BUILDING A QUESTION-ANSWER DATASET FOR VIETNAMESE PUBLIC ADMINISTRATIVE DOCUMENTS

Dinh-Dien La[▫], **Thi-Thanh Ha**[†], **Van-Khanh Tran**[†], **Trung-Nghia Phung**[†]
[▫] Department of Information and Communications, Ha Giang, Vietnam
[†] Thai Nguyen University of Information and Communication

***Tóm tắt*—The development of effective Chatbot for legal domains poses significant challenges due to the complexity, ambiguity, and specialized language inherent in legal texts. This paper introduces a comprehensive Question-Answer (QA) dataset specifically designed for Vietnamese public administrative documents. This dataset aims to serve as a standardized resource for fine-tuning deep learning models tailored for legal Chatbot. The primary goal is to enhance the Chatbots' capability to accurately address citizen inquiries regarding procedures in online public services. The dataset was constructed through a meticulous process involving the collection, preprocessing, and annotation of public administrative documents. We ensured a broad coverage of topics relevant to public services and crafted questions that reflect real queries. The dataset consists of 11,536 question-answer pairs divided into 11,334 pairs for the training set and 1,202 pairs for the test set. Our dataset contributes to the advancement of AI-driven public service solutions in Vietnam, providing a valuable resource for the research community to develop and refine legal Chatbot.**

***Từ khóa*—Vietnamese QA dataset, Legal Vietnamese dataset, Public service online.**

I. INTRODUCTION

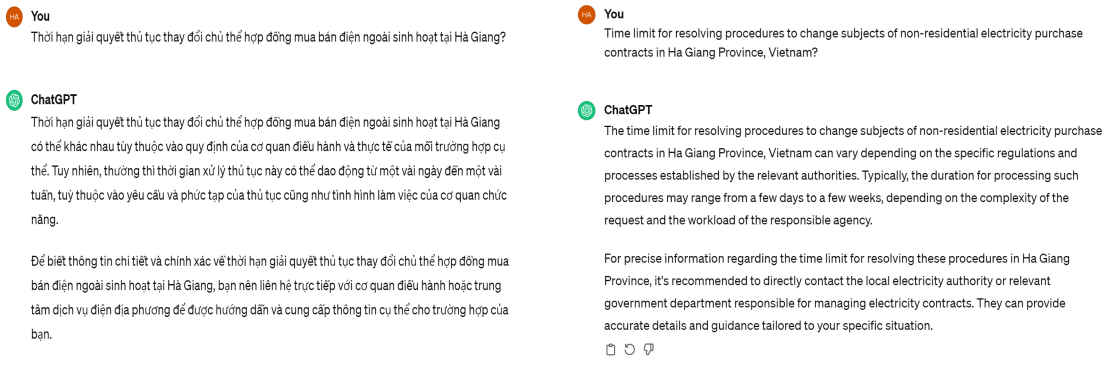
The issue of answering questions within the realm of online public services is akin to the challenges encountered in legal question-and-answer systems. Users engaging in administrative procedures through online public services often encounter numerous inquiries necessitating prompt responses. Consequently, there arises a need to establish a highly accurate question-and-answer system within the public service domain, enabling users to efficiently complete documents in a timely manner. Chatbot have increasingly become a vital tool in various domains, offering automated customer support, information retrieval, and user interaction. Specifically, in the context of Vietnamese public administrative documents, Chatbot can significantly enhance accessibility and efficiency by providing instant answers to citizens' inquiries regarding administrative procedures. However, the effectiveness of such Chatbot heavily depends on the quality and comprehensiveness of their underlying datasets.

Despite the advanced capabilities of large language models (LLMs) like ChatGPT, there remain significant limitations when applied to Vietnamese public administrative documents. These models often struggle with the specific terminology, context, and procedural nuances inherent in legal and administrative language. The lack of domain-specific data further exacerbates these challenges, resulting in less accurate and reliable responses from the Chatbot. The issue of question and answer concerning administrative procedures on the public service portal intersects with the legal domain. However, our survey indicates a lack of current research in this area. Additionally, existing question-answering systems, such as ChatGPT, yield unexpected outcomes when posed with inquiries within online public service domains. Figure.1 shows an example of a question about administrative procedures on online public services using chatGPT3.5. ChatGPT does not give the exact answer while the correct answer is "no more than 3 days from the date the customer applies"

Researchers have utilized machine learning and advanced NLP models to build QA systems [1], [2], [3], [4], [5], [6]. Existing legal datasets, while valuable, often fall short in several key areas. Many are not tailored to the specific needs of Vietnamese public administrative procedures and lack the necessary diversity to cover the full spectrum of potential user queries. Additionally, these datasets frequently miss out on the contextual subtleties and procedural complexities that are crucial for accurate information retrieval and user assistance.

Legal question-answering systems face a myriad of challenges. Firstly, there's Semantic Complexity: Legal documents are replete with intricate language structures, specialized terminology, and nuanced meanings, posing a significant challenge for NLP systems to accurately comprehend. Secondly, Ambiguity is pervasive: Legal texts often employ language open to multiple interpretations, demanding a nuanced understanding of context, which presents a hurdle for

Contact author: Thi-Thanh Ha,
Email: htthanh@ictu.edu.vn
Manuscript received: 8/2024, revised: 9/2024, accepted: 10/2024.



(a) in Vietnamese (b) in English

Hình 1: An example of a question about administrative procedures on online public services using ChatGPT 3.5 in both Vietnamese and English

NLP models. Thirdly, Variability in Legal Language adds complexity: Legal terminology, conventions, and styles vary across jurisdictions, periods, and individual documents, requiring NLP systems to robustly adapt to diverse legal contexts. Moreover, Large Document Sizes are a concern: Legal documents, including court cases and statutes, can be voluminous and contain extraneous information, necessitating NLP systems to efficiently sift through and extract relevant data for precise answers. Additionally, the Need for Domain Expertise is crucial: Legal inquiries often demand specialized knowledge, making it challenging for NLP systems to provide accurate responses without access to comprehensive legal databases and expert guidance. Furthermore, Inference and Reasoning are paramount: Legal questions frequently involve intricate reasoning based on precedent and case law, necessitating NLP systems to possess advanced logical reasoning capabilities beyond simple information retrieval. Lastly, Data Privacy and Security are critical considerations: Legal documents often contain sensitive information, underscoring the importance of robust data privacy measures in NLP systems to safeguard confidentiality. Overcoming these challenges requires continual research and development in NLP, focusing on enhancing natural language understanding, domain adaptation techniques, and crafting tailored knowledge representation models for legal texts. For this reason, current question-answering systems often have low accuracy. Sota in the English ILDC dataset, macro F1 is 77.8%, accuracy is 77.7% [7]. For Vietnamese legal domain data, the legal Textual Entailment Recognition in VLSP 2023 problem, the highest accuracy is 70%.

Building and curating legal datasets presents unique challenges. These include the need for meticulous data annotation, the difficulty in capturing the wide range of potential queries, and ensuring the data remains up-to-date with current laws and regulations. Moreover, the diversity of administrative procedures and the varying complexity of user inquiries add layers of complexity to the dataset creation process. Addressing these challenges is essential to improve the performance of Chatbot on Vietnamese public administrative documents. A well-constructed dataset can enhance the Chatbot's ability to provide accurate, relevant, and contextually appropriate responses, thereby improving user satisfaction and trust in automated public service tools.

This research aims to build a comprehensive Question -Answer dataset named VLPSO (Vietnamese Legal in Public Service Online) specifically designed for Chatbot dealing with Vietnamese public administrative documents. The dataset includes both a training set and a test set, characterized by its diversity across all areas of administrative procedures. It encompasses a wide range of questions, from simple to complex, covering various aspects of public administrative procedures.

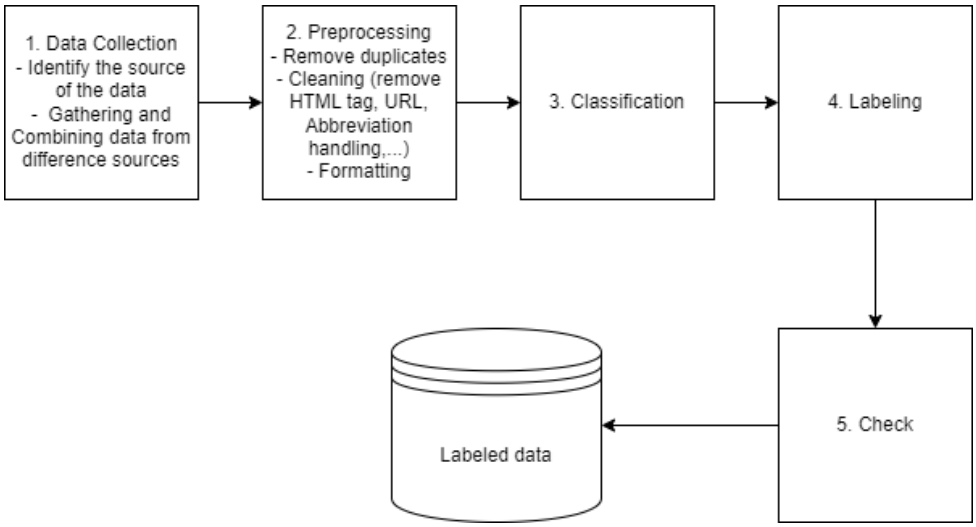
II. Related Works

In this section, we reviewed a few published legal domain datasets in English and other languages:

1) PrivacyQA: The dataset described in the EMNLP 2019 paper titled "Question Answering for Privacy Policies: Combining Computational and legal Perspectives"[8] is stored in this repository. Named PrivacyQA, this corpus comprises 1750 questions related to the contents of privacy policies, each accompanied by expert annotations. The primary objective of this initiative is to initiate the advancement of question-answering techniques within this domain, aiming to mitigate the impractical expectation that a significant portion of the population should be regularly reading numerous policies daily.

The data has been partitioned into a train and test set:

- The train set has 1,350 queries with an average length of 8.42 words and 185,200 segments (sen-



Hình 2: Process of building the VLPSO dataset

tences) with an average length of 22.72 words. Train- Label : Relevant, Irrelevant

- The test set has 400 queries with an average length of 8.56 words and 62150 segments with an average length of 23.14 words. Test- Label: Relevant, Irrelevant, None
- Relevant: Segment is relevant for query. Irrelevant: segment is irrelevant for query.

2) JEC-QA: The JEC-QA dataset[9] is compiled from the National Judicial Examination of China, comprising a total of 26,365 multiple-choice and multiple-answer questions. This dataset includes 26,365 questions; 105,460 options; 79,433 paragraphs. Its purpose is to anticipate responses based on the questions and pertinent articles provided. Effective performance on the JEC-QA necessitates proficiency in both retrieving information and providing accurate answers.

3) BSARD: The Belgian Legal Question Answering Dataset[10] (BSARD) comprises over 1,100 French native legal questions meticulously labeled by experienced jurists, alongside relevant articles sourced from a corpus of over 22,600 Belgian law articles in French.

This dataset consists 22,633 articles with an average length of 136.67 words. The training set consists of 886 questions with an average length of 14.95 words. The test set includes 222 questions with an average length of 15.84 words. The dataset encompasses 7 main categories and 50 subcategories.

4) CUAD: : The Contract Understanding Atticus Dataset (CUAD) [11] encompasses more than 13,000 labels derived from 510 commercial legal contracts. These labels were meticulously assigned with the guidance of experienced lawyers to discern 41 distinct types of legal clauses vital for scrutinizing contracts in diverse corporate dealings, including mergers and acquisitions. The dataset consists of 510 contracts and over 13,000 expert annotations across 41 label categories, distributed as follows: The training set comprises 408 contracts with 11,180 annotations; The test set comprises 102 contracts with 2,643 annotations.

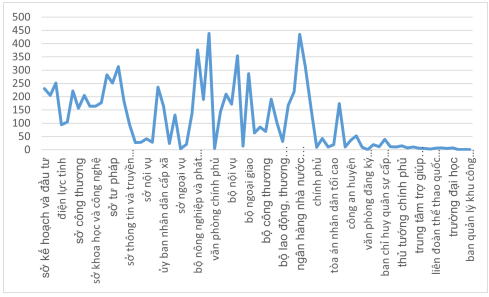
5) VLSP-LTER: The research investigates Vietnamese legal inquiries aimed at developing an automated question answering system, specifically concentrating on transportation law. It introduces a sequence labeling approach, and empirical findings demonstrate its capability to extract 18 categories of information with notable precision and recall. The analysis is based on a corpus comprising 1678 Vietnamese questions. Train set consists 76 statements with average length of 29.5 words; Test set has 140 statements with average length of 26.51 words and Legal passages includes 18 documents with 2256 articles with average length of 174.5 words

Datasets have the following limitations: Size constraints pose another challenge, with smaller datasets affecting the generalizability and performance of machine learning models. Furthermore, the domain specificity of certain datasets may limit their utility for researchers exploring different legal fields or tasks. These data sets are not yet available in the database of administrative procedure documents of online public services. Finally, the resource-intensive nature of working with forensic datasets, which requires domain expertise and computational resources, poses challenges for researchers who do not have access to that resource.

III. VLPSO: Vietnamese Legal in Public Service Online dataset

A. Steps to build the data set

In this section, we presented the process of building the VLPSO dataset (**Fig**.2):



Hình 3: The chart illustrates the number of questions-answers by group of public service providers

Step1 - Data collection: We collected data from two sources: First, we collected pairs of question-answer data from the national online public service site[1]. The second source is switchboard 1022 and on the online public service portal[2] of Ha Giang province. Furthermore, We also collected questions - answers pairs related to common public services from other provinces such as Bac Giang, Bac Ninh, Quang Ninh.

Step2 - Preprocessing: After completing step 1, we gathered two distinct sets of data. The first set comprises question-answer pairs extracted from the online public service portal website. The second set consists solely of questions sourced from public service websites across various provinces, obtained via phone inquiries. Both datasets underwent preprocessing steps, including the removal of duplicate questions, elimination of HTML tags and URLs, standardization of abbreviations, and conjoined words BeautifulSoup and NLTK tool (for example: "cap chung thu **socho** nguoi" to "cap chung thu **so cho** nguoi", English means Digital identity cards for people). Following this step, the initial dataset comprises 9,452 question-answer pairs about 4,591 administrative procedures accessible through the national public service portal. The dataset comprises various attributes, encompassing question, answer, associated administrative procedure, procedure code, related question, and the responsible agency. The second dataset consists of 3,833 question-answer pairs. This second dataset do not have information about related administrative procedures or any other information of questions. After this step, the total number of question-answer pairs from the two data sources is 13,285.

Step 3 - Classification: At this stage, questions are categorized based on the agency responsible for handling administrative procedures in the second dataset. The purpose of this step is that questions falling under the jurisdiction of any agency will be transferred to that agency to label, review and answer the corresponding questions.

[1]https://dichvucong.gov.vn/p/home/dvc-cau-hoi-pho-bien.html
[2]https://1022.hagiang.gov.vn/vi/phan-anh-kien-nghi/gui-phan-anh

Bảng I: Statistics about the question-answer pairs of VLPSO dataset

	Train	Test	All
Question-answer pairs	11,334	1,202	11,536
Average Question length in words	30	36	30.1
Average Answer length in words	174	131	173.2
Vocab size of Words	8792	2884	9152
Min words	6	8	6
Max words	5373	1108	5373

To categorized questions, we first standardized labels in first dataset. For instance, "state bank of vietnam" is normalized to "State Bank of Vietnam." Subsequently, similar labels are consolidated under a single label.

Following this grouping process, the number of agency labels were reduced from 1823 to 70. This consolidation aids in reviewing the content of both questions and answers in the subsequent phase. Questions and answers are then forwarded to the respective competent authorities for verification, ensuring the accuracy of responses to corresponding questions.

Next, a Support Vector Machine (SVC) model was used to classify 3,833 questions in the second dataset with labels. Subsequently, these questions send to a group of experts from provincial departments to review thoroughly. The SVC model parameters are set as follows: C=0.5, kernel='linear', max_iter=5000.

Questions were classified according to processing agencies, they are forwarded to the respective authorities for further review. Questions with provided answers were reviewed, while questions that do not have answers were moved to step 4.

Step 4 - Labeling answers: Professionals from departments undertakeed the review of questions and answers within the first dataset while concurrently addressing inquiries within the second dataset. During this phase, we utilized the Labelstudio tool to label. In the case of the second dataset, experts not only answer to queries but also assign labels to associated administrative procedures and procedure codes

Step 5 - Cross review: Each expert performed the annotation independently in step 4. Following this, the experts cross-checked their annotations. As a result, we identified 87% duplicate labels. We then removed question-answer pairs with inconsistent labels. After this process, we obtained 11,536 question-answer pairs with related administrative procedures.

B. Statistics

Following cross review, we acquired a grand total of 11,536 question-answer pairs from a pool of 14,942 administrative procedures associated with national and local public services across Ha Giang, Bac Giang, Bac Ninh, and Quang Ninh provinces. These questions consisted of real inquiries encountered throughout admin-

Hình 4: Results on the recall@k measure with two models BM25 and tf.idf

istrative proceedings, collected from various channels. Additionally, this dataset encompasses supplementary attributes, including the responsible agency, pertinent administrative procedures, and administrative procedure codes.

Finally, we divided the dataset into 2 sets: training set and test set. We manually selected 202 QnA samples to serve as the test set from the question-answer pairs of Ha Giang province. Question types of the dataset are in seven categories, including Reasoning Questions, Factoid Questions, Yes/No Questions, Multiple-choice Questions, and Questions involving multiple agencies, multiple relevant documents, and multiple relevant articles.

The statistics of the training (Train) and test (Test) sets of our dataset are described in table I. The average question length in words was 30.1, indicating that the questions were quite long. On the other hand, the average length of answers was significantly longer than questions in word (173.2). it means that anwers were typically detailed and long. Overall, this dataset provides a long and detailed source of information (Figure 3).

IV. Implemented Baselines

In this paper, we re-implemented the baseline models [12] on our dataset:

- Elastic Search (BM25): We used ElasticSearch[3] with BM25 as the similarity measure[12]
- Elastic Search(if.idf): Using TF-IDF instead of BM25, the current approach mirrors the previous one.
- Use cosine similarity measure on average word-embeddings using PhoBert[13]

Performance of baseline models were showed in Figure 4. The provided results present the call@k scores for the retrieval models BM25 and TF.IDF on different values of k, giving an assessment of their performance. In general, BM25 consistently outperforms TF.IDF

[3]https://elasticsearch.co/

Bảng II: Evaluation of BM25 with different top-k values

Method (top-k)	Precision	F2 score
BM25 (k=1)	0.6386	0.6386
BM25 (k=2)	0.542	0.6641
BM25 (k=3)	0.4991	0.6847
BM25 (k=5)	0.4624	0.6902
BM25 (k=10)	0.3809	0.6716

on all k values, demonstrating its superior ability to retrieve relevant documents. Even at lower k values, when retrieval is limited, BM25 maintains a higher recovery rate than TF.IDF. Overall, these findings highlight the overall effectiveness of BM25 compared to TF.IDF in document retrieval. The effectiveness of the BM25 model above, We evaluated the search task with Precision, recall and F2 measures (see table II)

BM25 outperforms TF.IDF due to its normalization of term frequencies and document lengths, scalability for large document collections, optimization for retrieval tasks, and consideration of document statistics. By normalizing term frequencies and document lengths, BM25 prevents biases towards longer documents and provides more robust relevance scoring. Additionally, its scalability and optimization for retrieval tasks make it well-suited for information retrieval in diverse document collections. Furthermore, BM25's consideration of document statistics allows for more nuanced relevance scoring, contributing to its superior performance over TF.IDF in many retrieval scenarios.

The relatively lower performance of **Cosine +PhoBERT** compared to BM25 and TF-IDF can be attributed to PhoBERT was not fine-tuned on the legal domain. This lack of domain-specific fine-tuning limits its effectiveness in handling the specialized language and nuances of legal texts, resulting in lower recall scores.

Next, we also evaluated the dataset on LLMs. The table III presents the results of answer generation on a test set, evaluated using BLEU, ROUGE-1, ROUGE-2, and ROUGE-L metrics. It compares the performance of ChatGPT3.5, Vistral without fine-tuning, and Vistral with fine-tuning on train set. ChatGPT3.5 achieves a BLEU score of 0.1221, a ROUGE-1 score of 0.3385, a ROUGE-2 score of 0.1298, and a ROUGE-L score of 0.2936. Vistral without fine-tuning shows slightly lower results with a BLEU score of 0.1105 and a ROUGE-L score of 0.2590. However, when fine-tuned, Vistral's performance improves significantly across all metrics, with the highest BLEU score of 0.1675, ROUGE-1 at 0.4269, ROUGE-2 at 0.2252, and ROUGE-L at 0.3590. This highlights the effectiveness of fine-tuning for better answer generation.

Bảng III: Results of Answer Generation on test set on BLEU, ROURGE-1, ROUGE-2, ROUGE-L

	Bleu	Rouge1	Rouge2	RougeL
ChatGPT3.5	0.122	0.339	0.130	0.294
Vistral not fine-tuning	0.111	0.311	0.115	0.259
Vistral+fine-tuning	0.168	0.427	0.225	0.359

V. CONCLUSIONS

In summary, we have built a legal dataset in the field of online public services, especially focusing on administrative procedure documents. The creation of this dataset is aimed at improving the efficiency of the Q&A system in this field, specifically to fine-tune LLM models in the data domain related to public administrative documents. In the future, we will further analyze the question-answer dataset. Besides, we will use this data set to evaluate the performance of the chatbot system on this data domain. Next, this dataset is also used as a data warehouse to apply RAG techniques on the Chatbot system

ACCKNOWLEGDE

This work was funded by basic level scientific research topic under project code T2024-07-05.

TÀI LIỆU THAM KHẢO

[1] H.-T. Duong and B.-Q. Ho, "A vietnamese question answering system in vietnam's legal documents," in *Computer Information Systems and Industrial Management*, K. Saeed and V. Snášel, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, pp. 186–197.

[2] R. Taniguchi and Y. Kano, "Legal yes/no question answering system using case-role analysis," in *New Frontiers in Artificial Intelligence*, S. Kurahashi, Y. Ohta, S. Arai, K. Satoh, and D. Bekki, Eds. Cham: Springer International Publishing, 2017, pp. 284–298.

[3] M.-Y. Kim, Y. Xu, and R. Goebel, "Applying a convolutional neural network to legal question answering," in *New Frontiers in Artificial Intelligence*, M. Otake, S. Kurahashi, Y. Ota, K. Satoh, and D. Bekki, Eds. Cham: Springer International Publishing, 2017, pp. 282–294.

[4] M.-Y. Kim, Y. Xu, Y. Lu, and R. Goebel, "Question answering of bar exams by paraphrasing and legal text analysis," in *New Frontiers in Artificial Intelligence*, S. Kurahashi, Y. Ohta, S. Arai, K. Satoh, and D. Bekki, Eds. Cham: Springer International Publishing, 2017, pp. 299–313.

[5] M.-Y. Kim, Y. Lu, and R. Goebel, "Textual entailment in legal bar exam question answering using deep siamese networks," in *New Frontiers in Artificial Intelligence*, S. Arai, K. Kojima, K. Mineshima, D. Bekki, K. Satoh, and Y. Ohta, Eds. Cham: Springer International Publishing, 2018, pp. 35–48.

[6] G. McElvain, G. Sanchez, D. Teo, and T. Custis, "Non-factoid question answering in the legal domain," in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR'19. New York, NY, USA: Association for Computing Machinery, 2019, p. 1395–1396. [Online]. Available: https://doi.org/10.1145/3331184.3331431

[7] A. Abdallah, B. Piryani, and A. Jatowt, "Exploring the state of the art in legal qa systems," 2023.

[8] A. Ravichander, A. W. Black, S. Wilson, T. B. Norton, and N. M. Sadeh, "Question answering for privacy policies: Combining computational and legal perspectives," *CoRR*, vol. abs/1911.00841, 2019. [Online]. Available: http://arxiv.org/abs/1911.00841

[9] H. Zhong, C. Xiao, C. Tu, T. Zhang, Z. Liu, and M. Sun, "Jec-qa: A legal-domain question answering dataset," in *Proceedings of AAAI*, 2020.

[10] A. Louis, G. Spanakis, and G. van Dijck, "A statutory article retrieval dataset in french," *CoRR*, vol. abs/2108.11792, 2021. [Online]. Available: https://arxiv.org/abs/2108.11792

[11] D. Hendrycks, C. Burns, A. Chen, and S. Ball, "CUAD: an expert-annotated NLP dataset for legal contract review," *CoRR*, vol. abs/2103.06268, 2021. [Online]. Available: https://arxiv.org/abs/2103.06268

[12] P. M. Kien, H.-T. Nguyen, N. X. Bach, V. Tran, M. L. Nguyen, and T. M. Phuong, "Answering legal questions by learning neural attentive text representation," in *Proceedings of the 28th International Conference on Computational Linguistics*, D. Scott, N. Bel, and C. Zong, Eds. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 988–998. [Online]. Available: https://aclanthology.org/2020.coling-main.86

[13] D. Q. Nguyen and A. T. Nguyen, "Phobert: Pre-trained language models for vietnamese," *CoRR*, vol. abs/2003.00744, 2020. [Online]. Available: https://arxiv.org/abs/2003.00744

XÂY DỰNG BỘ DỮ LIỆU HỎI ĐÁP TRÊN MIỀN CÁC VĂN BẢN HÀNH CHÍNH CÔNG VIỆT NAM

***Tóm tắt*:** Chatbot hỏi đáp về lĩnh vực pháp lý đặt ra nhiều thách thức do tính phức tạp trong lĩnh vực này. Trong bài báo này, chúng tôi trình bày tập dữ liệu Hỏi - Đáp (QA) được thiết kế riêng cho các văn bản hành chính công của Việt Nam. Tập dữ liệu này được sử dụng để tinh chỉnh các mô hình học sâu, đặc biệt là mô hình ngôn ngữ lớn LLM được thiết kế riêng cho Chatbot hỏi đáp trên miền dịch vụ công trực tuyến. Mục tiêu chính là nâng cao khả năng của Chatbot trong việc giải quyết chính xác các thắc mắc của công dân liên quan đến các thủ tục trong các dịch vụ công trực tuyến. Tập dữ liệu được xây dựng thông qua một quy trình tỉ mỉ bao gồm thu thập, xử lý trước và chú thích các tài liệu hành chính công. Chúng tôi đảm bảo phạm vi bao phủ rộng rãi các chủ đề có liên quan đến các dịch vụ công và các câu hỏi được thiết kế phản ánh các vấn đề quan tâm thực tế. Tập dữ liệu bao gồm 11.536 cặp câu hỏi-trả lời được chia thành 11.334 cặp cho tập huấn luyện và 202 cặp cho bộ kiểm tra. Tập dữ liệu của chúng tôi góp phần vào sự phát triển của các giải pháp dịch vụ công do AI thúc đẩy tại Việt Nam, cung cấp một nguồn tài nguyên có giá trị cho cộng đồng nghiên cứu để phát triển và tinh chỉnh Chatbot pháp lý.

***Từ khóa*:** Tập dữ liệu Hỏi đáp tiếng Việt, Tập dữ liệu Hỏi đáp trong lĩnh vực pháp lý, dịch vụ công trực tuyến.

Dinh-Dien La is a PhD student majoring in computer science, University of Information and Communications Technology, Thai Nguyen University. He is currently Deputy Director of the Department of Information and Communications of Ha Giang province, in charge of digital transformation. His research interests are data science, machine learning and deep learning in the domain of law, public administration. Email: ladien.it@gmail.com

Thi-Thanh Ha is received PhD degree in Information System at Ha Noi University of Science and Technology, Viet Nam. She obtained Bachelor Degree of Science in Applied Mathematics and Informatics from University of Natural Science, Vietnam National University in 2004. She now is a lecturer at Computer Science in Thai Nguyen University of Information and Communication Technology. Her researches are fields of deep learning in Natural Language Processing, Question Answering system and Chatbot.

Van-Khanh Tran received Ph.D. in Natural Language Processing from the Japan Advanced Institute of Science and Technology (JAIST), where his research focused on deep learning for natural language generation in spoken dialogue systems. He also holds a Master's degree in Information Technology from Manuel S. Enverga University Foundation, Philippines, and a Bachelor's degree in Computer Science from the University of Information and Communication Technology in Vietnam. He has held research positions at various institutions, including the Applied Artificial Intelligence Institute at Deakin University, Australia, and VinBigdata's Virtual Assistant Center, where he contributed to the development of NLP applications. He is currently an AI Research Scientist on the NLP team at FPT Smart Cloud's Generative AI (GenAI) Center, where he focuses on developing large language models and AI assistant ecosystems tailored for Vietnamese users. He also serves as the Deputy Head of the Institute of Applied Science and Technology. His research interests include natural language processing, large language models, and AI applications in the legal, healthcare, and finance domains.

Assoc. Prof. Trung-Nghia Phung received his Engineering degree in Electronics and Telecommunications from Hanoi University of Science and Technology (HUST) in 2002. He completed his Master of Science degree in Telecommunications from Vietnam National University –Hanoi (VNUH) in 2007 and his PhD degree in Information Science from Japan Advanced Institute of Science and Technology (JAIST) in 2013. He was Dean of Faculty of Electronics and Telecommunications, Head of Academic Affairs, and he has been Rector of Thai Nguyen University of Information and Communication Technology (ICTU). He has been a Vice President of Vietnam Club of Faculties-Institutes-Schools-Universities of ICT (FISU) and President of FISU Branch in the Northern Midlands, Mountains and Coastal Region of Vietnam. He was the recipient of the award for the excellent young researcher (Gloden Globe award) from Ministry of Science and Technology (MOST) of Vietnam in 2008. His main research interest lies in the field of interaction between signal processing and machine learning and he has published more than 70 research papers related to this field. He serves as a technical committee program member, organizing co-chair, program co-chair, track chair, section chair, editorial board member and reviewer of several conferences, journals and books. He is now an associate editor of Thai Nguyen University Journal of Science and Technology (ICT section).