Ensemble Learning Methods for Legal Processing Tasks in ALQAC 2022

Hau Nguyen Trung [1,2], Son Nguyen Truong [1]
[1] *Ho Chi Minh University of Science, VNU-HCM*
Ho Chi Minh City, Vietnam
[2]*Ho Chi Minh City Open University*
Ho Chi Minh City, Vietnam
20c11007@student.hcmus.edu.vn, hau.nt@ou.edu.vn, ntson@fit.hcmus.edu.vn

***Abstract*—Automated Legal Question Answering Competition is an annual competition to find the best solution to automatically answer legal questions based on well-known statute laws in the Vietnamese Language. In this paper, we will demonstrate how to solve the problems posed by ALQAC 2022, using BERT and its variants as a backbone network. In addition, we also study using tf-idf and BM-25 to rank the relevance of legal documents. At the same time, this publication also show how to enhance training data to solve the problem of limited training data.**

***Index Terms*—Deep Learning, Question Answering, Legal Text Processing**

I. INTRODUCTION

Automated Legal Question Answering Competition (ALQAC) is an annual competition to find the best solution to automatically answer legal questions based on well-known statute laws in the Vietnamese Language. In ALQAC 2022, the organizers posed questions to the participants based on the provisions of 4 Codes of Vietnam: Civil Code, Labor Code, Cybersecurity Code and Penal Code. The competition consists of two main tasks with the desire to promote the research community to develop legal support systems.

The Task 1 is legal document retrieval. With this task, models of receiving a request and having the task of extracting articles related to that request, these articles are all part of the 4 Codes of Vietnam. This is a very difficult but necessary task in practice for legal professionals and those in need of legal assistance. The Task 2 is legal question answering. The goal of Task 2 is to ask the model to give the exact answer extracted from the content of the rule in the output of Task 1. In general, the answers of Task 2 are divided into 2 main groups: true - false or answer extraction.

In order to conquer similar tasks in the past, some researchers have approached lexical-based methods, to find the relevance of word embedding vectors, these methods were also initially effective. But the predicted results are not so striking. Others use a combination of lexical and semantic approaches, the latter of which have clearly yielded remarkable predictive results.

Previous studies have demonstrated that the reduction of search space will be inversely proportional to the probability of correct prediction. Therefore, we expect that preprocessing to reduce the prediction space for both tasks before applying deep learning models will yield the expected results.

On that basis, in this article, we will conduct a survey, analysis and evaluation of text classification methods, deep learning models that effectively process both Vietnamese and foreign language data. From there, propose the appropriate models for the best results for both tasks.

II. RELATED WORKS

In this section, we will introduce the processing methods that have been effective on documents, especially Vietnamese documents. At the same time, this section also presents a selection of suitable methods for our proposal.

A. Task 1 - Legal Document Retrieval

Similar to this task, several effective approaches on Japanese legal documents were published. Accordingly, Lefoane et al. [1] took a lexical-based approach, they select candidates after ranking them based on tf-idf. Similar to this approach, but the combination of lexical and deep learning has yielded outstanding results for Tran et al. [2], they have demonstrated the effectiveness of deep learning networks in learning the semantics of legal documents. In recent years, the improvement of Transformer models has been widely adopted, which has brought a new approach to the research team, Nguyen et al. [3] has exploited these models well for classifying candidates to select suitable results, which has yielded positive results.

The complexity of the legal questions as well as the expanding space of the candidates have posed many new challenges. To be able to conquer these challenges, the research team of Nguyen et al. [4] improved on their previous models by combining tf-idf and BM-25 to rank candidates to find potential candidates. In addition, the application of the language model and variants of BERT [5] for semantic-based classification has yielded good results for their proposal.

In ALQAC 2021, Aleph [6] is the first research team that has adjusted the BERT language model for the legal question-and-answer system in Vietnamese and has the highest score in this competition. The authors built a candidate ranking model by perfecting their own VNLawBERT model [7] combined with a binary classifier. They perform labeling of negative samples by selecting the candidate closest to the expected patterns. The authors also believe that categorizing questions

into specific legal domains will result in improved reliability for this model.

B. Task 2 - Legal Question Answering

In this section, we will present two approaches to model training methods: true - false or answer extraction. This is also our recommendation for this task.

a) True - False: For retrievals with expected results belonging to the true/false binary classes, processed on Japanese legal texts, some research authors have used a similarity measurement on word-embedded texts [8] or used traditional linear classifiers [8]. However, these methods do not yield highly reliable results when the accuracy is approximately 56.25%. In contrast, most other authors use BERT [9] or variations thereof [10] as semantic text classification, which yields results with greater reliability when the accuracy can be as high as 72.32%.

In the publications on the legal support model of Vietnamese legal system, most of the authors use BERT as the backbone model and they also showed that this model also works well for understanding the semantics of legal documents in Vietnamese. The complexity of the problem as well as the limitation of training data forced the authors to use augment data [11] or take advantage of external data collection from websites [12] for training to improve the learning ability for the model is a necessity.

b) Answer extraction: is a component of the field Machine Reading Comprehension (MRC), the main task of the MRC is to automatically extract answers to questions based on a passage containing the answers. To evaluate the effectiveness of MRC models, Rajpurkar et al. [13] proposed the SQUAD dataset containing more than 100,000 question-answer pairs on more than 500 articles extracted from English Wikipedia. Models trained on this data set must predict the answer to a question based on a given passage. SQUAD has served as the basis for advanced models such as deep learning networks [14] or BERT-based [15] construction models.

The prominence of the SQUAD dataset has spurred the introduction of similar datasets in other languages, such as Chinese [16], French [17], and Korean [18]. Moreover, two data sets in Vietnamese were announced in 2020: UIT-ViQuAD [19] is a multi-field data set extracted from Vietnamese Wikipedia and UIT-ViNewsQA [20] is the other data set in the narrow field of health news. UIT-ViQuAD and UIT-ViNewsQA are the basis for the introduction of ViReader [21], a Vietnamese reading comprehension system using learning transfer.

III. METHOD AND DATASET

In this section, we will present our proposed methods for handling the two tasks of the competition. In parallel, the topic also shows the training data that we will use to train the models.

A. Task 1 - Legal Document Retrieval

In ALQAC 2022, we are provided with a corpus with 4 major Vietnamese Codes, each consisting of Articles of Law. Task 1's primary job is to predict the relevant Articles that will be used as the basis for answering a corresponding question posed. These Articles are often used by lawyers or legal experts in litigation. In this task, we will apply the methods of Nguyen et al. [4] on a Vietnamese legal dataset and build a classifier of questions to predict which Code the answer to that question will belong to in the 4 given Codes, this is the only difference which we will propose in this treatment.

The core problem posed in this task is to correctly predict one or more of the Articles that are relevant to the query based on a given large space of Articles, which was previously easily solved on a small space by simply combining the query with the Articles in pairs and forcing the model to predict which pairs are relevant to each other. However, this work will be difficult with a large space of Articles, since the combination of the query and each Article will create a very large number of pairs. Therefore, the removal of Articles that are completely unrelated to the query will help reduce the load on the predictive model, which was initially proposed by Nguyen et al. [4] handles this by ranking the relevance of the candidates and selecting only those with high rankings.

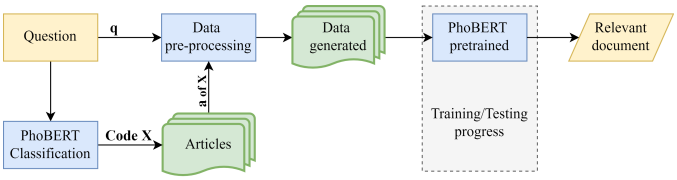

Figure 1. The general architecture of the Legal Document Retrieval system.

Nguyen et al.'s [4] original processing method consists of two main steps. First, the authors generated negative samples based on the data provided by ranking the candidates based on BM-25 combined with the cosine similarity (tf-idf vector) of the candidates with the query sentence. Then, the authors fine-tuned on the BERT pretrained model.

As we can see in Figure 1, we build the classifier with the sole purpose of reducing the space of candidates for each search for candidates related to the problem being retrieved. Based on our analysis of the data provided, most of the ALQAC 2022 queries with Related Articles only fall into a single Code, so we initially experimented with removing Articles from 3 out of 4 Codes are provided.

Accordingly, if we keep the same processing techniques as the method of Nguyen et al. [4] the model will have to search on 1378 candidates according to the problem of ALQAC 2022, but if we apply it before the input question classifier, the model only searches on a maximum of 689 candidates. We expect that this will improve the results for the next processing steps of the model as well as provide a new approach.

To handle this task, we exploited the entire dataset provided for training into our models. Specifically, we used 560 samples of provided data in an 8:2 ratio corresponding to the training

and validation phase for our classifier. In addition, to the stage of data generation and fine-tuned on PhoBERT pretrained we used the data size and scaling for training and validation as shown in Table I.

Table I
TASK 1 DATASET SIZE

Number of	question	potential candidate	sample
Training	3,004	150	450,600
Validation	751	150	12,650

B. Task 2 - Legal Question Answering

In Task 2, the classifier is again used by us as a means to divide the questions into two groups: true-false or answer extraction. We do this for the purpose of breaking down the work so that it is easy to handle them.

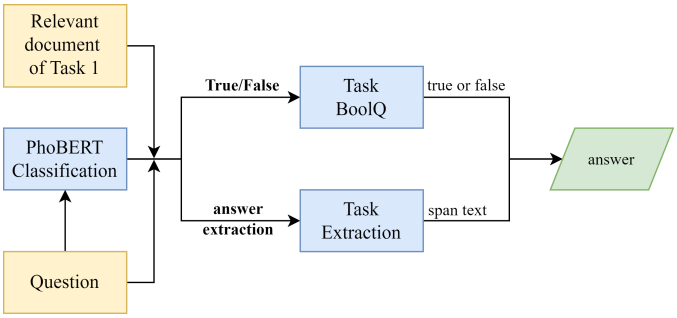

Figure 2. The general architecture of the Legal Question Answering system.

As we can see in Figure 2, we thread each question into the process pipeline, with the first pipeline consisting of questions of true or false form we will build a Bool Question (BoolQ) model to process for this part of the job. The other pipeline we will build a Extraction model to extract from the document that provides the answers. In both tasks, we use legacy document retrieval obtained from Task 1.

Task BoolQ: we use BERT's sentence classification task, which is Question-answering Natural Language Inference, to serve for our purpose to check between the statement in question and the material provided.

Task Extraction: we fine-tuning *bert-base-multilingual-cased-finetuned-squad* model, this is a multilingual (including Vietnamese) model that has been pre-trained on the XQuAD dataset [24] and published by Transformer. This model is built to retrieve answers based on the BERT language transformer.

To build the data set for this task, we combined the data set provided by ALQAC 2022 and extracted further data from the Thu Ky Luat's website [25], which is the website of a prestigious Vietnamese legal consultancy. For the question classifier, we perform the training and validation set size division as in Table II and the data samples are extracted and labeled by us as follows:

a) True - False question: Công ty không chi trả trợ cấp thôi việc, đúng hay sai? (The company does not pay severance pay, true or false?). Label: 1

b) Extraction question: Giữ giấy tờ gốc của người lao động phạt tối đa bao nhiêu tiền? (How much is the maximum fine if the company keeps the original documents of the employee?). Label: 0

Table II
THE SIZE OF THE DATASET FOR THE CLASSIFIER OF TASK 2

Number of	sample	true/false
Training	14,106	7,053
Validation	3,527	1,580

The scarcity of the training dataset for Task BoolQ forced us to find a solution to augment the data for this task. Accordingly, we will use the *VinAI-Translate* API [22] as a tool to convert the BoolQ dataset [23] from English to Vietnamese and we will perform a fine-tuning of a PhoBERT model on this dataset. In addition, we also perform additional data extraction from the Thu Ky Luat's website [25] as well as rule-based labeling to enhance the legal data for the second fine-tuning of the model. The specific size of the BoolQ dataset and the dataset we extracted are shown in Table III and some typical samples we have collected are shown in table IV.

Table III
THE SIZE OF THE DATASET FOR THE TASK BOOLQ

Number of	sample	training	validation
BoolQ dataset	9,427	6,157	3,270
Our dataset	8,633	6,907	1,726

Table IV
AN EXAMPLE OF THE EXTRACTED DATA SAMPLE FOR TASK BOOLQ

Question:	Công ty buộc nhân viên thôi việc khi đang mang thai, đúng hay sai? (The company forced employees to quit while pregnant, right or wrong?)
Passage:	Điều 7 nghị định 54 ngày 19-4-2005 của Chính phủ, người có thai, nghỉ thai sản, nuôi con dưới 12 tháng tuổi thuộc những trường hợp cơ quan, đơn vị không được cho thôi việc (Article 7 of Decree 54 dated April 19, 2005 of the Government, a person who is pregnant, on maternity leave, or raising a child under 12 months old falls under the circumstances in which the agency or unit is not allowed to quit their jobs.)
Label	False

To build a dataset for the model that handles questions in the form of answer extraction, we also combine the data provided with ALQAC 2022 and extract more data from Thu Ky Luat's website. The typical data samples we describe in Table VI and the specific dataset sizes we describe in Table V.

IV. EXPERIMENTS AND RESULTS

We present the experimental results in this section. Measurements including precision, recall, F1 score, F2 score, and accuracy were selected by ALQAC 2022 to evaluate learning models and rank teams.

Table V
THE SIZE OF THE DATASET FOR THE TASK EXTRACTION

Number of	sample	passage	question
Training	1,846	1,075	1,846
Validation	205	183	205

Table VI
AN EXAMPLE OF THE EXTRACTED DATA SAMPLE FOR TASK EXTRACTION

Question:	Ai có thẩm quyền thu hồi mã số REX? (Who has authority to revoke REX tokens?)
Passage:	...Việc thu hồi mã số REX do tổ chức tiếp nhận đăng ký mã số REX thực hiện. Tổ chức tiếp nhận đăng ký mã số REX lưu trữ thông tin thu hồi mã số REX trong vòng 10 năm kể từ ngày kết thúc của năm thu hồi mã số REX.(...The revocation of the REX code shall be carried out by the organization receiving the registration of the REX code. The organization receiving the REX code registration shall store information about the withdrawal of the REX code for 10 years from the end of the year of the REX code withdrawal.)
Answer start:	636
Answer:	tổ chức tiếp nhận đăng ký mã số REX (the organization receiving the registration of the REX code)

A. Task 1 - Legal Document Retrieval

In this task, we will conduct an experiment of the results based on the data set we presented in part III and fine-tuning on the PhoBERT pre-trained model. We have submitted 3 run specifically as follows:

- Run 1: experiment with the original architecture includes two steps of candidate ranking (combination of tf-idf and BM-25) and fine-tuning PhoBERT pre-trained.
- Run 2: experiment with the original architecture combined with the proposed classifier.
- Run 3: similar to Run 2, only the candidate with the highest final score is selected for each question.

The final experimental results are shown in detail in Table VII. According to the results table, the proposed classifier combined with the selection of the candidate with the highest final score achieved better results on the precision, F1 and F2 score scales. The final result of our method and the method of another team is having the same score on all 4 evaluation scales and leading the Task 1 ranking of ALQAC 2022.

Table VII
RESULT ON TASK 1

Run ID	Precision	Recall	F1 score	F2 score
origin	0.9333	0.9667	0.9444	0.9556
origin-classify	0.9333	0.9667	0.9444	0.9556
origin-classify-highest	**0.9667**	0.9667	**0.9667**	**0.9667**

B. Task 2 - Legal Question Answering

For Task 2, we train the models we use in turn on the dataset we propose. Once again we want to be more clear, in Task BoolQ we perform 2 steps of fine-tuning the model, the first step we fine-tuning the PhoBERT model on the BoolQ dataset [23] that has been translated into Vietnamese. Finally, we just fine-tuning on the dataset we extracted.

In addition to using the *VinAI-Translate* API [22] as a translation tool, we also use the Google Translate API to match the results. To observe the results, we also intentionally pressured Task Extraction to work harder by only grouping questions containing the final phrase as "đúng hay sai? (true or false?)" for Task BoolQ. The rest of the questions are similar in nature, such as those containing the final phrase "... đúng không? (... right?)", we expect Task Extraction to also handle them well.

For this task, we submitted 3 runs, and all 3 of these runs used the following backbone networks: PhoBERT classifiers of text without context (for question classifiers), PhoBERT classifiers text with context (for Task BoolQ) and BERT Question Answering (for Task Extraction):

- Run 1: use Google translate API and divide more work items for Task Extraction.
- Run 2: use Google translate API and divide work items evenly.
- Run 3: use Vin-AI translate API and divide more work items for Task Extraction.

As we can see in Table VII, there is no difference between Run 1 and Run 2 on the accuracy scale, even Run 1 has a higher exact match measurement than Run 2, which shows that the BERT Question Answering model is also capable of handling a small number of questions of the form bool question. In addition, we can realize that although it is not significant, google translate service still helps our model to have better learning results. The final result our method won in Task 2.

Table VIII
RESULT ON TASK 2

Run ID	Exact match	Precision	Recall	F1	Accuracy
Run 1	**0.4667**	0.7516	0.6805	0.6786	**0.6333**
Run 2	0.4000	0.7730	0.6497	0.6632	**0.6333**
Run 3	0.4000	0.7730	0.6497	0.6632	0.6000

V. CONCLUSION

In this paper, we tackled the problems posed by ALQAC 2022, using BERT and its variants as a backbone network. By the way, we also propose new approaches to training data shortage problems. We realized that a good application of training data augmentation techniques will help our methods improve reliability. In addition, the research to build a multitasking model that can handle many problems at the same time instead of dividing the jobs as we propose is also a new research direction for problems in the same field. These are necessary work and we will continue to experiment.

REFERENCES

[1] M. Lefoane, T. Koboyatshwene and L. Narasimhan, "KNN Clustering Approach to Legal Precedence Retrieval", Twelfth International Workshop on Juris-informatics (JURISIN), 2018 (System id: UBIRLED)

[2] V. Tran, S.T. Nguyen and M.L. Nguyen, "JNLP Group: Legal Information Retrieval with Summary and Logical Structure Analysis", Twelfth International Workshop on Juris-informatics (JURISIN), 2018 (System id: JNLP)

[3] H.T. Nguyen, H.Y.T. Vuong, P.M. Nguyen, T.D. Binh, Q.M. Bui, S.T. Vu, C.M. Nguyen, V. Tran, K. Satoh and M. L. Nguyen. 2020. JNLP Team: Deep Learning for Legal Processing in COLIEE 2020. arXiv preprint arXiv:2011.08071 (2020).

[4] H.T. Nguyen, P. M. Nguyen, T.H.Y. Vuong, Q.M. Bui, C.M. Nguyen, B.T. Dang, V. Tran, M.L. Nguyen and K. Satoh (2021). JNLP team: Deep learning approaches for legal processing tasks in coliee 2021. In: Proceedings of the COLIEE Workshop in ICAIL.

[5] J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[6] Q.H. Ngo, M.D.T. Nguyen, A.D. Nguyen, and Q.N.M. Pham. Aimelaw at alqac 2021: Enriching neural network models with legal-domain knowledge. Proceedings of the 1st Automated Legal Question Answering Competition. ALQAC'2021, 2021.

[7] C.C. Nguyen, S.T. Nguyen and M.L. Nguyen. VNLAWBERT: A Vietnamese legal answer selection approach using BERT language model. In 2020 7th NAFOSTED Conference on Information and Computer Science (NICS), pages 298–301. IEEE, 2020.

[8] Wehnert, S., Murugadas, V., Nandakumar, S., Saha, A., Khan, T.R. Urban, M. Luca, E.W.D.: Legal information retrieval and entailment detection: Hybrid approaches of traditional machine learning and deep learning. In: COLIEE (2020)

[9] H.T. Nguyen, H.Y.T. Vuong, P.M. Nguyen, B.T. Dang, Q.M. Bui, S.T. Vu, C.M. Nguyen, V. Tran, K. Satoh, M.L. Nguyen: JNLP team: Deep learning for legal processing in COLIEE 2020. In: COLIEE (2020)

[10] J. Rabelo, M.Y. Kim, R. Goebel: Application of text entailment techniques in COLIEE 2020. In: COLIEE (2020)

[11] Q.H. Ngo, M.D.T. Nguyen, A.D. Nguyen, and Q.N.M. Pham. Aimelaw at alqac 2021: Enriching neural network models with legal-domain knowledge. Proceedings of the 1st Automated Legal Question Answering Competition. ALQAC'2021, 2021

[12] T.T. Truong, C.C. Nguyen, N.M.H Bui, T.S. Nguyen and L.M. Nguyen. Apply bert-based models and domain knowledge for for automated legal question answering tasks at alqac 2021. Proceedings of the 1st Automated Legal Question Answering Competition. ALQAC'2021, 2021.

[13] P. Rajpurkar, J. Zhang, K. Lopyrev and P. Liang, SQuAD: 100,000+ Questions for Machine Comprehension of Text, Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016.

[14] A.W. Yu, D. Dohan, M.T. Luong, R. Zhao, K. Chen, 1014 M. Norouzi and Q.V. Le, QANet: Combining Local 1015 Convolution with Global Self-Attention for Reading 1016 Comprehension, International Conference on Learning 1017 Representations (2018).

[15] J. Devlin, M.W. Chang, K. Lee and K. Toutanova, BERT: 1019 Pretraining of Deep Bidirectional Transformers for Lan- 1020 guage Understanding, Proceedings of the 2019 Conference 1021 of the North American Chapter of the Association for 1022 Computational Linguistics: Human Language Technolo- 1023 gies, Volume 1 (Long and Short Papers), (2019).

[16] Y. Cui, T. Liu, W. Che, L. Xiao, Z. Chen, W. Ma, S. Wang and G. Hu, A Span-Extraction Dataset for Chinese Machine Reading Comprehension, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), (2019).

[17] M. d'Hoffschmidt, W. Belblidia, T. Brendle, Q. Heinrich and M. Vidal, FQuAD: French Question Answering Dataset, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings (2020).

[18] S. Lim, M. Kim and J. Lee, Korquad1.0: Korean qa dataset for machine reading comprehension, arXiv preprint arXiv:1909.07005, (2019)

[19] K. Nguyen, V. Nguyen, A. Nguyen and N. Nguyen, 1007 A Vietnamese Dataset for Evaluating Machine Reading 1008 Comprehension, Proceedings of the 28th International Con- 1009 ference on Computational Linguistics (COLING), (2020).

[20] K. Nguyen, T. Huynh, D.V. Nguyen, A.G.T. Nguyen and 1044 N.L.T. Nguyen, New Vietnamese Corpus for Machine Read- 1045 ing Comprehension of Health News Articles, arXiv preprint 1046 arXiv:2006.11138 (2020).

[21] K. Nguyen, Nh. D. Nguyen, P.N.T. Do, A.G.T. Nguyen, and N.L.T. Nguyen. 2021. Vireader: A wikipedia-basedvietnamese reading comprehension system usingtransfer learning. Journal of Intelligent FuzzySystems, pages 1–19.

[22] T.H. Nguyen, T.D.H. Nguyen, D. Phung, T.C.D. Nguyen, M.H. Tran, M. Luong, T.D. Vo, H.H. Bui, D. Phung and Q.D Nguyen. A Vietnamese-English Neural Machine Translation System. Proceedings of the 23rd Annual Conference of the International Speech Communication Association: Show and Tell (INTERSPEECH), 2022.

[23] Clark, Christopher and Lee, Kenton and Chang, Ming-Wei, and Kwiatkowski, Tom and Collins, Michael, and Toutanova, Kristina. BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions. NAACL, 2019.

[24] M. Artetxe and S. Ruder and D. Yogatam, On the cross-lingual transferability of monolingual representations, CoRR, abs/1910.11856, arXiv: 1910.11856 (2019)

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