KNOWLEDGE-ENHANCED DEEP NEURAL NETWORK FOR LEGAL JUDGMENT PREDICTION AND EXPLANATION

HE CONGQING

UNIVERSITI SAINS MALAYSIA

2025

KNOWLEDGE-ENHANCED DEEP NEURAL NETWORK FOR LEGAL JUDGMENT PREDICTION AND EXPLANATION

by

HE CONGQING

Thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosphy

September 2025

ACKNOWLEDGEMENT

I would like to express my sincerest appreciation to my supervisor, Associate Professor Dr. Tan Tien-Ping, for the continuous support of my PhD study and related research, for his patience, motivation, and immense knowledge. His guidance helped me throughout my research and the writing of this thesis.

Besides my supervisor, I would like to thank the rest of my thesis committee: Associate Professor Dr. Cheah Yu-N, Associate Professor Dr. Chan Huan Yong, and Associate Professor Dr. Gan Keng Hoon, for their insightful comments and encouragement, as well as for the challenging questions that prompted me to broaden my research from various perspectives.

My sincere thanks also go to the School of Computer Sciences, USM, for providing me with a conducive environment and the necessary facilities to complete my research.

I am deeply indebted to all my family for their love, support, and constant encouragement throughout my education and my life.

Finally, I would like to express my love and gratitude to my wife, Miss Duan Bohui, for her understanding, endless patience, and encouragement when it was most needed.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

RANGKAIAN NEURAL MENDALAM BERTAMBAH PENGETAHUAN UNTUK RAMALAN DAN PENJELASAN PENGHAKIMAN UNDANG-UNDANG

ABSTRAK

Ramalan Keputusan Undang-undang (LJP) adalah proses untuk meramalkan keputusan kes berdasarkan deskripsi fakta kes undang-undang. LJP bertujuan untuk meningkatkan kecekapan dan ketepatan prosedur undang-undang, menyediakan nasihat konsultasi undang-undang kepada profesional undang-undang dan orang awam. Baru-baru ini, pengenalan pengetahuan undang-undang untuk meningkatkan ramalan dan penjelasan keputusan undang-undang telah mendapat perhatian yang signifikan. Minat ini disebabkan oleh potensi pengetahuan undang-undang untuk memudahkan pemahaman maklumat penting dan menggantikan korpus latihan yang tidak mencukupi, dengan itu meningkatkan prestasi ramalan dan penjelasan keputusan undang-undang. Walau bagaimanapun, kaedah sedia ada untuk ramalan dan penjelasan keputusan undang-undang masih menghadapi beberapa cabaran utama, seperti bergantung kepada pengetahuan luar yang dianotasikan secara manual, risiko memperkenalkan pengetahuan undang-undang yang tidak tepat yang boleh mengelirukan ramalan keputusan, dan penjelasan yang tidak konsisten dengan fakta. Isu-isu ini boleh menyebabkan ramalan dan penjelasan keputusan undang-undang yang tidak tepat. Kajian ini bertujuan untuk menjembatani jurang tersebut dengan membangunkan rangka kerja JuriSim untuk meningkatkan prestasi dan kebolehjelasan LJP. Secara khusus, kami pertama kali mencadangkan penjanaan rasional dalam rangka kerja JuriSim dengan memperkenalkan rantaian acara sebagai pengetahuan tambahan. Ini meningkatkan keupayaan model untuk memberi tumpuan kepada peristiwa undang-undang yang penting ketika menghasilkan rasional, dengan itu meningkatkan keberkesanan penjelasan keputusan undang-undang. Kedua, kami mencadangkan mekanisme perhatian silang residu dwi yang mengintegrasikan pengetahuan tentang rasional dan peristiwa undang-undang dengan deskripsi fakta. Ini menangkap maklumat penting dari deskripsi fakta dan mengurangkan masalah pengetahuan yang tidak tepat yang mengelirukan ramalan keputusan

undang-undang, dengan itu meningkatkan prestasi LJP. Selanjutnya, kami memperkenalkan pengetahuan tentang pertuduhan undang-undang ke dalam rangka kerja JuriSim untuk lebih meningkatkan LJP dalam senario ekor panjang. Kami menggunakan pembelajaran rangsangan dan model bahasa besar untuk secara langsung menghasilkan pengetahuan elemen jenayah yang relevan untuk setiap kes. Pengetahuan ini diintegrasikan sebagai maklumat luaran ke dalam deskripsi fakta, dan kerugian entropi silang terkawal dicadangkan untuk mewakili hubungan antara pertuduhan dan artikel undang-undang. Kaedah yang dicadangkan telah diuji pada set data LAIC2021 untuk menilai prestasi penjanaan rasional dan ramalan keputusan undang-undang. Pendekatan ini menunjukkan peningkatan sebanyak 1.18% dalam Rouge-L, 0.93% dalam AVG-BLEU, dan 0.52% dalam BERTScore F1 untuk penjanaan rasional berbanding dengan model penjanaan rasional terkini. Di samping itu, pendekatan ini menunjukkan peningkatan sebanyak 8.1%, 2.07%, dan 9.25% dalam Macro-F1 untuk artikel undang-undang, pertuduhan, dan istilah hukuman, masing-masing, berbanding dengan model terkini. Kaedah yang dicadangkan juga telah dinilai pada set data CAIL-SMALL untuk LJP dalam senario ekor panjang, mencapai peningkatan sebanyak 0.75% dalam F1 untuk pertuduhan dan peningkatan 5.09% dalam F1 untuk artikel undang-undang dalam kes undang-undang ekor panjang. Secara keseluruhan, peningkatan yang signifikan yang dicapai oleh kaedah JuriSim yang dicadangkan menunjukkan keberkesanannya dalam ramalan dan penjelasan keputusan undang-undang.

KNOWLEDGE-ENHANCED DEEP NEURAL NETWORK FOR LEGAL JUDGMENT PREDICTION AND EXPLANATION

ABSTRACT

Legal Judgment Prediction (LJP) is a process for predicting judgment outcomes based on the fact description of a legal case. LJP aims to enhance the efficiency and accuracy of legal procedures, providing legal professionals and the general public with legal consultation advice. Recently, the introduction of legal knowledge to enhance legal judgment prediction and explanation has gained significant attention. This interest is due to the potential of legal knowledge to facilitate the understanding of crucial information and compensate for insufficient training corpora, thereby improving the performance of legal judgment prediction and explanation. However, existing methods for legal judgment prediction and explanation still face several key challenges, such as reliance on manually annotated external knowledge, the risk of introducing inaccurate legal knowledge that may mislead judgment predictions, and factual inconsistent explanations. These issues can lead to inaccurate legal judgment predictions and explanations. This study aims to bridge the gap by developing the JuriSim framework to enhance the performance and explainability of LJP. Firstly, we propose a rationale generation in the JuriSim framework by introducing event chains as auxiliary knowledge. This enhances the model's ability to focus on important legal events when generating rationales, thereby improving the effectiveness of legal judgment explanations. Secondly, we propose a dual residual cross-attention mechanism that integrates knowledge of rationales and legal events with the fact description. This mechanism captures crucial information from the fact descriptions and mitigates the issue of inaccurate knowledge misleading legal judgment predictions, thereby improving the performance of LJP. Furthermore, external knowledge, such as criminal elements and sub-task relationships, is incorporated into the JuriSim framework to further enhance LJP in long-tail scenarios. We use prompt learning and large language models to directly generate relevant criminal elements knowledge for each case. This knowledge is integrated as external information into the fact description, and a constrained cross-

entropy loss is proposed to represent the relationship between charges and law articles. The proposed method was tested on the LAIC2021 dataset to evaluate the performance of rationale generation and legal judgment prediction. The approach showed an improvement of 1.18% in Rouge-L, 0.93% in AVG-BLEU, and 0.52% in BERTScore F1 for rationale generation compared to the state-of-the-art rationale generation model. In addition, the approach showed improvements of 8.1%, 2.07%, and 9.25% in Macro-F1 for law articles, charges, and terms of penalties, respectively, compared to state-of-the-art models. The proposed method was also evaluated on the CAIL-SMALL dataset for LJP in long-tailed scenarios, achieving a 0.75% improvement in F1 for charges and a 5.09% improvement in F1 for law articles in tailed legal cases. Overall, the significant improvements achieved by the proposed JuriSim method demonstrate its effectiveness in legal judgment prediction and explanation.

CHAPTER 1

INTRODUCTION

1.1 Overview

The application of artificial intelligence (AI) in the legal domain has been studied for decades (Almuzaini & Azmi, 2023; Atkinson et al., 2020; Kort, 1957; Luo et al., 2017; Makridakis, 2017). Among various AI applications in the judicial process, legal judgment prediction (LJP) has attracted substantial attention in many countries, including but not limited to China (Hu et al., 2018; Le et al., 2023; Luo et al., 2017), the US (Katz et al., 2017), Europe (Chalkidis, Androutsopoulos, & Aletras, 2019), France (Sulea et al., 2017), Switzerland (Alali et al., 2021), India (Malik et al., 2021), Korea (Hwang et al., 2022), Arabic (Almuzaini & Azmi, 2023). Legal judgment prediction (LJP) is the task of automatically predicting the outcome of a court case, given the fact description of a legal case (Chalkidis, Androutsopoulos, & Aletras, 2019; Zhong et al., 2018). The LJP system can be used as a tool for case analysis and potentially for decision-making, thereby improving the efficiency of legal procedures. In real-world scenarios, LJP can automatically push case analyses, law articles, and judgment outcomes during the judges' case handling processes, providing them with unified and comprehensive trial standards and case handling guidelines. Furthermore, the system can automatically detect substantial discrepancies between judicial decisions and predicted outcomes, thus preventing major deviations in judicial standards.

Generally, LJP involves multiple tasks, including retrieval of relevant law articles, charge prediction, determination of penalty ranges, and generation of judgment rationales (J. Cui et al., 2023). As illustrated in Figure 1.1, LJP typically utilizes fact descriptions as input, and applies machine learning or deep learning techniques to extract features and predict judicial outcomes. In our research, we primarily focus on the Chinese criminal legal system due to the high-quality and large-scale datasets available in legal domain.

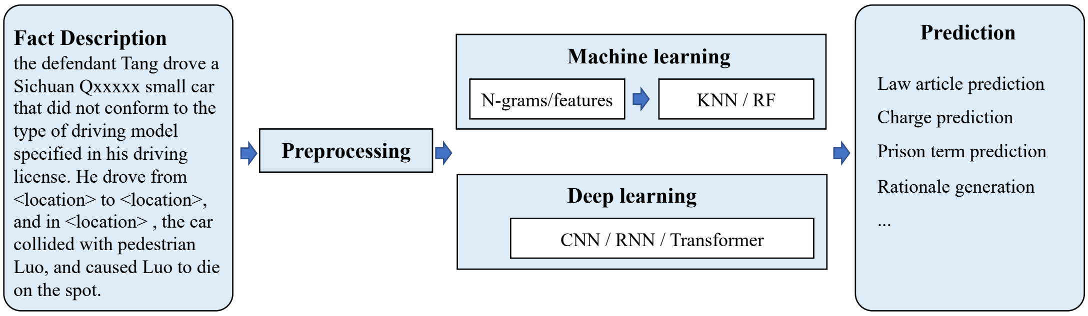

Figure 1.1: The general process of legal judgment prediction.

1.2 Introduction to Court Judgments

In real-world scenarios, legal systems are mainly classified into two types: civil law and common law. The civil law system, also known as the statutory law system, relies on statutory laws established and updated by the legislative branch of the government. This system is used in European and Asian countries, including China, France, Germany, Belgium, Switzerland, the Netherlands, Vietnam, and Thailand, as well as in Scandinavian and former Soviet countries. In contrast, the common law system is based on judicial decisions and precedents established by courts over time and is mainly used in Anglo-Saxon countries such as the United States, the United Kingdom, Australia, New Zealand, Canada, and India. Despite their significant differences in operation and theoretical concepts, both systems follow structured litigation procedures.

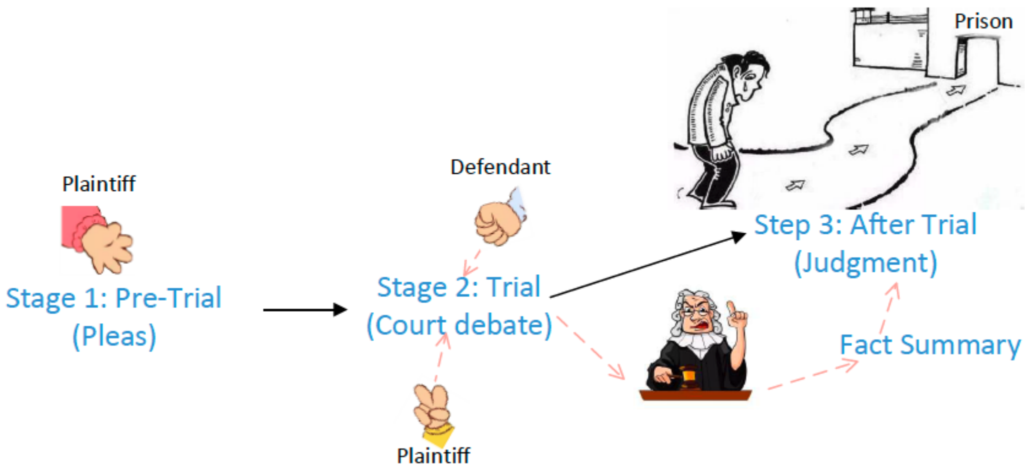

Figure 1.2: The litigation procedural stages of court judgment (J. Cui et al., 2023).

Figure 1.2 illustrates the litigation procedural stages of court judgment. In a real court setting, following the principle of 'no complaint, no trial,' the court's complaint and judgment are vital components in protecting the legal interests of the parties involved. Typically, the litigation process includes three stages: (i) Pre-Trial Stage: The plaintiff or appellant submits written materials to the court and appeals. (ii) Trial Stage: After the court agrees to hear the case, the involved parties, including the plaintiff, defendant, witnesses, and lawyers, present their arguments, focusing on the facts of the case. (iii) Post-Trial Stage: The judge issues a verdict based on a summary of the facts or the arguments presented during the second debate phase.

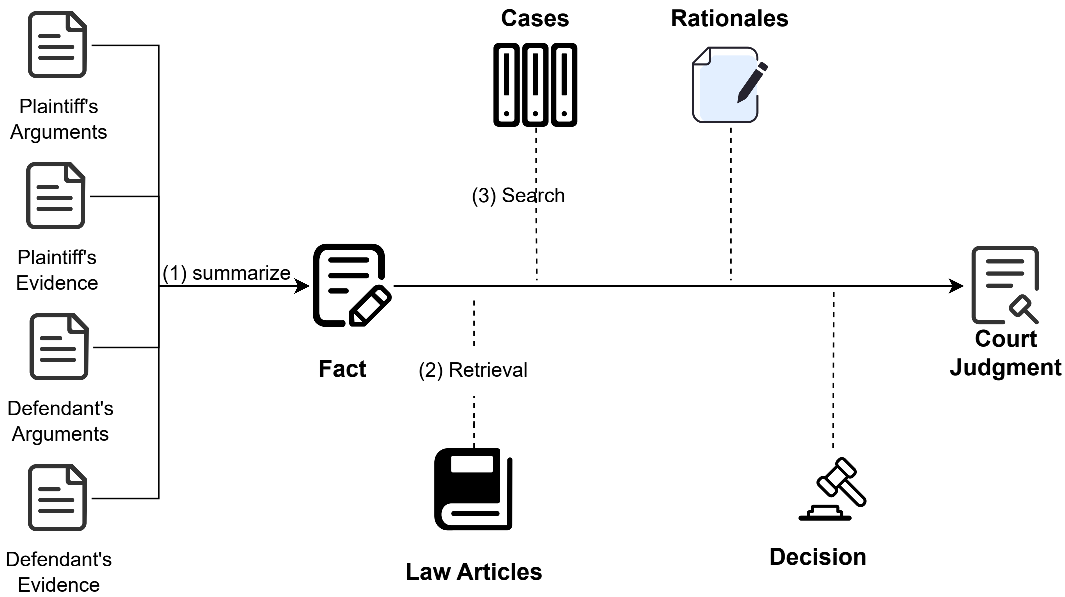

Figure 1.3: Overview of the process for court judgment decisions.

LJP mainly focuses on the post-trial stage, and we provide a detailed overview of this stage, as illustrated in Figure 1.3. In the judicial system, court judgments in general follow a structured decision-making process. First, the plaintiff, the defendant, or the public prosecution agency presents their arguments and supporting evidence to the court. The court then summarizes the case facts based on the presented evidence, highlighting key elements provided by the plaintiff or the prosecution. Next, the court reviews relevant law articles and similar cases that align with the established facts. Finally, the judge issues a judicial decision, determining charges, term of penalties, and compensation, accompanied by a detailed rationale explaining the application of legal principles and facts.

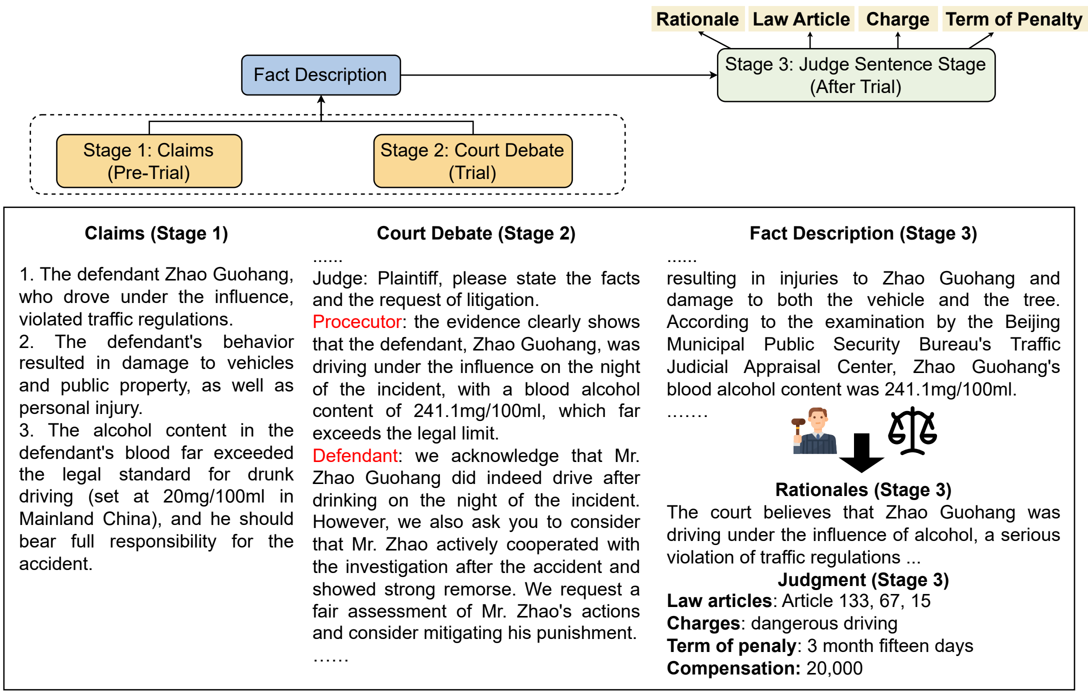

Figure 1.4: Illustration of the Chinese criminal legal case judgment process.

Then, we introduce the Chinese criminal legal case judgment process, as illustrated in Figure 1.4. The fact description is summarized by the pre-trial claims stage and the court debate stage, based on evidence and statements presented by the plaintiff or prosecution during litigation. According to the fact description, the judge delivers the court judgment by first providing rationales and relevant law articles, followed by the judgment results, including charges and the term of penalty.

In China, the legal jargon and terminology used in formal legal texts (e.g., law articles and judicial documents) are highly standardized at the national level. However, factual narratives may vary slightly across courts or judges. These variations affect the expression of details but do not alter the core legal terminology or the structure of formal legal reasoning.

The definition of concepts related to LJP is shown as follows:

Fact Description: The fact description is derived from evidence and statements presented by the plaintiff or prosecution during the litigation process, as depicted in Figure 1.4. Generally, the fact description involves content related to the suspect's actions, timing, location, and motives.

Law Articles: During the legal judgment process, judges search for the law articles relevant to the case, aiming to align the specific fact descriptions with applicable law articles, thereby making a decision.

Rationales: The rationale for a judgment is the detailed explanation and argumentation provided by a court when making its decision. It aims to clarify how the court makes decision based on the facts of the case and the applicable laws.

Judgment: The judgment includes a determination of whether the defendant is guilty of the charged crimes, along with the corresponding penalties. Charges refer to the final judgments given by the court, such as the crime of robbery or the crime of theft. The term of penalty refers to the duration of the punishment, as determined by the final judgment of the court.

1.3 Problem Statement

LJP has been studied extensively for decades, achieving notable advancements (Kort, 1960; Lauderdale & Clark, 2012; Segal, 1984; Ulmer, 1963). Earlier research mainly focused on the use of mathematical and statistical techniques to analyze legal issues (Kort, 1960). Inspired by the effectiveness of deep learning in natural language processing (NLP), researchers have begun incorporating deep neural networks to tackle challenges in LJP. Research demonstrates that deep neural networks enhance the accuracy of LJP over traditional machine learning techniques (Luo et al., 2017). However, LJP still confronts with three major challenges that affect its performance and explainability.

(1) **Explainability**: In the legal domain, explainability is fundamental for ensuring the transparency, reliability, and accountability of LJP systems. Explainability refers to the ability of a system to provide understandable and persuasive rationales for its decisions, allowing both legal professionals and non-experts to comprehend how and

why a particular judgment is made (Malik et al., 2021; Merrill, 1993; Yue, Liu, Wu, et al., 2021). In LJP, explainability is primarily achieved through model-generated rationales, which provide textual explanations for the reasoning behind the predicted judgments based on fact descriptions.

Research on explainable legal judgments can generally be categorized into two main approaches:

(i) Extraction-based methods: These approaches extract key sentences or phrases from legal documents as rationales. However, they often result in fragmented explanations that disrupt the overall logical coherence (Q. Li & Zhang, 2021; Wu et al., 2020; Ye et al., 2018; Yue, Liu, Wu, et al., 2021).

Fact Description:
The trial found that: Around 17:00 on June 17, 2015, the defendant Wang had a quarrel with the victim Chang, a stall owner, at a certain square in Yaodu District, Linfen City, Shanxi Province, and later Wang injured Chang. When the victim Chang Yi learned that his sister had and arrived at Drum Tower Square, he had a dispute with the defendant Wang who returned to look for his lost mobile phone, and Wang injured Chang Yi. According to the appraisal of the Public Security Judicial Appraisal Center directly under the Linfen City Public Security Bureau: Chang suffered a fracture of the left zygomatic arch, and the degree of his injury was minor injury level II; Chang Yi suffered a contusion of the soft tissue on the back of his left hand, of his injury was minor. After the incident, the defendant Wang reached a civil white compensation agreement with the victims Chang and Chang Yi. Wang compensated Chang and Chang Yi for their economic losses of 55,000 yuan in one lump sum, and has actually performed it, obtaining the forgiveness of Chang and Chang Yi. ...

Human-annotated Rationales:
The court believes that Wang intentionally injured others, causing one person to be slightly injured and one person to be minorly injured, his actions have constituted the crime of intentional injury. The defendant Wang pleaded guilty voluntarily in court, and after the incident, he was able compensate the victims for their economic losses and obtained the victims' forgiveness. The court will leniently punish him according to the law.

Rationales Generated by the BART Model:
The court believes that Wang intentionally injured others, causing one person to be slightly injured, his actions have constituted the crime of provoking trouble. The defendant Wang was able to truthfully confess his criminal facts after returning to the case, and his attitude of pleading guilty was relatively good, so he given a punishment according to the law. After the incident, the victims Chang and Chang Yi reached a civil compensation agreement and have actually performed it, obtaining the forgiveness of the victims. The court will leniently sentence him at its discretion according to the law.

Rationales Generated by the BART-C3VG Model:
The court believes that the defendant, Wang, intentionally injured one person, causing minor injuries, and his actions have constituted the crime of intentional injury. After committing the crime, Wang turned himself in and truthfully confessed his crime, which is considered surrender, and can be punished lightly according to law. After the incident, Wang reached a civil compensation agreement with the victims Chang and Chang Yi, and has actually performed it, obtaining the forgiveness of the victim, and can be sentenced lightly at discretion.

Figure 1.5: An example legal case with accurate (blue) and inaccurate (green) event descriptions highlighted.

(ii) Generation-based methods: Existing approaches (Wu et al., 2022) leverage sequence-to-sequence models or large language models (e.g., BART (Lewis et al., 2020), GPT-4 (OpenAI, 2023)) to generate rationales, providing intuitive explanations for both legal professionals and non-experts. However, these approaches typically treat rationale generation as a standard text-to-text task and tend to produce generic semantic content, lacking the capacity to capture critical legal details (Yue, Liu, Wu, et al., 2021). Consequently, the generated rationales may be inaccurate or even misleading. To address this limitation, the C3VG model first extracts key sentences from the fact description and then generates rationales based on these segments, which enhancing the model's capacity to capture essential information (Yue, Liu, Wu, et al., 2021). However, limiting the input to only these extracted sentences reduces the information available to the model, which may limit the quality of rationale generation. For example, as shown in Figure 1.5, both BART (Lewis et al., 2020) and BART-C3VG (Yue, Liu, Wu, et al., 2021) models produce rationales that mention only one injured person, while the actual case involved two victims. This case illustrates that both BART and BART-C3VG models failed to identify all victims involved, highlighting a limitation in capturing critical factual details. These inaccuracies may mislead legal professionals, affect explainability of legal cases, and potentially influence sentencing outcomes (Wu et al., 2022). To further assess this limitation, we conducted a manual evaluation comparing BART-generated rationales, BART-C3VG-generated rationales, and human-annotated rationales on a randomly sampled set of 100 cases. The results show that in approximately 40% of cases, the rationales generated by BART failed to accurately capture all key factual details or contained factual inaccuracies, while the proportion for BART-C3VG was around 25%. This quantitative analysis further demonstrates that factual inconsistency is a common issue in BART and BART-C3VG models.

(2) **The performance of LJP**: In LJP, fact descriptions often contain a lot of information that can obscure the core details relevant to judgment. Previous research has attempted to improve judgment performance by extracting crucial information (e.g., legal events or rationales) that capture criminal behaviors and consequences from fact description (Feng et al., 2022; Wu et al., 2022; Yue, Liu, Jin, et al., 2021). For instance,

Feng et al. (2022) extracted event information from case facts to predict judgments. However, obtaining accurate events is challenging because manual annotation is time-consuming and automatic extraction often introduces noise and errors that can degrade model performance. Similarly, Wu et al. (2022) incorporated rationales into fact representations using additive attention to capture criminal behaviors and consequences. Since rationales are also extracted or generated from fact descriptions, inaccuracies may be propagated by the additive attention mechanism, leading to erroneous legal judgments. Therefore, the main challenge is how to effectively integrate legal information from case facts to enhance LJP model performance.

(3) **The performance of LJP in long-tailed scenario**: In real-world scenarios, legal case datasets often display long-tailed label distributions, where some labels/categories have many sample data, while majority of labels/categories have very few sample data. For example, Figure 1.6 presents the distribution of charges/law articles in the CAIL-SMALL dataset. Considering the number of various charges, the top 50 charges cover 76.35% cases. On the contrary, the bottom 10 charges only cover 1.14% cases. Considering the number of various law articles, the top 50 law articles cover 84.6% cases. On the contrary, the bottom 10 law articles only cover 0.71% cases. Limited training data for these infrequent categories poses challenges for models to accurately learning the relationships between fact descriptions and the corresponding charges or law articles. This limitation restricts the model's performance. As shown in Figure 1.7, the accuracy of MPBFN (W. Yang et al., 2019) drops significantly for less frequent charges and law articles.

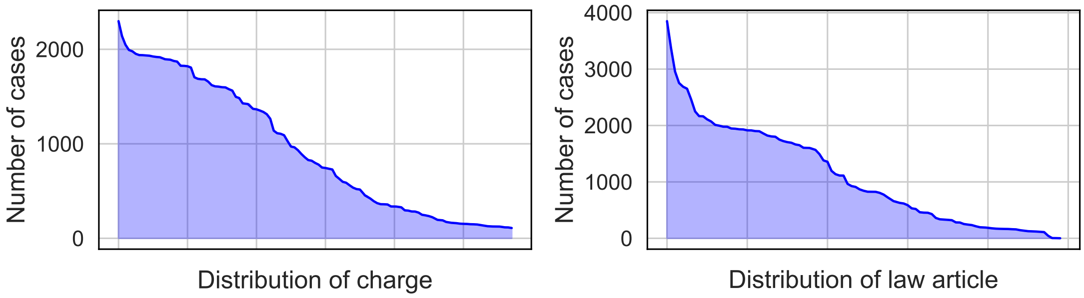

Figure 1.6: The distribution of charges and law articles on CAIL-SMALL dataset.

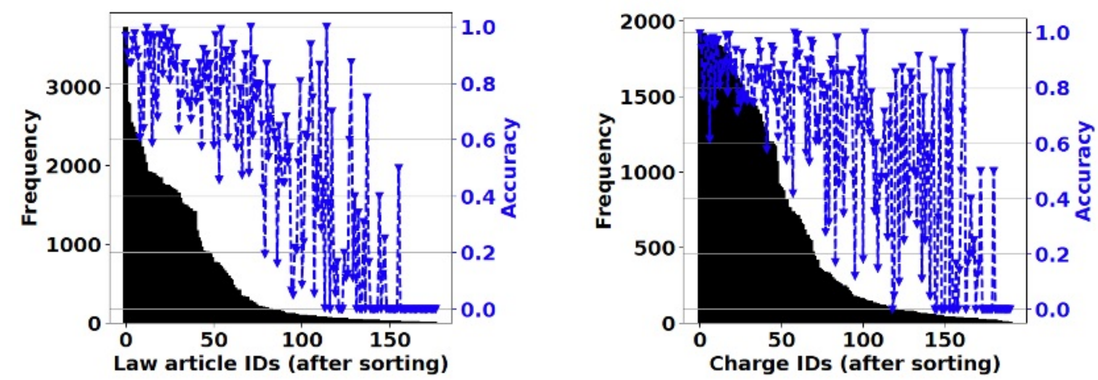

Figure 1.7: The prediction accuracy of MPBFN (W. Yang et al., 2019) on CAIL-SMALL dataset for each article and charge.

To improve LJP performance in long-tailed scenarios, existing studies have introduced external legal knowledge. For instance, Hu et al. (2018) utilized legal attributes as internal mappings between fact descriptions and charges. However, this method relies on manually crafted attributes, which require frequent expert updates and are difficult to scale as legal definitions change. D. Liu et al. (2022) tackled data imbalance by using similar and dissimilar cases to train on rare categories. However, accurately identifying similar cases is difficult in practice due to nuanced differences in facts and disputes (W. Li & Li, 2023). When irrelevant cases are retrieved, the model may learn incorrect patterns, reducing predictive accuracy. Thus, efficiently incorporating external legal knowledge to enhance LJP in long-tailed scenarios remains a key challenge.

1.4 Research Questions

The following research questions can be formed based on the research problems, the available literature, and the goal of improving the performance and explainability of legal judgment prediction:

1. How to generate factually consistent rationales for judicial decisions given case descriptions?

2. How to effectively integrate legal information, such as events and rationales, with fact descriptions to improve the performance of LJP?

3. How to incorporate external legal knowledge to enhance the performance of LJP in long-tailed legal cases?

1.5 Research Objectives

The main objective of this research is to enhance the performance and explainability of the legal judgment prediction.

1. To develop a method that improves the quality and factual consistency of rationale generation for each legal case.

2. To develop an approach for integrating legal information (such as legal events and rationales) with fact descriptions to enhance the performance of LJP.

3. To propose an approach that incorporate external legal knowledge to enhance the performance of LJP in long-tailed legal cases.

1.6 Research Scope

Legal systems and judicial procedures vary across different countries. This study was conducted on Chinese criminal legal cases only. In LJP, the main tasks include predicting law articles, charges,and term of penalty, as well as generating rationales for each criminal case.

Chinese criminal legal case have two features that attract researchers:

- **High-Quality and Large-Scale Datasets**: In recent years, the digitalization of the Chinese judicial system has progressed substantially, resulting in the large-scale accumulation of criminal judgment documents and related legal data. The data provides a foundation for machine learning and data analysis, effectively supporting the construction and training of models for predicting and explaining criminal legal judgments.

- **Policy Support and Research Resources**: The Chinese government and relevant judicial institutions highly prioritize the openness and effective utilization of legal data, implementing various policies to promote judicial data sharing and application. In addition, substantial resources from academic and technological communities have been dedicated to this area, producing a conducive research environment and yielding abundant research outcomes. Consequently, predicting and interpreting criminal law judgments is both highly feasible and practically valuable.

1.7 Contributions

The main contribution of this research is proposed an novel framework (dubbed JuriSim) to enhance the performance and explainability of LJP. The following are the sub contributions:

The first contribution focuses on improving the quality and factual consistency of rationale generation for judicial decisions. Unlike prior studies typically treat rationale generation as a standard text-to-text task, we propose the JuriSim-Rationale model, which is the first to jointly model event chain extraction and rationale generation within a unified multi-task learning method. By introducing event chains as an auxiliary task to guide the rationale generation process, our method enables the model to focus on essential legal events when generating rationales, thereby enhancing both the quality and factual consistency of the rationale generation.

The second contribution aims to enhance the performance of LJP by effectively integrating legal information, such as event chains and rationales, into fact descriptions. To achieve this, we develop the JuriSim-LJP model, which introduces a novel dual residual cross-attention mechanism. While cross-attention mechanism have been widely used in prior work, our proposed dual residual cross-attention mechanism enables the model to selectively capture essential information from fact descriptions and to effectively mitigate the influence of noisy or inaccurate event and rationale details, thus improving the performance LJP.

The third contribution addresses the critical challenge posed by the long-tailed distribution of legal cases through the incorporation of external legal knowledge into the LJP. Specifically, we integrate legal charge knowledge (i.e., criminal elements and relationships between charges and law articles) to alleviate data scarcity issues in long-tailed legal cases. To effectively fuse external legal knowledge, we propose two approaches: (1) The JuriSim-LT1 model employs a multi-cross attention network to dynamically identify and incorporate the most relevant legal charge knowledge into fact representations, thereby improving legal judgment prediction performance on long-tailed categories. (2) The JuriSim-LT2 model uses prompt-based learning with large language models to automatically generate case-specific criminal element knowledge, which is then integrated into the model via cross-attention and a constrained cross-entropy loss. This strategy further enhances the model's ability to handle rare cases and boosts overall performance in long-tailed scenarios.

1.8 Thesis Outline

This thesis is organized as follows:

Chapter One introduces the background and challenges of LJP, along with the problems associated with LJP, research motivation, objectives, and contributions.

Chapter Two provides a literature review and discusses related works that address the challenges in legal judgment prediction and explanation.

Chapter Three outlines the research methodology framework utilized in this thesis, detailing each component and its relationships. It explains the methods proposed for each stage and subsequently presents the benchmark datasets along with the evaluation metrics used.

Chapter Four investigates methods for explaining legal judgments by generating rationales. The chapter proposes a multi-task learning model that integrates both event chain and rationale generation.

Chapter Five presents a model with a dual residual cross-attention mechanism that integrates knowledge of rationales and legal events with fact descriptions.

Chapter Six introduces legal charge knowledge into legal cases and presents a multi-cross attention network to integrate the fact descriptions with legal charge knowledge. It then proposes a model that cooperates with large language models to integrate relevant legal charge knowledge with fact descriptions.

Chapter Seven concludes the thesis with a brief summary of the research, concise comments on the findings, and a discussion of possible future directions. Apart from the first and last chapters, each chapter begins with an introduction and concludes with a summary.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter reviews related works of legal judgment prediction and explainability. Section 2.2 presents an overview of the LJP, including its key features, significance, major research challenges, critical components, and standard processes. Section 2.3 introduces existing methods for legal knowledge graphs, domain-specific pre-trained language models, and presents state-of-the-art knowledge-enhanced LJP models. Section 2.4 reviews existing long-tailed learning methods. The chapter then discusses state-of-the-art methods for the explainability of legal judgments in Section 2.5.

2.2 Overview of Legal Judgment Prediction

LJP is an multidisciplinary research field integrating artificial intelligence, NLP, and legal studies. Its primary goal is to automatically predict judicial outcomes based on fact descriptions of real-world legal cases using AI system. In essence, LJP systems attempt to simulate a judge's decision-making by mapping a case's facts to a judgment outcome. This problem domain has significant practical importance: effective LJP models can assist judges and lawyers in analyzing cases, provide consultation to non-professionals, reduce legal costs, and improve access to justice. From a research perspective, LJP is challenging and significant because it requires natural language understanding, integration of domain-specific legal knowledge, and the ability to legal reasoning.

2.2.1 Key Tasks in LJP

The domain of LJP comprises several fundamental subtasks, each corresponding to a distinct component of the judicial decision-making process. The following summarizes the primary tasks typically addressed in LJP research, along with their definitions:

Violation Prediction: This task involves determining whether the facts of a case constitute a legal violation (Chalkidis, Androutsopoulos, & Aletras, 2019). In criminal law, this typically corresponds to a binary classification of guilty vs. not guilty, while in civil law it may indicate liable vs. not liable. For example, the model predicts whether a defendant is guilty of a specific crime or whether a particular legal right has been infringed. It should be noted that in some datasets, such as those containing only convicted cases from Chinese criminal law, this task may be omitted. However, it remains fundamental in the general framework of LJP.

Charge Prediction: This task aims to identify the specific criminal charge(s) applicable to a defendant, given the factual circumstances of the case (Hu et al., 2018). In criminal law, a "charge" denotes the formal legal accusation, such as theft, fraud, or assault. Charge prediction is typically formulated as a multi-class or multi-label classification problem, as a single case may involve one or more charges. The objective is for the model to accurately determine the appropriate charge category or categories based on the given fact description.

Law Article Prediction: This task involves predicting the relevant law articles or statutory applicable to a given case (H. Zhang et al., 2019). In many legal systems, particularly civil law jurisdictions such as China, it is necessary to reference specific law articles in judicial decisions. For instance, in a theft case, the model should determine which law article is applicable to the offense. Law article prediction is often formulated as a multi-label classification problem, as a case may involve multiple applicable law articles. This task is closely related to charge prediction, as each charge is typically associated with one or more law articles.

Judgment Outcome Prediction: In some LJP tasks, the main prediction target is the overall decision outcome of the court (Chalkidis, Androutsopoulos, & Aletras, 2019; Zhong et al., 2018). This may include outcomes such as the verdict (such as conviction vs. acquittal), or the result of an appeal (such as upheld vs. reversed, or approved vs. dismissed). For example, in European Court of Human Rights cases, the task involves predicting whether a violation of a human rights article has occurred. Similarly, other judicial system may require predicting whether an appeal is granted or rejected. This task is generally formulated as a classification problem with a limited set of outcome labels.

Term of Penalty / Sentence Prediction: This task focuses on determining the appropriate term of penalty or sentence following a conviction (Tan et al., 2020). In criminal cases, this typically involves predicting the time of imprisonment or other forms of punishment, such as fines. Sentence prediction can be formulated as a regression problem (predicting a specific term of penalty) or as a classification problem (predicting a category or range of term of penalty). This subtask is considered particularly challenging, as it requires not only legal reasoning but also quantitative analysis of law articles requirements. For instance, when a case involves multiple charges, the total sentence is not simply the sum of the individual penalties. Instead, it must comply with legal principles such as concurrent and consecutive sentencing as well as statutory limits.

Rationale Generation: This task focuses on generating a written explanation that demonstrates how the decision is derived from the case facts and the relevant law articles (Yue, Liu, Wu, et al., 2021). Effective rationale generation supports the transparency and explainability of LJP models.

2.2.2 Legal Documents Components

LJP is conducted on formal legal case documents, making it essential to clearly identify the key components of legal case documents and to understand their roles within the prediction tasks.

Fact Description: The fact description presents a summary of the case facts as established during legal proceedings. It typically details the time, location, parties involved, actions taken, resulting consequences, and relevant context. The fact description, authored by the court, serves as the basis for judicial decision-making and constitutes the input for LJP systems. For example, in criminal cases, the fact description outlines the defendant's alleged conduct, surrounding circumstances, and supporting evidence.

Law Articles: Law articles refer to the specific statutes or clauses that are relevant to the case. In judgment documents, the court cites relevant law articles that the defendant's behavior is found to have violated or that are used to determine the judgment outcome. Each charge typically corresponds to one or more law articles. For example, a charge of robbery may involve law articles 263 and 289. Accurate identification of applicable law articles is essential, as they constitute the legal foundation for the judgment. In LJP, predicting the relevant law articles is particularly important in civil law systems, where judgments require references to specific law articles.

Charge: The charge refers to the formal legal classification of the offense with which the defendant is accused or found guilty. In Chinese criminal law, charges are defined by the Criminal Law and encompass categories such as theft, fraud, and intentional injury. The charge maps the factual circumstances of the case to a specific legal category of crime. In judicial documents, after presenting the facts, the court typically states that "the defendant's conduct constitutes the crime (charge) according to law articles". In LJP tasks, charge prediction is a crucial objective, where the model assigns the appropriate charge to the fact description of the case.

Rationale: The rationale constitutes a critical component of legal judgment, providing a detailed explanation of how the court makes decision based on the facts of the case and the applicable law articles. The rationale explains the reasoning process by which the judge determines that the facts satisfy the legal criteria, thereby justifying the outcome. While automated LJP research has focused on predicting outcomes rather than generating rationales, the rationale remains essential in actual judicial practice. Recent research has increasingly incorporated rationale generation or leveraged rationales to enhance the transparency and explainability of model predictions. For instance, a rationale may state: "The court finds that Wang intentionally injured others, resulting in one person sustaining minor injuries and another sustaining slight injuries; therefore, his actions constitute the crime of intentional injury.” Incorporating rationales is essential for improving the explainability of LJP systems, which is fundamental for fostering trust and credibility in automated legal decision-making.

Judgment Outcome: The judgment outcome represents the final decision rendered by the court. In criminal cases, this includes both the verdict (such as guilty, not guilty or charges) and the sentence imposed (e.g., the term of imprisonment or other penalties). In civil cases, the outcome may indicate which party prevails and specify any remedies granted. Typically, the outcome is stated in judgment documents, such as “The court finds the defendant guilty of X and sentences them to Y years of imprisonment”. For LJP tasks, the outcome can be decomposed into structured elements, including the predicted charges, applicable law articles, and the term of penalty. In some LJP studies, the term “judgment result” collectively refers to these predicted components.

Typically, a standard court judgment document, particularly in civil law systems such as China’s, consists of several essential components: (a) the fact description, (b) the court’s rationale or reasoning, (c) references to relevant law articles, (d) the identified charges, and (e) the final judgment outcome, such as the charges and term of penalty.

2.2.3 Legal Judgment Process

A typical legal judgment process involves several key steps, each corresponding to fundamental components in both judicial practice and LJP tasks. The following outlines the main stages of this process:

Fact Description: The fact description is derived from evidence and statements presented by the plaintiff or prosecution during the litigation process. In LJP, the fact description serves as the foundation for subsequent analysis.

Law Article: Based on the established facts, the judge determines the applicable law articles by identifying which law articles are relevant and assessing whether the facts satisfy the conditions required for an offense. For example, the judge may consider whether the facts constitute theft, robbery, or not guilty. This step corresponds to both the violation determination (i.e., whether a legal violation has occurred) and the identification of law articles and charges within LJP tasks. In judicial practice, the judge typically first determines whether a crime has been committed. If so, the appropriate charge is identified and the relevant law article is cited. If no law articles applies, the case results in a verdict of not guilty.

Rationale: During the judgment process, the judge provides a rationale that explains how the decision was reached. Rationale is an essential component of judicial reasoning, serving to ensure transparency and to justify the judgment outcome. In LJP research, rationale corresponds to the explainability of model predictions.

Judgment Decision: The judgment decision encompasses both the determination of the appropriate charge(s) and the corresponding term of penalty. After determining the applicable charges, the judge delivers the final judgment, which includes both the verdict (guilty or not guilty) and the sentence. The sentencing process follows legal guidelines and takes into account factors such as the severity of the offense and any mitigating or aggravating circumstances. The judge assigns a term of penalty, such as the length of imprisonment. This step can involve complex reasoning, especially in cases involving multiple charges, where the aggregation of term of penalties must be consistent with relevant law articles rather than being a straightforward sum of individual term of penalty. In LJP, the judgment decision corresponds to predicting both the charge(s) and the term(s) of penalty for each case.

2.2.4 Challenges in LJP

Despite notable advancements in recent years, LJP research still confronts several persistent challenges.

Complexity of Legal Language: Legal texts are characterized by specialized vocabulary, formal language, and complex syntactic structures, distinguishing them from the general texts. For instance, a legal document might state, "The defendant's conduct constituted intentional injury as defined in Law Article 234 of the Criminal Law," whereas a general text would express the same idea as, "He hurt someone on purpose." In this example, the legal text provides a precise definition of the offense and cites the relevant law articles, while the general text is more vague and informal. Similarly, a legal statement such as, "The defendant is exonerated due to lack of mens rea," stands in contrast to the general texts, "He didn't mean to do it, so he is not guilty". The term "mens rea" illustrates a legal concept of criminal intent, which has no direct equivalent in general language. These examples highlight that legal documents often involve legal definitions, citations of law articles, and domain-specific concepts. Accurately understanding and modeling this linguistic complexity remains a critical challenge in the development of LJP systems.

Domain Knowledge: Legal documents are composed in formal and domain-specific language, characterized by complex structures and specialized terminology. Effective LJP models must be able to comprehend this legal language and frequently benefit from the integration of external legal knowledge, such as law article definitions or precedents. Incorporating domain-specific knowledge remains challenging, since general NLP models lack legal expertise and training data alone may not capture all legal nuances. Recent work addresses this by introducing legal knowledge via rule-based methods or pre-training on legal texts.

Explainability: In the legal domain, explainability is fundamental to establishing the trustworthiness of predictive models. This is especially important, as obscure "black box" models are often distrusted by legal practitioners. A primary challenge in LJP research is generating human-interpretable justifications for model outputs, such as indicating which facts or legal arguments influenced the prediction or providing automatically generated rationale text as part of the output. The absence of adequate explainability can hinder the real-world adoption of LJP systems.

Ethical and Fairness: The deployment of AI in the legal domain raises critical ethical and fairness concerns. Training data biases may be learned and amplified by LJP systems, resulting in discriminatory or unfair outcomes. Ensuring fairness is therefore a significant challenge. Furthermore, LJP systems should serve as decision-support tools rather than replace human judgment, to protect legal rights and procedural fairness.

2.3 Legal Knowledge Modeling

Traditional rule-based approaches, dictionary-based models, and traditional machine learning techniques have limitations in capturing language semantics and complex linguistic structures, especially in legal domain (Medvedeva et al., 2020; Sulea et al., 2017).

To model legal knowledge methods, current approaches can be broadly categorized into three types: (1) legal knowledge graphs, (2) domain-specific pre-trained language models, (3) knowledge-enhanced models.

Legal knowledge graphs focus on constructing structured representations of legal entities and their relationships, serving as external knowledge resources for downstream tasks.

Domain-specific pre-trained language models (PLMs) aim to capture the linguistic and semantic patterns of legal texts through large-scale pre-training. PLMs can be applied independently for text representation, or the outputs can be further used in downstream models.

Knowledge-enhanced models refer to methods that integrate legal knowledge (such as domain rules, logical constraints, or features derived from legal texts) into the neural modeling process to improve performance. These approaches are not limited to the use of knowledge graphs or PLMs and can incorporate legal knowledge through explicit rule injection, data augmentation, prompt-based guidance, or other domain-specific strategies.

Although knowledge-enhanced models may incorporate knowledge graphs or PLMs, our categorization is based on the main purpose and methodology of each approach. Specifically, we distinguish the three types according to whether their primary goal is to construct structured knowledge (knowledge graphs), to learn representations from legal texts (PLMs), or to inject legal knowledge into the model's reasoning or prediction process (knowledge-enhanced models), regardless of whether they utilize particular external resources.

2.3.1 Legal Knowledge Graphs

The construction of a legal knowledge graph based on legal cases is beneficial for various legal tasks. S. Chen et al. (2019) introduced a legal knowledge graph and proposed a legal graph network for charge prediction. They defined two types of entities and four types of relationships for the legal knowledge graph. The entities include criminal charges and keywords, extracted from the charges' definitions. The relationships include: (1) Adjacency, which describes the adjacency relation among keywords in the charges' definitions. (2) Co-Group, which describes the relation among charges that have similar characteristics. Charges in the same law articles belong to the same group. (3) Co-Occurrence, which describes the co-occurrence relation among charges. Charges violated at the same time for the same fact description have this relationship. (4) Component-to-Charge, which describes the relationship between keywords and charges. Based on the legal knowledge graph, this method utilizes the HIN2VEC network (Fu et al., 2017) for legal knowledge graph network representation learning. Both legal knowledge graph representations and fact representations are fed into an attention-based neural network for charge prediction. Similarly, Gao et al. (2023) further explored the use of legal knowledge graphs for charge prediction. They proposed a fine-grained legal knowledge graph, defining 19 categories of legal entities and 11 categories of legal relationships. Then, based on annotated corpora, they developed a judicial long-text triplet extraction model to extract entities and relationships related to criminal behavior. The NetworkX tool[1] is utilized to construct a legal knowledge graph. To integrate legal knowledge graph into charge prediction, the model extracts key elements and case subgraphs by matching the fact description with the legal knowledge graph. Subsequently, TextCNN (Kim, 2014) and Graph Attention Network (Veličković et al., 2017) are used to encode the key elements and case subgraphs, respectively. Finally, the representations of key elements, case subgraphs, and fact descriptions are concatenated and used for charge prediction.

On the other hand, Tang et al. (2020) presented a system developed for constructing a legal knowledge graph using a semi-automatic approach. Using the Harvard Caselaw dataset as an example, the system employs NLP tools, such as SpaCy and LexNLP, for entity and relationship extraction. These triples form the basis of the legal knowledge graph. Additionally, the system utilizes Neo4j to manage the legal knowledge graph. Jain et al. (2022) aimed to construct a knowledge graph from the Indian Legal domain corpus using a framework that involves entity extraction, relation extraction, and triple construction. The preprocessing stage utilized GATE software tools for language processing, including POS tagging, sentence splitting, and tokenization, to prepare the text for entity and relation extraction. Then, the entity extraction stage used GATE software to identify key entities within the legal texts, such as Bench, Case Type, Court, and Court Decision. They also utilized NyOn which provides a predefined schema or structure reflecting the typical components and relationships found in Indian Supreme Court judgments, to aid in the extraction process. After entities were identified, the next phase focused on determining the relationships between them. Relations such as hasCourtOfficials, hasParties, and documentType were annotated based on their occurrences within the text, leveraging the structure provided by the NyOn ontology for guidance. The core of the knowledge graph construction involved creating RDF (Resource Description Framework) triples that represent the legal knowledge in a structured form. These triples, following a subject-predicate-object format, were then stored in a triple store, allowing for complex querying and further applications like

[1] `https://github.com/networkx/networkx.`

question answering and legal judgment analysis.

Legal knowledge graphs focus on constructing structured representations of legal entities and their relationships, serving as external resources for various downstream tasks. The main advantages and limitations are summarized as follows: legal knowledge graphs provide structured representations that support explainable legal AI. However, constructing and maintaining high-quality legal knowledge graphs requires substantial domain expertise, manual annotation, and ongoing updates to reflect evolving legal standards. Consequently, their scalability and adaptability may be limited, especially in dynamic or data-scarce legal domains.

2.3.2 Domain-specific Pre-trained Language Models

2.3.2(a) Traditional Legal Neural Network Models

In recent years, deep learning has gained attention in NLP tasks. Word2Vec (Mikolov, Chen, et al., 2013) is a technique used for generating word embeddings. In legal domain, it involves pre-trained word embeddings from legal corpora, learning the specific semantic relationships of vocabulary within this domain. This means that the unique vocabularies are converted into word vector representations. For example, Chalkidis and Kampas (2019) trained legal word embeddings using the Word2Vec approach on a large corpus of legal texts from various sources. These embeddings can be utilized in future studies for tasks such as LJP and legal information extraction (Le et al., 2020).

2.3.2(b) Legal Pre-trained Language Models

Pretrained language models (PLMs) such as BERT (Devlin et al., 2019a) has achieved impressive performance in several NLP tasks. Various studies have explored the use of PLMs to enhance the comprehension of legal documents and improve performance in downstream legal tasks.