63%	**54.05%**	52.94%
Others	42.86%	**45.95%**	38.24%

Table 10: Performance of each class on BERT under two strategies

Category	ELECTRA-Base	ELECTRA-Base with Upsampling	ELECTRA-Base with EasyEnsemble
Politics	**75.42%**	73.19%	72.42%
Economy	72.32%	**77.51%**	75.14%
Society	68.62%	**73.26%**	69.44%
Military	42.86%	**59.46%**	55.00%
Environment	38.10%	**54.55%**	45.45%
Culture	**47.62%**	44.44%	35.20%
Technology	41.38%	**46.67%**	45.71%
Others	**37.96%**	36.99%	24.24%

Table 11: Performance of each category on ELECTRA under two strategies

6. Conclusion

In this paper, we fill the gaps in the scarcity of pre-trained language models and open-source text classification datasets for the Lao language. We pre-train four Lao language models and evaluate the model performances on the part-of-speech tagging task and news classification task. We release our models and datasets to the community, hoping to facilitate the future development of Lao NLP applications. Given the class imbalance problem and the small size of the classification dataset, we will further expand the dataset size in the future.

7. Acknowledgement

This work is supported by the National Natural Science Foundation of China (No. 61572145), the Major Projects of Guangdong Education Department for Foundation Research and Applied Research (No. 2017KZDXM031) and National Social Science Foundation of China (No. 17CTQ045). The authors would like to thank the anonymous reviewers for their valuable comments and suggestions.

8. Bibliographical References

Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171-4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Cui, Y., Che, W., Liu, T. Qin, B. Wang, S., and Hu, G. (2019). Revisiting Pre-Trained Models for Chinese Natural Language Processing. In *Findings of the Association for Computational Linguistics: EMNLP 2020,* pages 657-668, Online, November. Association for Computational Linguistics.

de Vries, W., van Cranenburgh, A., Bisazza, A., Caselli, T., Noord, G. V., and Nissim, M. (2019). BERTje: A Dutch BERT Model. *arXiv preprint*, arXiv:1912.09582.

Vu Xuan, S., Vu, T., Tran, S., and Jiang, L. (2019). ETNLP: A Visual-Aided Systematic Approach to Select Pre-Trained Embeddings for a Downstream Task. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1285-1294, Varna, Bulgaria, September. INCOMA Ltd.

Martin, L., Muller, B., Ortiz Suárez, P. J., Dupont, Y., Romary, L., de la Clergerie, É., Seddah, D., and Sagot, B. (2020). CamemBERT: a Tasty French Language Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203-7219, Online, July. Association for Computational Linguistics.

Nguyen, D. Q. and Tuan Nguyen, A. (2020). PhoBERT: Pre-trained language models for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1037-1042, Online, November. Association for Computational Linguistics.

Conneau, A. and Lample G. (2019). Cross-lingual Language Model Pretraining. In the *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, Vancouver, Canada. Curran Associates, Inc.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440-8451, Online, July. Association for Computational Linguistics.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2021). mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online, June. Association for Computational Linguistics.

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 preprint*, arXiv: 1907.11692.

Vilavong S. and Huy, K. P. (2015). Comparison on some machine learning methods for Lao text categorization. *International Journal of Computer Science and Telecommunications*, 2(7):8-13.

Chen, Z., Zhou, L., Li, X., Zhang J., and Huo, W. (2020). The Lao Text Classification Method Based on KNN. *Procedia Computer Science*, 166:523-528.

Yang, B., Zhou, L., Yu, Z., and Liu, L. (2016). Research on Semi-supervised Learning Based Approach for Lao Part of Speech Tagging. *Computer Science*, 43(9):103-106.

Wang, X., Zhou, L., Zhang, J., Zhou F., and Guo, J. (2019). Research on the Fusion of Semi-supervised Lao Part of Speech Tagging and Word Prediction. *Journal of Chinese Computer Systems*, 40(12):2500-2505.

Rabiner, L. R. and Juang, B. H. (1986). An introduction to hidden Markov models. *IEEE ASSP Magazine*, 3(1):4-16.

Wang, X., Zhou, L., Zhang J., and Zhou, F. (2019). A Multi-task Lao Part-of-Speech Tagging Method Fusing Structural Features of Word. *Journal of Chinese Information Processing*, 33(11):39-45.

Caruana, R. (1993). Multitask learning: A knowledge-based source of inductive bias. In *Proceedings of the Tenth International Conference on International Conference on Machine Learning*, pages 41-48, Amherst, MA, USA, June. Morgan Kaufmann Publishers Inc.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. In *the 3rd International Conference on Learning Representations (ICLR 2015)*, San Diego, CA, USA, May.

Tang, W., Zhou, L., and Zhang, J. (2021). On Part-of-speech Tagging of Lao by Integrating Fine-grained Word Features. *Journal of Chinese Computer Systems*, 40(12).

Clark, K., Luong, M., V. Le, Q., and D. Manning, C. (2020). ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators, In *the 8th International Conference on Learning Representations (ICLR 2020)*, Online, May.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is All you Need. In *the 31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA. Curran Associates, Inc.

Wang, L., Lin, X., and Lin, N. (2021). Research on Pseudo-label Technology for Multi-label News Classification. In the *16th International Conference on Document Analysis and Recognition*, pages 683-698, Online, September. Springer.

Heinzerling, B. and Strube, M. (2019). BPEmb: Tokenization-free Pre-trained Subword Embeddings in 275 Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Yang, Z., Chen, H., Zhang, J., Ma, J., and Chang, Y. (2020). Attention-based Multi-level Feature Fusion for Named Entity Recognition. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 3594-3600, Online. International Joint Conferences on Artificial Intelligence Organization.

Xu, Y., Wu, J., and Zhou, Z. (2009). Exploratory Undersampling for Class-Imbalance Learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 39(2):539-550.

Rajagede, R. A. and Hastuti, R. P. (2021). Stacking Neural Network Models for Automatic Short Answer Scoring. In *Proceedings of the 5th International Conference on Information Technology and Digital*, Yogyakarta, Indonesia, November. IOP Publishing.

Suárez, P. J. O., Sagot, B., and Romary, L. (2019). Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures. In *Proceedings of the 7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*.

Fu, Y., Chen, J., Lin, N., Huang, X., Qiu, X., and Jiang, S. (2022). Yunshan Cup 2020: Overview of the Part-of-Speech Tagging Task for Low-resourced Languages. *arXiv preprint*, arXiv: 2204.02658.