VIETNAM NATIONAL UNIVERSITY, HANOI
UNIVERSITY OF ENGINEERING AND TECHNOLOGY

VUONG THI HAI YEN

MODELING AND LEARNING TEXTUAL AND STRUCTURAL RELATIONS FOR DEEP LEGAL INFORMATION RETRIEVAL

DOCTOR OF PHILOSOPHY IN INFORMATION TECHNOLOGY DISSERTATION

Hanoi, 2024

Abstract

With the recent advances in digitalization and digital transformation, legal professionals can now easily access a huge volume of online legal materials. This is extremely important because judges and lawyers frequently need to find relevant legal information when they are working on a new legal case, performing legal research, case analysis, court preparation, giving legal advice to a client, developing a defense strategy, or making decison on a current case. However, the larger a legal database is, the more difficult for them to find relevant materials manually. In addition, legal documents like statutory law, case law or contract are normally lengthy and complex, consisting of multiple parts, chapters, sections, articles, and so on. Therefore, building an intelligent and automated legal information retrieval (IR) system is significant to improve and accelerate their legal process and workflow. Generally, this thesis aims to propose different legal IR methods and solutions based on an in-depth understanding of the nature and characteristics of legal data as well as the complexity of legal IR problems.

Accordingly, two major issues we need to consider carefully in this study are legal materials and legal IR problems. Legal materials are diverse, consisting of many different types of documents like constitution, statutory law, regulation, decision, case law, court document, contract, legal notice, patent, trademark, and so on. Among them, we focus on two main types of legal texts – statutory law and case law – because working on all types of legal materials is too broad and goes beyond the scope of the thesis. Regarding legal IR problems, this study focuses on three major IR tasks: (i) case law retrieval; (ii) statutory – case law retrieval; and (iii) IR–based legal question answering. The first task locates and returns case law documents from a case law database that relate and entail the decision of an input legal case. The second task retrieves statutory laws from a statutory law database that are relevant to a query case. And the third task seeks and returns statutory law articles that are likely to contain answers to a given legal question.

The three legal IR problems stated above are much more challenging than traditional IR for general-domain texts. The concept of relevancy in these tasks is no longer

about keyword or topic matching. The similarity between legal texts requires the understanding of legal arguments and logical reasoning that are far beyond the lexical or topical comparison. In addition, while working with legal data, we realized that legal language is rigorous and complicated. Legal documents are normally lengthy and heavily rely on domain-specific terminologies, jargons, and linguistic nuances. Furthermore, there is a complex graphical structure hidden in any legal dataset that results from frequent mentions, citations, references within and between legal materials. Also, the style and content of legal documents highly depend on the domain and the legal system of each country. And one more important issue is that annotated data is limited because labeling for legal data requires a lot of human effort and domain expertise. All of these reasons are both the challenges as well as the motivations behind our study.

The main objective of this thesis is to enhance the performance and accuracy of the three legal IR problems by making the most of textual and structural relations in the legal data. First, we propose a supporting model that encodes both the lexical and legal relations at different levels of granularity to deal with the case law retrieval problem. In addition, we introduce a method to automatically create a large weak-labeling dataset to overcome the limitation of labeled data. Second, a heterogeneous legal knowledge graph was defined and constructed to leverage the statutory–case relationships in the statutory – case law retrieval. Third, the thesis presents a novel approach that builds an article reference network to uncover both local and long-range dependencies between legal articles to enhance the performance of the IR–based legal question answering. Moreover, throughout the thesis, we propose appropriate deep learning architectures to encode the textual and structural characteristics of legal data and combine them with powerful pre-trained language models to enhance the overall performance of the three IR problems. Besides the technical contributions, the literature review, the analysis, and discussions throughout this thesis would provide a deeper and clearer understanding of the nature and the limitations in legal NLP in general and in legal IR in particular. It would also be a potential reference for future studies in the field, particularly for low-resource language like Vietnamese.

Keywords: statutory law, case law, legal case, deep legal information retrieval, legal question answering, case law retrieval, statutory – case law retrieval, IR–based legal question answering, legal case entailment, supporting model, weakly labeled data, relevancy, textual relation, structural relation, legal knowledge graph, article reference network, pre-trained language model.

Chapter 1

Introduction

1.1 The Deep Legal Information Retrieval Problems

IR and QA for legal texts are any tasks related to retrieving information relevant to an input query or finding a correct answer to an input question. There are also various types of legal documents. Therefore, we can have different ways to define IR and QA tasks. However, as mentioned in the previous subsection, in the scope of this study, we only work with two main types of legal materials: statutory law and case law. Accordingly, we will limit three primary IR and QA problems for these two types of legal documents in this thesis. Those problems are:

(i) Case law retrieval;

(ii) Statutory – case law retrieval;

(iii) IR–based legal question answering.

All of our proposed ideas and methods as well as our technical contributions in this thesis are around these three IR and QA problems. We will address here these problems in more detail for a precise understanding of what they are since we will encounter them frequently throughout this thesis.

Case Law Retrieval

With the recent advances in digitalization and digital transformation, judges and lawyers can now easily access a huge volume of online legal materials. However, the

larger number of legal documents is, the more difficult to find most relevant case laws. Therefore, developing an automated case law retrieval system will significantly accelerate and improve the performance of the judge's and lawyer's workflow. The case law retrieval problem was defined to meet this need. This problem consisits of two phases (or two sub-tasks) that are the Task 1 (The Legal Case Retrieval) and Task 2 (The Legal Case Entailment) of the COLIEE competition, respectively. Figure 1.1 shows the two phases and the logical flow of the case law retrieval problem.

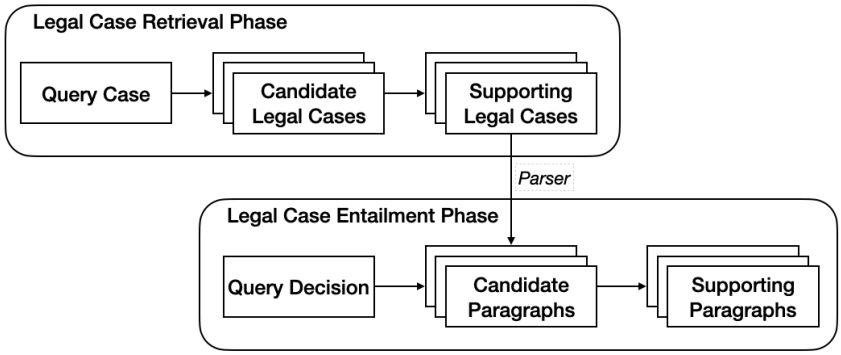


Figure 1.1: The logical flow of the case law retrieval problem

The legal case retrieval phase: Let C be the space of all possible legal cases and case laws and let $C \subset \mathbf{C}$ be a corpus of case laws (i.e., a case law database). Given an input query case $c_q \in \mathbf{C}$. The query c_q is normally a new legal case that a judge or a lawyer is currently working on. The aim of this phase is to locate and retrieve a set of all relevant case laws $C^r = \{c_1^r, c_2^r, \ldots, c_k^r\} \subset C$ that support the decision of c_q. In the legal domain, these supporting cases $c_1^r, c_2^r, \ldots, c_k^r$ are also called "noticed cases". Technically, this case retrieving phase can be expressed as the following mapping:

$$f_{case\_retrieval}(c_q, C) \rightarrow C^r \tag{1.1}$$

The legal case entailment phase: Given a triplet including the input query case c_q, a decision d_q of the query case c_q, and the list of all supporting cases C^r returned from the previous phase. Let P^r be the set of all text paragraphs being segmented from a given supporting case $c^r \in C^r$. The aim of this phase is to identify a set of supporting paragraphs $P^e = \{p_1^e, p_2^e, \ldots, p_l^e\} \subset P^r$ that entail the decision d_q of the query case c_q. Technically, this case retrieving phase can be expressed as the following mapping:

$$f_{case\_entailment}(c_q, d_q, P^r) \rightarrow P^e \tag{1.2}$$

The entailment relation between two legal text paragraphs is similar to the concept

of textual entailment in natural language understanding and inference. That is the relationship between two text segments where one (called *the hypothesis*) can be inferred or implied by the other (called *the text* or *premise*). In other words, if the text is true, then the hypothesis is likely to be true as well. Both the supporting in the first phase and the entailment in the second phase are complicated relations that are based on legal and logical reasoning. They are much deeper and go beyond the normal concept of relevancy in traditional IR that is merely based on the lexical and topical proximity. That is why we call and consider these tasks as **deep legal information retrieval** problems.

Statutory – Case Law Retrieval

Statutory – case law retrieval aims to find relevant statutory texts in addition to case laws. Its objective is to locate both legal statutes and case law pertinent to a particular legal inquiry, offering a holistic legal resource encompassing both legislative frameworks and judicial precedents.

Let S be a statutory law corpus (i.e., a database of statutory laws). Given an input query case c_q (normally, c_q is the new legal case that judges and lawyers are currently working on), the aim of this problem is to locate and retrieve all statutory laws $S^r = \{s^r_1, s^r_2, \ldots, s^r_k\}$ from the corpus S that are most relevant to the query case c_q. This can be expressed as the following mapping:

$$f_{law\_retrieval}(c_q, S) \rightarrow S^r \tag{1.3}$$

This problem will be described and discussed in more detail and the solution will be proposed in Chapter 3 of the thesis.

IR–based Legal Question Answering

Let A be a corpus (i.e., a database) of statutory law articles. Given a question q about any legal issues that can be covered by the corpus A, the goal of this problem is to find the most relevant statutory articles $A^r = \{a^r_1, a^r_2, \ldots, a^r_k\}$ from the corpus A that are most likely to contain the answers to the input question q. This can be expressed as the following mapping:

$$f_{statute\_retrieval}(q, A) \rightarrow A^r \tag{1.4}$$

1.2 Research Questions and Objectives

Based on the research challenges and motivations, as well as what have been done in the previous studies and what remain unsolved, this thesis addresses the following research questions:

- **Q1**: How will complex and lengthy legal documents be processed and represented? How to formulate and learn the textual legal relations and similarity between legal texts at different levels of granularity (case, paragraph, decision . . .) to enhance the relevancy and accuracy for the IR and QA problems?
- **Q2**: How to overcome the limitation of annotated legal data in the legal IR and QA problems? How can we have more labeled data in this domain to improve the retrieval performance?
- **Q3**: How can we represent and learn the structural rations that are the graphical connections among legal texts (e.g., local and long-range references) and the links among legal entities (e.g., courts, cases, laws, domains) to help enhance the performance of the IR and QA problems?
- **Q4**: How to integrate and leverage the legal textual and structural characteristics with powerful deep learning models (including pre-trained language models) to improve the performance of the IR and QA problems?

The overall goal of this thesis is to enhance the performance and efficiency of the legal IR and QA problems in different ways. Technically, we have three concrete objectives as folows:

- **O1**: Proposing new approaches and models to enhance the effectiveness of the legal IR and QA problems.
- **O2**: Leveraging and making the most of the nature and characteristics of legal data (i.e., both the textual and structural legal relations) to boost the performance of the three legal IR problems stated in Section 1.1: case law retrieval, statutory – case law retrieval, and IR–based legal question answering.
- **O3**: Proposing suitable methods to combine and integrate the textual and structural features of legal data with powerful deep learning models (including pre-trained language models) to further improve the efficiency of the legal IR and QA tasks.

1.3 Contributions

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 short 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 model architectures introduced in this Experimental research methods design and conduct 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 the thorough 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.

All in all, the disseration makes three major 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 in Chapter 2. 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 short 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 in Chapter 3.

- 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 in Chapter 4. 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.

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.

Chapter 2

Supporting Relation Model for Case Law Retrieval

Case law retrieval is the task of locating truly relevant case laws given an input query case. Unlike information retrieval for general texts, this task includes two phases (*case law retrieval* and *case law entailment*) and is much harder due to a number of reasons. First, both the query and candidate cases are long documents that consist of several paragraphs. This makes it difficult to model them with representation learning that usually has restriction on input length. Second, the concept of *relevancy* in this domain is defined based on the legal relation that goes beyond the lexical or topical relevance. This is a real challenge because normal text matching will not work. Third, building a large and accurate case law dataset requires a lot of effort and expertise. This is obviously an obstacle to creating enough data for training deep retrieval models. In this chapter, we propose a novel approach called *supporting model* that can deal with both phases. The underlying idea is the case-case supporting relation as well as the paragraph-paragraph and the decision-paragraph matching strategy. In addition, we propose a method to automatically create a large weak-labeling dataset to overcome the lack of data. The experiments showed that our solution has achieved the state-of-the-art results for both case retrieval and case entailment phases.

2.1 Case Law Supporting Relation

The case-case supporting relation does not only involve similar situations. The supporting cases can be mentioned and cited to support the query case law. According

to our observation, a case law s is a noticed case of a query case qc, which does not mean that all parts of s support qc. In other words, if there are only some paragraphs in s that support some decisions in qc, we can conclude that s support qc. Therefore, we introduct a supporting case concept for the case law supporting based on the supportive component. The long-text case law is splited into paragraph-like components and the supporting relation on the component level instead of focusing on the support relationship in the case law unit like in the previous studies. Figure 2.1 illustrates an example of our supportive component. This is the part of supporting relation graph between a query case and a candidate case.

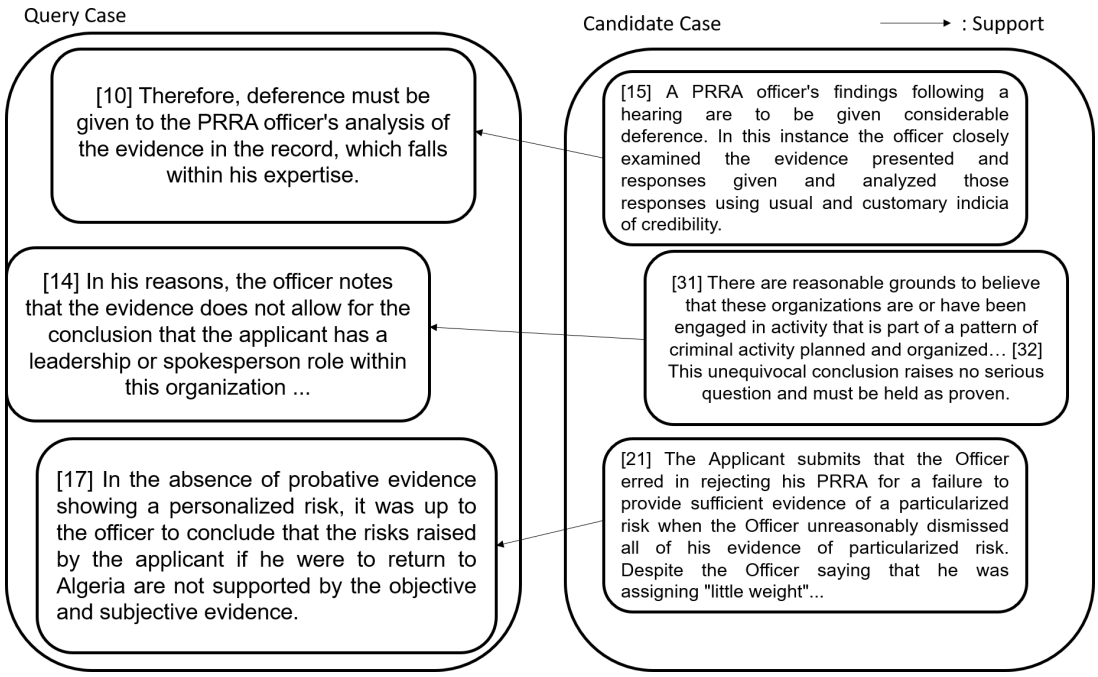
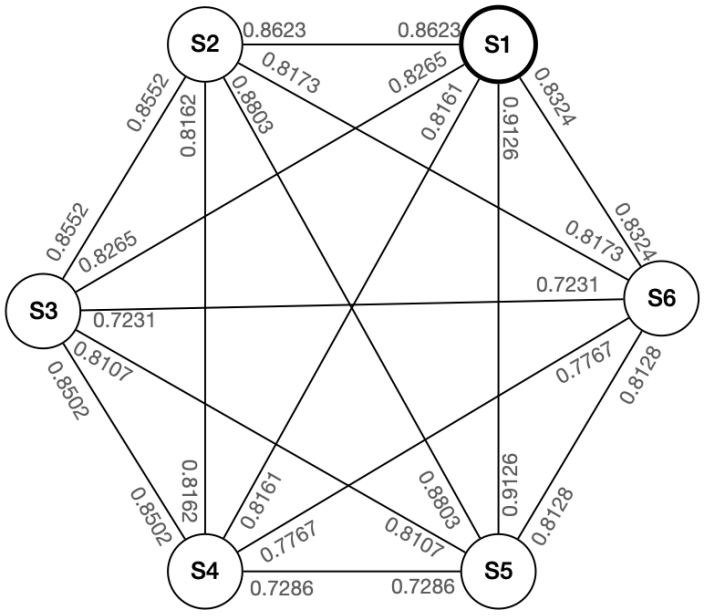


Figure 2.1: Example of supporting component extraction between a query case and a candidate case

Similarly, case law paragraphs are typically structured in an argumentative form, presenting legal arguments with clarity, precision, and logical coherence. Each paragraph focuses on a specific legal issue or specific legal point; utilizing logic, evidence, and specific citations to elucidate the issue or viewpoint presented. There is thematic unity within each legal paragraph, ensuring that the narrative is presented clearly. Figure 2.2 shows an example of supporting relation graph among sentences in the case law paragraph.



Case IMM-2683-96, paragraph 4:
S1: The applicant alleges that the Board used standard form "boiler-plate" reasons and therefore denied the applicant a fair hearing.
S2: Key passages in the Board's reasons are identical or virtually identical to two other decisions, Jafari v. Minister of Citizenship & Immigration , Board
S3: In all three cases involving Iranian refugee claimants, Mr. Jack Davis was the presiding Board member.
S4: Jafari was heard three days after the case at bar.
S5: The respondent in turn argues that the Board clearly made an independent decision.
S6: The identical passages are nothing more than digests of the law on such legal questions as credibility and the documentary evidence

Figure 2.2: An example of supporting relation graph among sentences in the case law paragraph, each sentence in the paragraph is represented as a vertice, edges are semantic similarity between the sentences. S1 is topic sentence in this example.

2.2 Supporting Relation in Case Law Retrieval

With the recent advances in digitalization and digital transformation, lawyers can now easily access a huge volume of online legal materials. However, the larger number of legal documents is, the more difficult to find most relevant case laws that assist the lawyer's court preparation. Thus, developing an automated law retrieval system is significant to accelerate the lawyer's workflow.

Legal information extraction and entailment (COLIEE) is an annual competition for researchers to tackle the problems of information retrieval, extraction, and reasoning in the legal domain. One of the main challenges in the competition is the case law task. The data for this task is based on The Federal Court of Canada case law provided by vLex Canada[1].

A case law is typically a collection of previous legal conclusions written by courts. A lawyer can find relevant case laws and use appropriate conclusions to support the decision in the current case. The case law can vary in structure, the components may not be the same in all cases, which requires significant effort in processing. It is even

[1] https://ca.vlex.com/

difficult for trained lawyers to read, scan and find truly relevant case laws from a large legal database. Case law retrieval is, therefore, a complicated task that have a number of challenges as follows:

Challenge 1: Both the query and supporting cases are extremely long texts which contain around 3000 words on average.

The long query is a challenge in the retrieval task. Both representation learning and matching learning methods have limitations in processing lengthy documents. It is challenging to learn representation for long text in a limited vector space. Constructing and aggregating long documents in matching learning is also a difficult problem.

Challenge 2: The definition of relevance in the legal domain is quite different from the general definition of topical relevance.

In the legal scenario, relevant cases are those that can support the decision of a new case, which usually have similar situations and appropriate regulations. It is crucial to identify the supportive relationship between case laws. This relationship is far beyond the topical and lexical relevancy. Matching between the query case and candidate cases, between the query decision and supporting cases becomes much more difficult in comparison with general text retrieval.

Challenge 3: Creating a large and accurate dataset for the case law task requires much effort and expert knowledge in the legal domain. The lack of labeled data is an obstacle to training and evaluation of large deep neural models.

In this study, we propose a deep learning approach with a supporting model for case law retrieval called SM-BERT-CR to deal with the above challenges. We propose a supporting case concept for the case law retrieval phase based on our supportive component in the case law supporting relation (**Challenges 1 and 2**). The relation between supporting paragraphs and a given decision in the case law entailment phase is similar to the relation between paragraphs in supporting cases and a query case in the retrieval phase.

Denoting a support relation as $support(a, b)$ (a supports b), the case law retrieval and entailment tasks are formalized as follows:

Case law retrieval phase :Let C be the space of all possible legal cases and case laws and let $C \subset$ C be a corpus of case laws (i.e., a case law database). Given an input query case $c_q \in$ C. The query c_q is normally a new legal case that a judge or a lawyer is currently working on, the task is to extract a set of supporting cases $C^r = \{c_i^r \mid c_i^r \in$

$C \wedge support(c_i^r, c_q)\}$. We assume that a candidate case C_i^r supports the query case c_q if and only if there are one or more paragraphs in s which support a decision in c_q:

$$support(c_i^r, c_q) \iff \exists p_j \in c_i^r \wedge \exists p_k \in c_q : support(p_j, p_k)$$

Case law entailment phase: Given a triplet including the input query case c_q, a decision d_q of the query case c_q, and the list of all supporting cases C^r returned from the previous phase. Let P^r be the set of all text paragraphs being segmented from a given supporting case $c^r \in C^r$, the task is to identify a set of entailing paragraphs $P^e = \{p_i^e \mid p_i^e \in P^r \wedge support(p_i^e, d_q)\}$.

The previous works usually tackle finding the supportive relationship between query-case/fragment and candidate-case/candidate-paragraph indirectly through similarity measures. Unlike previous studies, we build a supporting model to predict the supportive relationship directly (**Challenge 2**). Inspired by the success of the pre-trained language model BERT on a wide variety of natural language processing tasks, we adapt the BERT model to build our supporting model in case law tasks.

Besides the supporting model, we also exploit multiple similarity measurements such as lexical similarity (keyword matching) and semantic similarity (context matching). Although lexical similarity and semantic similarity are quite different from each other, they can be combined and complementary. The lexical similarity can be obtained by matching word by word with some alteration such as stemming, stopword removal, lemmatization, etc. A higher score in lexical similarity can show high matching between two documents, but with low lexical similarity, it does not mean that these documents do not have any relation. Thus, we combine the supporting model with the lexical model in our case law retrieval system.

To tackle the challenge of lacking labeled data, we use some heuristics to automatically construct the training dataset about the supporting relationship in case law called a weak-labeled supporting dataset (**Challenge 3**). This dataset is constructed based on our supporting relation graph in the case law paragraph that a paragraph contains a decision sentence and the remaining sentences support this decision sentence. Moreover, we assume that the decision sentence is the topic sentence in the candidate paragraph. To identify the decision sentence in the candidate paragraph, we apply the TextRank algorithm - a graph-based ranking model for automatic sentence extraction. The introduction of this dataset can reduce the dependency of neural models on labeled data.

Chapter 3

Knowledge Graph for Statutory – Case Law Retrieval

In this chapter, e develop a novel approach to a knowledge graph encompassing case law documents and relevant legislation to improve legal information organization and retrieval. Our method involves data collection, entity extraction, and graph construction. The constructed heterogeneous graph connects courts, cases, domains, and laws, significantly enriching information provided by retrieval systems. Our approach demonstrates potential in case analysis, legal recommendations, and decision support, providing valuable insights and resources for the legal domain.

3.1 Legal Knowledge Graph

A legal knowledge graphrepresents structured legal information in a graph format, capturing relationships between legal entities such as statutes, regulations, cases, and concepts. This graph-based representation enables a more comprehensive understanding of legal domains by organizing and connecting disparate legal data points. By modeling legal knowledge as interconnected nodes and edges, legal knowledge graphs facilitate various tasks, including legal research, information retrieval, and decision support systems.

In case law and statute law presentation, knowledge graphs provide structured frameworks for organizing and analyzing legal information derived from cases and statutes, respectively. Utilizing graph theory, these knowledge graphs represent legal entities, such as cases, statutes, regulations, legal domains, and relationships, as nodes and edges.

The method of constructing a knowledge graph serves as a suitable tool for identifying and representing the relationships between case laws and relevant laws. Knowledge graphs can effectively depict vast amounts of knowledge with semantic meaning, facilitating easy access and structured querying. These knowledge graphs are designed in a user-friendly manner, catering to non-expert users such as lawyers, judges, scholars, etc., enabling them to easily utilize and explore the information. Moreover, knowledge graphs can be applied to enhance various downstream tasks in the legal domain such as information retrieval, question-answering, classification, and more.

3.2 Vietnamese Legal Case Knowledge Graph Definition

A case law archived on the website of the Vietnam Supreme People's Court consists of two parts: metadata and the case document. Figure 3.1 illustrates the structure and content of a case law. The metadata contains basic information about the case, including the case number, case name, type of case, etc. The body of the case law document comprises four sections: the Introduction, the Content of the case, the Court's Judgment, and the Court's Decision. The description of each part is shown in Table 3.1.

Table 3.1: The description of a law document.

Part	Description
Introduction	details of case, court, defendant, plaintiff, related parties (e.g full name, date of birth, address of parties)
Content of the case	opinions of case, court, defendant, plaintiff, related parties
Court's judgment	Opinions, analysis of the court
Court's decision	Decisions of the court based on above parts

We construct the Vietnamese legal case knowledge graph based on a heterogeneous graph, which is can have nodes and edges of different types. A heterogeneous graph $G = (V, E)$ contains an entity set V and a relation set E with an entity type mapping function $f : V \rightarrow A$ and a relation type mapping function $g : E \rightarrow R$. A and R denote the sets of entity types and relations types, where $|A| + |R| > 2$.

Particularly, we define 4 types of entity based on the characteristic of the Vietnamese case law. Case node, which embeds information about each judgment/trial that

Bản án số: 577/2022/HC-PT ngày 28/07/2022 **a**
Tên bản án: Phạm Đăng M kiện UBND TP PR-TC
Đối tượng khởi kiện: QĐ hành chính, hành vi hành chính về quản lý đất đai […]
Cấp xét xử: Phúc thẩm
Loại án: Hành chính
Tòa án xét xử: TAND cấp cao tại TP Hồ Chí Minh
Áp dụng án lệ: Không
Đính chính: 0
Thông tin về vụ án: Không chấp nhận yêu cầu kháng cáo của người khởi kiện ông Phạm Đăng M [...]

1. Mở đầu: **b**
- Thành phần Hội đồng xét xử phúc thẩm gồm có [...]
- Thư ký phiên tòa [...]
[...]
2. Nội dung vụ án:
Theo đơn khởi kiện, biên bản đối thoại và tại phiên tòa người đại diện theo ủy quyền của người khởi kiện ông Lê Văn H trình bày: [...]
3. Nhận định của tòa án:
Sau khi nghiên cứu các tài liệu có trong hồ sơ vụ án đã được thẩm tra tại phiên tòa và căn cứ vào kết quả tranh tụng, ý kiến của đại diện Viện kiểm sát, các quy định pháp luật, Hội đồng xét xử nhận định: [...]
4. Quyết định:
Căn cứ khoản 1 Điều 241 của Luật tố tụng Hành chính năm 2015 [...]

Figure 3.1: The structure of a case law (a: meta-data, b: case content)

is currently in effect. Domain node, which embeds information about crimes, types of disputes and decisions. Court node embeds information about every court's name and level in the juridical system. Law node contains the name of specific law/code of law

There are a total of 3 types of relations between entities. Decide relation between courts and cases, indicating the relationship of a particular court hearing the trial. Belong-to relation between cases and domains, indicating the relationship of a particular domain and subdomain under which the case falls. Based-on relation between cases and laws, indicating the relationship of a particular judgment or decision that has referenced a set of laws/codes of law to support its verdict.

In a heterogeneous graph, two entities can be connected via different paths. Formally, these path are called meta-paths. A meta-path P is defined in the form of $A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \dots \xrightarrow{R_k} A_{k+1}$, which presents a composite relation $R = R_1 \circ R_2 \circ \cdots \circ R_k$ be-

tween A_1 and A_{k+1}, where $\circ$ denotes the composition operator on relations. Two case laws can be connected via different meta-paths, e.g. Case-Court-Case (CCC) or Case-Domain-Case (CDC). Different meta-paths describe semantic relationships in different views. For instance, the CCC path means these cases were judged by the same court, while the CDC path denotes that they belong to the same domain.

3.3 Statutory – Case Law Retrieval Model

Along with the development of technology, the volume of digital documents has significantly increased, especially in the legal field. This advancement has made it easier to search for and access legal information more efficiently. Legal documents are often lengthy, structured, and presented in a specific writing style. Effectively harnessing this data largely depends on how it is organized and standardized. In the legal domain, particularly in case law documents, one can find information about the cases, court decisions, and laws related to those cases. Although the information is available, retrieving legal information can be complex, especially when dealing with specific case law or investigating a particular case law as a legal expert. The desired information may need to be searched for from various sources and approached in different ways.

Regarding the information extraction, we have successfully identified entities and connections within the unstructured legal texts to populate a diverse graph. Beside helping to enhance the performance of the statutory – case law retrieval task, this method also facilitates a wide range of other applications in the legal field, including case analysis, legal guidance, and decision-making support. The baseline model, using unsupervised learning techniques and the knowledge graph, showed promising outcomes in recognizing pertinent laws for a specific case law. The future research can concentrate on refining information extraction, incorporating advanced graph-based learning approaches, and broadening the knowledge graph's range for enhanced performance and wider utility.

Chapter 4

Article Reference Network for IR-based Legal Question Answering

The increasing complexity of statute law has led to a growing demand for efficient and effective retrieval methods. This chapter presents a novel approach to statute law retrieval that utilizes reference networks to uncover connections between laws. By presenting laws as a network of references, our method allows users to quickly identify relevant laws and navigate the intricate web of legal documents. The key point is that the reference network can encode both internal and external legal relations, helping to integrate both the local relevancy and the long-range dependencies into the final retrieval model. We evaluate the performance of our approach using a large corpus of statute law documents and demonstrate that it outperforms existing retrieval methods. Our approach can contribute to the development of AI-assisted legal research tools, making it easier for legal practitioners to find relevant laws and precedents. Furthermore, by uncovering hidden connections between laws, our method can assist in identifying inconsistencies and gaps in the legal system, ultimately improving its effectiveness and reliability.

Additionally, this chapter synthesizes models that represent relationships within the legal domain to address the problem of Vietnamese legal document question-answering.

4.1 The Article Reference Relation Network

Legal documents are characterized by substantial length and a stringent organizational structure, typically partitioned into various hierarchical levels such as parts, chapters, sections, articles, and clauses with the article level being the predominant and

widely employed tier.

Internal reference: Within the context of legal documents, successive articles in a chapter frequently exhibit a proximate relationship in terms of content or through direct references using co-referential terms , we named it as *internal reference*.

External reference: Additionally, it is common for articles to make references to antecedent articles within the same or even different legal documents, , we named it as *external reference*.

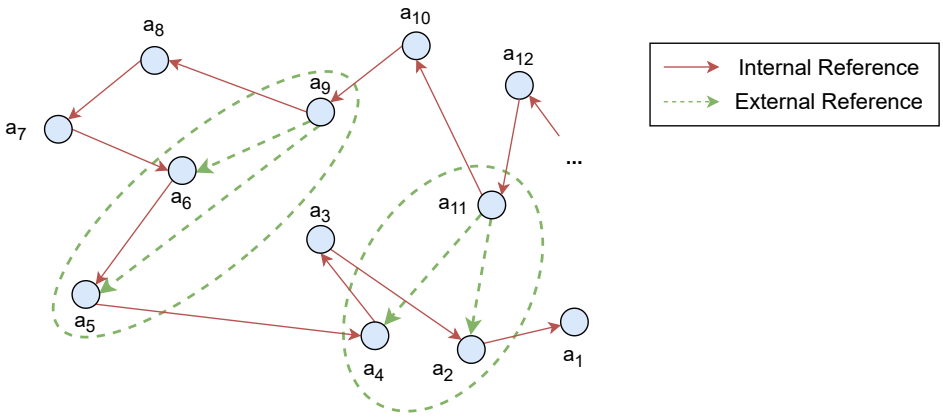


Figure 4.1: Illustration of reference relations between articles

In practice, legal documents encompass a substantial volume of reference relations, and disregarding these relations results in a significant loss of information. In this study, we propose the construction of a knowledge graph that could capture reference relations within legal documents. The legal reference relation graph was constructed based on a heterogeneous graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ as shown in Figure 4.1. The nodes in the graph are legal articles $\mathcal{V} = \{a_1, a_2, ..., a_N\}$. There are a total of two types of relations of entities, including the internal reference edges from a_i are $\mathcal{E}_{a_i}^{IN} = \left\{ in_{a_{i-1}}^{a_i} | a_i, a_{i-1} \in \mathcal{V} : InSameChapter(a_i, a_{i-1}) = 1 \right\}$ and the external reference edges from a_i are $\mathcal{E}_{a_i}^{EX} = \{ex_1^{a_i}, ex_2^{a_i}, \ldots, ex_{n_i}^{a_i}\}$.

4.2 Reference Network for IR-based Legal Question Answering

In an increasingly complex and highly specialized world, legal professionals are required to navigate vast arrays of statutory laws that are constantly evolving and in-

creasing in volume. The task of identifying relevant laws from a large corpus is not only laborious but also crucial for legal reasoning, legislative drafting, litigation, and legal scholarship. The shift towards digitized legal documents has driven the demand for efficient and effective law retrieval systems that can aid legal professionals in this endeavor. Traditional legal research methods, predominantly reliant on manual search or simple keyword-based searches, are often insufficient to cope with the intricate nature of legal texts. The structure of legal documents is characterized by a network of references, where laws cite other laws, creating a web of interdependent statutes. Understanding and navigating these interdependencies is essential for comprehensive legal analysis. However, due to the sheer volume and complex language of legal texts, manually tracking these references can be a daunting and error-prone task. The advent of information retrieval (IR) technology, coupled with recent advancements in natural language processing (NLP) and machine learning, has spurred the development of IR systems that can process large volumes of texts to find relevant information. However, the specific challenges posed by legal texts, such as domain-specific language, the necessity for high precision, and the importance of context and inter-document references, require tailored and more sophisticated solutions.

In this chapter, we introduce a novel law retrieval approach that utilizes the concept of reference networks to enhance the retrieval process. Our method capitalizes on the observation that legal statutes are not isolated entities; rather, they function within a network of references, with laws often citing other laws. By treating laws as nodes within a reference network, we can explore the direct and indirect connections between statutes, thereby enabling more effective identification of relevant laws. We propose an architecture that incorporates information from cited laws to enrich the representation of a given law, thus capturing both the content and the context of the references. This approach represents a significant departure from traditional document retrieval techniques that typically rely on content similarity alone. By considering the reference network, our system is better equipped to understand the legal context and relevance of documents, enabling it to yield more accurate retrieval results.

Recent studies related to law retrieval have employed various neural network techniques, such as CNNs, LSTMs, attention mechanisms, and graph neural networks, to achieve remarkable results in the legal domainThe previous approaches mainly focus on content-based similarity and may not fully capture the complex web of references within legal documents. In contrast, our approach aims to tackle this issue by harnessing the power of reference networks, especially making the most of both the internal (i.e., local)

relevancy and the external (i.e., long-range) dependencies to enhance the final retrieval model. Through the comprehensive evaluations of a large corpus of statute law documents, we demonstrate that our method outperforms existing retrieval methods in terms of relevance and efficiency. Additionally, we discuss the potential contributions of our model to the development of AI-assisted legal research tools, which can streamline the legal discovery process.

We delineate the methodology employed for leveraging correlations among legal articles to construct a data representation aimed at enhancing the outcomes of the legal retrieval task. The comprehensive structure of the model is depicted in Figure **??**. First, we present a comprehensive overview of the legal article retrieval problem. Following this, we introduce various symbols, knowledge graph structures, and the methodology involved in their construction process. Ultimately, we elucidate the architecture and training process of a model that integrates a graph representation of legal relations with pre-trained language models.

Legal article retrieval is one of the most traditional and common in the field of legal text processing. Let A be a corpus (i.e., a database) of statutory law articles. Given a question q about any legal issues that can be covered by the corpus A, the system aims to retrieve a subset $A^r \subset A$ that every article $a_i^r \in A^r$ semantically related or support to a given query q (legal question or statement). The problem can be described as follows:

$$\textit{Relevance}(q, a_i^r) = \begin{cases} 1 & \text{if } a_i^r \text{ is semantically related to } q \\ 0 & \text{otherwise} \end{cases} \tag{4.1}$$

$$A^r = \{a_i^r \in A : \textit{Relevance}(q, a_i^r) = 1\} \tag{4.2}$$

4.3 Vietnamese Legal Question Answering

4.3.1 General Architecture

Our proposed end-to-end article retrieval-based question-answering system architecture is demonstrated in Figure 4.2. The system comprises three primary phases: preprocessing, training, and inference phase, which work together to provide accurate and efficient responses to user queries.

The Preprocessing Phase A database consisting of individual articles is generated

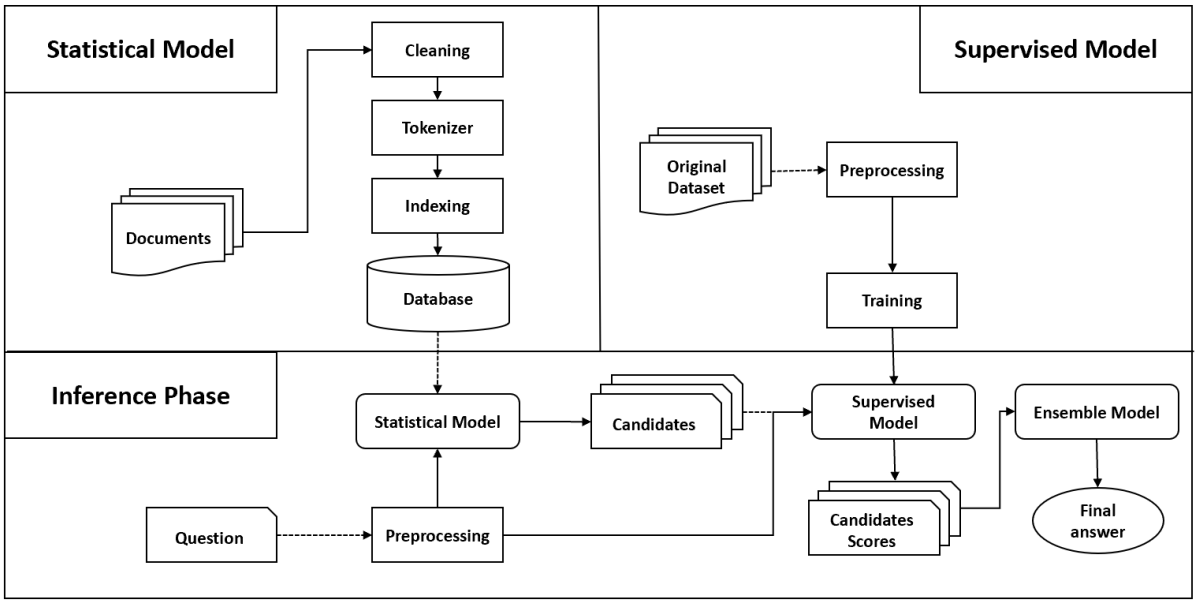


Figure 4.2: The pipeline of the end-to-end article retrieval-based QA system

by processing the Vietnamese civil law documents. The resulting article-level database enables easy access and retrieval of specific information contained within the documents.

The Training Phase A supervised machine learning model is developed to rank the articles related to the input question. This model uses training data to learn patterns and relationships within the articles and applies this knowledge to provide accurate rankings of relevant articles.

Inference phase Inference phase refers to the process of generating a response for a new input question. This phase typically involves applying a trained machine learning model to the input question and selecting the most appropriate response from a set of potential answers.

The proposed methods to improve the performance for the task of legal question answering for Vietnamese using language models through weak labelling and reference network. By demonstrating the effectiveness of this method through experiments, we have verified the hypothesis that improving the quality and quantity of datasets is the right approach for this problem, especially in low-resource languages like Vietnamese. The results of our work can provide valuable insights and serve as a reference for future attempts to tackle similar challenges in low-resource legal question answering.

Conclusions

Summary of the Results and Contributions

The dissertation conducted a systematic and thorough study of the legal retrieval and question answering tasks, that are two of the most critical and challenging problems in legal NLP. According to the research challenges, motivations, and objectives addressed in Chapter 1, the disssertation have presented the problem statement, formulation, and proposed the use of various types of legal characteristics (i.e., features) as well as introduced several deep model architectures to integrate those features in order to enhance the performance of the three IR and QA tasks. All in all, the dissertation has the following important results and contributions:

• In order to leverage and make the most of the nature and characteristics of legal data to boost the performance of the three main IR and QA tasks addressed in this dissertation (i.e., the research objective - O2), we have introduced the supporting model (in Chapter 2) that helps to integrate the supporting relations at different levels of granularity (i.e., case-case, paragraph-paragraph, and decision-paragraph) for the case law retrieval problem. In addition to the legal textual features, structural or graph-based features are also really useful for the legal IR and QA tasks. We therefore defined and constructed a heterogeneous knowledge graph consisting of legal case documents and relevant legislative materials in order to improve the legal information organization and the statutory – case law retrieval task (in Chapter 3). The knowledge grahp links cases, courts, domains, and laws to enrich graph-based features and therefore help to improve the retrieving performance significantly. In Chapter 4, we proposed the use of a reference network to enhance the performance of the legal question answering problem. The reference network captures both the local citations and the long-range (global) dependencies among legal articles in order to uncover potential links that help to locate and retrieve truly relevant articles that cannot be found by traditional lexical matching, by using synonyms, or even by text embedding methods.

• In addition to the uncovering and utilizing legal characteristics, this dissertation attempted to introduce suitable deep learning based architectures for the IR and QA tasks (i.e., the research objectives O1 and O3). These model architectures help (i) leverage the existing resources, methods, and models (e.g., powerful pre-trained language models) and (ii) learn better representations and integration of legal textual and structural characteristics. These model improvements result in the further inhancement of the effificiency and performance of the tasks. Technicall, Chapter 2 proposed the SM-BERT-CR architecture, a supporting model for the case law retrieval task. In Chapter 3, the construction of the heterogeneous graph involves data collection, entity extraction, and graph construction using NLP techniques. Our approach demonstrates its potential in the statutory – case law retrieval task and other downstream tasks such as case analysis, legal recommendations, and decision support, providing valuable insights and resources for the legal domain. Chapter 4 proposed the reference network model to address the legal document question answering task. The local and global reference links in the network were embedded using powerful pre-trained models and then incorporated into the final question answering model to improve the efficiency.

• Moreover, the experimental results in this dissertation are competitive with the state-of-the-art results, in which some models perform better than the previous work. In Chapter 2, the supporting model, i.e., SM-BERT-CR, achieved F_1 scores of 0.6060 and 6528 for the case law retrieval phase on the COLIEE 2019 and 2020 datasets, respectively. These outcomes are only 2 percentage points less than the state-of-the-art results even we did not use training data for this phase. In the second phase (i.e., legal entailment), this model has achieved very high results with F_1 of 0.7253 and 0.6753 on the COLIEE 2019 and 2020 datasets, respectively. These results are significantly higher (around 6 percentage points) than the runner-up team. In the statutory – case law retrieval task (Chapter 3), our knowledge graph-based method attained an F_1 score of 0.503, much higher than the baseline ($F_1 = 0.288$) that did not utilize the knowledge graph. In Chapter 4, the reference network-based method gave significant results on both COLIEE 2019 and 2020 datasets with F_2 scores of 0.7648 and 0.8266. Additionally, we also built an end-to-end for the Vietnamese legal document QA task and achieved the highest results on several Vietnamese datasets.

• Besides the techincal contributions, the analysis and discussions throughout this dissertation would help provide a better understanding of legal texts and processing problems, present the advancements and remaining challenges of legal NLP in general and legal IR and QA in particular. This study would also suggest the future legal IR and

QA research directions, especially inspiring and stimulating further studies in legal NLP for low-resource languages like Vietnamese.

The Limitations of the Dissertation

Although the dissertationn has attempted to leverage various legal features and introduced different deep learning based architectures to enhance the efficiency and performance of the three IR and QA tasks, there are still several limitations and remaning issues that can be done better.

First, in the first phase of the legal case retrieval task, the SM-BERT-CR model has identified and retrieved supporting cases for a given query from the entire case law corpus based on both the textual proximity and legal relation. However, in legal domain, a real supporting case is called a "noticed case" which is assumed to be relevant to the query case by lawyers. This normally causes an inconsistency in the data. As result, the first phase of the task normally retrieves more relevant cases than needed. This is still an issue that can be improved more in further studies. Second, the information in the legal knowledge graph built in Chapter 3 has not been fully exploited. This is partly because the knowledge graph was defined and constructed to cover many other downstream tasks in legal NLP. Finally, the proposed models in this dissertation still require high power computing systems to train and inference due to both the complexity of the models as well as the use of pre-trained language models. This sill needs to be improved for practical applications.

The Future Direction

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.

List of Publications

[VTHY1] **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)

[VTHY2] **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)

[VTHY3] **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)

[VTHY4] **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 2023. Lecture Notes in Computer Science*. Springer, Cham. (In press, Scopus)

[VTHY5] 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)

[VTHY6] 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)