VNLawBERT: A Vietnamese Legal Answer Selection Approach Using BERT Language Model

Chieu-Nguyen Chau
University of Science, Ho Chi Minh city, Vietnam.
chauchieunguyen@gmail.com

Truong-Son Nguyen
Faculty of Information Technology, University of Science, Ho Chi Minh city, Vietnam.
Vietnam National University, Ho Chi Minh city, Vietnam.
ntson@fit.hcmus.edu.vn

Le-Minh Nguyen
Japan Advanced Institute of Science and Technology, Japan.
nguyenml@jaist.ac.jp

Abstract— **Recently, with the development of NLP (Natural Language Processing) methods and Deep Learning, there are several solutions to the problems in question answering systems that achieve superior results. However, there are not many solutions to question-answering systems in the Vietnamese legal domain. In this research, we propose an answer selection approach by fine-tuning the BERT language model on our Vietnamese legal question-answer pair corpus and achieve an 87% F1-Score. We further pre-train the original BERT model on a Vietnamese legal domain-specific corpus and achieve a higher F1-Score than the original BERT at 90.6% on the same task, which could reveal the potential of a new pre-trained language model in the legal area.**

Keywords— natural language processing, question answering, answer selection, language model, legal document.

I. INTRODUCTION

Asking questions about laws is a crystal clear need in any country, but it is not easy since an enormous number of laws have been enacted over the last decades; furthermore, understanding laws requires certain knowledge in the legal domain. Therefore, building a question-answering system in the legal domain is an essential need. It not only helps a normal person to find an answer to their question based on current legal documents but also helps lawyers in their work.

A question-answering system consists of several parts, one of them is the answer selection which aims to choose the best relevant candidates among retrieved documents by measuring the relevance between a question and each retrieved document. Besides that, modern language models have proved to give a great contextual representation of words in sentences. Their impressive results on downstream tasks like sentence pair classification yield a promising approach to measure the relevance in the answer selection task.

BERT[1] is a language model that was pre-trained on a large general corpus and achieved state-of-the-art in several NLP tasks. It is interesting to investigate the effect of using BERT and a sentence pair classification task to the answer selection problem in the Q-A system, especially in Vietnamese. Therefore, we introduce VNLawBERT, an approach to select relevant candidates by fine-tuning BERT using our question-answer pair dataset. Additionally, we further pre-train BERT on a legal domain-specific corpus to achieve higher performance. Our contributions are:

- We construct a training and hand-annotated testing dataset for the Vietnamese answer selection task.
- We propose a solution to the answer selection problem in the Vietnamese legal question answering system.
- We compile a large corpus of text that represents the Vietnamese legal documents.
- We achieve a higher performance model (VNLawBERT), evaluate and compare it with the original BERT model on the Vietnamese answer selection task.

II. RELATED WORKS

In this section, we will present some existing Q-A (Question-Answering) systems, especially in Vietnamese and answer selection methods.

Q-A systems are divided into two types: knowledge-based and retrieval-based.

A knowledge-based system tends to build a huge graph with linked entities. Dai Quoc Nguyen, Dat Quoc Nguyen, Son Bao Pham[2] introduced an ontology-based Q-A system for the Vietnamese language, it includes a question analysis module and an answer extraction module, their experiment results were promising, they achieved an accuracy of 95% in Question Analysis module and 70% in Answer Retrieval module.

On the other hand, retrieval-based systems try to retrieve relevant documents and extract the answer from those documents. Huu-Thanh Duong, Bao-Quoc Ho[3] proposed a Q-A system for Vietnamese legal documents. They applied similarity calculation to select and extract the answer from the relevant documents retrieved from Lucence. They achieved a precision of approximately 70% in the experiment. However, the answer selection method in this system relies on calculating similarity scores using tf-idf. Therefore it can not capture the contextual relationship between words, our approach address this problem using a contextual language model like BERT.

Jamshid Mozafari, Afsaneh Fatemi, Mohammad Ali Nematbakhsh[4] made use of the BERT language model and proposed their answer selection method. The result was pretty high, it proved that a pre-trained language model is an essential tool in NLP tasks such as answer selection.

In the evolution of NLP, traditional language models like word2vec[5][6] tried to convert words token into vectors in a non-contextual way, in which a word is represented as a single vector in the vocabulary. This is not suitable in some cases, for example, the words "bank" in the sentence "My bank was robbed" and in "I am sitting at the bank of the river" have the same vector representation. Recent unsupervised pre-trained language models like BERT, ElMo[7], XLNet[8] address this problem by contextually embed each word token based on its surroundings. In this paper, we will focus on BERT, a language model pre-trained on general text BooksCorpus and English Wikipedia, which significantly improve the performance of many NLP tasks such as sentence-pair classification, question-answering, language inference. Also, many studies have shown that further pre-training BERT or completely pre-training the model from scratch using domain-specific corpora can significantly improve the performance of it. SciBERT[9], BioBERT[10], ClinicalBERT[11], FinBERT[12] are examples, they all performed better than the original BERT in a domain-specific task.

To the best of our knowledge, our work is the first to propose an answer selection approach using BERT language model for the Vietnamese legal Q-A system.

III. BACKGROUND AND DATASETS

A. Background & Problem statement

1) Vietnamese legal documents: In this section, we present an overview of Vietnamese legal documents and their structure. Legal documents in Vietnam are divided into the following categories:

- Constitution
- Code (of Law)
- Ordinance
- Order
- Resolution
- Joint Resolution
- Decree
- Decision
- Circular
- Joint Circular
- Directive

A legal document content has different levels include: chapter, section, article, paragraph, point. Each legal document has its own validity. When the government enacts an update of a certain document, the existing one will be expired or partially expired.

Lawyers give their advice based on articles of valid or partial valid documents, they typically quote some articles from legal documents and conclude an answer to the question or a situation.

2) Question-Answering system and Problem Statement: In this research, we focus on the retrieval-based question-answering system, its architecture consists of four parts: Question Processing, Document Retrieval, Answer Selection, and Answer Extraction. The question processing part detects the question's type, generates a query from the question. The document retrieval part takes that query and retrieve relevant documents. Those documents are then evaluated by the answer selection part to pick the best relevant candidates. Finally, the answer extraction part processes the candidates to find the exact answer to the question. Fig. 1 shows the general architecture of a retrieval-based Q-A system.

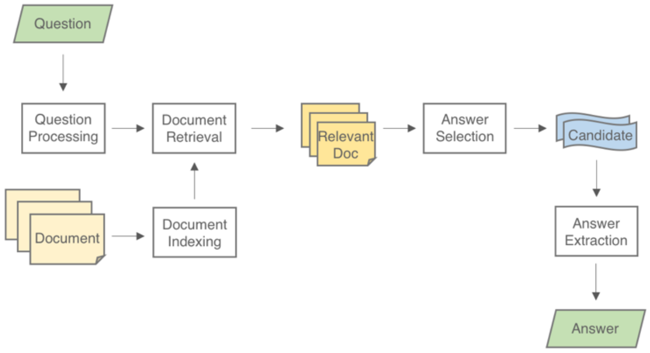

Fig. 1 The general architecture of a Retrieval-based question-answering system.

To address the answer selection problem, we aim to select relevant candidates among the retrieved documents, a retrieved document can be articles, passages, or sentences of a legal document. To this end, we compile a classification task with an input consists of two sequences represent a question and a retrieved document, the output is a confirmation whether the document is a candidate for the question or not.

B. Question - Answering Classification Dataset

To build the dataset for our answer selection task, we need real-world legal questions and relevant answers. We choose Thu Ky Luat's website[13] to extract the data, Thu Ky Luat is a law consulting company that provides lawyer's advice to user's questions. We ran an extractor using Scrapy[14] and acquired about 250,000 question-answer pairs in 27 domains shown in Table I.

Table I. QUESTION-ANSWER DOMAINS LIST

No.	Name of domain	English name of domain
1	Doanh nghiệp	Enterprise
2	Đầu tư	Investment
3	Thương mại	Commerce
4	Xuất nhập khẩu	Import and export
5	Tiền tệ-Ngân hàng	Monetary-Banking
6	Thuế-Phí-Lệ Phí	Tax-Fee-Charge
7	Chứng khoán	Stock
8	Bảo hiểm	Insurance
9	Kế toán-Kiểm toán	Accounting-auditing
10	Lao động-Tiền lương	Labor-Salary
11	Bất động sản	Real estate
12	Dịch vụ pháp lý	Legal service
13	Sở hữu trí tuệ	Intellectual property
14	Bộ máy hành chính	Bureaucracy
15	Vi phạm hành chính	Administrative violation
16	Trách nhiệm hình sự	Criminal responsibility
17	Thủ tục Tố tụng	Procedures
18	Tài chính nhà nước	State financial
19	Xây dựng-Đô thị	Construction-Urban
20	Giáo dục	Education
21	Tài nguyên-Môi trường	Resources-Environment
22	Thể thao-Y tế	Sports-Health
23	Quyền dân sự	Civil rights
24	Văn hóa-Xã hội	Sociocultural
25	Công nghệ thông tin	Information technology
26	Giao thông-Vận tải	Transportation
27	Lĩnh vực khác	Other domains

Those questions and answers are used to create the training and testing datasets for our answer selection task. With each question, we pair it with the correct answer as a candidate (labeled as "1") and other answers as non-candidates (labeled as "0"). The total number of tokens in a question and a candidate/non-candidate is less than or equal to 512.

The training dataset is automatically generated. For non-candidate examples, we use Elasticsearch[15] to find a sequence from our question-answer data that has similar content to the candidate. In this dataset, each question has one candidate and two non-candidates.

To make the evaluation accurate, the testing dataset is handpicked by us to make sure non-candidates are correctly labeled. In our testing dataset, each question has one candidate and four non-candidates.

The sizes of training and testing datasets are described in Table II. An example of our datasets is shown below:

1) Candidate example

- Question: Vi rút máy tính là gì? (What is a computer virus?)
- Candidate: Căn cứ pháp lý: Điều 4 Luật Công nghệ thông tin 2006 Vi rút máy tính là chương trình máy tính có khả năng lây lan, gây ra hoạt động không bình thường cho thiết bị số hoặc sao chép, sửa đổi, xóa bỏ thông tin lưu trữ trong thiết bị số. (Article 4 of the Law on Information Technology 2006 Computer virus means a computer program capable of spreading or causing abnormal operation of digital equipment or copying, modifying or deleting information stored in digital equipment.)
- Label: 1

2) Non-candidate example

- Question: Vi rút máy tính là gì? (What is a computer virus?)
- Non-candidate: Căn cứ pháp lý: Điều 2 Luật phòng, chống nhiễm vi rút gây ra hội chứng suy giảm miễn dịch mắc phải ở người (HIV/AIDS) 2006 HIV dương tính là kết quả xét nghiệm mẫu máu, mẫu dịch sinh học của cơ thể người đã được xác định nhiễm HIV. (Article 2 of Law on Prevention and Control of Human Immunodeficiency Syndrome (HIV / AIDS) 2006 HIV positive means the result of an HIV infection confirmed test of a human blood sample or biological fluid.)
- Label: 0

Table II. QUESTION-ANSWER DATASET SIZE

	Number of questions	Number of domains	Number of examples	Disk size
Training	68,174	27	204,522	266 MB
Testing	350	27	1,750	2.8 MB

C. Pre-train Dataset

We also prepare a dataset used to further pre-train BERT in order to give the model more legal domain-specific knowledge. Legal documents should be accurate and come from a truthful source; to this end, we choose the Vietnam Legal Documents National Database's website[16] to extract legal documents. We obtain 23,254 valid or partial valid legal documents and create a 320 MB cased dataset.

IV. METHODOLOGY

In this section, we describe the methods which are applied in fine-tuning BERT on our answer selection task and further pre-training it using the datasets in Section III. We use a Colab Pro instance along with a Cloud TPU from Google to experiment with our methods.

A. Fine-tuning BERT for answer selection task

The answer selection task's goal is to define a sequence is or contains the answer to the question or not. We compile a sentence-pair classifier that uses BERT as the initial checkpoint and fine-tune it with our dataset.

We use the last checkpoint from the BERT-Base, Multilingual Cased. We set a maximum sequence length of 512 tokens, a batch size of 128, a learning rate of 2e-5, and train over three epochs.

We found there is small randomness in the results, so we run the process three times using the same model configurations and calculate the average result.

B. Pre-training BERT with legal data

We further pre-trained BERT with our legal documents dataset to give the model more knowledge in the Vietnamese legal domain. From the last checkpoint of Multilingual Cased BERT-Base, we make use of the original MLM (Masked Language Model) and NSP (Next Sentence Prediction) task of BERT to further pre-train the model with our legal documents dataset.

We set a maximum sequence length of 512 tokens, a batch size of 128, a learning rate of 2e-5 (recommended by BERT document), and train over 20 epochs (181765 steps) in 2 days to achieve a new pre-trained model called VNLawBERT. Then we compare the result of our new model with BERT-Base on our answer selection task.

V. EXPERIMENTS AND RESULTS

We present the results of the models in this section. We use precision, recall, and F1-score as our metrics, all metrics are calculated on positive results

A. Results

The result in Table III indicates that BERT can perform quite good to classify candidate and non-candidate with the F1-Score about 87%.

However, it also shows that the legal domain-specific model VNLawBERT performs even better than the BERT-Base in all metrics, especially improving F1-Score by 3.5% from the BERT-Base.

It is because the context of words in legal documents is way more different than the context in a general domain corpus like Wikipedia. The model needs further pre-training process to understand the contexts of legal problems.

Table III. PERFORMANCE OF FINE-TUNED BERT-Base AND VNLawBERT

	Precision	Recall	F1-Score
BERT-Base	0.804	0.952	0.872
VNLawBERT	**0.860**	**0.958**	**0.906**

B. Additional experiments

In this paper, we also fine-tune VNLawBERT models on different training datasets to discover the effect of training size and the question's domain on the result.

1) Effect of the question's domain: Since our testing dataset consists of questions from several domains, we hypothesize that training on a multi-domains dataset makes the model perform better than training on a dataset of one or two domains. We test our hypothesis in this section.

We fine-tune the models on a training dataset which only consists of all the examples of one domain. In this case, we use "Thủ tục tố tụng" (Procedures) and "Thuế-Phí-Lệ Phí" (Tax-Fee-Charge) domains since they contribute the least amount of examples in our testing dataset, we obtain 40,000 examples.

We fine-tune another model on a different dataset which has the same size, constructed from questions in all domains. The performance of our answer selection task in Table IV proves that the model with multi-domains knowledge performs better.

Table IV. PERFORMANCE OF 2-DOMAINS VNLawBERT AND MULTI-DOMAINS VNLawBERT

	Precision	Recall	F1-Score
2-domains VNLawBERT	0.709	0.960	0.816
VNLawBERT	**0.749**	0.960	**0.841**

2) Effect of the training size: In this experiment, we evaluate the model with different dataset sizes to explore a sufficient number of examples needed to train the model.

We use the methods described in Section III to build three datasets of different sizes: 20%, 52%, 100% of our training dataset respectively.

The result is shown in Table V indicates that the more examples we have, the more accurate the model is. In our experience, using more than 204,000 examples does not improve the performance of the model. Therefore, a training dataset of 204,000 examples (8,000 examples from each domain) is sufficient for the model to perform at its best.

Table V. PERFORMANCE OF VNLawBERT WITH DIFFERENT TRAINING SIZES

	% of our training dataset	Precision	Recall	F1-Score
40,500 examples	20	0.749	0.960	0.841
107,000 examples	52	0.818	0.976	0.890
204,000 examples	100	**0.860**	0.958	**0.906**

VI. CONCLUSION

In this paper, we address the answer selection problem by fine-tuning BERT language model on our question-answer dataset. We also reveal the potential of a new domain-specific model for the legal area since our VNLawBERT model outperforms the original BERT model in our answer selection task. With this research, we hope researchers can experiment the model with other tasks in the legal domain such as Named Entity Recognition, Reference Extraction, Question Classification to build a legal domain-specific language model, which is also our future work.

ACKNOWLEDGMENT

This research is funded by the University of Science, VNU-HCM under grant number CNTT 2020-14.

REFERENCES

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, in Proceedings of NAACL, pages 4171-4186, 2018.

[2] Dai Quoc Nguyen, Dat Quoc Nguyen, Son Bao Pham, “A Vietnamese Question Answering System”, International Conference on Knowledge and Systems Engineering, 2009.

[3] Huu-Thanh Duong, Bao-Quoc Ho, “A Vietnamese Question Answering System in Vietnam’s Legal Documents”, IFIP International Conference on Computer Information Systems and Industrial Management, 2016.

[4] Jamshid Mozafari, Afsaneh Fatemi, Mohammad Ali Nematbakhsh, “BAS: An Answer Selection Method Using BERT Language Model”, 2019.

[5] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, “Efficient Estimation of Word Representations in Vector Space”, Proceedings of the International Conference on Learning Representations (ICLR 2013), 2013.

[6] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, “Distributed Representations of Words and Phrases and their Compositionality”, Advances in neural information processing systems 26, 2013.

[7] Matthew Peters et al., “Deep Contextualized Word Representations”, ”, in Proceedings of NAACL, 2018.

[8] Zhilin Yang et al., “XLNet: Generalized Autoregressive Pretraining for Language Understanding”, Conference on Neural Information Processing Systems (NeurIPS 2019), 2019.

[9] Iz Beltagy, Kyle Lo, Arman Cohan, “SCIBERT: A Pretrained Language Model for Scientific Text”, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019.

[10] Jinhyuk Lee et al., “BioBERT: a pre-trained biomedical language representation model for biomedical text mining”, Bioinformatics, Volume 36, Issue 4, 15 February 2020, Pages 1234–1240, 2019.

[11] Kexin Huang, Jaan Altosaar, Rajesh Ranganath, “ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission”, The ACM Conference on Health, Inference, and Learning, 2020.

[12] Yi Yang, Mark Christopher Siy UY, Allen Huang, “FinBERT: A Pretrained Language Model for Financial Communications”, 2020.

[13] Thu Ky Luat's website [Online] https://nganhangphapluat.thukyluat.vn/

[14] Scrapy [Online] https://scrapy.org/

[15] Elasticsearch [Online] https://www.elastic.co/

[16] Vietnam Legal Documents National Database’s website [Online] http://vbpl.vn/pages/portal.aspx