Multi-task Legal Judgement Prediction Combining a Subtask of the Seriousness of Charges

Zhuopeng Xu, Xia Li*, Yinlin Li, Zihan Wang, Yujie Fanxu and Xiaoyan Lai
Guangzhou Key Laboratory of Multilingual Intelligent Processing
School of Information Science and Technology
Guangdong University of Foreign Studies,Guangzhou,China
zhuopengxu@126.com,ksyzformy@126.com,zihanwang0703@126.com,
yujiefanxu@126.com,avaxiaoyan@126.com
xiali@gdufs.edu.cn

Abstract

Legal Judgement Prediction has attracted more and more attention in recent years. One of the challenges is how to design a model with better interpretable prediction results. Previous studies have proposed different interpretable models based on the generation of court views and the extraction of charge keywords. Different from previous work, we propose a multi-task legal judgement prediction model which combines a subtask of the seriousness of charges. By introducing this subtask, our model can capture the attention weights of different terms of penalty corresponding to the charges and give more attention to the correct terms of penalty in the fact descriptions. Meanwhile, our model also incorporates the position of defendant making it capable of giving attention to the contextual information of the defendant. We carry several experiments on the public CAIL2018 dataset. Experimental results show that our model achieves better or comparable performance on three subtasks compared with the baseline models. Moreover, we also analyze the interpretable contribution of our model.

1 Introduction

Legal Judgement Prediction (LJP) aims to predict charge, law article and terms of penalty automatically based on the fact descriptions of the criminal cases. It can be used to help the court's judgement and provide legal guidance and assistance to the public.

In recent years, different methods have been proposed to improve the performance of legal judgement prediction task. Some previous studies need to design features manually (Katz et al., 2014; Lin et al., 2012; Liu and Hsieh, 2006; Liu et al., 2015) and some of neural network based models extract features automatically and achieve significant improvements (Liu et al., 2019; Ye et al., 2018; Zhong et al., 2018). However, there are still some challenging problems, including the improvement of the performance and the enhancement of the interpretability of the terms of penalty prediction.

For the improvement of the performance in terms of penalty prediction, previous studies use multi-task and joint learning to obtain the sharing information among different subtasks. Zhong et al. (2018) propose a Directed Acyclic Graph structure with topological relations to capture the information attribution among three subtasks, which effectively improve the problem of insufficient fine-grained in LJP. For the enhancement of the interpretability, different solutions are proposed to the problem. Ye et al. (2018) propose a Seq2Seq model to formulate legal judgement prediction task as a natural language generation problem. Their model take fact descriptions and charge labels as input and outputs the court's view. The outputs are used as an auxiliary information for practical judgement. Liu et al. (2019) propose a multi-task learning model to incorporate charge keywords extracted by TF-IDF and TextRank. Their model has a good interpretability by introduced the keyword information.

Although different methods are proposed for the above two problems, we argue that some of the knowledge are known in legal judgement prediction task and can be incorporated into the model for improving the performance and the interpretability of the prediction results.

*Corresponding author: xiali@gdufs.edu.cn

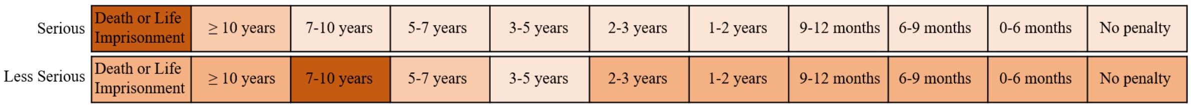

Figure 1: Attentions of different terms of penalty for charge of ”murder” generated by the proposed subtask. As can be seen, term of death or life imprisonment and terms of 7-10 years are paid more attention for serious murder and less serious murder respectively.

The first one is the seriousness of charges. Actual judgement procedure tells us that the final decision of the terms of penalty is largely determined by the seriousness of the case, which depends on the case fact descriptions and the terms of penalty definition described in the article corresponding to the charge of the case. Inspired by the actual judgement procedure, we propose to design a subtask of the seriousness of charges which is determined by the charge and the terms of penalty for the task of legal judgement prediction. According to the scope of legal terms of penalty, we can easily divide a fact description into two categories: serious and less serious. Detailed descriptions and examples are given in Section 3.3. The new subtask is used to obtain attentions of different terms of penalty according to serious and less serious predicted by the subtask and let the model pay more attention to those important terms of penalty for the corresponding fact descriptions. As an example with predicted charge as “murder” which is shown in Fig.1., we can see that our model captures more attention on “Death or life Imprisonment” with predicted serious label and on “7-10 years” with predicted less serious label, which is useful for the model selecting the right terms of penalty.

The second one is the defendant information which is known in the case fact descriptions. Previous studies focuses on the fact descriptions only (eg., just using text words), ignoring the importance of the context information of the defendant. To this end, we propose to incorporate the position of the defendant into the model. By introducing the defendant position-aware embedding for the fact descriptions, we can capture more context information of the defendant which is helpful for the prediction of subtasks. The main contributions of our work are as follows:

1)We propose a multi-task legal judgement prediction model combining a subtask of the seriousness of charges. By introducing this subtask, our model improves the performance and the interpretability of the terms of penalty prediction in LJP.

2)Based on the importance of defendant in the fact descriptions, we propose to incorporate the position information of the defendant into the model, making it capable of giving attention to the relevant context information of the defendant.

3)We carry several experiments on the CAIL2018 dataset. We will show that our proposed model achieves a better or comparable performance in all subtasks than the baseline models. We also give a discussion of our model’s interpretability in terms of penalty prediction.

2 Related work

Legal judgement prediction task usually includes three subtasks: charges prediction, law articles recommendation and terms of penalty prediction. We will review the work of legal judgement prediction from single-task based models and multi-task based models.

2.1 Single-task based Legal Judgement Prediction Models.

In the models of single-task based legal judgement prediction, the core perspective is to use different encoding method to represent the fact descriptions more correctly. Luo et al. (2017) propose an attention-based neural network with two hierarchical encoding structures to jointly model the fact descriptions and the top k relevant law articles. Their model achieves good performance for those simple cases, which indicates that the hierarchical encoding structure and introducing of law articles effectively improve the result of charge prediction. Hu et al. (2018) propose an attribute-attentive charge prediction model. They incorporate the fact descriptions attributed by attention mechanism with the original text. Their model performs well in few-shot charges and confusing charge pairs. Ye et al. (2018) propose a labelconditioned Seq2Seq model with attention mechanism. The model take the fact descriptions and charge labels as input and formulates legal judgement prediction as a natural language generation problem. Their model can automatically generate court views and give a better interpretability of the prediction. In order to improve the terms of penalty prediction, Chen et al. (2019) regard term prediction as a kind of regression problem. By introducing charge labels and using a structure of Deep Gating Network (DGN), their model achieves good results for the terms of penalty prediction.

2.2 Multi-task based Legal Judgement Prediction Models.

Most of above models are proposed for single task such as charge prediction or terms of penalty prediction. However, judge's actual judgement procedure tells us that different subtasks are often related with each other, like charge is related with law and charge is also related to the terms of penalty. To this end, different multitask based learning models are proposed to obtain the relationship information of different subtasks. Zhong et al. (2018) propose a topological multitask learning framework for three subtasks of law articles, charges, and the terms of penalty. They formalized the dependencies among these subtasks as a Directed Acyclic Graph for neural network learning. Their model improves the problem of insufficient fine-grained of legal judgement prediction task. Yang et al. (2019) propose a multi-perspective bi-feedback network with the word collocation attention mechanism. Liu et al. (2019) propose a multi-task learning framework for legal judgement prediction. They use charge keywords extracted by TF-IDF and Text Rank as auxiliary information and use a hierarchical structure to decode the fact descriptions. Their model shows good interpretability because of the introduced charge keywords. Wang et al. (2019) propose a hybrid attention model which combines the improved hierarchical attention network (iHAN) and the deep pyramid convolutional neural network (DPCNN) by ResNet. Their model achieves a good performance for the subtask of the terms of penalty. Xu et al. (2020) take advantages of a novel graph neural network to distinguish confusing law articles and improve the capacity of the encoding of the fact descriptions. Zhong et al. (2020) propose a model based on reinforcement learning, which can visualize the prediction process and give interpretable judgements by giving a process of QA judgement. Their model greatly improves the interpretability of legal judgement prediction task.

This paper focuses on multi-task legal judgement prediction. Different from previous studies, our work focuses on the scope of legal terms of penalty of the different seriousness in the law. We introduce the seriousness of charges as a subtask into the model. By introducing this subtask, it is expected that the prediction of the terms of penalty can obtain improvements not only on the performance but also on the interpretability of the prediction. In addition, in order to make a better judgement to the defendant, our model also combines the defendant's position information in the model.

3 Proposed Model

3.1 Architecture of Our Model

In this paper, we propose a multi-task legal judgement prediction model combining a subtask of the seriousness of charges, which consists of two parts. The first part is encoding layer, in which a defendant position-aware context information is incorporated into the fact descriptions representation. The second part is decoding layer, in which a subtask of the seriousness of charges is introduced to help obtain the attention of different terms of penalty corresponding to the predicted charge. Our model is shown in Fig.2. In the following sections, we will introduce embedding of the defendant's position information in section 3.2 and describe our design of subtask of the seriousness of charges in section 3.3. Section 3.4 will describe model training and prediction.

3.2 Embedding of the Defendant's Position Information

3.2.1 Design of the Defendant's Position Information.

In order to obtain the context information of defendant in the fact descriptions, we use the relative position of each word to the defendant as an indicator to represent the context information of the defendant

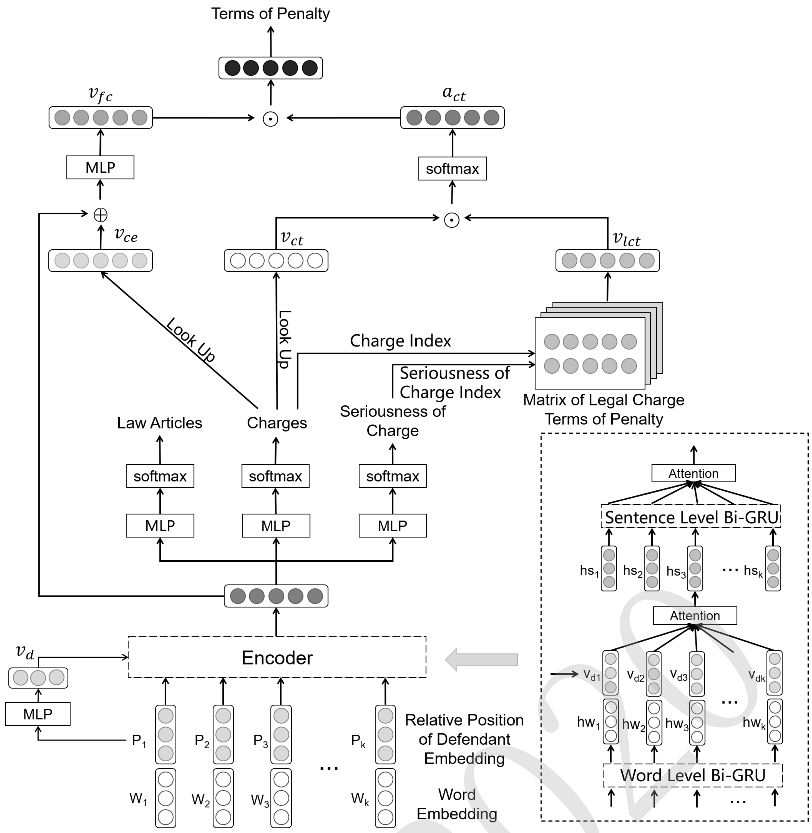

Figure 2: The whole architecture of our model.

He Mou	hold	chopper	run after and cut	the hurt	Zheng MouMou	the defendant	He MouMou	in charge of	drive	pick up
贺某	持	砍刀	追砍	被害人	郑某某	被告人	贺某某	负责	驾车	接应
7	6	5	4	3	2	1	0	1	2	3

Figure 3: An example of relative position of each word to the defendant in a sentence.

in a sentence. For example, as show in Fig.3., the defendant is “贺某某(He MouMou)” whose position is set to 0, the word “驾车(drive)” whose position is 2 and the word “追砍(run after and cut)” whose position is 4. We can see that the action “驾车(drive)” is more relative to the defendant than that of “追砍(run after and cut)”. By incorporating the position, the model can learn to focus more on the action ”驾车(drive)” than the action ”追砍(run after and cut)”. This kind of defendant position-aware fact descriptions representation has a better expression of the context information of defendant.

3.2.2 Defendant’s Position-aware Fact Descriptions Encoding.

For a given fact descriptions, we formulate it as $d = \{s_1, ..., s_n\}$, in which $s_i \in \mathbb{R}^{L_w \times m}$ is the representation of vectorization of i-th sentence, m is the dimension of the word vector, L_w is the maximum length of a sentence. For sentence s_i, it is formulated as $s_i = \{w_{i1}, \ldots, w_{ik}\}$ represented by k words, in which $w_{ik} \in \mathbb{R}^m$ is the representation of word embedding vector. The relative position of defendant in document d is formulated as $p = \{sp_1, \ldots, sp_n\}$, in which $sp_i \in \mathbb{R}^{L_s \times n}$ is the representation of vectorization of i-th relative position of defendant of sentence formulated as $sp_i = \{wp_{i1}, \ldots, wp_{ik}\}$, $wp_{ik} \in \mathbb{R}^n$ is the representation of vectorization of k-th position in i-th sentence, n is the dimension of the vector of relative position of defendant.

As shown in Fig.2., we employ a structure of hierarchical attention network (Yang et al., 2016) to encode the fact descriptions. Firstly, we encode each word in each sentence on word level by employing Bi-GRU network with attention. We then obtain the hidden representation of each sentence. Secondly, we encode each sentence of a document on sentence level by employing Bi-GRU network with attention, and then obtain the hidden representation of the document.

For word level encoding, the new word representation is obtained by concatenating the word embedding vector and the relative position of defendant vector. We formulate sentence consist of the new word representation as $s = \{x_1, \ldots, x_k\}$, in which x_k is obtained by w_k and wp_k formulated as $x_k = [w_k; wp_k]$. The w_k and wp_k represent the representation of k-th word in sentence s and the representation of the relative position of defendant of k-th word respectively. Then, we input the representation of sentence s into a word level Bi-GRU network, and then obtain the hidden output of sentence s formulated as $hw = \{hw_1, hw_2, \ldots, hw_k\}$. At t time stamp, we concatenate the hidden output of the forward and backward GRU unit formulated as $hw_t = [\overrightarrow{h_t}, \overleftarrow{h_t}]$.

3.2.3 Defendant's Position-aware Attention Enhancing.

We combine the relative position of defendant vector into word level attention so that the hidden output of each GRU unit in sentence s can better capture the information of position of defendant. Firstly, we employ a multilayer perceptron to obtain the vector v_{dj} which represent the information of position of defendant in each unit in sentence s. Then, we concatenate the hw and v_{dj}. Employing a one-layer MLP, we obtain the new hidden output u_h. Finally, we obtain the hidden representation H_s of sentence s after obtaining the attention aw_t of the new hidden output u_h via softmax function. The W_w and b_w are the parameter of hidden layer projection, u_w is word level context vector. The calculation formula is shown in equations $(1) \sim (4)$.

$$v_{dj} = MLP(sp_j) \tag{1}$$

$$u_h = tanh(W_w[hw, v_{dj}] + b_w) \tag{2}$$

$$aw = softmax(u_h^T u_w) \tag{3}$$

$$H_s = \sum_t aw_t hw_t \tag{4}$$

For sentence level encoding, we input each representation H_s of sentences into a sentence level Bi-GRU network with attention, and then obtain the final hidden representation v_f of fact descriptions.

3.3 Design of the Subtask of the Seriousness of Charges

Based on the definition of terms of penalty, we divide each charge into two categories: serious and less serious. Then we annotate each charge with two legal terms of penalty vectors, which have the same dimension with the prediction of terms of penalty subtask. We also annotate all the samples with the seriousness of chargs, then we can carry a new subtask of the seriousness of charges in the model.

3.3.1 Tagging Rules.

First of all, we manually annotate the legal terms of penalty vectors of the two categories with serious and less serious. The tagging rules are as follows: when the legal terms of penalty is less serious, according to the actual terms of penalty described in law articles, we set the vector of less serious category of the corresponding charge. If there is no distinction between the seriousness of the charge of legal terms of penalty, the vectors of the corresponding serious and less serious legal terms of penalty are set as the same. When a charge includes several seriousness such as less serious, serious, very serious, etc, we combine the serious and more serious parts as the serious category.

Then, we annotate each sample with the label of seriousness. Given a sample, we can determine its corresponding range of legal terms of penalty based on the charge label, if the corresponding range is serious, the seriousness label of the sample is annotated 'serious'; if the corresponding range is less serious, then the seriousness label is annotated 'less serious'. A special case is that if the terms of penalty label is not within the scope of the serious and less serious, we will still annotate it as 'serious'.

3.3.2 Example Demonstration.

In order to better illustrate our annotation rules, we give an example of tagging for the legal terms of penalty vector tagging of a specific charge. As shown in Fig.4., take "故意伤害罪(intentional assault)"

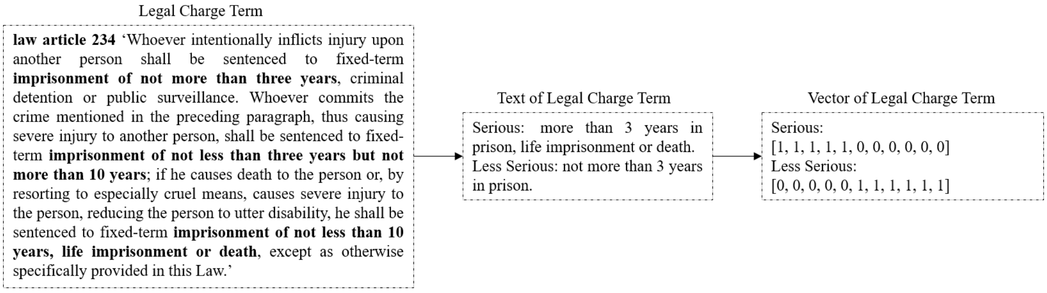
Legal Charge Term
law article 234 'Whoever intentionally inflicts injury upon another person shall be sentenced to fixed-term imprisonment of not more than three years, criminal detention or public surveillance. Whoever commits the crime mentioned in the preceding paragraph, thus causing severe injury to another person, shall be sentenced to fixed-term imprisonment of not less than three years but not more than 10 years; if he causes death to the person or, by resorting to especially cruel means, causes severe injury to the person, reducing the person to utter disability, he shall be sentenced to fixed-term imprisonment of not less than 10 years, life imprisonment or death, except as otherwise specifically provided in this Law.'
Text of Legal Charge Term
Serious: more than 3 years in prison, life imprisonment or death.
Less Serious: not more than 3 years in prison.
Vector of Legal Charge Term
Serious:
[1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0]
Less Serious:
[0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1]

Figure 4: An example of legal charge terms of penalty.

as an example, according to the definition of the corresponding law article 234, we firstly divide the legal terms of penalty into the following three categories: 'less serious': fixed-term imprisonment of not more than three years, criminal detention or public surveillance; 'serious': fixed-term imprisonment of not less than three years but not more than 10 years; 'very serious': fixed-term imprisonment of not less than 10 years, life imprisonment or death. Based on our classification of seriousness, we combine the corresponding legal terms of penalty range of "serious" and "very serious", and the final "serious" legal terms of penalty text is: "fixed-term imprisonment of not less than 3 years, life imprisonment or death"; the "less serious" legal terms of penalty text is "fixed-term imprisonment of not more than three years". Then according to the 11 categories of the subtask of terms of penalty prediction, the corresponding legal terms of penalty range vectors are generated, the serious category vector of the legal terms of penalty of 'intentional assault' is [1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0], and the less serious category vector is [0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1].

3.4 Model Training and Prediction

As shown in section 3.2, after getting the final hidden representation of the fact descriptions v_f ,we employ three different multilayer perceptrons to obtain the decoding vector of law articles, charges, seriousness of charge respectively. Then, as shown in Equation (5), we input them into softmax function to get prediction results $\hat{y}_1$, $\hat{y}_2$ and $\hat{y}_3$ of three subtasks. As shown in Equations (6) and (7), the index vector of the corresponding charges and seriousness of charge was obtained by the prediction results of charges and seriousness of charge respectively.

$$\hat{y}_k = softmax(MLP_k(v_f)), k = 1, 2, 3 \tag{5}$$

$$i_{charge} = argmax(\hat{y}_2) \tag{6}$$

$$i_{seriousness} = argmax(\hat{y}_3) \tag{7}$$

According to the index of the prediction results of charges and seriousness of charge, we obtain legal charge term vector v_{lct}. Similar to word embedding, we obtain charge term vector v_{ct} and charge embedding vector v_{ce} in weight matrices W_{ct} and W_{ce} respectively which have different dimensions and perform joint learning in the model. Then, as shown in Equation (8), we calculate charge term attention weight a_{ct} via charge term vector v_{ct} and legal charge term vector v_{lct}.

$$a_{ct} = softmax(v_{ct} \odot v_{lct}) \tag{8}$$

After concatenating the final hidden representation v_f and charge embedding vector v_{ce}, we input it into a multilayer perceptron. Then, we obtain the vector v_{fc} which is the fusion of fact descriptions and charges, as shown in Equation (9).

$$v_{fc} = MLP([v_f, v_{ce}]) \tag{9}$$

Finally, as shown in Equation (10), we do a hadamard product of v_{fc} and a_{ct} and obtain the final decoding vector v_t of terms of penalty. Then, as shown in Equation (11), we input the vector v_t into softmax function to get prediction results $\hat{y}_4$.

$$v_t = v_{fc} \odot a_{ct} \tag{10}$$

$$\hat{y}_4 = softmax(v_t) \tag{11}$$

In the training process, we use cross-entropy loss function as the loss function of our model. After calculating each cross-entropy loss for each subtask, we sum each loss of different subtasks as the total loss. As shown in Equation (12), i represents the i-th subtask, Y_i represents the total number of classes of the i-th subtask, and j represents the j-th class.

$$loss_{total} = -\sum_{i=1}^{4}\sum_{j=1}^{Y_i} y_{i,j} log(\hat{y}_{i,j}) \tag{12}$$

4 Experiments

4.1 Dataset

We use the CAIL2018[1] (Xiao et al., 2018) dataset to be evaluated in this paper. Similar to the work of Zhong et al. (Zhong et al., 2018), we do some relevant preprocess on the datasets. Firstly, we filter out the crime data that contained multiple charges and multiple relevant law articles. Secondly, we remove the crime data with charges appeared less than 100 times in the datasets. Finally, similar to the work of TOPJUDGE (Zhong et al., 2018), we divide the terms of penalty into 11 non-overlapping intervals. The detailed information of the CAIL2018 are shown in Table 1.

Dataset	Amount	Subtasks	Amount
Training Set	101513	Charges	119
Testing Set	26731	Law Articles	103
Validation Set	10818	Terms of penalty	11

Table 1: Statistical information of the CAIL2018 dataset.

4.2 Compared Models

In order to compare on three subtasks, we built a multi-task implementation on those not designed for multi-task baseline models. We use Bi-LSTM, TextCNN (Kim, 2014) and Hierarchical Attention Networks (HAN) (Yang et al., 2016) as three different structures to encode the fact descriptions. For HAN structure, we employ a word level of Bi-GRU network with attention and a sentence level of Bi-GRU network with attention to encode the fact descriptions. We employ three different multilayer perceptrons for multitask prediction for these three baselines. We use TOPJUDGE (Zhong et al., 2018) and Few-Shot (Hu et al., 2018) as our another compared models based on their multi-task joint learning and additional auxiliary information design. For the Few-Shot model, we also employ three different multilayer perceptrons for multitask prediction.

4.3 Experimental Setting

In our experiment, we use THULAC (Li and Sun, 2009) for word segmentation. We use skip-gram (Mikolov et al., 2013) for pre-training of all fact descriptions and get a pre-trained 200-dimensional matrix of word vectors. For the position of defendant, we embed each position into a 100-dimensional vector and perform joint training in the model. For the CNN-based and Bi-LSTM-based models in the baselines, we set the maximum document length to 512 words. For the HAN-based models, the maximum sentence length is set to 100 words, the maximum document length is set to 15 sentences. The unit dimension of hidden layer is set to 256, and the output dimension of each level vector is set to 256.

[1] https://github.com/china-ai-law-challenge/CAIL2018

In our model, we embed the maximum sentence length of the relative position of defendant with a blank vector. For training, we use the Adam optimizer to control stochastic gradient descent. The learning rate of the optimizer set to 0.001, the batch size set to 128, and the epoch set to 16. We select the model that performed best on the validation set, and report the results on the testing set.

4.4 Experimental Results

Similar to previous work, we use accuracy (Acc.), macro-precision (MP), macro-recall (MR) and macro-F1 (F1) as metrics in this paper, the final experimental results are shown in Table 2. As LJP task is a multi-label classification task, and there is an extremely unbalanced phenomenon among various categories in the CAIL2018 dataset, we mainly focus on the comparison of the results of macro-F1.

Tasks	Law Articles				Charges				Terms of penalty			
Metrics	Acc.	MP	MR	F1	Acc.	MP	MR	F1	Acc.	MP	MR	F1
Bi-LSTM	79.33	76.45	77.11	75.20	81.69	81.26	**81.77**	80.38	39.66	31.99	29.34	28.26
Text CNN	76.77	74.21	73.13	71.43	82.38	81.20	78.16	78.32	37.85	32.49	27.78	27.79
HAN	81.08	76.85	77.48	76.05	81.97	80.89	81.90	80.37	41.07	31.25	30.71	28.40
TOPJUDGE	**82.11**	76.14	75.82	75.01	82.40	79.48	79.21	78.29	40.04	32.74	30.45	29.59
Few-Shot	79.59	75.62	74.97	73.97	83.33	82.22	80.42	80.56	40.33	30.88	**33.38**	**30.65**
Our model	81.04	**78.43**	**77.27**	**76.49**	**84.47**	**82.42**	81.46	**81.14**	**41.96**	**34.89**	31.11	30.45

Table 2: Experimental results of our model and baselines.

Firstly, we compare our model with Bi-LSTM, TextCNN and HAN models. As shown in Table 2, we can see our model achieves the best macro-F1 value in all three subtasks. And it shows that our model performs great results especially in the subtask of terms of penalty. Our model is 30.45% which is 2.19% higher than Bi-LSTM, 2.66% higher than TextCNN, 2.05% higher than HAN. The results prove the effectiveness of the subtask of seriousness of charge introduced in our model.

Secondly, we compare our model with TOPJUDGE model which is also a multi-task LJP model. As shown in Table 2, our model also achieves better performance in all three subtasks. Our model increases by 1.48% on law articles prediction subtask, 2.85% on charges prediction subtask and 0.86% on terms of penalty prediction subtask. This result shows that our model is ascending to a certain extent on three subtasks compared with the TOPJUDGE model.

Finally, we compare our model with the Few-Shot model which also uses an auxiliary information to help improve the performance of a subtask. We can see that our model increases by 2.52% on law articles prediction, 0.58% on charges prediction, and decreases by 0.2% on terms of penalty prediction which is comparable with the Few-Shot model. The results indicate that the overall performance of our model can be improved on the basis of improving term prediction results.

4.5 Ablation Studies

In order to analyze the influence of each part of our model, several ablation experiments are conducted in this paper. We remove four parts from our model to see the influences: 1)We remove the word level attention calculated by the position of defendant which is named as w/o drp_att. 2)We remove the whole part of using position of defendant, named as w/o drp_pos+drp_att, which means that the model only judges with the fact descriptions. 3)We remove the subtask of the seriousness of charges which is named as w/o seriousness to see the influence of the subtask to the whole model. 4)We remove the whole part of relative position of defendant and the subtask of the seriousness of charges, which means that the fact descriptions is only encoded by hierarchical attention networks and predicted by multitask learning, and the model is named as w/o drp_both+seriousness. The results of different parts of ablation studies are shown in Table 3.

As shown in Table 3, when we remove the subtask of seriousness of charge (w/o seriousness), the macro-F1 of the subtask of terms of penalty prediction is reduced by 2.04%, which shows that the introducing of the subtask of seriousness of charge can significantly improve the result of the terms of penalty

Tasks	Law Articles		Charges		Terms of Penalty		Seriousness of Charge	
Metrics	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Our Model	81.04	76.49	84.47	81.14	41.96	30.45	87.18	80.09
w/o drp_att	81.71	76.3	83.49	81.31	41.63	29.95	86.88	80.05
w/o drp_pos+drp_att	81.68	75.53	83.28	81.02	41.66	29.84	86.87	79.57
w/o seriousness	81.28	76.59	83.87	81.51	41.41	28.41	/	/
w/o drp_both+seriousness	81.08	76.05	81.97	80.37	41.07	28.40	/	/

Table 3: Results of ablation experiments.

prediction. In addition, according to Table 3, we can see that the decoder with the introducing of the subtask of seriousness of charge is the most effective part in all additional components.

We also can see that after embedding the position information of defendant, the prediction results of charges and law articles can be improved. Moreover, compared with embedding the position information of defendant, using position information of defendant to improve word level attention can further improve the performance of the model in three subtasks. When we remove all the position information of defendant, the macro-F1 of the subtask of law articles prediction will decrease by 0.96%. This result shows that the position information of defendant can mainly improve the result of law articles prediction. In the end, when we combine the position information of defendant and the subtask of the seriousness of charges into a model, the performances of all three subtasks are improved.

4.6 Interpretability Analysis

In order to analyze the interpretability of our model, we choose a representative case to illustrate how the design of the subtask of the seriousness of charges can be improved in interpretability of the prediction of terms of penalty.

As shown in Fig.5., given the fact descriptions, previous method will predict and give the terms of penalty directly without any auxiliary information. While in our model, firstly, we will preliminarily predict the prediction results of charges, law articles and the seriousness of charge. With the predicted charge and the seriousness of the charge, our model can determine the range of legal charge terms of penalty, this is important and useful for the judge and the public to get the auxiliary information of the terms of the penalty. Finally, the model outputs the prediction result of the terms of penalty. Compared with previous direct prediction process of terms of penalty, the prediction process of our model has a better interpretability of the prediction.

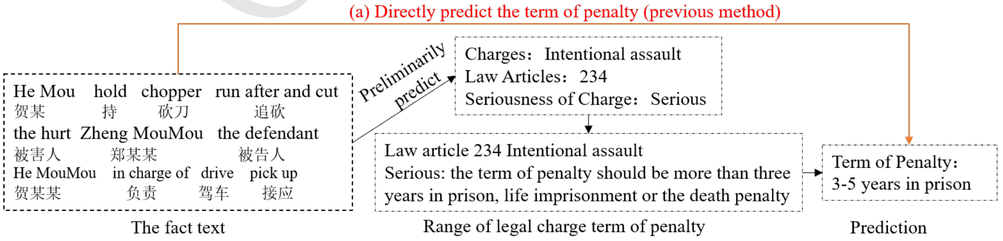
(a) Directly predict the term of penalty (previous method)
He Mou hold chopper run after and cut
贺某 持 砍刀 追砍
the hurt Zheng MouMou the defendant
被害人 郑某某 被告人
He MouMou in charge of drive pick up
贺某某 负责 驾车 接应
The fact text
Preliminarily predict
Charges：Intentional assault
Law Articles：234
Seriousness of Charge：Serious
Law article 234 Intentional assault
Serious: the term of penalty should be more than three years in prison, life imprisonment or the death penalty
Range of legal charge term of penalty
Term of Penalty：
3-5 years in prison
Prediction

Figure 5: Terms of penalty prediction process of our model compared with previous method.

5 Conclusion

In this paper, we propose to design and combine a subtask of the seriousness of charges for multi-task legal judgement prediction. Evaluations demonstrate the effectiveness of our model on charge prediction, law article recommendation and the terms of penalty prediction, indicating that the introduced subtask of the seriousness of charges and the sufficient encoding of the fact descriptions for the defendant are useful. Our model also shows the good interpretability on the task of terms of penalty prediction. In the future, we will explore a better method to incorporate the contextual information of defendant and investigate the usefulness of different subtasks for multi-task legal judgement prediction.

Acknowledgements

This work is supported by National Natural Science Foundation of China (No. 61976062), the Science and Technology Program of Guangzhou (No. 201904010303) and Special Funds for the Cultivation of Guangdong College Students' Scientific and Technological Innovation (No. pdjh2020a0197).

References

Huajie Chen, Deng Cai, Wei Dai, Zehui Dai, and Yadong Ding. 2019. Charge-based prison term prediction with deep gating network. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 6361–6366.

Zikun Hu, Xiang Li, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. 2018. Few-shot charge prediction with discriminative legal attributes. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 487–498.

Daniel Martin Katz, Michael J Bommarito Ii, and Josh Blackman. 2014. Predicting the behavior of the supreme court of the united states: A general approach. *Plos One*, 12(4).

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1746–1751.

Zhongguo Li and Maosong Sun. 2009. Punctuation as implicit annotations for chinese word segmentation.

Wan-Chen Lin, Tsung-Ting Kuo, Tung-Jia Chang, Chueh-An Yen, Chao-Ju Chen, and Shou-de Lin. 2012. Exploiting machine learning models for Chinese legal documents labeling, case classification, and sentencing prediction in Chinese. In *Proceedings of the 24th Conference on Computational Linguistics and Speech Processing (ROCLING 2012)*, pages 140–141.

Chao-Lin Liu and Chwen-Dar Hsieh. 2006. Exploring phrase-based classification of judicial documents for criminal charges in chinese. In *Foundations of Intelligent Systems*, pages 681–690.

Yi-Hung Liu, Yen-Liang Chen, and Wu-Liang Ho. 2015. Predicting associated statutes for legal problems. *Inf. Process. Manag.*, 51(1):194–211.

Zonglin Liu, Meishan Zhang, Ranran Zhen, Zuoquan Gong, Nan Yu, and Guohong Fu. 2019. Multi-task learning model for legal judgment predictions with charge keywords. *Journal of Tsinghua University(Science and Technology)*, 59(7):497.

Bingfeng Luo, Yansong Feng, Jianbo Xu, Xiang Zhang, and Dongyan Zhao. 2017. Learning to predict charges for criminal cases with legal basis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2727–2736.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Computer Science*.

Wenguan Wang, Yunwen Chen, Hua Cai, Yanneng Zeng, and Huiyu Yang. 2019. Judicial document intellectual processing using hybrid deep neural networks. *Journal of Tsinghua University(Science and Technology)*, 59(7):505.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, and Jianfeng Xu. 2018. CAIL2018: A large-scale legal dataset for judgment prediction. *CoRR*, abs/1807.02478.

Nuo Xu, Pinghui Wang, Long Chen, Li Pan, Xiaoyan Wang, and Junzhou Zhao. 2020. Distinguish confusing law articles for legal judgment prediction. *CoRR*, abs/2004.02557.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489.

Wenmian Yang, Weijia Jia, Xiaojie Zhou, and Yutao Luo. 2019. Legal judgment prediction via multi-perspective bi-feedback network. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 4085–4091.

Hai Ye, Xin Jiang, Zhunchen Luo, and Wenhan Chao. 2018. Interpretable charge predictions for criminal cases: Learning to generate court views from fact descriptions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1854–1864.

Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun. 2018. Legal judgment prediction via topological learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3540–3549.

Haoxi Zhong, Yuzhong Wang, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. Iteratively questioning and answering for interpretable legal judgment prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:1250–1257.