VIETNAM NATIONALUNIVERSITY, HANOI
VNUUNIVERSITY OF ENGINEERING AND TECHNOLOGY

SOCIALISTREPUBLIC OF VIETNAM
Independence – Freedom – Happiness

INFORMATION ON DOCTORAL THESIS

1. Full name : Vuong Thi Hai Yen 2. Sex: Female
3. Date of birth: 21/08/1994 4. Place of birth: Hanoi
5. Admission decision number:776/QĐ-CTSV Dated 31/07/2019
6. Changes in academic process:
146/QĐ-ĐT Dated 11/03/2022
920/QĐ-ĐHCN Dated 21/10/2022
7. Official thesis title: Modeling and learning textual and structural relations for deep legal information retrieval
8. Major: Information system 9. Code: 9480104.01
10. Supervisors:
Supervisor: Assoc.Prof. Phan Xuan Hieu
VNU University of Engineering and Technology
Co-supervisor: Prof. Nguyen LeMinh
Japan Advanced Institute of Science and Technology
11. Summary of the **new findings** of the thesis:

This dissertation offers significant contributions in different aspects: the learning representation of legal features, the data agumentation, the definition and creation of the legal knowledge graph, the uncovering and usage of graphical relationships, and the graph-inspired deep learning model integration. First, the dissertation focuses on exploring and representing the legal relations between texts at different levels of granularity to deal with lengthy documents as well as make the most of both lexical and complex logical relations into a so-called *supporting model* to solve the case law retrieval task. Second, we propose a weak-labeling strategy to overcome the shortage of annotated data and improve the retrieval efficiency. Third, we define and create a *heterogeneous knowledge graph* of different types of legal entities to boost the performance of the statutory -- case law retrieval problem. We also define and build *a reference network* that captures and uitilizes the graphical connections or relationships among legal texts to enhance the performance of the question answering task. Moreover, throughout this

dissertation, we propose deep model architectures to smoothly integrate both legal textual and structural characteristics of the legal data to improve the performance of the IR and QA models. The proposedmodel architecturesdesigned and conducted experiments to validate the accuracy and effectiveness of the proposed models in the dissertation demonstrate better performance compared to the current benchmarks, with some achieving unparalleled results on established data collections. Performance enhancement demonstrated through thethorough experiments, analysis, evaluation elucidate the effectiveness of the proposed approaches and methods. Finally, the analysis and discussions throughout this work would help provide a deeper understanding of legal texts and processing problems, present the advancements and remaining limitations of legal NLP in general and legal IR and QA in particular; and would also suggest the future legal IR and QA research directions, especially for low-resource languages like Vietnamese.

The dissertation makes threemain contributions:

- We study the supporting relation in the legal texts, and propose an approach called *supporting model* that can deal with both the retrieval and the entailment phases in the case law retrieval task. The underlying idea is the case-case, the paragraph-paragraph as well as the decision-paragraph supporting relations to enhance the relevancy for legal text retrieval. Additionally, based on the supporting relation, we also propose a method to automatically create a large weak-labeling dataset to overcome the shortage of annotated data.
- We propose and construct *a heterogeneous knowledge graph* encompassing different types of legal entities (case law, courts, statutory laws, and legal domains) to improve legal information organization and retrieval in the statutory - case law retrieval task.
- We study the citation, reference relationships between the legal articles and propose *a reference network* approach to enhance the performance of the legal document question answering task. Embedding and encoding the local references and the global (long-range) dependencies among legal articles into deep pre-trained language models make the final QA model more robust and accurate. Also, by uncovering hidden connections between laws, our method can assist in the identification of inconsistencies and gaps in the legal system, ultimately improving its effectiveness and reliability.

12. Practical applicability:

This PhD dissertation contributes to both the scientific and practical areas. The dissertation presents a comprehensive overview of legal NLP for legal document IR and

QA. It also provides insights into the characteristics of legal documents and the relationships among them. Additionally, the methods of representation, architectural designs of models, and the procedural steps for training and evaluating these models are elaborately described within this dissertation.

13. Further research directions:

The future study will explore and improve the proposed method in a number of directions. First, continue to enhance methods for addressing problems related to the length and complexity of legal documents. Second, the efficiency of integrating the legal relations into the legal document IR and QA tasks suggests that we can extend our methods with larger and more sophisticated legal knowledge presentation, i.e., in terms of both scale and diversity. Expand research on logical representation in legal documents to improve accuracy for retrieval tasks in particular and legal NLP in general. Additionally, we can try larger pre-trained language models, especially models specialized for each particular language. Finally, developing solutions and models for legal IR and QA from various perspectives to serve various types of users including lawmakers, judges, plaintiffs, defendants, and non-expert users.

14. Thesis-related publications:

[1] Yen Thi-Hai Vuong, Quan Minh Bui, Ha-Thanh Nguyen, Thi-Thu-Trang Nguyen, Vu Tran, Xuan-Hieu Phan, Ken Satoh, and Le-Minh Nguyen. "SM-BERT-CR: a deep learning approach for case law retrieval with supporting model." Artificial Intelligence and Law 31, no. 3 (2023): 601-628. (SCIE, ISI/Q1 journal)

[2] Thi-Hai-Yen Vuong, Hai-Long Nguyen, Tan-Minh Nguyen, Hoang-Trung Nguyen, Thai-Binh Nguyen, and Ha-Thanh Nguyen. "NOWJ at COLIEE 2023: Multi-task and Ensemble Approaches in Legal Information Processing." The Review of Socionetwork Strategies (2024): 1-21. (ESCI, WoS journal)

[3] Thi-Hai-Yen Vuong, Hoang Minh-Quan, Tan-Minh Nguyen, Hoang-Trung Nguyen, and Ha-Thanh Nguyen. "Constructing a Knowledge Graph for Vietnamese Legal Cases with Heterogeneous Graphs." In 2023 15th International Conference on Knowledge and Systems Engineering (KSE), pp. 1-6. IEEE, 2023. (Scopus conference)

[4] Thi-Hai-Yen Vuong, Ha-Thanh Nguyen, Quang-Huy Nguyen, Le-Minh Nguyen, Xuan-Hieu Phan. "Improving Vietnamese Legal Question-Answering System based on Automatic Data Enrichment". In JSAI-isAI 2022. Lecture Notes in Computer Science, Springer. (In press, Scopus conference)

[5] Hai-Long Nguyen, Thai-Binh Nguyen, Tan-Minh Nguyen, Ha-Thanh Nguyen, and Hai-Yen Thi Vuong. "Vlh team at alqac 2022: Retrieving legal document and extracting

answer with bert-based model." In 2022 14th International Conference on Knowledge and Systems Engineering (KSE), pp. 1-6. IEEE, 2022. (Scopus conference)
[6] Nguyen, Hai-Long, Dieu-Quynh Nguyen, Hoang-Trung Nguyen, Thu-Trang Pham, Huu-Dong Nguyen, Thach-Anh Nguyen, Thi-Hai-Yen Vuong and Ha-Thanh Nguyen. "NeCo@ ALQAC 2023: Legal Domain Knowledge Acquisition for Low-Resource Languages through Data Enrichment." In 2023 15th International Conference on Knowledge and Systems Engineering (KSE), pp. 1-6. IEEE, 2023. (Scopus conference)

Date: ……..	Date: ……..
Signature: ……	Signature: ……
Full name: ……	Full name: ……