Abstract—In this paper, we propose deep learning based methods for addressing the problems of legal text processing in the Automated Legal Question Answering Competition (ALQAC 2021). The competition consists of three challenging tasks based on well-known statute laws in Vietnamese and Thai Language. We participated in three tasks related to the Vietnamese statute law, including the legal document retrieval (Task 1), the legal textual entailment (Task 2), and the legal question answering (Task 3). In Task 1, we combine semantic and lexical scores to identify relevant articles. In Tasks 2&3, we fine-tune pretrained models for the Vietnamese language in order to classify proper labels. The experimental results demonstrate both the difficulties and potentials associated with these approaches.

Index Terms—Legal Document Processing, Deep Learning, Kodiak Team

I. INTRODUCTION

The Automated Legal Question Answering Competition (ALQAC) is a competition designing to develop methods for legal text processing. The competition uses two main types of data: Vietnamese statute law and Thai statute law. In ALQAC 2021, there are three tasks in total, with Task 1, 2, and 3 being retrieval, entailment, and question answering challenges, respectively. The first task requires identifying the set of articles in a law corpus that support a legal statement’s decision. Given a set of relevant articles, the second task designs to determine the entailment of a legal question. In a simple way, the third task could be a combination of the two above tasks, in that the system should determine if a legal question is true or false.

Legal information processing is a problem that has received a lot of attention from the natural language processing community. In the legal document retrieval, there were a large number of studies approaching this problem by adopting lexical similarity models [1]–[3], such as BM25. Besides that, various approaches have been combined the lexical similarity model and the deep learning model [4]–[6]. Their experiments demonstrate that combining a deep learning model and a lexical similarity model produces better results than using lexical similarity alone. Regarding the legal textual entailment, approaches based on deep learning are gaining traction. In particular, to learn various semantic fragments which represent the correctness of a statement, Rabelo et al. [7] use a decomposable attention model, which is a simple neural architecture for natural language inference. Besides, there are multiple studies [8], [9] also apply BERT [10] for this task. In comparison to the above two tasks, the legal question answering is the most challenging, as the model must learn how to perform both tasks: retrieval and entailment. Nguyen et al. [4] have examined the effectiveness of two-stage (pre-training and fine-tuning) learning for this task. Specifically, to reduce ambiguity and increase the performance in legal text processing, the study proposed ParaLaw Nets, a pretrained model family using sentence-level cross-lingual information.

The amount of data provided by the organizer makes it challenging to develop good models using the common deep learning approaches. As a result, we use pretrained models from problems that have much more data and then fine-tune them for Tasks 2 and 3. In Task 1, we have used the combination of lexical score and semantic score to filter correct candidates. With the above approaches, the Kodiak team achieves competitive results in ALQAC 2021.

The rest of the paper is organized as follows. The next section describes task descriptions, a brief analysis of the dataset, and evaluation metrics. Section 3 details the proposed methods for each task. Section 4 presents the experimental results of the proposed approaches, and we draw some conclusions in Sect. 5.

II. TASK DESCRIPTION

A. Task 1. Legal Document Retrieval

In legal technical terms, a legal article is regarded as ”relevant” to a statement/question if the article implies the statement’s correctness. The legal document retrieval task (Task 1) focuses on filtering relevant articles $S_1, S_2, \ldots S_N$ from the well-known statute laws in the Vietnamese Language to a legal question $Q$, such that the utilization of these articles in combination with the statement can entail the statement rightness (as Yes/No).

B. Task 2. Legal Document Entailment

The legal textual entailment task requires determining whether a legal question $Q$ is true or false given the content of a set of articles, $S_1, S_2, \ldots S_N$ relevant to $Q$. The task’s aim is
to build Yes/No question answering systems for legal queries, by entailing from the relevant articles.

C. Task 3. Legal Question Answering

The legal question answering task requires determining whether a legal question $Q$ is true or false. The task’s goal is to construct Yes/No question answering systems for legal queries. In essence, Task 3 is similar to Task 2, except that the task doesn’t limit the articles, and the model needs to extract knowledge from the whole dataset or any other source to obtain correct labels.

D. Dataset

In ALQAC 2021, a short analysis is indicated in Table I. The data of tasks are drawn from well-known statute laws in the Vietnamese language. We can see that every single article has more than 170 tokens long on average, the number of paragraphs over 6. In addition, almost all queries have only one relevant article.

<table>
<thead>
<tr>
<th>TABLE I</th>
<th>DATASET ANALYSIS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Train</td>
</tr>
<tr>
<td>Average Word/Article</td>
<td>179</td>
</tr>
<tr>
<td>Average Paragraph/Article</td>
<td>6.48</td>
</tr>
<tr>
<td>Maximum Word</td>
<td>1226</td>
</tr>
<tr>
<td>Maximum Paragraph</td>
<td>139</td>
</tr>
<tr>
<td>Number of Query</td>
<td>412</td>
</tr>
<tr>
<td>Average relevant article</td>
<td>1.017</td>
</tr>
</tbody>
</table>

E. Evaluation Measure

In Task 1, the evaluation methods are precision, recall, and F2-measure, as follows:

$$
Precision_i = \frac{Q_{(TP)i}}{Q_{(TP+FP)i}}
$$  
(1)

$$
Recall_i = \frac{Q_{(TP)i}}{Q_{(TP+FN)i}}
$$  
(2)

$$
F_{2i} = \frac{5 \times Precision_i \times Recall_i}{4 \times Precision_i + Recall_i}
$$  
(3)

$$
F2 = \text{average of } (F_{2i})
$$  
(4)

Where $Q_{(TP)i}$ denotes the number of correctly retrieved articles of query $i^{th}$, $Q_{(TP+FP)i}$ is the number of retrieved articles of query $i^{th}$, and $Q_{(TP+FN)i}$ is the number of relevant cases (paragraphs) of query $i^{th}$.

In Tasks 2&3, both tasks are evaluated with accuracy measure, as follow:

$$
Accuracy = \frac{N_{TRUE}}{N_{ALL}}
$$  
(5)

Where $N_{TRUE}$ denotes the number of queries that were correctly confirmed as true or false, and $N_{ALL}$ is the number of all queries.

III. METHODS

A. Task 1. Legal Document Retrieval

We approach Task 1 in three different ways. We return a ranked list of articles where the most similar one is placed on the highest position, followed by any additional ones determined by the difference $\partial$ between the current score and the highest score. If the difference is less than a specified threshold $H$, we continue to return more articles until the threshold is reached or a maximum number of retrieved articles $R$ is discovered. We utilize the macro-average of precision, recall, and the F-2 metric to evaluate our model.

Approach 1 - Lexical Matching: BM25 [11] is a bag-of-words retrieval technique that ranks a set of documents based on the query keywords included within each document, regardless of the inter-relationship between them. To begin, we normalize all text data by eliminating punctuation and substituting text for digits. We then compute the BM25 scores of all the articles for every statement using Rank-BM25

Approach 2 - Combination of Semantic Searching and Lexical Matching: Semantic search [12] is used to increase search accuracy by understanding the search query’s content. In contrast to traditional approaches, which look for texts exclusively based on lexical matching, the semantic search looks for synonyms as well. The idea behind semantic search in this task is to embed all paragraphs in the legal corpus. At search time, the query is embedded into the same vector space and the closest embedded paragraphs are found. We make use of BERT-Base-Multilingual-Uncased [10] and PhoBERT [13] (Bidirectional Encoder Representation from Transformer) models to embed the query and the paragraphs in the corpus. After that, we compute the cosine similarity between the vector of the query with all vectors of paragraphs. Finally, we select articles with the highest paragraph scores.

Evaluating the effectiveness of semantic search is challenging, however, we are the first to show that the combination of lexical models and semantic models can get promising results on legal documents in the Vietnamese language. Specifically, we use both of lexical score and semantic score as union score with a hyper-parameter $\alpha$:

$$
\text{union score} = \alpha \times \text{semantic score} + (1 - \alpha) \times \text{lexical score}
$$

B. Task 2. Legal Document Entailment

To achieve texture entailment, we take three different approaches based on fine-tuning the pretrained models. We

1https://pypi.org/project/rank-bm25/
divided the provided dataset into 90:10 training and development subsets. Following that, the development set is used to determine the optimal model for each approach.

**Approach 1 - Multilingual-BERT Fine-tuning:** Transformer [14] and two-stage (pre-training + fine-tuning) learning have gained considerable interest in recent years because of their ability to leverage pre-trained contexts. Therefore, in this approach, we use Multilingual-BERT [10] and fine-tune it with a downstream sentence pair classification task (Fig. 2). To be specific, we generate an input text pair by utilizing all of sentences in a relevant article. If the total number of input tokens exceeds the length limitation (512), the sequence is symmetrically truncated. On the training subset, the model is trained by optimizing the cross-entropy loss. We perform an end-to-end way to update all of the parameters. For each question, we applied the textual entailment with new data named BM25 extra data that contains the given relevant articles and two top articles returned by the BM25 algorithm. The labels of the two additional articles are based on the label of the relevant articles.

**Approach 2 - Multilingual-BERT Fine-tuning with Asymmetric Truncation:** The main difference from Approach 1 is that we truncate the text asymmetrically if the input tokens exceed the length limitation. We observe that the majority of questions in the training data are no greater than 128 tokens in length, while the lengths of articles differ a lot. In this case, when the candidate article is too long, the symmetric truncation in Approach 1 can lead to severe information loss in the input sequence. As a result, we are motivated to create a new way for truncating the article content.

In detail, we split each article into multiple parts of no more than 384 tokens in length. Each part has multiple completed sentences, which means we break the article into parts at the end of sentences. Following that, we count the number word of each part overlap with the question and choose the part with the highest value. The model structure and training procedure are identical to those in Approach 1. The experiment shows that the model converges faster and produces a better result on the development set when using the asymmetric truncation strategy.

**Approach 3 - PhoBERT Fine-tuning with Asymmetric Truncation:** The main difference from Approach 2 is that we use the pretrained model PhoBERT [13] instead of the Multilingual-BERT. Because the maximum sequence length for the input to PhoBERT is 256 and the majority of questions are no greater than 128 tokens, we set max tokens in each article part to 128 in this run.

**C. Task 3. Legal Question Answering**

To complete this task, we simply combine the results from the two preceding tasks: Task 1 and Task 2. We attempted three runs of this challenge with varying settings. Specifically, we combine Task 1’s Approach 2 with three approaches in Task 2. The result on the development set surprised us in that the accuracy of Task 2 and Task 3 were not significantly different.
### IV. Experiments and Results

#### A. Task 1. Legal Document Retrieval

To begin, we performed an experiment to determine the most suitable value for the hyperparameter $\alpha$ in Approach 2 of this task (Table II). We evaluated our approach using relevant articles from Task 2’s training data as the development set for this task. We discovered that $\alpha = 0.5$ is a more reasonable value for both pretrained models: PhoBERT and Multilingual BERT, based on experimental results on the development set.

<table>
<thead>
<tr>
<th>Settings ($\alpha$)</th>
<th>Multilingual BERT</th>
<th>PhoBERT</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha = 0.25$</td>
<td>0.7814</td>
<td>0.7855</td>
</tr>
<tr>
<td>$\alpha = 0.5$</td>
<td><strong>0.7872</strong></td>
<td><strong>0.7919</strong></td>
</tr>
<tr>
<td>$\alpha = 0.75$</td>
<td>0.7763</td>
<td>0.7733</td>
</tr>
</tbody>
</table>

### Table III

<table>
<thead>
<tr>
<th>Team ID</th>
<th>Run ID</th>
<th>F2</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kodiak</td>
<td>Run 3</td>
<td>0.7955</td>
<td>3</td>
</tr>
<tr>
<td>Kodiak</td>
<td>Run 2</td>
<td>0.7919</td>
<td>4</td>
</tr>
<tr>
<td>Kodiak</td>
<td>Run 1</td>
<td>0.7855</td>
<td>5</td>
</tr>
<tr>
<td>AimeLaw</td>
<td>Run 2 &amp; 3</td>
<td>0.8061</td>
<td>2</td>
</tr>
<tr>
<td>AimeLaw</td>
<td>Run 1</td>
<td>0.7842</td>
<td>6</td>
</tr>
<tr>
<td>Dat N</td>
<td>Run 1</td>
<td>0.7128</td>
<td>7</td>
</tr>
<tr>
<td>Aleph</td>
<td>Run 1</td>
<td>0.8807</td>
<td>1</td>
</tr>
</tbody>
</table>

Table III shows the results on the test data in Task 1, the legal document retrieval task. From the table, we can see that Run 3, which combines semantic understanding and lexical matching, achieves the best performance among our submitted runs (“Kodiak”). By comparing the result of Run 3 to that obtained using the lexical matching method (Run 1), it is clear that developing semantic neural models improves the performance in the legal case retrieval. This result implies that semantic understanding models and lexical matching models concentrate on distinct aspects of legal texts and hence complement one another in the task of legal document retrieval. For the settings of using the same approach, the PhoBERT based model (Run 3) performs better than the multilingual-BERT based model (Run 2). We assume that for short text (less than 256 tokens), the performance of the PhoBERT pre-trained on a specific language performs better than the multilingual-BERT pre-trained on multi-languages.

Generally, our best run ranked 3rd out of all submitted runs for this task. The highest-ranked run (Run 1 from the “Aleph” team) does outperform all other runs significantly and is worth further discussion.

#### B. Task 2. Legal Document Entailment

As shown in Table IV, our second run, which makes use of asymmetric truncation during multilingual-BERT fine-tuning, yields the best performance among our submitted runs, with a nontrivial improvement. This result demonstrates the effectiveness of asymmetric truncation. To be specific, when creating the input pairs, the asymmetric truncation aims to preserve the information included in the article content. As a result, the model can learn useful information for entailing decisions. Unfortunately, our third run seems to be less effective in this task. We interpret it from several perspectives. For one thing, given the PhoBERT model’s default setting, a sequence cannot exceed 256 tokens; hence, the fine-tuned model may omit critical content necessary to entail the decision. For another, given the limited data scale, the division of training and development set is constant in our experiments. As a result, any bias in the data division process may mislead the output of the model. In summary, our results are ranked 2nd according to accuracy, and the top-1 runs are two teams “AimeLaw” and “Aleph”.

### Table IV

<table>
<thead>
<tr>
<th>Team ID</th>
<th>Run ID</th>
<th>Accuracy</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kodiak</td>
<td>Run 2</td>
<td>0.6818</td>
<td>2</td>
</tr>
<tr>
<td>Kodiak</td>
<td>Run 1</td>
<td>0.6364</td>
<td>4</td>
</tr>
<tr>
<td>Kodiak</td>
<td>Run 3</td>
<td>0.5568</td>
<td>6</td>
</tr>
<tr>
<td>AimeLaw</td>
<td>Run 1</td>
<td>0.6989</td>
<td>1</td>
</tr>
<tr>
<td>AimeLaw</td>
<td>Run 2</td>
<td>0.6761</td>
<td>3</td>
</tr>
<tr>
<td>VIVN</td>
<td>Run 1</td>
<td>0.5795</td>
<td>5</td>
</tr>
<tr>
<td>Aleph</td>
<td>Run 1</td>
<td>0.6989</td>
<td>1</td>
</tr>
</tbody>
</table>

### C. Task 3. Legal Question Answering

As discussed in the methodology section, Task 3 is a simple combination of Tasks 1 and 2. In comparison to the best result for Task 2, the best result for Task 3 is significantly different. We assume that the retrieved relevant articles in Task 1 have a significant impact on the result of this task. The highest-ranked run (Run 1 from the “Aleph” team) does outperform the others significantly. Especially, their highest performance in Task 3 is superior to their best performance in Task 2. This is worth further discussion.

### V. Conclusion and Future Work

In this paper, we introduce our methods for three ALQAC 2021 law tasks: legal document retrieval, legal textual entail-
ment, and legal question answering. In the legal document retrieval task, we investigate both semantic understanding and lexical matching methods to filter correct candidates. The combination of two types of methods improves each one, which implies that semantic understanding models and lexical matching models are complementary in this task. We are ranked 3rd in Task 1, and our best run is ranked third overall among all submitted runs. In the legal case entailment and legal question answering task, the less amount of data is the most challenge to build an effective deep learning model. Therefore, we approach these tasks by fine-tuning pre-trained models from problems that have much more data. We also discovered that minimizing information loss in the input sequence can be greatly beneficial in these tasks. Our best run ranks 2nd among all runs in Task 2 and 3rd among all in Task 3. For future directions, if more Vietnamese legal documents become available, pre-training a BERT in the legal domain may help improve performance on these tasks.

**References**


