Apply Bert-based models and Domain knowledge for for Automated Legal Question Answering tasks at ALQAC 2021

Truong-Thinh Tieu
Faculty of Information Technology,
University of Science,
Vietnam National University,
Ho Chi Minh city, Vietnam
thinhtieu1107@gmail.com

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

Nguyen-Minh-Hoang Bui
Faculty of Information Technology,
University of Science,
Vietnam National University,
Ho Chi Minh city, Vietnam
buinm.hoang@gmail.com

Truong-Son Nguyen
Faculty of Information Technology,
University of Science,
Vietnam National University,
Zalo Group – VNG Corporation
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*— With robust development in NLP (Natural Language Processing) methods and Deep Learning, there are a variety of solutions to the problems in question answering systems that achieve extraordinary results. In this paper, we describe our approach using at the Automated Legal Question Answering Competition (ALQAC) 2021. In this competition, we achieved the first prize of all tasks with the scores of 88.07%, 71.02%, 69.89% in Task 1, Task 2 and Task 3 respectively.**

Keywords*— *alqac 2021, natural language processing, question answering, answer selection, legal domain, language model, deep learning

I. Introduction

To widen knowledge and improve the research community in terms of systems engineering, the 13th IEEE International Conference on Knowledge and Systems Engineering is organized. The main purpose of the ALQAC 2021 [1] competition is to bring together researchers and students to share research results and make contributions in the field. The conference of the KSE series will be held in Bangkok and be co-organized by Sirindhorn International Institute of Technology, Thammasat University (SIIT) and Artificial Intelligence Association of Thailand (AIAT).

Question and answer in the law documents is an ambitious task due to the sophistication and topic varies of law document systems especially in Vietnam. Providing an accurate answer to a legal question is a challenging task and requires expert knowledge. Legal documents are often long and complex and take time to read as well as indicate the correct answer. Therefore, using an information retrieval system combined with BERT [2] pre-train on the amount of legal data from the Internet is the feasible combination to get a more correct answer. We use VNLawBERT [3], an approach to select relevant candidates by fine-tuning BERT [2] using our question-answer pair dataset. Additionally, we further pre-train BERT [2] on a legal domain-specific corpus to achieve higher performance.

The accuracy of the answer set provided by the system currently is affected by legal documents data shortage or lack of preference data observed from question asker and answer from experts with profound knowledge. Several approaches have been proposed to solve this issue. In this study, we enhance the performance of question and answering systems. The contributions of this study are as follows:

- We proposed a legal document retrieval to get relevant articles for Task 1.
- We enriched the training dataset for Task 2 and Tak 3 by compiling more data from external sources.
- We applied VNLawBERT [3] model to enhance the performance from BERT-Base [2] for legal domain tasks.

II. Dataset

Automated Legal Question Answering Competition (ALQAC) 2021 [1] provides 2 legal data files to support the calculation process are law dataset and statement dataset.

A. Law dataset

This is a dataset containing information about laws and its articles. This dataset has 16 laws and 2279 articles of laws in file, here is an example of an article:

Table I. An example of a law article

Law id	45/2019/QH14
Article id	50
Article text	Thẩm quyền tuyên bố hợp đồng lao động vô hiệu. Tòa án nhân dân có quyền tuyên bố hợp đồng lao động vô hiệu. (Authority to declare the labor contract invalid. The People's Court has the right to declare an employment contract invalid.)

B. Statement dataset

This is a dataset containing a list of statements and its relevant articles along with a label representing the truth and falsehood of the statements. This dataset has 412 statements and 419 relevant articles, in which 182 statements with label True, 230 labels False, here is an example of a statement:

Table II. An example of a statement and relevant article

Question id	q-189
Statement	Chiếm đoạt tài sản lớn hơn 500.000.000 đồng sẽ bị phạt tù dưới 5 năm. (Appropriating property greater than 500,000,000 VND will be punished by imprisonment for less than 5 years.)
Label	False
Relevant law id	100/2015/QH13
Relevant article id	170

III. OUR PROPOSED SOLUTION

A. Task 1 - Legal Document Retrieval

1) Requirement:

Task 1 required is finding all law articles related to the statement. An article is considered as relevant to the statement if the statement rightness can be entailed by the article.

2) Methodology

We proposed a model consisting of two parts: the Retrieval part and the Selection part.

From an input statement, the Retrieval part queries a list of candidate articles from the indexed law articles using Elasticsearch [4]. The top 20 results are sent to the Selection part to select the relevant articles of the statement.

In the Selection part, we built an article ranking model by fine-tuning VNLawBERT [3] with our dataset. VNLawBERT [3] is a domain-specific BERT [2] model for Vietnam legal tasks proposed by Chieu-Nguyen Chau, Truong-Son Nguyen, and Le-Minh Nguyen in 2020.

The input of the model is a pair of the statement and each candidate article. The output is a value that indicates whether a candidate article is relevant to the statement or not, along with a relevance score from 0 to 1.

Then we select the article that has the highest relevance score given by the model. In case there is another article that has the same score, both articles will be selected. If the selected article has a score lower than 0.7, we consider the Selection part's result to be unreliable. Therefore, we discard the result of the Selection part and choose the best candidate article from the result of the Retrieval part.

3) Data processing

To generate the training dataset for the model in the Selection part, we pair a statement from the statement dataset of Section II with its relevant articles to form an example with the label True. Then, for its statement we create 4 more examples with the label False. In order to create a labeled False example, we paired the statement with a different article which has the closest content with the statement's relevant article.

With 412 statements, we create 2060 pairs of sentences, randomly select 40 pairs of sentences to use for testing data. Here is an example of a pair sentences dataset:

Table III. An example of a pair sentences dataset

Statement	Thời hạn tạm trú tối đa là hai năm nếu không gia hạn. (The maximum temporary stay is two years if not extended.)
Relevant article	Thời hạn của giấy phép lao động Thời hạn của giấy phép lao động tối đa là 02 năm, trường hợp gia hạn thì chỉ được gia hạn một lần với thời hạn tối đa là 02 năm. (The term of the work permit The maximum term of a work permit is 2 years, in case of extension, it can only be extended once with a maximum term of 2 years.)
Label	False

B. Task 2 - Legal Textual Entailment

1) Requirement:

Task 2's goal is to construct Yes/No question answering systems for legal queries, by entailment from the relevant articles. Based on the content of legal articles, the system should answer whether the statement is true or false.

2) Methodology

We proposed a law entailment model using a sentence pair classification task to solve this task. The model accepts an input consisting of a statement and a relevant article, the output is a binary value that indicates whether the rightness of a statement is Yes or No regarding its relevant article.

To train this model, we construct a statement-article pair dataset along with its Yes-No label from the given data set described in Section II. Because of the lack of examples of the given dataset, we collected an external dataset to enlarge the training dataset. We also used VNLawBERT [3] as our pre-trained model for Task 2.

3) Data processing

Most of the legal documents, statements and question & answer pairs are represented as HTML format and not all websites have APIs provided. Some websites refuse to provide any APIs, some propose RSS feeds, so they cause limitations on data use and training model. In that case, building a small application to crawl data is necessary to eliminate law data shortage.

Firstly, we collect all feasible URLs for data crawling, the correct format must contain three pieces of information: a question or a statement, a statement label true and false are extracted from the answer, an answer for a given question and relevant law article, explanation. Secondly, those data were crawled from HocLuat.vn [5] and ThuKyPhapLy.com [6] with Nodejs and Cheerio [7] library, after got all data, we displayed it as following format:

Table IV. An example of an external dataset

Statement	Người nước ngoài phạm tội trên máy bay của Việt Nam khi máy bay đó đang hoạt động trên không phận quốc tế thì không bị coi là phạm tội trên lãnh thổ Việt Nam. (Foreigners who commit crimes on a Vietnamese aircraft while that aircraft is operating in international airspace shall not be considered a crime in the Vietnamese territory.)
Explanation	Bộ luật hình sự Việt Nam hiện nay còn có khái niệm lãnh thổ mở rộng, tức là lãnh thổ theo giác độ chủ quyền quốc gia về phương diện pháp lý. (The current Vietnamese penal code also has the concept of extended territory, that is, territory from the perspective of national sovereignty in legal terms.)
Label	False

From the data provided in Section II, each pair of sentences will be created in the format of a statement, relevant article and label as in statement dataset. In addition, with data collected from external sources, the structure is similar: a statement, an explanation and True/False label.

From the statement dataset of Section II generated 419 pairs of sentences (25% split, about 100 sentences for test data), combined with sentence pairs created from external sources, we got 2986 pairs of sentences for the training and testing process. Below is an example of a dataset:

Table V. An example of a pair sentences dataset

Statement	Trong mọi trường hợp việc truy cứu trách nhiệm hành chính không cần xét đến thực tế là hậu quả đã xảy ra hay chưa xảy ra. (In all cases, the prosecution of administrative liability does not need to take into account the fact that the consequences have occurred or have not occurred.)
Relevant article	Vì vi phạm hành chính là vi phạm cấu thành hình thức nên có đủ hành vi cấu thành vi phạm hành chính mà không cần hậu quả xảy ra. Hậu quả chỉ là tình tiết để lựa chọn hình thức và mức độ xử phạt. (Since an administrative violation is a formal violation, there are enough acts to constitute an administrative violation without consequences. Consequences are just circumstances to choose the form and level of sanction.)
Label	True

C. *Task 3 - Legal Question Answering*

1) Requirement:

This task is to answer Yes/No for the given legal statement, in particular indicating whether its rightness is true or false. This question answering is a concatenation of Task 1 and Task 2.

2) Methodology

We proposed two methods to solve Task 3 are combining Task 1 with Task 2 method, and Single Sentence Classification method:

- In the first method, we used Task 1 to retrieve the article with the highest relevance score for the input statement. After that, we used Task 2 to entail the rightness of the statement with respect to its relevant article.
- In the Single Sentence Classification method, we built a legal statement rightness classification model by fine-tuning VNLawBERT [3] model with our training dataset. The dataset consists of statements and their rightness labels. To enrich the dataset, we add clauses in the law dataset from the given dataset in Section II as labeled True examples.

3) Data processing

An example of a single sentence dataset will be created with the format including 1 statement sentence and corresponding label. With 3 data sources, each will be created as follows:

- Statement and external dataset: statement, corresponding label
- Law dataset: the statement is created by dividing the sub-clauses in the law and labeling it as True

From the statement dataset in Section II, we create 412 pairs of sentences (25% split, about 100 sentences to data test), combined with data obtained from law dataset and external sources, we have 9820 sentences for training. Here is an example of a single dataset:

Table VI. An example of a single sentence dataset

Statement	Chiếm đoạt tài sản lớn hơn 500.000.000 đồng sẽ bị phạt tù dưới 5 năm. (Appropriating property greater than 500,000,000 VND will be punished by imprisonment for less than 5 years.)
Label	False

IV. EXPERIMENTS AND RESULTS

We fine-tuned our models over 20 epochs using the configurations of a maximum sequence length of 512 tokens, a batch size of 32, a learning rate of 2e-5. The experiment result was calculated by the average of 3 tries to eliminate the randomness of the model.

A. Task 1 - Legal Document Retrieval

The evaluation metric is the Macro F2-Score. It is the average of F2-Score from each query. The F2-Score formula for each query is as follow:

$$Precision_i = Precision\ of\ query\ i^{th}$$

$$Recall_i = Precision\ of\ query\ i^{th}$$

$$F2_i = \frac{5 \times Precision_i \times Recall_i}{4 \times Precision_i + Recall_i}$$

$$Macro\ F2 = average\ of\ (F2_i)$$

We chose the method of just using the Retrieval part of our model, selecting the top 1 result from Elasticsearch [4] as our baseline model. We created another model to compare by replacing the model we used in the Selection part with the Multilingual BERT-Base [2] model.

The results in Table VII showed that our solution outperformed the baseline model by 12.25% and was 3.08% higher than using the BERT [2] model in the Selection part.

Table VII. Experiment results of Task 1

	Precision	Recall	Macro F2
Elasticsearch	0.733	0.695	0.701
BERT-Base	0.825	0.788	0.792
Our model	**0.863**	**0.821**	**0.823**

B. Task 2 - Legal Textual Entailment

In the experiment of Task 2, we chose the BERT-Base [2] as our baseline model and the Accuracy as our evaluation metric. From the result shown in Table VIII, our model outperformed the BERT-Base [2] model and achieved an accuracy of 69%.

Table VIII. Experiment result of Task 2

	Precision	Recall	Accuracy
BERT-Base	0.583	0.480	0.613
Our model	**0.659**	**0.627**	**0.686**

We made another experiment to evaluate the effectiveness of training epoch numbers on our model. We set the batch size of 32 and experimented with three epoch numbers: 5, 10, and 20 epochs. Our model reached the best performance with 20 epochs with an accuracy of 69%, described in Table IX. From our experience, using a higher epochs number would not give a better performance.

Table IX. Experiment the effectiveness of epoch number

	Precision	Recall	Accuracy
5 epochs	0.594	0.476	0.625
10 epochs	0.585	0.526	0.622
20 epochs	**0.659**	**0.627**	**0.686**

C. Task 3 - Legal Question Answering

We made two experiments following the methods described in part C of Section III. We noticed an improvement in the accuracy of the method combining Task 1 and Task 2 compared to the single classification method. It has an accuracy of 69.5% compared to 59.2% of the latter.

Table X. Experiment results of Task 3

	Precision	Recall	Accuracy
Single sentence classification	0.550	0.478	0.592
Combining Task 1 and Task 2	**0.614**	**0.643**	**0.695**

D. Competition Result

Table XI shows the competition result at ALQAC 2021 [1] of our team (Aleph) compared to the other top teams. We got the highest score for all tasks, with a Macro F2 of 88.07% for Task 1, 69.89% Accuracy for Task 2, and 71.02% Accuracy for Task 3.

Table XI. ALQAC 2021 competition results

	Task 1	Task 2	Task 3
Aleph	**88.07%**	**69.89%**	**71.02%**
AimeLaw	80.61%	69.89%	64.77%
Kodiak	79.55%	68.18%	62.50%
Dat N	71.28%		64.77%

V. CONCLUSION

In this paper, we use pre-trained models BERT [2] and Elasticsearch [4] and extract datasets on the HocLuat.vn [5] and ThuKyPhapLy.com [6] website to enrich legal data. Our approach is that utilizes the results of previous tasks to generate more data for the training process, creating extra pairs of sentences from related articles. We experimented on competition's datasets for legal questions and answers in which datasets given by competition. Experimental results have shown that our approach with integrated VnLawBERT [3] gave better results.

REFERENCES

[1] ALQAC [Online] https://iisi.siit.tu.ac.th/KSE2021/front/show/call-for-alqac

[2] 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.

[3] Chieu-Nguyen Chau, Truong-Son Nguyen, and Le-Minh Nguyen, "VNLawBERT: A Vietnamese Legal Answer Selection Approach Using BERT Language Model", in Proceeding of 2020 7th NAFOSTED Conference on Information and Computer Science (NICS), page

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

[5] HocLuat.vn [Online] https://hocluat.vn

[6] ThuKyPhapLy.com [Online] https://thukyphaply.com

[7] Cheerio [Online] https://github.com/cheeriojs/cheerio.git