IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 21 July 2023, accepted 5 August 2023, date of publication 14 August 2023, date of current version 21 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3305259

RESEARCH ARTICLE

Knowledge-Enriched Multi-Cross Attention Network for Legal Judgment Prediction

CONGQING HE[1], (Student Member, IEEE), TIEN-PING TAN[1], XIAOBO ZHANG[1], AND SHENG XUE[2]

[1]School of Computer Sciences, Universiti Sains Malaysia, Penang 11800, Malaysia
[2]Guangdong Research Institute, China Telecom Corporation Ltd. Research Institute, Guangzhou 510006, China

Corresponding author: Tien-Ping Tan (tienping@usm.my)

ABSTRACT Legal judgment prediction (LJP) automatically predicts the judgment results of a legal case based on its fact description, which has excellent prospects in judicial assistance systems and consultation services for the public. Most previous studies either focused on enhancing LJP's performance while ignoring the issue of confusing charges and law articles, or only used law articles to improve the judgment of confusing verdicts, resulting in the limited model performance. This paper introduces legal charge knowledge as a type of knowledge to enhance the representation of fact descriptions and incorporates it into deep neural networks. We then propose a Knowledge-enriched Multi-Cross Attention Network (KEMCAN) to improve LJP's performance, and resolve legal cases involving confusing charges and law articles. Specifically, a cross-attention mechanism is proposed to model the relationship between legal charge knowledge and fact description in a unified model. The experimental results demonstrate that our model outperforms the state-of-the-art methods on two real-world datasets, achieving an average improvement of 3.95% in macro-F1 for charge prediction and 1.98% for law article prediction.

INDEX TERMS Legal judgment prediction, legal charge knowledge, multi-cross attention, confusing charges and law articles.

I. INTRODUCTION

Legal judgment prediction (LJP) automatically predicts the judgment results of a legal case based on its fact description by utilizing machine learning techniques. The development of LJP aims at helping legal professionals improving work efficiency, also guiding people unfamiliar with the legal process and jargon. A comprehensive LJP typically encompasses several sub-tasks, including charge prediction, law article prediction, prison sentence prediction, and court view generation [1], [2], [3], [4], [5], [6]. In detail, charge prediction and law article prediction stand out as the most pivotal tasks within LJP, as they lay the foundation for subsequent prison sentence prediction and court view generation. Thus, this study focuses on these two critical aspects of LJP.

Figure 1 presents two legal cases, each legal case has a fact description, and the corresponding charge and law article. A legal charge is a formal accusation of criminal activity, whereas a legal article is a legal rule in a legal document. LJP aims to predict these charges and law articles based on the given fact description. The tasks of charge prediction and law article prediction are formalized as text classification problems and studied since [7]. Early research involved the application of mathematical models to predict judgment outcomes [8], [9], [10], [11]. However, the emergence of deep learning within the domain of natural language processing has prompted researchers to employ deep neural networks to tackle LJP problems [12]. Notably, studies indicate that deep learning neural networks significantly enhance the precision of LJP in comparison to conventional machine learning approaches [1], [12], [13].

One of the challenges for LJP is predicting the charges and legal articles for legal cases that have similar fact descriptions but different charges and legal articles, which are referred to as confusing legal cases [1], [14]. These cases typically have a high error rate due to the similarity in fact descriptions,

The associate editor coordinating the review of this manuscript and approving it for publication was Derek Abbott.

Case 1:

Fact Description:

On November 22, 2004, the criminal suspects Guo Xinpo, Wu Hongbiao, Zhang Yongqing, and Zhang Yongjun planned to snatch and rob depositors in front of the Tiexi branch of the Industrial and Commercial Bank of China located at the intersection of Zhongzhou Road and Qingfeng Street in Anyang City. At 10:50 a.m. that day, when depositors Zhao Fucheng and Zhao Xiurong took money out of the bank and were about to leave by car, Guo Xinpo, Wu Hongbiao, and Zhang Yongjun rushed up to snatch them, and openly carried out the robbery when the snatch failed, ...

Law Articles:

Article 263 Whoever robs public or private property by violence, coercion, or other methods shall be sentenced to fixed-term imprisonment of not less than three years but not more than 10 years and shall also be fined; .....

Charges:

Crime of Robbery

Case 2:

Fact Description:

The People's Procuratorate of Jiangyang District, Luzhou City accused the defendant, Zhu xx, of stealing the gold necklace the victim was wearing on his neck and fleeing the scene at victim's unpreparedness. The above happened at the sidewalk of the Great World Parking entrance on Zhiping Road, Jiangyang District, Luzhou City. It was identified that the gold necklace of the defendant Zhu XX was worth xx yuan. ..... The public prosecution agency accused the defendant, Zhu xx, for the purpose of illegal possession, and seizing other people's property with the relatively large amount at others' unpreparedness. His behavior violated the provisions of the Criminal Law of the People's Republic of China. Appeal to this court for a sentence.

Law Articles:

Article 267 Whoever robs a relatively large amount of public or private property or who robs it multiple times shall be sentenced to fixed-term imprisonment of not more than three years, criminal detention .....

Charges:

Crime of Forcible Seizure

FIGURE 1. Two similar legal cases and their respective fact descriptions, law articles, and charges are presented. The text in red is important points in a fact description that decide the judgment outcomes.

making it difficult to obtain accurate prediction results. For instance, Figure 1 presents two cases with similar fact descriptions, but with subtle differences in whether violence and coercion were used in the crime. It can be observed from the fact description of Case 1 that the defendant used a knife to slash the victim's head after the robbery failed. Therefore, the defendant is involved in the crime of robbery. The key to solving this problem is capturing the small but significant text in the fact description to determine the relevant charges and legal articles for the case. There are few attempts to solve this problem. Hu et al. proposed a multi-task learning model that considers ten distinct legal attributes (e.g., violence, death) for each charge [1]. The model can predict both the attributes and the charges simultaneously. However, this approach is limited by the requirement for expert-annotated attributes, which reduces its generalizability to other legal domains. Xu et al. developed a graph neural network capable of extracting the differences among similar law articles and mining the similarities between fact descriptions and law articles [14]. However, this approach heavily relies on legal articles to distinguish confusing charges, which limits its scope of application. Differentiating between misleading law articles with highly similar representations can be challenging. Li et al. proposed an approach to enhance the prediction of legal articles and prison terms by incorporating defendant persona and law articles as knowledge in case-specific semantic representations [15]. Nevertheless, the effectiveness of this approach in resolving confusing charges is limited due to the inherent limitations of defendant persona in understanding fact descriptions of behavior.

This study proposes the integration of legal charge knowledge into the LJP model to enhance its predictive capabilities for cases with confusing charges and law articles. The legal charge knowledge provides a comprehensive understanding of legal charges, including their definitions, subjective and objective elements of crime, subjects and objects of crime, and legal basis, which can be used to enhance the representation of fact descriptions. For example, in cases of confusing charges between *the crime of robbery* and *the crime of forcible seizure*, the LJP model can utilize the legal charge knowledge to capture the subtle differences between them, such as the requirement of violence or coercion in *the crime of robbery*. By incorporating this knowledge, the LJP model can achieve heightened accuracy in predicting charges and pertinent legal articles, thereby enhancing its overall predictive performance.

In the paper, we introduce a novel approach called Knowledge-enriched Multi-Cross Attention Network (KEMCAN), aimed at improving the accuracy of predicting legal judgments. First, KEMCAN utilizes a cross-attention mechanism to integrate the fact description and legal charge knowledge. This enables the model to leverage domain-specific prior knowledge, thereby enhancing its comprehension of legal texts. Second, the model aligns sentences within the fact description and legal charge knowledge, identifying the utmost pertinent knowledge for each sentence. Subsequently, this relevant knowledge is integrated into the sentence's representation, thereby amplifying comprehension.

This study was carried out on Chinese criminal law and the main contributions of this paper are presented as follows:

- We introduce the use of legal charge knowledge to enhance the comprehension of legal text. By incorporating knowledge about legal charges, including their definition, subjective and objective elements, and legal basis, the proposed methods improve the representation of fact descriptions and enable better differentiation between confusing charges and legal articles.
- We propose a novel approach called KEMCAN, which utilizes a cross-attention mechanism to integrate the fact description and legal charge knowledge. This enables the model to leverage domain-specific prior knowledge, thereby enhancing its comprehension of legal texts.
- We conduct extensive experiments on two real-world datasets to evaluate KEMCAN, and compare the results with other state-of-the-art methods.

The subsequent sections of this paper are structured as follows: Section II provides an overview of related work on legal judgment prediction and attention mechanisms. In Section III, we introduce a knowledge-enriched multi-cross attention network. Section IV outlines the experimental dataset, evaluation metrics, baseline methods, and experimental setting. Next, Section V presents the experimental results and provides a discussion. In Section VI, we present the conclusion and discuss future work. Lastly, Section VII discusses the limitations of this research.

II. RELATED WORK

Legal judgment prediction (LJP) has been studied for decades and has achieved a lot of progress. Early approaches used conventional machine learning with limited success [8], [9], [10], [11], and deep learning methods have achieved state-of-the-art results in recent years [1], [12].

A. CONVENTIONAL MACHINE LEARNING METHODS

Earlier studies mainly focused on the application of mathematical and statistical methodologies to analyze legal tasks [8], [9], [10], [11]. However, these studies were constrained by a scarcity of datasets and limited labels.

Subsequent advancements in machine learning have motivated researchers to apply machine learning techniques in the field of LJP. For instance, Liu et al. introduced a case-based reasoning (CBR) system that classified cases into 12 charges using a blend of pre-defined crime rules and the k-nearest neighbor algorithm (KNN) [16]. To distinguish between similar cases more effectively, Lin et al. focused on robbery and intimidation criminal-related cases [7]. They used Liblinear and Logistic model tree to predict the charges and sentences of the cases, using a set of 21 legal factor labels for feature engineering. Katz et al. proposed a time-evolving random forest classifier to predict the justice vote (Affirm, Reverse, Other) and case outcomes of the Supreme Court of the United States by a self-developed feature engineering [17]. Similarly, Sulea et al. extracted word unigrams and word bigrams from fact description, and used an ensemble system with multiple SVM classifiers to predict the case's legal area and ruling [18]. Medvedeva et al. constructed features using n-grams and TF-IDF and employed SVM as a classifier to predict whether a case involved violations or non-violations [19].

B. DEEP LEARNING BASED METHODS

Recently, there has been an increase in the interest among researchers in applying neural networks to solve LJP tasks. The deep learning approaches can be divided into single-task learning and multi-task learning, depending on whether the approach simultaneously solves one or many tasks.

1) SINGLE TASK LEARNING

Luo et al. introduced a hierarchical attention-based network that aims to enhance charge prediction performance by jointly learning the relationship between fact descriptions and relevant legal articles [12]. To distinguish confusing charges and few-shot charges based on fact description, Hu et al. proposed a multi-task learning network to learn the tasks of attributes and charges simultaneously, by introducing ten representative attributes of charges (i.e., violence, death) [1]. Le et al. devised a self-attentive capsule network to capture the representation of fact descriptions and introduced the focal loss to alleviate the problem of imbalanced charges [2], [20]. Li et al. proposed a law article de-duplication attention neural network for charge prediction by incorporating fact description and relevant legal articles [21].

2) MULTI-TASK LEARNING

Multi-task learning in LJP is the joint modeling of tasks such as charge prediction and law article prediction. Zhong et al. simulated the decision-making rationale of human judges, capturing interdependencies between tasks encompassing law articles, charges, and penalty terms [22]. To this end, they introduced a multi-task learning framework aimed at simultaneously predicting law articles, charges, and penalty terms. Similarly, Yang et al. also addressed the interrelationships between law article, charge, and penalty terms in LJP. They proposed a multi-perspective bi-feedback multi-task learning framework (MPBFN for short) for LJP [23]. Wang et al. modeled charge prediction and law article prediction as a tree-shape hierarchical structure with the parent label of charges and children label of law articles, and proposed

a hierarchical matching network by fusing the hierarchical structure and semantics of labels [24]. Li et al. proposed an approach to enhance the prediction of legal articles and prison terms by incorporating defendant persona and law articles as knowledge in case-specific semantic representations [15]. This method employed BiGRU-based sequence encoders to generate attention vectors for the facts-channel, articles-channel, and personas-channel, and then used a dynamic mechanism to fuse information from each channel. Xu et al. proposed a legal article distillation-based attention network by mining similarities between fact descriptions and legal articles, to solve the problem of confusing charges and law articles [14]. Considering the fact description has different impacts on LJP subtasks (i.e., charge prediction, law article prediction, and term of penalty prediction.), Yue et al. presented a circumstance-aware framework, utilizing the outputs of inter-mediate subtasks to separate the fact description [13]. In addition, they employed a label-embedding method to incorporate the semantics of charge labels and law article labels into fact descriptions to generate more expressive fact representations for clarifying confusing charges and law articles. Yang et al. proposed a multi-task legal judgment prediction framework, MVE-FLK, which utilizes a multi-view encoder to fuse legal keywords and jointly model multiple subtasks in legal judgment prediction [25]. Zhang et al. proposed a supervised contrastive learning framework for legal judgment prediction (LJP) to improve the accuracy of predicting judgment results in legal cases [26].

In spite of the advancements achieved by prior methods, they have ignored the valuable legal charge knowledge that is inherent to LJP. To fill this gap, we leverage legal charge knowledge obtained from online resources and propose a multi-cross attention mechanism to improve the performance of LJP.

C. ATTENTION MECHANISMS

Attention mechanisms have gained widespread adoption across various NLP tasks, including machine translation, text classification, and text summarization.

Vaswani et al. proposed the Transformer model, which utilizes self-attention mechanisms to capture long-range dependencies in text [27]. Devlin et al. developed BERT, a pre-trained language model that leverages self-attention mechanisms to achieve state-of-the-art performance on various NLP tasks [28]. Besides, Chen et al. used a dual attention mechanism to integrates representations of short texts with prior knowledge from external sources to improve classification accuracy [29]. Ying et al. utilized a multi-modal cross-attention network to jointly model the inter-modality and intra-modality relationships of image regions and text fragments, a multi-level encoding network to model and jointly learn the abundant multi-level semantics of the multi-modal content, and a fake news classification network to classify each post on social multimedia as fake or real news [30]. Luong et al. explored two effective approaches, global and local attention, for improving neural machine translation (NMT) by selectively focusing on parts of the source sentence during translation [31].

In legal domain, Hu et al. proposed a multi-task learning model for charge prediction that employs an attribute-based attention mechanism to jointly learn attribute-free and attribute-aware fact representations [1]. Li et al. presented a multichannel attentive network for LJP. The proposed framework, MANN, utilizes BiGRU-based sequence encoders to generate attention vectors for facts-channel, articles-channel, and personas-channel. Moreover, MANN incorporates a dynamic mechanism to generate context attention vectors, which are guided by other channels [15]. Li et al. used a hierarchical Bi-GRU encoder with word collocation attention mechanism to generate fact embeddings and introduced a difference aggregation mechanism among similar law articles for extracting effective distinguishable features [21]. LADAN employs a graph neural network and a attention mechanism to distinguish confusing law articles and extract discriminative features from fact descriptions [14]. The attention mechanism is utilized to extract distinguishable features from fact descriptions by attentively exploring the differences among similar law articles.

This paper introduces an attention mechanism that models the relationship between legal charge knowledge and fact description within a unified framework. This mechanism enhances the model's ability to capture the alignment between the fact descriptions and the legal charge knowledge, identify the utmost pertinent knowledge for each sentence, and integrate this relevant knowledge into he sentence's representation.

III. METHODOLOGY

We present a novel approach, the Knowledge-enriched Multi-Cross Attention Network (KEMCAN) for LJP. KEMCAN takes a fact description as input and assumes there is only one relevant law article and legal charge as output [22]. The architecture of KEMCAN is depicted in Figure 2 and comprises four modules: the `Fact Representing`, `Knowledge Representing`, `Multi-Cross Attention`, and `Output Layer`. The `Fact Representing` module extracts sentence vectors from the fact description. The `Knowledge Representing` module takes the legal charge knowledge as input, and obtains the knowledge representation through sentence embedding. Next, the `Multi-Cross Attention` module employs a cross-attention mechanism to merge the fact representation and the knowledge representation. These representations are then fed into a fully connected layer. Finally, the `Output Layer` predicts the charge and law article simultaneously. The details of different components are described in the following sections.

A. LEGAL CHARGE KNOWLEDGE CONSTRUCTION

In this subsection, we introduce legal charge knowledge to LJP. The intuition of the approach is that a judge determines

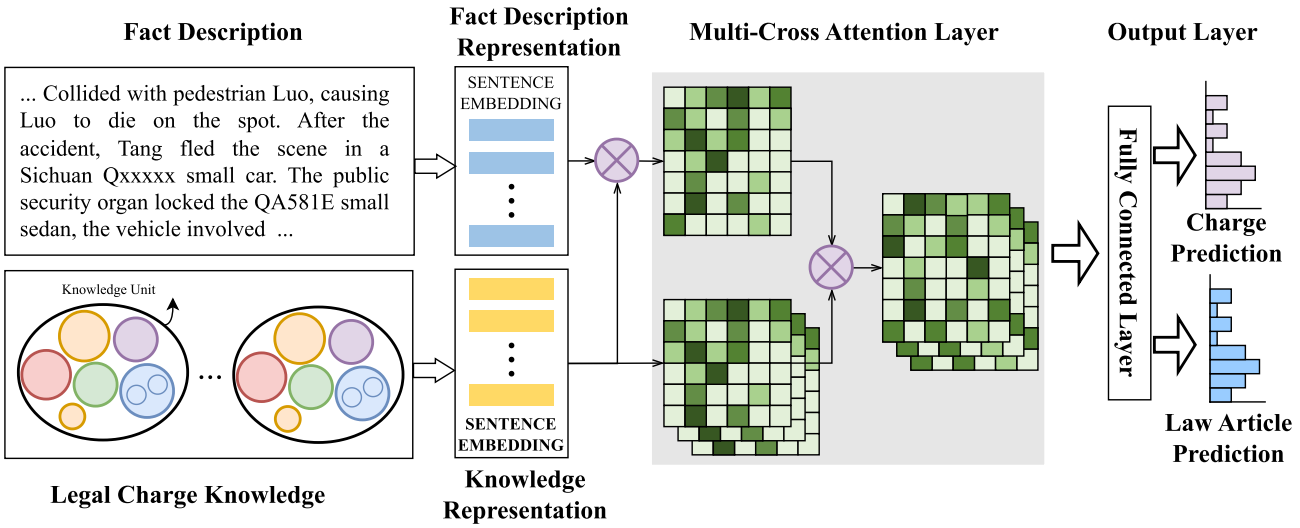


FIGURE 2. The overall framework of the proposed KEMCAN: (i) It takes the fact description of a legal case and the constructed legal charge knowledge as inputs. (ii) The fact description is processed by the "FACT DESCRIPTION REPRESENTATION" module to generate sentence embeddings, whereas the legal charge knowledge is processed by the "KNOWLEDGE REPRESENTATION" module to obtain knowledge embeddings. (iii) The multi-cross attention layer, depicted by the gray box, is employed to fuse the sentence embeddings and knowledge embeddings, and learn the alignment between sentences and knowledge. (iv) These features are then passed through a fully connected layer to calculate the probabilities of charges and law articles respectively.

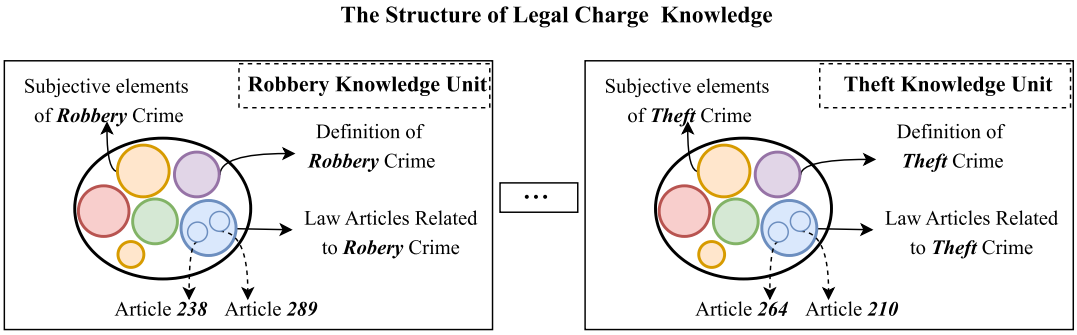


FIGURE 3. The legal charge knowledge composes of knowledge units. Each knowledge unit is described as a six-tuple that consists of the definition, subjective elements of crime, subject of crime, objective elements of crime, object of crime, and legal basis.

whether a party is guilty of a specific charge and has violated a legal article, based on the information about the legal charge. Legal charge knowledge is defined as a six-tuple, consisting of definition, subjective elements of the crime, subject of the crime, objective elements of the crime, object of the crime, and legal basis. The detailed explanation is shown in Table 1. Notably, a legal basis has the capacity to encompass several interconnected law articles, with each individual law article being linked to multiple knowledge units.

We manually collected legal charge knowledge for 129 knowledge units from an online Chinese law repository.[1] Specifically, we utilized a web crawler to extract texts that explain the details of legal charge knowledge, and then applied regular expressions to extract the six-tuple elements corresponding to the charges, with a maximum of 64 tokens for each element. Two examples of knowledge unit for the Crime of Robbery and Crime of Theft are shown in Figure 3. The proposed model captures comprehensive features in legal charge knowledge that enable the distinction of similar charges, such as the crime of robbery and the crime of theft. For instance, the KEMCAN considers factors such as the presence of violence or other criminal means during the crime, and the direct or illegal occupation of public or private property, etc. Therefore, incorporating knowledge units enhances the model's predictive performance.

B. PROBLEM FORMULATION

Each legal case is comprised of a fact description and the corresponding judgment outcome, including the legal charge and law article. The LJP system aims to predict both the charge and law article associated with each case. Following [22],

[1] https://china.findlaw.cn/zuiming/12_729.html

TABLE 1. The descriptions of elements in legal knowledge architecture.

Notation	Description
Definition	The concept of crime is a high-level generalization of the nature or main characteristics of a specific crime.
Subjective elements of crime	The psychological attitude of the criminal subject to the harmful behavior and the harmful result of the victim, including intention, negligence and purpose.
Subject of crime	The subject of crime refers to a person who has reached the legal age, has the ability to be responsible, and commits acts that endanger society. The unit can also become the subject of some crimes.
Objective elements of crime	The objective elements of crime refers to the objective external manifestations of criminal activities, which mainly include harmful behavior, harmful results, causality, etc.
Object of crime	The object of crime refers to the social relationship protected by the criminal law of our country and violated by the criminal act.
Legal basis	Legal basis are the basic elements of normative legal documents.

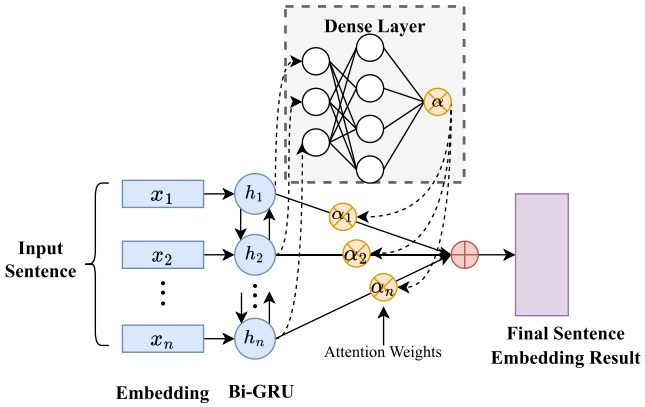


FIGURE 4. The structure of Sentence Embedding. The input to the Sentence Embedding is a sentence, denoted as $S = (\mathbf{x_1}, \mathbf{x_2}, \cdots, \mathbf{x_n})$, which is initialized with word embeddings. (i) The word embedding vectors are then fed into a Bi-GRU to generate hidden representations, denoted as $\mathbf{h} = (\mathbf{h_1}, \mathbf{h_2}, \cdots, \mathbf{h_n})$. Refer to Eq. (2), Eq. (3). (ii) The hidden representations $\mathbf{h}$ from the sequence are passed through a dense layer to generate attention weights, denoted as $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_n)$. Refer to Eq. (4), Eq. (5). (iii) The vector $\mathbf{r}$ is computed as the weighted average of $\mathbf{h}$, with weights determined by α. Refer to Eq. (6).

we assume that each case has only one relevant law article and charge.

Formally, we represent a set of $\mathcal{N}$ legal cases as $\mathcal{T} = (\mathcal{F}_i, y_i^{\texttt{charge}}, y_i^{\texttt{article}})_{i=1}^{\mathcal{N}}$, where $\mathcal{F}_i$ is the fact description for case i, $y_i^{\texttt{charge}}$ is the legal charge for case i, and $y_i^{\texttt{article}}$ is the law article for legal case i. During training, the model $\mathcal{M}(\{\mathcal{F}_i, \mathcal{G}\}_{i=1}^{\mathcal{N}})$ learns to predict the judgment results, specifically the predicted charges $\hat{y}_i^{\texttt{charge}}$ and law articles $\hat{y}_i^{\texttt{article}}$.

In other words, we have the following mapping:

$$\mathcal{M}(\{\mathcal{F}, \mathcal{G}\}) \Rightarrow (\hat{y}_{\texttt{charge}}, \hat{y}_{\texttt{article}}) \tag{1}$$

For example, as shown in Case 1 of Figure 1, the $\mathcal{M}$ model predicts the charge as Crime of Robbery and the law article as Article 263.

C. FACT DESCRIPTION REPRESENTATION

The `Fact Description Representation` module employs Sentence Embedding to convert each sentence of the fact description into corresponding sentence vectors. As depicted in Figure 4, this module specifically uses a Bidirectional Gated Recurrent Unit (Bi-GRU) to extract contextual information for each sentence. A Bi-GRU consists of a forward GRU, denoted as $\overrightarrow{f}$, which reads the sentence $\mathcal{S}i$ from **xi1** to $\mathbf{x_{in}}$, and a backward GRU, denoted as $\overleftarrow{f}$, which reads in the reverse direction, from $\mathbf{x_{in}}$ to $\mathbf{x_{i1}}$.

$$\overrightarrow{\mathbf{h_{it}}} = \overrightarrow{f}(\mathbf{x_{it}}), t \in [1, n], \tag{2}$$

$$\overleftarrow{\mathbf{h_{it}}} = \overleftarrow{f}(\mathbf{x_{it}}), t \in [n, 1]. \tag{3}$$

The two output hidden states $\overrightarrow{\mathbf{h_{it}}}$ and $\overleftarrow{\mathbf{h_{it}}}$ are concatenated $\mathbf{h_{it}} = [\overrightarrow{\mathbf{h_{it}}}, \overleftarrow{\mathbf{h_{it}}}]$ subsequently.

Considering that not all words are equally important in a sentence and contribute equally to the sentence vector, the attention mechanism is introduced in this subsection to aggregate weights to the sentence embedding, as follows [32].

$$\mathbf{u_{it}} = \tanh(\mathbf{W_t} \cdot \mathbf{h_{it}} + \mathbf{b_{it}}) \tag{4}$$

$$\alpha_{it} = \frac{\exp(\mathbf{u_{it}})}{\sum_t \exp(\mathbf{u_{it}})} \tag{5}$$

$$\mathbf{r_i} = \sum_t \alpha_{it} \cdot \mathbf{h_{it}} \tag{6}$$

The vector $\mathbf{h_{it}}$ is passed through a dense layer to derive a hidden representation $\mathbf{u_{it}}$, where $\mathbf{W_t}$ and $\mathbf{b_t}$ represent the weight matrix and bias, respectively. Then a normalized importance weight α_{it} is obtained using a softmax function, which measures the significance of the word. After that, the sentence vector $\mathbf{r_i}$ is computed as the weighted sum of the bidirectional GRU hidden states.

D. KNOWLEDGE REPRESENTATION

The `Knowledge Representation` module encodes legal charge knowledge as vectors. Formally, a legal charge knowledge $\mathcal{G} = (G_1, G_2, \cdots, G_l)$ consisting of l knowledge units, and each knowledge unit $G_i = (E_1, E_2 \cdots, E_k)$, where $k = 6$ in this paper. Each element E_i is represented as $E_i = (x_{i1}, x_{i2} \cdots, x_{im})$, where x_{ij} denotes the jth word in element E_i.

Each element in legal charge knowledge is then encoded as a vector using the same sentence embedding applied in the `Fact Description Representation` module. Then, the same attention mechanism is employed to assign different weights to the hidden output states, generating the sentence vector $\mathcal{V} = \{\mathbf{v_1}, \mathbf{v_2}, \cdots, \mathbf{v_k}\}$. Refer to equation (4)–(6).

E. MULTI-CROSS ATTENTION LAYER

While the Fact Description Representation module and Knowledge Representation module have vectorized the fact description and legal charge knowledge respectively, they do not explore the relationship between elements in legal charge knowledge and sentences in the fact description.

Although the Fact Description Representation module and Knowledge Representation module have vectorized the fact description and the legal charge knowledge respectively, the relationship between elements in legal charge knowledge and the sentence in fact description is not explored. This section introduces a unified model that utilizes our proposed multi-cross attention mechanism to model both legal charge knowledge $\mathcal{G}$ and fact description $\mathcal{F}$.

Specifically, each knowledge unit G_i and the fact description $\mathcal{F} = (\mathcal{S}_1, \mathcal{S}_2, \cdots, \mathcal{S}_T)$ are fed into a cross-attention mechanism (Abbreviated as CoAtt) to compute the attention weight. Subsequently, a max pool layer is utilized to get the most relevant features.

$$\texttt{CoAtt}(\mathcal{G}, \mathcal{F}) = [\texttt{CoAtt}(G_1, \mathcal{F}), \cdots, \texttt{CoAtt}(G_i, \mathcal{F}), \cdots, \texttt{CoAtt}(G_l, \mathcal{F})] \tag{7}$$

$$\mathcal{O} = \texttt{MaxPool}(\texttt{CoAtt}(\mathcal{G}, \mathcal{F})) \tag{8}$$

1) CROSS-ATTENTION MECHANISM

The cross-attention mechanism takes a fact description $\mathcal{F}$ and a knowledge unit G_i as the input, where the fact description and knowledge unit are represented by the Fact Description Representation module and Knowledge Representation module respectively, to obtain the sentence vectors of the fact description $\mathcal{R} = \{\mathbf{r_1}, \mathbf{r_2}, \cdots, \mathbf{r_n}\}$ and the element vectors of knowledge unit $\mathcal{V} = \{\mathbf{v_1}, \mathbf{v_2}, \cdots, \mathbf{v_k}\}$. Then $\mathcal{R}$ and $\mathcal{V}$ are passed into a inner product attention to calculate the similarity.

$$\omega_{ij} = f(\mathbf{r_i}, \mathbf{v_j}), i \in [1, n], j \in [1, k] \tag{9}$$

$$\beta_{ij} = \frac{\exp(\omega_{ij})}{\sum_{i=1}^{n} \exp(\omega_{ij})} \tag{10}$$

Here, the scoring function $f(\cdot)$ is the inner product function. β_{ij} represents the weight of attention from the element E_j in the knowledge unit to the sentence $\mathcal{S}_i$ in the fact description.

We applied the scaled dot-product attention mechanism to measure the relative importance of each element to the knowledge unit as follows [27]:

$$\boldsymbol{\rho} = \texttt{softmax}\frac{(\mathcal{V} \cdot \mathbf{W_1})(\mathcal{V} \cdot \mathbf{W_2})^T}{\sqrt{d_b}} \tag{11}$$

Here $\boldsymbol{\rho}$ denotes the attention weight of the knowledge unit G_i. The weight matrices $\mathbf{W_1} \in \mathbb{R}^{d \times d_b}$ and $\mathbf{W_2} \in \mathbb{R}^{d \times d_b}$ are randomly initialized and updated during training.

After that, $\boldsymbol{\beta}$ and $\boldsymbol{\rho}$ are combined by the inner product $f(\cdot)$ to obtain the final attention weight $\boldsymbol{\gamma}$.

$$\boldsymbol{\gamma} = f(\boldsymbol{\beta}, \boldsymbol{\rho}) \tag{12}$$

The final representation $\mathbf{q}$ is computed as the inner product between the sentence representation $\mathbf{r_i}$ in the fact description and the final attention weight $\boldsymbol{\gamma}$.

$$\mathbf{q} = \texttt{softmax}(f(\boldsymbol{\gamma}, \mathbf{r_i})) \tag{13}$$

F. OUTPUT LAYER

The final representation $\mathcal{O}$ is fed through a linear layer with a ReLU activation function, yielding the fully feed-forward representations $\mathcal{O}_{\texttt{FC}}$. Subsequently, the obtained representations $\mathcal{O}_{\texttt{FC}}$ are passed through another linear layer with a softmax activation function, enabling the calculation of predicted probabilities for both charges and law articles.

$$\hat{y}_{\texttt{charge}} = \texttt{softmax}(\mathbf{W_c}\mathcal{O}_{\texttt{FC}} + \mathbf{b_c}), \tag{14}$$

$$\hat{y}_{\texttt{article}} = \texttt{softmax}(\mathbf{W_a}\mathcal{O}_{\texttt{FC}} + \mathbf{b_a}). \tag{15}$$

$\mathbf{W_c}$, $\mathbf{b_c}$, $\mathbf{W_a}$ and $\mathbf{b_a}$ are the trainable parameters. Taking the loss sum of all sub-tasks obtained by cross-entropy loss functions as the overall prediction loss:

$$\mathcal{L} = \mathcal{L}_{\texttt{charge}} + \mathcal{L}_{\texttt{article}} \tag{16}$$

IV. EXPERIMENTS

A. DATASET

The experiments were conducted with publicly available datasets from the Chinese AI and Law challenge (CAIL2018) [33]. CAIL2018[2] is a large-scale Chinese legal dataset for criminal judgment prediction, consisting of the CAIL-SMALL dataset and CAIL-BIG dataset. CAIL-SMALL and CAIL-BIG are the exercise stage dataset and the first stage dataset in CAIL2018, respectively. Each legal case within the dataset consists of a fact description and judgment results, including the charge, law article, and term of penalty. For a fair comparison with existing state-of-the-art methods [13], [14], the data preprocessing pipeline applied is consistent with the earlier works.

Table 2 presents the statistical analysis of the two datasets after preprocessing. Furthermore, the distribution of charge labels and law article labels is visualized in Figure 5 and Figure 6, respectively. The figures demonstrate that the distribution of charge labels and law article labels are highly imbalanced.

TABLE 2. Statistics of the legal document dataset.

Dataset	CAIL-SMALL	CAIL-BIG
Training Dataset	108 619	1 593 982
Test Dataset	26 120	185 721
Charges	99	118
Law Articles	115	129
Avg. words in legal cases	173	149
Avg. Sentences in legal cases	6	5

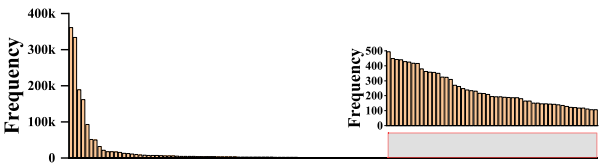


FIGURE 5. The distribution of charge labels in CAIL-BIG dataset.

[2]http://wenshu.court.gov.cn/

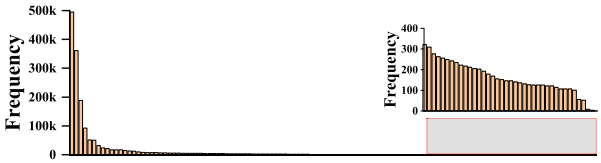


FIGURE 6. The distribution of law articles labels in CAIL-BIG dataset.

B. METRICS

The Charge prediction and law article prediction are imbalanced multi-class classification tasks. Therefore, we utilized accuracy (Acc), macro-precision (MP), macro-recall (MR), and macro-F1 (MF) to evaluate our proposed model.

C. BASELINE METHODS

In this subsection, two types of representative baselines were applied to measure the performance of LJP model. The KEMCAN method was compared with conventional text classification methods before, and with several state-of-the-art LJP based methods consequently.

- SVM+word2vec is a text classification method that utilizes Word2Vec to represent word features, and SVM for classification [34], [35],
- LSTM-MTL employs a two-layer LSTM with a max-pooling operation to encode fact descriptions, and a softmax as a classifier [36].
- HAN is a strong baseline for text classification which uses two levels of attention mechanisms at the word-level and sentence-level representations [37].
- FLA is an attention-based neural network that models the interplay between fact descriptions and the relevant laws [12].
- TOPJUDGE is a topological multi-task learning model that captures subtask relationships in LJP [22].
- Few-Shot is an attribute-aware model that leverages charge attributes to improve the fact representation in order to relieve the difficulty of confusing charges [1].
- LADAN is an attention-based model that uses a graph distillation operator to distinguish confusing verdicts with the learning of legal articles [14].
- NeurJudge employs a label embedding technique to incorporate the semantics of labels into fact descriptions to distinguish confusing verdicts problems [13].
- BERT is used to produce contextualized word embeddings [28]. It has demonstrated superior performance on various natural language processing (NLP) tasks. Given that our experiments involve Chinese datasets, we adopted the Chinese BERT model trained by as our baseline method [38].
- BERT-Crime is a variant of BERT that has been pre-trained using crime data [39].

D. EXPERIMENTAL SETUP

The fact description was segmented into sentences using symbols, with a maximum of 16 sentences. Each sentence was tokenized using THULAC [40], with a maximum of 64 tokens per sentence. The legal charge knowledge was tokenized using THULAC, with a maximum of 64 tokens per element. The training dataset was randomly split into 90% for training and 10% for validation. Our proposed models were implemented using TensorFlow[3] and trained on NVIDIA 3090 GPU.

A word embedding model with a dimension of 200 was pre-trained from the CAIL-SMALL/BIG training dataset using the Word2Vec algorithm [34]. The Bi-GRU models in the Fact Description Representation and Knowledge Representation modules had hidden units of size 200. The Adam optimizer was used for training [41], with a batch size of 128 and a learning rate of 10^{-3}. For methods based on BERT, we employed a pre-trained Chinese BERT model developed by Cui et al. [38], and configured it to allow a maximum of 16 sentences, and each sentence was limited to 128 tokens. During the experiments, we repeated the model runs five times for each dataset and calculated the average values of the model results. The model was trained for 30 epochs, and the latest model was evaluated on the validation set at each epoch. The model with the highest accuracy was saved during training.

V. RESULT AND ANALYSIS

In this section, we reported the main experimental results of both the baseline models and the proposed method for charge prediction and law article prediction. We subsequently conducted a more detailed analysis of the models' performance by investigating the charges and law article labels in long-tailed learning. Finally, we evaluated the performance of each module, namely the `Knowledge Representation` module, `Multi-Cross Attention Layer` and different loss functions.

A. MAIN EXPERIMENTAL PERFORMANCE

1) PERFORMANCE ON CHARGE PREDICTION

Table 3 shows the charge prediction results of both baseline approaches and the proposed approach on CAIL-SMALL and CAIL-BIG datasets. The first three rows after the heading show the results of conventional text classification methods. The five subsequent rows show the results of the state-of-the-art LJP-based methods, while the result of the proposed `KEMCAN` method is presented in the last row.

`KEMCAN` significantly outperforms all baseline approaches, achieving 1.4% and 6.5% higher MF on the CAIL-SMALL and CAIL-BIG datasets, respectively, compared to the state-of-the-art `NeurJudge`. In particular, the following conclusions may be drawn from the findings: (1) All deep learning-based models perform better than `SVM+word2vec`. One probable explanation is that `SVM+word2vec` approach fails to capture the intricate interactions between fact descriptions and labels. (2) Traditional deep learning text classification methods such as `LSTM-MTL`,

[3] www.tensorflow.org

TABLE 3. Charge prediction results on both CAIL-SMALL dataset and CAIL-BIG dataset, where the best result for each metric is highlighted in bold.

	CAIL-SMALL dataset				CAIL-BIG dataset			
Methods	Acc	MP	MR	MF	Acc	MP	MR	MF
SVM+word2vec [35]	83.37	80.78	77.30	78.25	92.09	82.26	65.28	69.06
LSTM-MTL [36]	84.84	81.40	79.16	79.88	92.48	79.42	74.25	75.48
HAN [37]	85.09	83.18	80.08	79.95	93.02	81.60	76.53	76.53
FLA [12]	84.72	83.71	73.75	75.04	93.01	76.56	72.75	72.94
TOPJUDGE [22]	86.48	84.23	78.39	80.15	93.19	79.44	75.52	75.50
Few-Shot [1]	88.15	87.51	80.57	81.98	93.24	80.59	76.62	76.89
LADAN [14]	88.28	86.36	80.54	82.11	93.26	81.21	77.65	77.60
NeurJudge [13]	89.92	87.76	86.75	86.96	95.57	85.57	78.81	80.54
KEMCAN (ours)	**90.57**	**88.90**	**87.75**	**88.20**	**96.32**	**88.99**	**83.75**	**85.76**

TABLE 4. Law articles prediction results on both CAIL-SMALL dataset and CAIL-BIG dataset, where the best result for each metric is highlighted in bold.

	CAIL-SMALL dataset				CAIL-BIG dataset			
Methods	Acc	MP	MR	MF	Acc	MP	MR	MF
SVM+word2vec [35]	84.17	80.74	75.96	77.09	92.62	77.92	61.03	64.29
LSTM-MTL [36]	85.50	80.12	77.73	78.51	93.15	76.96	71.44	71.75
HAN [37]	85.56	82.97	79.22	79.62	93.31	78.72	72.75	73.50
FLA [12]	85.63	83.46	73.83	74.92	93.51	74.94	70.40	70.70
TOPJUDGE [22]	87.28	85.81	76.25	78.24	93.24	74.24	71.19	70.40
Few-Shot [1]	88.44	86.76	77.93	79.51	93.74	78.51	73.79	74.18
LADAN [14]	88.78	85.15	79.45	80.97	93.27	75.10	72.04	71.26
NeurJudge [13]	90.37	87.22	85.82	86.13	95.58	82.01	77.05	78.05
KEMCAN (ours)	**90.65**	**88.04**	**85.97**	**86.87**	**96.37**	**84.87**	**79.13**	**81.27**

HAN yield promising results. However, most LJP-based methods perform better, indicating certain limitations of traditional text classification methods in addressing LJP tasks. (3) Both Few-Shot and KEMCAN introduced legal charge knowledge to model the relationship between fact description and charges. However, The results of KEMCAN are better than Few-Shot, suggesting that KEMCAN better models the proposed relationship between fact description and knowledge. (4) In comparison to FLA and LADAN, which utilize legal articles as auxiliary knowledge, KEMCAN achieves better performance, using legal charge knowledge as auxiliary knowledge. This can be attributed to its superior modeling of the relationship between fact description and legal charge knowledge. (5) On the CAIL-BIG dataset, the MF of the evaluated methods is lower compared to the CAIL-SMALL dataset, while the accuracy is higher. This discrepancy can be attributed to the imbalanced classes in the CAIL-BIG dataset.

2) PERFORMANCE ON LAW ARTICLE PREDICTION

Table 4 presents the law article prediction results obtained using various models on two datasets. The proposed KEMCAN model consistently outperforms all other methods on both datasets. In comparison to the state-of-the-art NeurJudge, KEMCAN achieves higher MF in law article prediction, with improvements of 0.74% and 3.22% on the CAIL-SMALL and CAIL-BIG datasets, respectively.

The following conclusions can be drawn from the Table 4: (1) The performance of NeurJudge and KEMCAN on the CAIL-SMALL dataset is comparable, but KEMCAN yields superior results on the CAIL-BIG dataset. This indicates that our method, which incorporates legal charge knowledge and designs an attention fusion mechanism, has proven effective. (2) FLA, LADAN, and KEMCAN all incorporate knowledge from legal domains to enhance legal fact description representation. However, KEMCAN achieves better results, attributed to its superior modeling of the relationship between fact description and legal charge knowledge. FLA, a two-stage approach, filters out a significant portion of irrelevant law articles, leading to error propagation and thus, inferior performance. By introducing legal charge knowledge as auxiliary information, KEMCAN unifies the modeling of fact descriptions and legal charge knowledge, resulting in improved performance over FLA. (3) FLA, LADAN, and KEMCAN outperform TOPJUDGE in law article prediction. This suggests that incorporating legal knowledge into the deep learning model is more beneficial for law article prediction than merely mining the associations between labels. (4) Similar to charge prediction results, the MF of the evaluated methods on the CAIL-BIG dataset was lower than that on the CAIL-SMALL dataset for law article prediction. However, accuracy was higher on the CAIL-BIG dataset. This could be due to the more imbalanced law article labels on the CAIL-BIG dataset, resulting in poorer prediction results for many difficult labels.

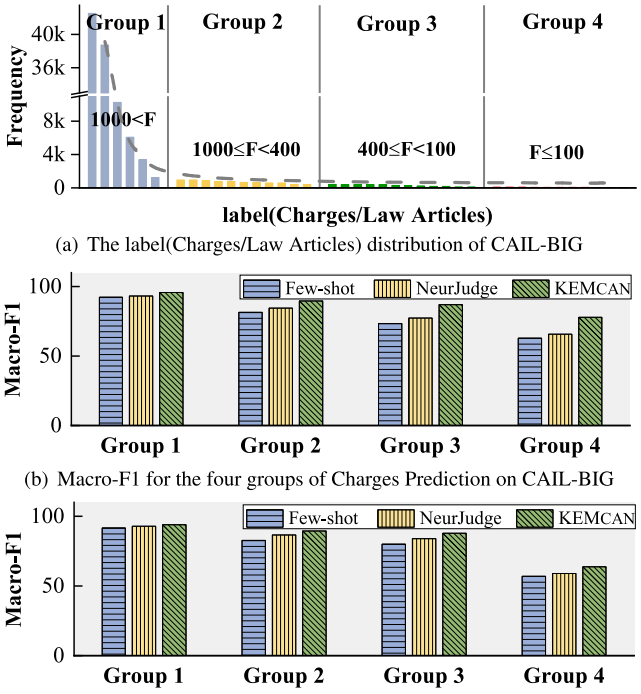
B. PERFORMANCE ON BERT-BASED METHODS

To further validate the effectiveness of our BERT-based model, we conducted a comparative experiment against alternative models. Due to the large size of the CAIL-BIG dataset and the time-intensive training process, we conducted the experiment using the CAIL-SMALL dataset. The experimental results are shown in Table 5. Importantly, the BERT-based model outperformed the Word2Vec+GRU model, demonstrating the superior effectiveness of pre-trained models.

TABLE 5. Charge prediction and law articles prediction results on CAIL-SMALL dataset (BERT-based Methods).

Methods	Charges Prediction		Law Articles Prediction	
BERT	90.68	87.69	90.81	86.06
BERT-Crime	91.26	87.81	91.30	85.70
BERT-NeurJudge	92.74	90.60	92.60	88.33
BERT-NeurJudge+	92.91	90.89	92.64	88.75
BERT-KEMCAN	**93.22**	**91.10**	**93.06**	88.74

Although BERT model showed promising performance in LJP tasks, our proposed BERT-KEMCAN model outperforms its performance, validating the effectiveness of our proposed model.



(a) The label(Charges/Law Articles) distribution of CAIL-BIG

(b) Macro-F1 for the four groups of Charges Prediction on CAIL-BIG

(c) Macro-F1 for the four groups of Law Article Prediction on CAIL-BIG

FIGURE 7. (a) The figure splits the charges/law articles to 4 groups depending on the frequency of the label available in CAIL-BIG test dataset. (b) and (c) compare the Macro-F1 performance of different models in different frequency groups.

C. EFFECTS IN LONG-TAILED SCENARIOS

As shown in Figure 7(b) and Figure 7(c), the performance gap between NeurJudge and KEMCAN increases from the high-frequency group to the low-frequency group. The MFs of KEMCAN are 18.26% and 8.23% higher in charge prediction and law article prediction respectively when compared with the NeurJudge. In addition, compared with the Few-Shot that focuses on low-frequency charges task, the MFs of KEMCAN are 23.90% and 12.11% higher on charge prediction and law article prediction respectively. In conclusion, the results show that the KEMCAN performs better than NeurJudge and Few-Shot on the long-tailed problem. This also suggests that our model demonstrates superior performance in handling hard samples.

TABLE 6. Comparison of the variants on CAIL-SMALL dataset.

	Charges Prediction		
Methods	MP	MR	MF
KEMCAN w/o Knowledge & MCA	83.18	80.08	79.95
KEMCAN w/o MCA	85.68	82.03	82.98
KEMCAN	**88.90**	**87.75**	**88.20**
	Law Articles Prediction		
Methods	MP	MR	MF
KEMCAN w/o Knowledge & MCA	82.97	79.22	79.62
KEMCAN w/o MCA	85.43	82.29	83.33
KEMCAN	**88.04**	**85.97**	**86.87**

D. ABLATION STUDY

We conducted two ablation experiments to evaluate the contribution of Knowledge Representation module and Multi-Cross Attention Layer, aiming to demonstrate the significance of these modules within KEMCAN.

(1) KEMCAN w/o Knowledge & MCA: In this variant, the Knowledge Representation module and Multi-Cross Attention Layer are removed from KEMCAN.

(2) KEMCAN w/o MCA: In this variant, only the Multi-Cross Attention Layer is removed from KEMCAN. The output from the Knowledge Representation module is passed through max pooling and then combined with the output of the Fact Description Representation to obtain the final representation.

Table 6 presents the experiment results. When the Knowledge Representation module and the Multi-Cross Attention Layer are removed from KEMCAN, this alteration significantly degrades performance. In comparison to KEMCAN, the MF drops by 9.35% and 8.34% for charge prediction and law article prediction tasks, respectively. Conversely, when only the Multi-Cross Attention Layer is removed, the MF decreases by 5.92% for charge prediction and 4.07% for law article prediction in comparison to KEMCAN. This highlights the crucial roles of legal charge knowledge and the Multi-Cross Attention Layer within the KEMCAN framework.

TABLE 7. Comparison of variant loss function on CAIL-SMALL dataset.

		Charges Prediction		
Methods	Acc	MP	MR	MF
KEMCAN + CE	90.57	88.90	87.75	88.20
KEMCAN + FL	**90.87**	**89.99**	**87.80**	**88.71**
KEMCAN + DL	without convergence			
		Law Articles Prediction		
Methods	Acc	MP	MR	MF
KEMCAN + CE	90.65	88.04	**85.97**	86.87
KEMCAN + FL	**90.77**	**89.03**	85.75	**86.88**
KEMCAN + DL	without convergence			

E. LOSS FUNCTION

Considering the task contains an extremely imbalanced label (charges and law articles) set, we exploited the balancing

loss functions for LJP and investigated the impacts of Cross-entropy(CE), Focal Loss (FL), Dice Loss (DL) on KEMCAN [20], [42]. The FL and DL are calculated as:

$$\text{Focal Loss} = -\sum_{j}^{C}(1-p_j^{\gamma})\log(p_j) \tag{17}$$

$$\text{Dice Loss} = \sum_{j}^{C}(1-\frac{2(1-p_j)^{\alpha}p_j y_j+\gamma}{(1-p_j)^{\alpha}p_j+y_j+\gamma}) \tag{18}$$

The performances of KEMCAN with different loss functions based on CAIL-SMALL datasets are shown in Table 7. We can find that KEMCAN+FL outperforms KEMCAN+CE by 1.09% in term of MP, and almost achieves similar results on MR. One possible reason is that the focal loss function focused on learning hard examples, and alleviated the difficult sample problem in the CAIL-SMALL dataset. We guess KEMCAN+DL does not converge at fixed epochs may be caused by that DL aims at optimizing MF. With the extreme low-frequency label distribution of CAIL-SMALL, it is easy for the model to focus on these labels with low-frequency, resulting in the slow convergence of KEMCAN+DL.

VI. CONCLUSION AND FUTURE WORK

In this paper, we introduce legal charge knowledge as a kind of knowledge to enhance the representation of fact description, and proposed a Knowledge-enriched Multi-Cross Attention Network (KEMCAN) to improve the performance of legal judgment prediction and solve legal cases involving confusing charges and law articles. Specifically, a cross-attention mechanism is proposed to model the relationship between legal charge knowledge and fact description in a unified model. By incorporating legal charge knowledge, the model can better capture the nuances and complexities of legal cases, leading to improved performance in charge prediction and law article prediction tasks. The experiments conducted on the CAIL-SMALL and CAIL-BIG datasets demonstrate the superiority of KEMCAN compared to conventional text classification methods and state-of-the-art LJP models.

In the future, we can explore the following directions: (1) Advancing knowledge integration techniques by transitioning from vector-based methods to logic-based approaches. This involves formalizing legal knowledge into logical rules and incorporating them through logical reasoning for more effective knowledge fusion. (2) Building on the foundation of logic-based knowledge representation, future research can explore the application of this approach in prompt-based learning.

VII. LIMITATIONS

One limitation of the study is that the experimental results were exclusively conducted on Chinese criminal law datasets only, raising questions about the generalizability of the proposed KEMCAN model to other legal systems or domains. Additionally, the study focused solely on charge prediction and law article prediction, neglecting other crucial tasks such as prison sentence prediction and court view generation. Future research should encompass evaluating the performance of KEMCAN across different legal systems and incorporating additional tasks into the model.

REFERENCES

[1] Z. Hu, X. Li, C. Tu, Z. Liu, and M. Sun, "Few-shot charge prediction with discriminative legal attributes," in *Proc. 27th Int. Conf. Comput. Linguistics*, 2018, pp. 487–498.

[2] Y. Le, C. He, M. Chen, Y. Wu, X. He, and B. Zhou, "Learning to predict charges for legal judgment via self-attentive capsule network," in *Proc. ECAI*, 2020, pp. 1802–1809.

[3] H. Zhong, J. Zhou, W. Qu, Y. Long, and Y. Gu, "An element-aware multi-representation model for law article prediction," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 6663–6668.

[4] H. Chen, D. Cai, W. Dai, Z. Dai, and Y. Ding, "Charge-based prison term prediction with deep gating network," in *Proc. Conf. Empirical Methods Natural Lang. Process., 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 6362–6367. [Online]. Available: https://aclanthology.org/D19-1667

[5] S. Li, H. Zhang, L. Ye, S. Su, X. Guo, H. Yu, and B. Fang, "Prison term prediction on criminal case description with deep learning," *Comput., Mater. Continua*, vol. 62, no. 3, pp. 1217–1231, 2020.

[6] Q. Li and Q. Zhang, "Court opinion generation from case fact description with legal basis," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 17, pp. 14840–14848.

[7] W.-C. Lin, T.-T. Kuo, T.-J. Chang, C.-A. Yen, C.-J. Chen, and S.-D. Lin, "Exploiting machine learning models for Chinese legal documents labeling, case classification, and sentencing prediction," *Int. J. Comput. Linguistics Chin. Lang. Process.*, vol. 17, no. 4, pp. 49–68, 2012. [Online]. Available: https://aclanthology.org/O12-5004

[8] F. Kort, "Predicting supreme court decisions mathematically: A quantitative analysis of the 'right to counsel' cases," *Amer. Political Sci. Rev.*, vol. 51, no. 1, pp. 1–12, Mar. 1957, doi: 10.2307/1951767.

[9] S. S. Ulmer, "Quantitative analysis of judicial processes: Some practical and theoretical applications," *Law Contemp. Problems*, vol. 28, no. 1, pp. 164–184, 1963.

[10] R. Keown, "Mathematical models for legal prediction," *Comput./LJ*, vol. 2, pp. 829–862, Jan. 1980.

[11] J. A. Segal, "Predicting supreme court cases probabilistically: The search and seizure cases, 1962–1981," *Amer. Political Sci. Rev.*, vol. 78, no. 4, pp. 891–900, Dec. 1984.

[12] B. Luo, Y. Feng, J. Xu, X. Zhang, and D. Zhao, "Learning to predict charges for criminal cases with legal basis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 2727–2736. [Online]. Available: https://aclanthology.org/D17-1289

[13] L. Yue, Q. Liu, B. Jin, H. Wu, K. Zhang, Y. An, M. Cheng, B. Yin, and D. Wu, "NeurJudge: A circumstance-aware neural framework for legal judgment prediction," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2021, pp. 973–982.

[14] N. Xu, P. Wang, L. Chen, L. Pan, X. Wang, and J. Zhao, "Distinguish confusing law articles for legal judgment prediction," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 3086–3095. [Online]. Available: https://aclanthology.org/2020.acl-main.280

[15] S. Li, H. Zhang, L. Ye, X. Guo, and B. Fang, "MANN: A multichannel attentive neural network for legal judgment prediction," *IEEE Access*, vol. 7, pp. 151144–151155, 2019.

[16] C.-L. Liu, C.-T. Chang, and J.-H. Ho, "Case instance generation and refinement for case-based criminal summary judgments in Chinese," *J. Inf. Sci. Eng.*, vol. 20, no. 4, pp. 783–800, 2004.

[17] D. M. Katz, M. J. Bommarito, and J. Blackman, "A general approach for predicting the behavior of the supreme court of the United States," *PLoS ONE*, vol. 12, no. 4, Apr. 2017, Art. no. e0174698, doi: 10.1371/journal.pone.0174698.

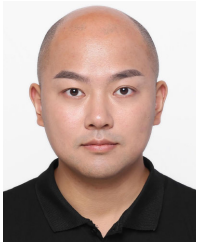
[18] O.-M. Sulea, M. Zampieri, S. Malmasi, M. Vela, L. P. Dinu, and J. van Genabith, "Exploring the use of text classification in the legal domain," 2017, *arXiv:1710.09306*.

[19] M. Medvedeva, M. Vols, and M. Wieling, "Using machine learning to predict decisions of the European court of human rights," *Artif. Intell. Law*, vol. 28, no. 2, pp. 237–266, Jun. 2020.

[20] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.
[21] D. Li, Q. Zhao, J. Chen, and D. Zhao, "ADAN: An intelligent approach based on attentive neural network and relevant law articles for charge prediction," *IEEE Access*, vol. 9, pp. 90203–90211, 2021.
[22] H. Zhong, Z. Guo, C. Tu, C. Xiao, Z. Liu, and M. Sun, "Legal judgment prediction via topological learning," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Brussels, Belgium, 2018, pp. 3540–3549. [Online]. Available: https://aclanthology.org/D18-1390
[23] W. Yang, W. Jia, X. Zhou, and Y. Luo, "Legal judgment prediction via multi-perspective bi-feedback network," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 4085–4091, doi: 10.24963/ijcai.2019/567.
[24] P. Wang, Y. Fan, S. Niu, Z. Yang, Y. Zhang, and J. Guo, "Hierarchical matching network for crime classification," in *Proc. 42nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2019, pp. 325–334.
[25] S. Yang, S. Tong, G. Zhu, J. Cao, Y. Wang, Z. Xue, H. Sun, and Y. Wen, "MVE-FLK: A multi-task legal judgment prediction via multi-view encoder fusing legal keywords," *Knowl.-Based Syst.*, vol. 239, Mar. 2022, Art. no. 107960.
[26] H. Zhang, Z. Dou, Y. Zhu, and J.-R. Wen, "Contrastive learning for legal judgment prediction," *ACM Trans. Inf. Syst.*, vol. 41, no. 4, pp. 1–25, Oct. 2023.
[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. U. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30. Long Beach, CA, USA: Curran Associates, 2017, pp. 1–12.
[28] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423
[29] J. Chen, Y. Hu, J. Liu, Y. Xiao, and H. Jiang, "Deep short text classification with knowledge powered attention," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 6252–6259.
[30] L. Ying, H. Yu, J. Wang, Y. Ji, and S. Qian, "Multi-level multi-modal cross-attention network for fake news detection," *IEEE Access*, vol. 9, pp. 132363–132373, 2021.
[31] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 1412–1421. [Online]. Available: https://aclanthology.org/D15-1166
[32] C. Raffel and D. P. W. Ellis, "Feed-forward networks with attention can solve some long-term memory problems," 2015, *arXiv:1512.08756*.
[33] C. Xiao, H. Zhong, Z. Guo, C. Tu, Z. Liu, M. Sun, Y. Feng, X. Han, Z. Hu, H. Wang, and J. Xu, "CAIL2018: A large-scale legal dataset for judgment prediction," 2018, *arXiv:1807.02478*.
[34] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," 2013, *arXiv:1310.4546*.
[35] J. A. K. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Process. Lett.*, vol. 9, no. 3, pp. 293–300, Jun. 1999.
[36] H. Zhang, L. Xiao, Y. Wang, and Y. Jin, "A generalized recurrent neural architecture for text classification with multi-task learning," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 2873–2879.
[37] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, 2016, pp. 1480–1489.
[38] Y. Cui, W. Che, T. Liu, B. Qin, and Z. Yang, "Pre-training with whole word masking for Chinese BERT," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 3504–3514, 2021.
[39] H. Zhong, Z. Zhang, Z. Liu, and M. Sun, "Open Chinese language pre-trained model zoo," THUNLP, Beijing, China, Tech. Rep. 1, 2019. [Online]. Available: https://github.com/thunlp/openclap
[40] Z. Li and M. Sun, "Punctuation as implicit annotations for Chinese word segmentation," *Comput. Linguistics*, vol. 35, no. 4, pp. 505–512, 2009, doi: 10.1162/coli.2009.35.4.35403.
[41] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[42] X. Li, X. Sun, Y. Meng, J. Liang, F. Wu, and J. Li, "Dice loss for data-imbalanced NLP tasks," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 465–476. [Online]. Available: https://aclanthology.org/2020.acl-main.45

CONGQING HE (Student Member, IEEE) received the B.S. degree from the School of Computer Science and Technology, Hefei Normal University, Hefei, China, in 2016, and the M.S. degree from the College of Computer Science and Electronic Engineering, Hunan University, Changsha, China, in 2019. He is currently pursuing the Ph.D. degree with the School of Computer Science, Universiti Sains Malaysia, Penang, Malaysia. His research interests include natural language processing and legal intelligence.

TIEN-PING TAN received the Ph.D. degree from Université Joseph Fourier, France, in 2008. He is currently an Associate Professor with the School of Computer Sciences, Universiti Sains Malaysia. His research interests include automatic speech recognition, machine translation, and natural language processing.



XIAOBO ZHANG received the bachelor's degree in computer science from Nanchang University and the M.Sc. degree in internet computing from Abertay University. He is currently pursuing the Ph.D. degree in natural language processing with the School of Computer Science, Universiti Sains Malaysia. He is a Lecturer with the Jiangxi Vocational College of Finance and Economics. His research interests include named entity recognition, knowledge graph, and question answering.

SHENG XUE received the B.S. degree in communication engineering from the North University of China, Taiyuan, China, in 2017, and the M.S. degree in optical engineering from South China Normal University, Guangzhou, China, in 2020. He is currently an Engineer with Research Institute of China Telecom Corporation Ltd., Guangzhou. His research interests include the analysis of cloud and network operation business and optical sensors.