7-2021

An efficient transformer-based model for Vietnamese punctuation prediction

Hieu TRAN

Cuong V. DINH

Hong Quang PHAM
Singapore Management University, hqpham.2017@phdis.smu.edu.sg

Binh T. NGUYEN

Citation

TRAN, Hieu; DINH, Cuong V.; PHAM, Hong Quang; and NGUYEN, Binh T.. An efficient transformer-based model for Vietnamese punctuation prediction. (2021). *Advances and trends in artificial intelligence: International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2021, July 26-29, Kuala Lumpur, Virtual*. 47-58.
Available at: https://ink.library.smu.edu.sg/sis_research/7102

An Efficient Transformer-Based Model for Vietnamese Punctuation Prediction

Hieu Tran[2,3,4], Cuong V. Dinh[1], Quang Pham[5], and Binh T. Nguyen[2,3,4(✉)]

[1] Dublin City University, Dublin, Ireland
[2] AISIA Research Lab, Ho Chi Minh City, Vietnam
ngtbinh@hcmus.edu.vn
[3] University of Science, Ho Chi Minh City, Vietnam
[4] Vietnam National University, Ho Chi Minh City, Vietnam
[5] Singapore Management University, Singapore, Singapore

Abstract. In both formal and informal texts, missing punctuation marks make the texts confusing and challenging to read. This paper aims to conduct exhaustive experiments to investigate the benefits of the pre-trained Transformer-based models on two Vietnamese punctuation datasets. The experimental results show our models can achieve encouraging results, and adding Bi-LSTM or/and CRF layers on top of the proposed models can also boost model performance. Finally, our best model can significantly bypass state-of-the-art approaches on both the novel and news datasets for the Vietnamese language. It can gain the corresponding performance up to 21.45% and 18.27% in the overall F1-scores.

Keywords: Transfer learning · Transformer models · Punctuation prediction

1 Introduction

In different languages, punctuation is a collection of symbols designating a sentence structure to slow down, remark, or expose emotion. Using punctuation marks is an essential step in writing to make each sentence or paragraph easy to read and understand. In many formal and informal texts, the misuse of punctuation marks frequently happens due to the lack of knowledge in grammars or human mistakes [1,2]. Along with developing Automatic Speech Recognition (ASR) systems, speech transcripts may not have any punctuation marks. Accordingly, selecting relevant punctuation marks to transcribed text is vital to ensure one can correctly understand the text.

Punctuation prediction is an indispensable problem in multiple languages. Some studies treat this problem as a sequence labeling task and settle it using neural networks in recent years. There have been various approaches to punctuation prediction in different languages (English, Chinese, and Slovenia) in recent years, achieving particular accomplishments. Previous studies utilize both statistic models and traditional deep neural networks to settle the missing punctuation

H. Fujita et al. (Eds.): IEA/AIE 2021, LNAI 12799, pp. 47–58, 2021.
https://doi.org/10.1007/978-3-030-79463-7_5

marks problem [3–6]. One of the earliest works on the English language is in [7], where the authors presented the stacked RNNs model to learn more hierarchical aspects. It can be followed by a layer-wise multi-head attention mechanism to focus on the relevant contexts at each time step and capture the features directly from each hierarchical level. Makhija and colleagues [8] utilized the BERT language model to capture contextualized word embeddings, fed into a hybrid of BiLSTM and CRF layer after that. The authors evaluate the model's effectiveness on the IWSLT2012 English dataset, and these models achieve an overall F1-score of 81.4%, which is higher than the previous models' score. Fang et al. [9] tried using the same architecture for the Chinese punctuation prediction task. The experiment of this model shows the pre-trained language model also accomplish on Chinese characteristics. Furthermore, this work indicates that the CRF layer does not make performance increase clearly, so it has little effect on this problem. Wang and co-workers [10] addresses the punctuation prediction problem like machine translation instead of sequence labeling. The authors use Transformer architecture and two softmax layers: the label softmax and word softmax. The combination of word sequence information and labeling information significantly improves compared to the previous models.

In previous Vietnamese studies, Quang et al. [1] apply the CRF model and a set of appropriate features for solving the punctuation prediction problem. To inspect the model's effectiveness, the authors ask some volunteers to insert punctuation marks into a small dataset and compare them with their model. The model has achieved approximately human performance. Thuy and colleagues [2] investigate the deep neural networks Bi-LSTM model in two novels and news dataset. To enhance the model for capturing more complex data structures, the authors add an attention mechanism on top of the Bi-LSTM model. It makes the model can focus on particular syllables in the past while predicting the current punctuation mark. This study also benchmarks the traditional CRF method by replacing the softmax layer. Cross-Entropy loss is not suitable for the different distribution of punctuation marks. This work proposes to use the focal loss that can give more weights to rare classes in the data. The experiment results show the combination between the Bi-LSTM model and the attention mechanism outperforms the others. Besides that, the CRF method is not very useful in this problem.

More recently, the release of Transformer [11] architecture is the inspiration for some robust models such as BERT [12], ELECTRA [13], and XLM-RoBERTa [14]. They have dramatically improved the state-of-the-art results on various downstream natural language tasks. These pre-trained models learn useful contextual representations from massive unlabeled datasets using self-supervised pre-training objectives, such as masked language modeling. It can predict the original masked word based only on its context from a masked input sentence. These advantages motivated us to apply them for our punctuation prediction task.

This paper contributes to using the pre-trained Transformer models such as BERT, ELECTRA, and XLM-RoBERTA on two large-scale Vietnamese novel

and news datasets. We also extend our proposed method by incorporating a BiLSTM layer and a CRF layer compared with the previous work. We consider adding a BiLSTM and/or CRF layer on top of the pre-trained models to captures semantics and long-range dependencies in the input sentence. Results show that the transfer learning method is useful in the punctuation prediction problem.

The rest of this paper can be organized as follows. Section 1 briefly introduces the punctuation prediction problem and discusses all related works of the problem. Section 2 describes the model architecture of punctuation prediction, and Sect. 3 presents experimental results on two Vietnamese datasets. The paper ends with the conclusion and further work.

2 Methodology

2.1 Problem Formulation

Similar to the previous studies [1,2,5], we represent the punctuation prediction task as a sequence labeling problem and find an appropriate model for this task. It is worth noting that we label each word by its immediately following punctuation. In this paper, we consider six standard punctuation marks in the Vietnamese language: the period (.), the comma (,), the colon (:), the semicolon (;), the question mark (?), the exclamation mark (!), and space. We use the label O to indicate that a given the word is not followed by any punctuation.

For instance, let us consider the following sentence in the Vietnamese language.[1]

> Năm ngoái, dù có doanh số bán quảng cáo sụt giảm nghiêm trọng trong khoảng thời gian đầu năm, Facebook đã chứng kiến doanh thu phục hồi vào những tháng sau đó nhờ nhu cầu tiếp cận khách hàng tăng cao của các doanh nghiệp nhỏ.
>
> (Last year, despite a dramatic decline in ad sales in the early part of the year, Facebook saw its revenue rebound in the following months, thanks to rising demand for customer access by small businesses.)

Typically, one can label the above paragraph as follows:

> Năm/O ngoái/Comma dù/O có/O doanh/O số/O bán/O quảng/O cáo/O sụt/O giảm/O nghiêm/O trọng/O trong/O khoảng/O thời/O gian/O đầu/O năm/Comma Facebook/O đã/O chứng/O kiến/O doanh/O thu/O phục/O hồi/O vào/O những/O tháng/O sau/O đó/O nhờ/O nhu/O cầu/O tiếp/O cận/O khách/O hàng/O tăng/O cao/O của/O các/O doanh/O nghiệp/O nhỏ/Period

Remarkably, the word case information is not available for the punctuation prediction problem since all the words are lower case during the data processing step.

[1] https://vnexpress.net/facebook-hay-google-manh-hon-4226827.html.

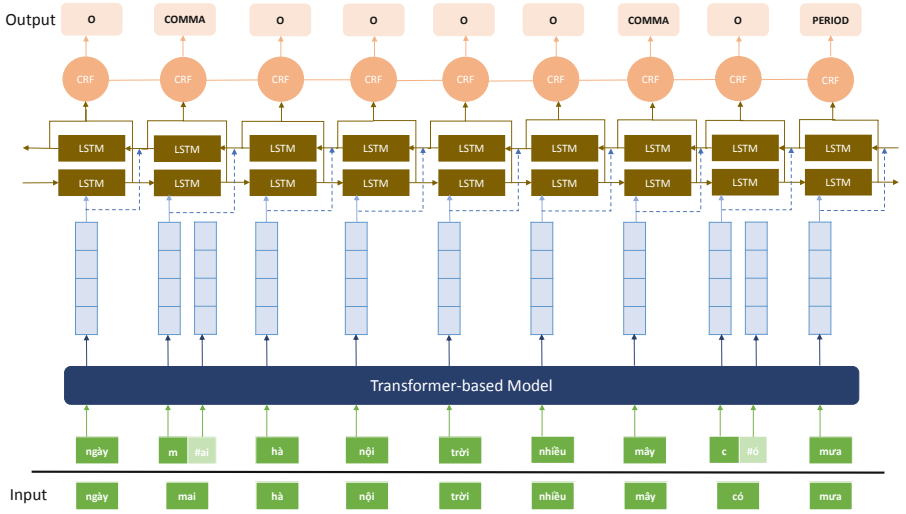

Fig. 1. The first proposed neural network architecture by combining LSTM and CRF. The input sentence means "Hanoi is cloudy and rainy tomorrow".

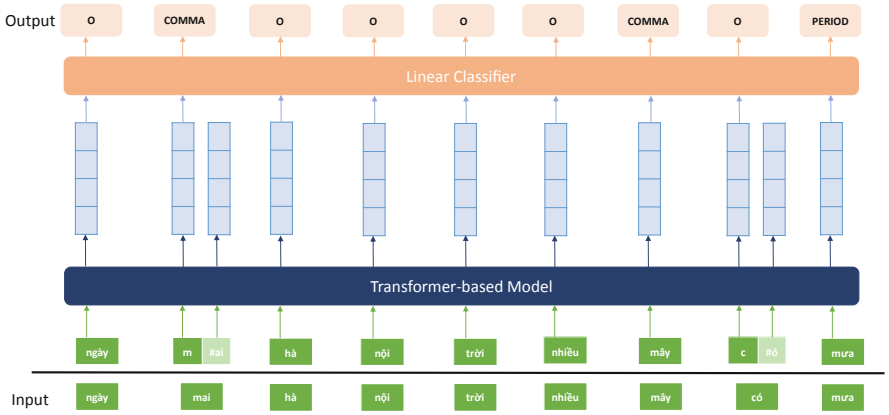

Fig. 2. The second proposed standalone neural network architecture (without using LSTM and CRF). The input sentence means "Hanoi is cloudy and rainy tomorrow".

2.2 Our Proposed Techniques

Data Processing. The dataset described in Sect. 3.1 is not segmented into sentences. Consequently, we split the data into multiple paragraphs with a maximum length of 128 words but still have to ensure that they start after the end punctuation mark. We also replace all numbers in the dataset with <NUM> token, and this helps to reduce the influence of different numbers. It is important to note that all words in this dataset are lowercase, and some special characters are excluded.

To fine-tune the Transformer-based models such as BERT and ELECTRA, one needs to insert two particular tokens, [CLS] and [SEP], into the input. The [CLS] is encoded, including all representative information of the whole input sentence. Meanwhile, the use of [SEP] token is to separate different sentences of an input. We only need to insert the [SEP] to the end of every input data. The advantage of these models is that they can handle words that are not part of the vocabulary. The tokenizer takes the input sentence and relies on the vocabulary to decide whether to keep the whole word or slice it into subwords, containing the first subword and subsequent subwords starting with the # (sharp) symbol. As every word in each sentence has its label, we need to assign the corresponding labels from the first subword to subsequent subwords. Finally, we convert the new input to the sequence of indices with the same length. We also pad the sequences with the [PAD] token if their length is less than a given threshold.

Feature Extraction. In the text classification tasks, one usually trains a model to predict the CLS token encoded, including all representative information of the whole input sentence. Nevertheless, in the sequence labeling tasks, we can feed the last hidden state, which encloses all words of each input sentence's hidden ones into later layers. We also include the subsequent subwords inside the last hidden state to take the hidden state's first subword to represent the whole word and use it for prediction.

Models. More recently, pre-trained language models have shown to be useful in learning common language representations by utilizing a large amount of unlabeled data. These models have achieved great results in many downstream natural language processing tasks. In what follows, we describe our chosen architectures for implementation in this paper:

- BERT stands for Bidirectional Encoder Representations from Transformers. The model contains multiple bidirectional Transformer encoders, and each encoder layer is composed of a multi-head self-attention. However, BERT Encoder slightly differs from the canonical Transformer, which uses a GELU [15] activation rather than the standard RELU. BERT uses Word Piece [16] to tokenize the input sentence into tokens.
  Each token is represented by the sum of the token embedding, segment embedding, and positional embedding. All these things make it possible for BERT to learn contextualized word representations. BERT uses masked language modeling objective and next sentence prediction for pre-training. In the masked language modeling task, some percentage of tokens can be selected at random as the masked tokens and then predicting only these tokens. In the next sentence prediction task, BERT treats this task as a binary classification and indicate whether the second sentence is the following sentence of the first sentence.
- ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately) uses a new pre-training approach that is substituted token

detection. Instead of replacing the input tokens with masked tokens, this method corrupts the input by replacing some tokens with incorrect tokens. And then, the model tries to determine which tokens from the original input have been replaced. This setup requires two Transformer models in pre-training: a generator and a discriminator. After pre-training, one can remove the generator and keep the discriminator for fine-tuning in downstream tasks. ELECTRA also uses a Word Piece tokenizer like BERT.
- XLM-RoBERTa is the first multilingual model that can outperform traditional monolingual pre-trained models. This model is the combination of XLM [17] and RoBERTa [18]. Specifically, XLM-RoBERTa relies on masked language modeling objective and cross-lingual language modeling objective without next sentence pre-training objective. Unlike BERT, XLM-RoBERTa uses the Sentence Piece model [19] for tokenizing the input sentence instead of using Word Piece tokenizer. The Sentence Piece implements two segmentation algorithms, including the byte-pair-encoding and unigram language model. It is an effective way to tackle the out-of-vocab problems.

We examine to stack a Bidirectional Long Short-Term Memory network (BiLSTM) and Conditional Random Field (CRF) layer on the top of output representations.

- BiLSTM was proposed to deal with the vanishing gradient problem encountered by traditional RNNs. A typical LSTM layer consists of three gates, input gate, output gate, and forget gate. These gates control how the information in a sequence of input comes into, stores in, and leaves the network. With the bidirectional term, this layer looks at the sequence from left to right in the forward stage and observes from right to left in the backward stage. The advantage of this layer is it can captures long-distance dependencies across the sequences from both past and future contexts.
- CRF tries to learn the conditional probability over label sequences given a particular observation sequence, rather than a joint probability over both label and observation sequences. Instead of directly using cross-entropy loss, we pass the outputs from the previous layer through the CRF layer to compute all possible classes' probability or the negative log-likelihood loss. This layer helps to choose labels based on both past and current dependencies.

The summary of the proposed model can be illustrated in Figs. 1 and 2.

Performance Metrics. We measure the performance of our models on precision, recall, and F1-score. Following the previous work, we only consider six punctuation marks, including comma, period, colon, semicolon, question mark, and exclamation. The best model is the model that dominates in the average micro F1-score.

3 Experiments

In this section, we compare the effectiveness of a multilingual pre-trained model and a monolingual pre-trained model. We make use of multilingual BERT[2] and XLM-RoBERTa,[3] both models trained on large-scale data, in 100 languages. For the monolingual model, we choose the viELECTRA model, which is released by [20], this model trained on 60 GB of Vietnamese texts. All these models contain 12 Encoder layers, 768 hidden units, and we use Adam optimizer with a learning rate of 5e-5 and weight decay of 0.01. In the Bi-LSTM layers, we stack two layers, and the number of unit cells in each of the LSTM layers is half of the pre-trained model hidden size. The results in Table 2 and Table 3 are selected from the highest scores in 15 epochs. We also provide all the source codes[4] for reproducing experiments.

3.1 Datasets

Our models are tested by performing experiments on the two large-scale Vietnamese datasets, as described by Quang et al. [1]. The novel dataset contains 111,601 sentences on the training set and 44,081 sentences on the test set. Meanwhile, the size of the news dataset is over three times the novel dataset. It contains 440,866 sentences on the training set and 145,768 on the test set. Comma and period are punctuation marks with the most percentage of occurrence in both datasets, but the others are significantly different. Specifically, the number of colon and semicolon marks is so small, and they have only a few tens or a few hundred in the novel dataset. In the news dataset, the smallest number of punctuation marks are semicolons and exclamation marks. See Table 1 for more information about punctuation mark distributions.

3.2 Results

We compare our models with the combination of BiLSTM and Attention from previous work [2] that achieves the highest overall f1-score in both Vietnamese datasets.

In the novels dataset, mBERT, XLM-RoBERTa, and viELECTRA improve the overall F1-score by 12.01%, 18.56%, and 21.12%, respectively. Because of the poverty of semicolon marks in both training and test set, our proposed models fail to predict the semicolon, leading to zero correction even when these models are pre-trained on a massive corpus.

[2] https://github.com/google-research/bert/blob/master/multilingual.md.
[3] https://github.com/facebookresearch/XLM.
[4] https://github.com/heraclex12/VN-Punc-Pretrained-LMs.

Table 1. The number of punctuation marks on both training and test sets in Vietnamese novels and news datasets.

	Novel		News	
Punctuation	Training	Test	Training	Test
COMMA	50909	21231	482435	160472
PERIOD	66519	29643	419580	138967
COLON	742	1153	32177	10728
QMARK	14899	5271	13902	4468
EXCLAM	30183	9167	7384	2333
SEMICOLON	48	43	5675	2045

There are many grammatical errors in the news dataset, and existing foreign words that lead to the performance improvements are slightly different. The performance of the "mBERT" is improved further by 13.49%. On the other hand, the performance improvement of both XLM-RoBERTa and viELECTRA respectively fall to 16.75% and 18.29%. Compared with the novels dataset, the distribution of punctuation marks is more balanced. Thus, our models can predict all the punctuation marks well, though the performance still needs further improvement.

As shown in Tables 2 and 3, all our models outperform the previous model. Especially, fine-tuning viELECTRA reaches the highest performance compared to other fine-tuning models. However, there are a few observations contrary to our expectations. All our additional layers could not obtain the expected outcomes, with less than 1% improvement in the overall F1-score. It indicates that the influence of dependencies in the BiLSTM layer and CRF layer is not much. Finally, the combination of the "viELECTRA" and CRF layer achieves 71.97% in overall F1-score, which is the best performance in the Vietnamese novels dataset, and 21.45% higher in absolute difference from the previous work. On the other hand, the combination of viELECTRA, BiLSTM, and CRF dominates the Vietnamese news benchmark. It achieves 80.98% in overall F1-score and outperforms the previous model 18.31% in an absolute score.

3.3 Discussion

Experimental results show that the transfer learning method is effective. All fine-tuning models outperform the previous model on both the Vietnamese novels dataset and news dataset. Our models employ robust architecture to learn different aspects of the language in both left and right contexts and allow our models to focus on essential words in the sentence. Another reason is our models benefit from the weights that we have learned on a lot of training data. Both of the above reasons make our models decide the meaning of the words based on the context, instead of getting a particular meaning. We assume that incorporating a Bi-LSTM layer or a CRF layer can boost the performance, but it is not

Table 2. The performance of our proposed models in Vietnamese novel dataset. We consider six punctuation marks in our experiments and compare the proposed models with the state-of-the-art method for the punctuation prediction task in the Vietnamese language [2].

Model	Avg	,	.	:	?	!	;
BiLSTM+Att [2]	56.52	56.10	55.86	21.43	70.34	52.09	0.00
	45.67	38.45	47.33	0.95	65.60	54.30	0.00
	50.52	45.63	51.24	1.81	67.89	53.18	0.00
mBERT	62.26	60.47	63.47	28.57	75.44	55.64	0.00
	62.80	56.13	67.93	3.82	75.77	61.90	0.00
	62.53	58.22	65.63	6.73	75.61	58.60	0.00
+CRF	62.74	59.45	65.28	29.22	75.23	55.81	0.00
	62.51	57.49	66.31	3.90	75.92	61.78	0.00
	62.62	58.45	65.79	6.89	75.57	58.64	0.00
+LSTM	62.92	59.83	64.93	22.86	75.82	56.52	0.00
	62.17	56.87	66.12	0.69	75.26	62.17	0.00
	62.54	58.31	65.52	1.35	75.54	59.21	0.00
+LSTM+CRF	63.31	60.49	64.53	25.00	77.23	58.13	0.00
	62.03	55.98	66.41	1.47	75.47	62.06	0.00
	62.67	58.15	65.46	2.78	76.34	60.03	0.00
XLM-RoBERTa	69.72	67.21	73.12	40.15	80.96	60.01	0.00
	68.44	64.33	72.21	4.6	78.03	68.65	0.00
	69.08	65.74	72.66	8.25	79.47	64.04	0.00
+CRF	69.36	66.53	72.64	40.57	80.55	60.65	0.00
	68.94	65.15	72.40	7.46	79.74	68.35	0.00
	69.15	65.83	72.52	12.6	80.14	64.27	0.00
+LSTM	69.34	66.72	76.62	35.96	81.33	59.90	0.00
	68.67	64.70	72.31	5.55	79.34	68.22	0.00
	69.01	65.70	72.47	9.62	80.32	63.79	0.00
+LSTM+CRF	70.34	66.53	73.83	34.92	80.46	63.01	0.00
	68.15	65.78	71.50	3.82	79.76	64.58	0.00
	69.23	66.15	72.65	6.88	80.11	63.78	0.00
viELECTRA	71.84	69.66	75.79	41.28	81.18	61.15	0.00
	71.44	67.19	75.43	**7.81**	81.01	71.24	0.00
	71.64	68.40	75.61	**13.13**	81.10	65.81	0.00
+CRF	**72.17**	**70.29**	**75.97**	**48.41**	80.95	61.27	0.00
	71.77	**68.39**	75.32	5.29	**81.45**	**71.26**	0.00
	71.97	**69.32**	75.65	9.54	**81.20**	**65.89**	0.00
+LSTM	71.27	68.53	74.67	32.52	81.75	62.20	0.00
	71.06	67.49	75.70	5.81	79.91	67.79	0.00
	71.17	68.00	75.18	9.86	80.82	64.87	0.00
+LSTM+CRF	71.66	68.02	74.46	34.31	**82.44**	**65.47**	0.00
	71.44	68.31	**77.10**	5.55	79.45	64.40	0.00
	71.55	68.16	**75.76**	9.56	80.92	64.93	0.00

Notes: The 1st line in each row is Precision, the 2nd line is Recall, and the 3rd one is F1-score.

Table 3. The performance of our proposed models in Vietnamese news dataset. We consider six punctuation marks in our experiments and compare the proposed models with the state-of-the-art method for the punctuation prediction task in the Vietnamese language [2].

Model	Avg	,	.	:	?	!	;
BiLSTM+Att [2]	69.63	68.30	72.09	61.54	61.01	35.71	29.25
	56.97	52.42	68.13	29.87	51.30	7.50	4.92
	62.67	59.32	70.06	40.22	55.73	12.40	8.43
mBERT	77.48	73.50	83.06	64.25	68.67	**42.95**	36.09
	74.89	69.01	85.92	49.58	65.35	11.62	13.89
	76.16	71.18	84.46	55.97	66.97	18.29	20.06
+CRF	77.58	73.47	83.32	64.80	68.36	41.96	36.98
	74.61	68.66	85.64	49.00	66.05	12.09	15.55
	76.06	70.98	84.46	55.80	67.18	18.77	21.89
+LSTM	77.79	73.60	83.53	65.12	68.08	42.57	36.97
	74.63	68.91	85.57	48.04	65.44	11.66	10.61
	76.17	71.18	84.54	55.29	66.74	18.30	16.49
+LSTM+CRF	77.78	73.33	83.86	65.19	69.57	40.72	36.49
	74.66	69.18	85.32	48.31	64.28	12.13	11.69
	76.19	71.20	84.59	55.50	66.82	18.69	17.70
XLM-RoBERTa	80.61	76.70	86.22	67.64	74.96	44.97	38.56
	78.27	72.90	88.47	54.19	72.09	14.96	18.63
	79.42	74.75	87.33	60.18	73.50	22.45	25.12
+CRF	80.48	76.54	86.31	66.09	75.96	42.89	36.47
	78.40	73.01	88.48	56.28	72.07	14.62	18.78
	79.42	74.73	87.38	60.79	73.96	21.80	24.79
+LSTM	80.89	76.73	86.72	67.88	75.12	45.49	43.45
	78.29	73.21	88.24	54.73	71.08	13.84	14.77
	79.57	74.93	87.47	60.60	73.05	21.23	22.04
+LSTM+CRF	80.83	76.58	86.72	68.67	75.59	45.45	40.58
	78.28	73.31	88.17	53.68	71.24	14.36	13.69
	79.54	74.91	87.44	60.62	73.35	21.82	20.48
viELECTRA	81.86	77.78	88.06	68.18	**76.94**	41.70	37.53
	80.07	**75.42**	89.16	59.13	**74.40**	18.52	20.24
	80.96	**76.59**	88.60	63.33	**75.65**	25.65	26.30
+CRF	81.82	77.73	88.06	68.46	76.06	42.27	36.40
	80.08	75.37	**89.22**	**59.29**	74.10	**19.46**	20.15
	80.94	76.53	88.64	63.55	75.07	**26.65**	25.94
+LSTM	82.07	78.01	88.23	68.76	76.65	40.69	36.49
	79.85	75.06	89.19	58.28	73.16	17.70	20.54
	80.95	76.51	88.71	63.08	74.87	24.67	26.28
+LSTM+CRF	**82.12**	**78.06**	**88.29**	**69.09**	75.97	40.92	**38.29**
	79.87	74.98	89.19	59.23	73.93	18.17	**22.05**
	80.98	76.49	**88.74**	**63.78**	74.93	25.18	**27.99**

Notes: The 1st line in each row is Precision, the 2nd line is Recall, and the 3rd one is F1-score.

significant. The possible reason is that the pre-trained models already consist of deep networks, making the architecture more complicated is a redundant option.

4 Conclusion

This paper has conducted extensive experiments to investigate the various pre-trained Transformer-based models for the Vietnamese language's punctuation prediction problem and consider two categories of models: monolingual and multilingual. The experimental results show that the monolingual Transformer-based models are better than the multilingual models, and our proposed models outperform the previous works for the Vietnamese language. It can demonstrate that using a transfer-learning method can provide high efficiency in the punctuation prediction task. Besides, we also stack LSTM and/or on top of the pre-trained models and achieve promising results. Our best model dominates other approaches in both datasets. We can conduct an overall F1-score of 21,45% on the novel dataset and 18.27% on the news dataset.

For future work, we plan to expand two Vietnamese datasets to distribute classes more balanced. It can help our model learn better in the minority class. We also aim to analyze our proposed model's robustness on other informal texts that are more challenging in the real world. Instead of treating the missing punctuation mark problem as a sequence labeling task, we can deal with this problem as the punctuation restoration task, using another approach like machine translation. In this approach, the input sentence does not contain any punctuation marks, and then the model generates and outputs that sentence with the correct punctuation marks.

References

1. Pham, Q.H., Nguyen, B.T., Cuong, N.V.: Punctuation prediction for vietnamese texts using conditional random fields. In: Proceedings of the Tenth International Symposium on Information and Communication Technology, SoICT 2019, pp. 322–327 (2019)
2. Pham, T., Nguyen, N., Pham, Q., Cao, H., Nguyen, B.: Vietnamese punctuation prediction using deep neural networks. In: Chatzigeorgiou, A., et al. (eds.) SOFSEM 2020. LNCS, vol. 12011, pp. 388–400. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-38919-2_32
3. Beeferman, D., Berger, A., Lafferty, J.: Cyberpunc: a lightweight punctuation annotation system for speech. In: Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 2, pp. 689–692, May 1998 (1998)
4. Huang, J., Zweig, G.: Maximum entropy model for punctuation annotation from speech (2002)
5. Lu, W., Tou Ng, H.: Better punctuation prediction with dynamic conditional random fields, pp. 177–186 (2010)
6. S. Peitz, M. Freitag, A. Mauser, and H. Ney, "Modeling punctuation prediction as machine translation," in IWSLT, 2011

7. Kim, S.: Deep recurrent neural networks with layer-wise multi-head attentions for punctuation restoration. In: ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7280–7284 (2019)
8. K. Makhija, T. Ho, and E. Chng, "Transfer learning for punctuation prediction," in 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp. 268–273, 2019
9. Fang, M., Zhao, H., Song, X., Wang, X., Huang, S.: Using bidirectional LSTM with Bert for Chinese punctuation prediction. In: 2019 IEEE International Conference on Signal, Information and Data Processing (ICSIDP), pp. 1–5 (2019)
10. Wang, F., Chen, W., Yang, Z., Xu, B.: Self-attention based network for punctuation restoration. In: 2018 24th International Conference on Pattern Recognition (ICPR), pp. 2803–2808 (2018)
11. Vaswani, A., et al.: Attention is all you need (2017)
12. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding (2019)
13. Clark, K., Luong, M.-T., Le, Q.V., Manning, C.D.: Electra: pre-training text encoders as discriminators rather than generators (2020)
14. Conneau, A., et al.: Unsupervised cross-lingual representation learning at scale (2020)
15. Hendrycks, D., Gimpel, K.: Bridging nonlinearities and stochastic regularizers with Gaussian error linear units, ArXiv, vol. abs/1606.08415 (2016)
16. Wu, Y., et al.: Google's neural machine translation system: bridging the gap between human and machine translation (2016)
17. Lample, G., Conneau, A.: Cross-lingual language model pretraining (2019)
18. Liu, Y., et al.: Roberta: a robustly optimized bert pretraining approach (2019)
19. Kudo, T., Richardson, J.: 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, (Brussels, Belgium), pp. 66–71. Association for Computational Linguistics, November 2018 (2018)
20. The, V.B., Thi, O.T., Le-Hong, P.: Improving sequence tagging for vietnamese text using transformer-based neural models (2020)