An Empirical Study of Using Pre-trained BERT Models for Vietnamese Relation Extraction Task at VLSP 2020

Pham Quang Nhat Minh
Aimesoft JSC
Hanoi, Vietnam
minhpham@aimesoft.com

Abstract

In this paper, we present an empirical study of using pre-trained BERT models for the relation extraction task at the VLSP 2020 Evaluation Campaign. We applied two state-of-the-art BERT-based models: R-BERT and BERT model with entity starts. For each model, we compared two pre-trained BERT models: FPTAI/vibert and NlpHUST/vibert4news. We found that NlpHUST/vibert4news model significantly outperforms FPTAI/vibert for the Vietnamese relation extraction task. Finally, we proposed an ensemble model that combines R-BERT and BERT with entity starts. Our proposed ensemble model slightly improved against two single models on the development data and the test data provided by the task organizers.

1 Introduction

The relation extraction task is to extract entity mention pairs from a sentence and determine relation types between them. Relation extraction systems can be applied in question answering (Xu et al., 2016), detecting contradiction (Pham et al., 2013), and extracting gene-disease relationships (Chun et al., 2006), protein-protein interaction (Huang et al., 2004) from biomedical texts.

In VLSP 2020, the relation extraction task is organized to assess and advance relation extraction work for the Vietnamese language. In this paper, we present an empirical study of BERT-based models for the relation extraction task in VLSP 2020. We applied two state-of-the-art BERT-based models for relation extraction: R-BERT (Wu and He, 2019) and BERT with entity starts (Soares et al., 2019). Two models use entity markers to capture location information of entity mentions. For each model, we investigated the effect of choosing pre-train BERT models in the task, by comparing two Vietnamese pre-trained BERT models: NlpHUST/vibert4news and FPTAI/vibert (Bui et al., 2020). In our understanding, our paper is the first work that provides the comparison of pre-trained BERT models for Vietnamese relation extraction.

The remainder of this paper is structured as follows. In Section 2, we present two existing BERT-based models for relation classification, which we investigated in our work. In Section 3, we describe how we prepared datasets for the two BERT-based models and our proposed ensemble model. In Section 4, we give detailed settings and experimental results. Section 5 gives discussions and findings. Finally, in Section 6, we present conclusions and future work.

2 BERT-based Models for Relation Classification

In the following sections, we briefly describe BERT model (Devlin et al., 2019), problem formalization, and two existing BERT-based models for relation classification, which we investigated in this paper.

2.1 Pre-trained BERT Models

The pre-trained BERT model (Devlin et al., 2019) is a masked language model that is built from multiple layers of bidirectional Transformer encoders (Vaswani et al., 2017). We can fine-tune pre-trained BERT models to obtain the state-of-the-art results on many NLP tasks such as text classification, named-entity recognition, question answering, natural language inference.

Currently, pre-trained BERT models are available for many languages. For Vietnamese, in our understanding, there are three available pre-trained BERT models: PhoBERT (Nguyen and Nguyen, 2020), FPTAI/vibert (Bui et al., 2020), and NlpHUST/vibert4news[1]. Those models are differ-

[1] vibert4news is available on https://huggingface.co/NlpHUST/vibert4news-base-cased

ent in pre-training data, selected tokenization, and training settings. In this paper, we investigated two pre-trained BERT models including FPTAI/vibert and NlpHUST/vibert4news for the relation extraction task. Investigation of PhoBERT for the task is left for future work.

2.2 Problem Formalization

In this paper, we focus on the relation classification task in the supervised setting. Training data is a sequence of examples. Each sample is a tuple $\mathbf{r} = (\mathbf{x}, \mathbf{s}_1, \mathbf{s}_2, y)$. We define $\mathbf{x} = [x_0 ... x_n]$ as a sequence of tokens, where $x_0 =$ [CLS] is a special start marker. Let $\mathbf{s}_1 = (i, j)$ and $\mathbf{s}_2 = (k, l)$ are pairs of integers such that $0 < i \leq j \leq n, 0 < k \leq l \leq n$. Indexes of $\mathbf{s}_1$ and $\mathbf{s}_2$ are start and end indexes of two entity mentions in $\mathbf{x}$, respectively. y denotes the relation label of the two entity mentions in the sequence $\mathbf{x}$. We use a special label OTHER for entity mentions which have no relation between them. Our task is to train a classification model from the training data.

2.3 R-BERT

In R-BERT (Wu and He, 2019), for a sequence $\mathbf{x}$ and two target entities e_1 and e_2 which specified by indexes of $\mathbf{s}_1$ and $\mathbf{s}_2$, to make the BERT module capture the location information of the two entities, a special token '$' is added at both the beginning and end of the first entity, and a special token '#' is added at both the beginning and end of the second entity. [CLS] token is also added to the beginning of the sequence.

For example, after inserting special tokens, a sequence with two target entities "Phi Sơn" and "SLNA" becomes to:

"[CLS] Cầu thủ $ Phi Sơn $ đã ghi bàn cho # SLNA # vào phút thứ 80 của trận đấu ."

The sequence $\mathbf{x}$ with entity markers, is put to a BERT model to get hidden states of tokens in the sequence. Then, we calculate averages of hidden states of tokens within the two target entities and put them through a tanh activation function and a fully connected layer to make vector representations of the two entities. Let H'_0, H'_1, H'_2 be hidden states at [CLS] and vector representations of e_1 and $e2$. We concatenate three hidden states and add a softmax layer for relation classification. R-BERT obtained 89.25% of MACRO F1 on the SemEval-2010 Task 8 dataset (Hendrickx et al., 2010).

2.4 BERT with Entity Start

We applied the BERT model with entity starts (hereinafter, referred to as BERT-ES) presented in (Soares et al., 2019) for Vietnamese relation classification. In the model, similar to R-BERT, special tokens are added at the beginning and end of two target entities. In experiments of BERT-ES for Vietnamese relation classification, different from (Soares et al., 2019), we used entity markers '$' and '#' instead of markers '[E1]', '[/E1]', '[E1]', and '[/E2]'. We did not add [SEP] at the end of a sequence. In BERT-ES, hidden states at the start positions of two target entities are concatenated and put through a softmax layer for final classification. On SemEval-2010 Task 8 dataset, BERT-ES obtained 89.2% of MACRO F1.

3 Proposed Methods

In this work, we applied R-BERT and BERT-ES as we presented in Section 2 for Vietnamese relation extraction, and proposed an ensemble model of R-BERT and BERT-ES. In the following sections, we present how we prepared data for training BERT-based models and how we combined two single models: R-BERT and BERT-ES.

3.1 Data Preprocessing

Relation extraction data provided by VLSP 2020 organizers in WebAnno TSV 3.2 format (Eckart de Castilho et al., 2016). In the data, sentences are not segmented and tokens are tokenized by white spaces. Punctuations are still attached in tokens.

According to the task guideline, we consider only intra-sentential relations, so sentence segmentation is required in data preprocessing. We used VnCoreNLP toolkit (Vu et al., 2018) for both sentence segmentation and tokenization. For the sake of simplicity, we just used syllables as tokens of sentences. VnCoreNLP sometimes made mistakes in sentence segmentation, and as the result, we missed some relations for those cases.

3.2 Relation Sample Generation

From each sentence, for training and evaluation, we made relation samples which are tupes $\mathbf{r} = (\mathbf{x}, \mathbf{s}_1, \mathbf{s}_2, y)$ as described in Section 2. Since in the data, named entities with their labels are provided, a simple way of making relation samples is generating all possible entity mention pairs from entity mentions of a sentence. We used the label OTHER for entity mention pairs that lack relation

No.	Relation	Arguments	Directionality
1	LOCATED	PER - LOC, ORG – LOC	Directed
2	PART–WHOLE	LOC – LOC, ORG – ORG, ORG-LOC	Directed
3	PERSONAL–SOCIAL	PER – PER	Undirected
4	AFFILIATION	PER – ORG, PER-LOC, ORG – ORG, LOC-ORG	Directed

Table 1: Relation types permitted arguments and directionality.

between them. All entity mentions pairs that are not included in gold-standard data are used as OTHER samples.

In the annotation guideline provided by VLSP 2020 organizers, there are constraints about types of two target entities of relation types as shown in Table 1. Thus, we consider only entity mention pairs whose types satisfy those constraints. In training data, sometimes types of two target entities do not follow the annotation guideline. We accepted those entity pairs in making relation samples from provided train and development datasets. However, in processing test data for making submitted results, we consider only entity pairs whose types follow the annotation guideline.

Since the relation PERSONAL-SOCIAL is undirected, for this type, if we consider both pairs (e_1, e_2) and (e_2, e_1) in which e_1 and e_2 are PERSON entities, it may introduce redundancy. Thus, we added an extra constraint for PER-PER pairs that e_1 must come before e_2 in a sentence.

In the training data, we found a very long sentence with more than 200 relations. We omitted that sentence from the training data because that sentence may lead to too many OTHER relation samples.

3.3 Proposed Ensemble Model

In our work, we tried to combine R-BERT and BERT-ES to make an ensemble model. We did that by calculating weighted averages of probabilities returned by R-BERT and BERT-ES. Since in our experiments, BERT-ES performed slightly better than R-BERT on the development set, we used weights 0.4 and 0.6 for R-BERT and BERT-ES, respectively.

4 Experiments and Results

We conducted experiments to compare three BERT-based models on Vietnamese relation extraction data: R-BERT, BERT-ES, and the proposed ensemble model. We also investigated the effects of two Vietnamese pre-trained BERT models on the performance of models.

Relation	Train	Dev
LOCATED	507	304
PART-WHOLE	1,016	402
PERSONAL-SOCIAL	101	95
AFFILIATION	756	489
OTHER	23,904	13,239
Total	26,284	14,529

Table 2: Label distribution of relation samples generated from train and dev data.

Hyper-Parameters	Value
Max sequence length	384
Training epochs	10
Train batch size	16
Learning rate	2e-5

Table 3: Hyper-parameters used in training models.

4.1 Data

The provided training dataset contains 506 documents, and the development dataset contains 250 documents. After data preprocessing and relation sample generation, we obtained relations with label distributions shown in Table 2.

4.2 Experimental Settings

In development, we trained models on the training data and evaluated models on the development data. However, to generate results on the provided test dataset, we trained BERT-based models on the dataset obtained by combining the provided training dataset and the development dataset.

Table 3 shows hyper-parameters we used for training models. We trained all models on a single 2080 Ti GPU.

We used MICRO F1 and MACRO F1 of four relation labels which do not include the label OTHER as evaluation measures.

Model	Pre-trained BERT Model	MACRO F1	MICRO F1
R-BERT	NlpHUST/vibert4news	0.6392	0.7092
R-BERT	FPTAI/vibert	0.596	0.6736
BERT-ES	NlpHUST/vibert4news	**0.6439**	0.7101
BERT-ES	FPTAI/vibert	0.5976	0.6822
Ensemble Model	NlpHUST/vibert4news	0.6412	**0.7108**
Ensemble Model	FPTAI/vibert	0.6029	0.6851

Table 4: Evaluation results on dev dataset.

Model	MACRO F1	MICRO F1
R-BERT	0.6294	0.6645
BERT-ES	0.6276	0.6696
Ensemble Model	**0.6342**	**0.6756**

Table 5: Evaluation results on test dataset.

4.3 Results

Table 4 shows the evaluation results obtained on the development dataset. We can see that using NlpHUST/vibert4news significantly outperformed FPTAI/vibert in both MICRO F1 and MACRO F1 scores. BERT-ES performed slightly better than R-BERT. The proposed ensemble model is slightly improved against R-BERT and BERT-ES in terms of MICRO F1 score.

Table 5 shows the evaluation results obtained on the test dataset. We used NlpHUST/vibert4news for generating test results. Table 5 confirmed the effectiveness of our proposed ensemble model. The ensemble model obtained the best MACRO F1 and the best MICRO F1 score on the test data among the three models.

4.4 Result Analysis

We looked at details of precision, recall, and F1 scores for each relation type on the development data. Table 6 shows results of the ensemble model with vibert4news pre-trained model. PERSONAL-SOCIAL turned out to be a difficult label. The proposed ensemble obtained a low Recall, and F1 score for that label. The reason might be that the relations of PERSONAL-SOCIAL are few in the training data while the patterns of PERSONAL-SOCIAL relations are wider than other relation types.

5 Discussion

In experiments, we compared the effects of two pre-trained BERT models: NlpHUST/vibert4news and FPTAI/vibert on relation extraction. The two pre-trained models have the same BERT architecture (BERT base model) but are different in chosen tokenizers, vocabulary size, pre-training data, and training procedure. Table 7 shows a comparison of the two models.

FPTAI/vibert was trained on 10GB of texts collected from online newspapers while NlpHUST/vibert4news was trained on 20GB of texts in the news domain. FPTAI/vibert used subword tokenization, and vocabulay of FPTAI/vibert was modified from mBERT while tokenization of vibert4news is based on syllables.

We come up with some reasons why using NlpHUST/vibert4news significantly outperformed FPTAI/vibert for Vietnamese relation extraction.

- Pre-training data used to trained vibert4news is much larger than FPTAI/vibert.
- Tokenization used in NlpHUST/vibert4news is based on syllables while FPTAI/vibert used subwords and modified the original vocabulary of mBERT. We hypothesize that syllables which are basic units in Vietnamese are more appropriate than subwords for Vietnamese NLP tasks.

Due to the time limit, we did not investigate PhoBERT (Nguyen and Nguyen, 2020) which used word-level corpus to train the model. As future work, we plan to compare vibert4news that uses syllable-based tokenization with PhoBERT that uses word-level/subword tokenization for Vietnamese relation extraction.

	Precision	Recall	F1
AFFILIATION	0.7615	0.744	0.7528
LOCATED	0.7053	0.7007	0.7030
PART – WHOLE	0.65	0.8085	0.7206
PERSONAL - SOCIAL	0.6136	0.2842	0.3885

Table 6: Precision, Recall, F1 for each relation type on the dev dataset.

	FPTAI/vibert	vibert4news
Data size	10GB	20GB
Data domain	News	News
Tokenization	Subword	Syllable
Vocab size	38168	62000

Table 7: Comparison of NlpHUST/vibert4news and FPTAI/vibert.

6 Conclusion

We have presented an empirical study of BERT-based models for relation extraction task at VLSP 2020 Evaluation Campaign. Experimental results show that the BERT-ES model which uses entity markers and entity starts obtained better results than the R-BERT model, and choosing an appropriate pre-trained BERT model is important for the task. We showed that pre-trained model NlpHUST/vibert4news outperformed FPTAI/vibert for Vietnamese relation extraction task. In future work, we plan to investigate PhoBERT (Nguyen and Nguyen, 2020) for Vietnamese relation extraction to understand the effect of using word segmentation to the task.

References

The Viet Bui, Thi Oanh Tran, and Phuong Le-Hong. 2020. Improving sequence tagging for vietnamese text using transformer-based neural models. *arXiv preprint arXiv:2006.15994*.

Richard Eckart de Castilho, Éva Mújdricza-Maydt, Seid Muhie Yimam, Silvana Hartmann, Iryna Gurevych, Anette Frank, and Chris Biemann. 2016. A web-based tool for the integrated annotation of semantic and syntactic structures. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 76–84, Osaka, Japan. The COLING 2016 Organizing Committee.

Hong-Woo Chun, Yoshimasa Tsuruoka, Jin-Dong Kim, Rie Shiba, Naoki Nagata, Teruyoshi Hishiki, and Jun'ichi Tsujii. 2006. Extraction of gene-disease relations from medline using domain dictionaries and machine learning. In *Biocomputing 2006*, pages 4–15. World Scientific.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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. Association for Computational Linguistics.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38, Uppsala, Sweden. Association for Computational Linguistics.

Minlie Huang, Xiaoyan Zhu, Yu Hao, Donald G Payan, Kunbin Qu, and Ming Li. 2004. Discovering patterns to extract protein–protein interactions from full texts. *Bioinformatics*, 20(18):3604–3612.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. PhoBERT: Pre-trained language models for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1037–1042.

Minh Quang Nhat Pham, Minh Le Nguyen, and Akira Shimazu. 2013. Using shallow semantic parsing and relation extraction for finding contradiction in text. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1017–1021.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30:5998–6008.

Thanh Vu, Dat Quoc Nguyen, Dai Quoc Nguyen, Mark Dras, and Mark Johnson. 2018. VnCoreNLP: A Vietnamese natural language processing toolkit. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 56–60, New Orleans, Louisiana. Association for Computational Linguistics.

Shanchan Wu and Yifan He. 2019. Enriching pre-trained language model with entity information for relation classification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 2361–2364. ACM.

Kun Xu, Siva Reddy, Yansong Feng, Songfang Huang, and Dongyan Zhao. 2016. Question answering on freebase via relation extraction and textual evidence. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2326–2336.