Charge Prediction with Legal Attention

Qiaoben Bao[1,2,3], Hongying Zan[1], Peiyuan Gong[1], Junyi Chen[1],
and Yanghua Xiao[4]

[1] School of Information Engineering, Zhengzhou University, Henan, China
mbaoqiaoben@outlook.com, iehyzan@zzu.edu.cn, gongpeiyuan1@163.com,
junyichen_ch@sina.com

[2] CETC Big Data Research Institute Co., Ltd., Guiyang, China

[3] Big Data Application on lmproving Government Governance Capabilities National
Engineering Laboratory, Guiyang, China

[4] School of Computer Science, Fudan University, Shanghai, China
shawyh@fudan.edu.cn

Abstract. Charge prediction aims to predict the corresponding charges for a specific case. In civil law system, human judges will match the facts with relevant laws, and the final judgments are usually made in accordance with relevant law articles. Existing works either ignore this feature or simply model the relationship using multi-task learning, but neither make full use of relevant articles to assist the charge prediction task. To address this issue, we propose an attentional neural network, LegalAtt, which uses relevant articles to improve the performance and interpretability of charge prediction task. More specifically, our model works in a bidirectional approach: First, it uses the fact description to extract relevant articles; In return, the selected relevant articles assist to locate key information from the fact description, which helps improve the performance of charge prediction. Experimental results show that our model achieves the best performance on the real-world dataset compared with other state-of-the-art baselines. Our code is available at https://github.com/nlp208/legal_attention.

Keywords: Charge prediction · Text classification · Civil law system

1 Introduction

The automatic charge prediction task takes fact description as input and predicts the corresponding charges for a specific case. This task plays a crucial role in legal assistance system. For example, this technique makes it easier for users without legal knowledge to conduct legal consultations, and it also provide reference information for people in legal field to simplify their work.

As an important task in the field of intelligent justice, charge prediction has a long history of research. Most existing works regard charge prediction as a text classification task. Liu et al. [7,8] attempt to use k-Nearest Neighbor (KNN) combined with word-level features and phrase-level features to predict corresponding charges. Lin et al. [6] manually designed a variety of factor labels

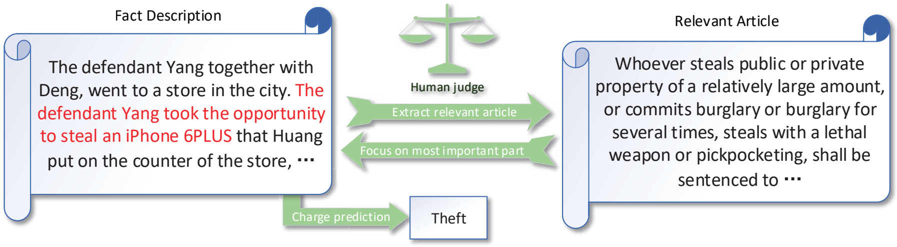
Fact Description
The defendant Yang together with Deng, went to a store in the city. The defendant Yang took the opportunity to steal an iPhone 6PLUS that Huang put on the counter of the store, ···
Human judge
Extract relevant article
Focus on most important part
Charge prediction
Theft
Relevant Article
Whoever steals public or private property of a relatively large amount, or commits burglary or burglary for several times, steals with a lethal weapon or pickpocketing, shall be sentenced to ···

Fig. 1. Charge prediction procedure in civil law system.

for charge prediction. Şulea et al. [12] propose a classification system based on Support Vector Machine (SVM), which is applied to the data of French Supreme Court. These works heavily rely on manually designed features, which is time-consuming and thus cannot be applied to large-scale dataset directly.

In recent years, neural networks have achieved great success on many natural language processing (NLP) tasks, such as text classification [4,16], machine translation [10,14] and so on. Inspired by these works, researchers begin to use neural networks to model the charge prediction task. Luo et al. [9] propose an attentional neural network to jointly model charge prediction task and relevant article extraction task. Jiang et al. [2] use reinforcement learning mechanism to output the predicted charge as well as rationales. Hu et al. [1] manually design 10 different attributes to improve the performance on few-shot charges. Zhong et al. [17] focus on the dependencies among subtasks of legal judgement prediction, and propose a topological multi-task learning framework.

Although many efforts have been made in charge prediction, we still faces many challenges:

Multi-Label Cases: In the real scenario, cases are complex and diverse, which may involve multiple different laws and charges. This requires the model to have the ability of predicting multi-label charges and make full use of information from different labels. But many existing works only focus on single label cases [1,2,17].

Interpretability: One obvious difference between legal domain tasks and other domain tasks is that users not only care about the results, but also want to know the legal basis for the predicted results. In charge prediction task, it's more convincing if the model output relevant legal basis for making such a decision, or tell us which part of the fact description leads to such a result. As illustrated in Fig. 1, in civil law system, human judges first use fact description to extract relevant articles. Then the key information in fact description is matched with relevant articles, and the final judgements are made accordingly. Methods like [9,17] simply take advantage of multi-task learning, ignoring the interpretability between related tasks.

In order to solve these problems, we propose an attentional neural network to predict charges using knowledge from relevant articles. In this framework, we first use the fact description to predict the relevant articles. Then the extracted relevant articles are used to focus on the most important part of the fact description and assist the final charge prediction task. Our model simulates the charge prediction process in the civil law system, making full use of the information from different relevant law articles. Experimental results show that our model outperforms other state-of-the-art charge prediction models and text classification models on the real-world dataset. We also analyze the attention from relevant articles, and prove that our model can utilize the extra knowledge from relevant articles. In attention mechanism, relevant articles pay more attention to the key information in fact description, which explains why the model makes the final decision and improves the interpretability compared with previous works [9,17].

The main contributions of this paper can be summarized as follows:

(1) We propose an attentional model based on relevant articles for charge prediction in civil law system, and achieve the best results on the real-world dataset.

(2) Our model focuses on the multi-label attributes of legal tasks, which better reflect the real situation.

(3) Our model has a better performance in interpretability and provide more legal basis for charge prediction task through attention mechanism.

2 Related Work

2.1 Text Classification

Text classification is a classical task in NLP, which aims at categorizing documents based on their specific representation on different topics, sentiment, etc. Kim [4] proposes a Convolutional Neural Network (CNN) based model with different window sizes for text classification. Tang et al. [13] regard document as a set of sentences, so a two-level structure is proposed to learn the representation at the level of word and sentence respectively. Yang et al. [16] then use a two-level attention mechanism based on [13]. Johnson and Zhang [3] propose a deep CNN model using down sampling without increasing the number of feature maps, which effectively takes care of the model complexity with more hidden layers.

2.2 Charge Prediction

Charge prediction mainly focuses on predicting the corresponding charges for an input case. With the development of machine learning methods, researches begin to formalize charge prediction as a text classification task. Many works [7,8,6] use KNN to classify cases by taking shallow information from fact description or using manually designed features. Şulea et al. [12] use SVM combined with N-gram features to build a charge prediction system. These works take a small amount of charges as input and need manual feature extraction, which only obtain the superficial features of legal text, thus making it hard to generalize.

In recent years, advances in neural networks help us simplify many NLP tasks [4,16,10]. Inspired by these works, more and more researchers use neural network to model charge prediction and related tasks. Luo et al. [9] propose an attention-based model to jointly model the charge prediction task and the relevant article extraction task. Our model shares similar ideas with them, that is, relevant articles can benefit the performance of charge prediction. But they only use fact description attention to extract relevant articles, which cannot make full use of the knowledge of the relevant articles and lacks interpretability. Zhong et al. [17] pay more attention to the hierarchical relationships between subtasks of legal judgement prediction, and model the dependency relationships between different tasks by using directed acyclic graph (DAG). Hu et al. [1] manually design ten features for charge prediction, resulting in significant improvements on few-shot charges. However, with the increase of the number of charges, more features need to be introduced, which leads to the limitation of the model extensibility. Jiang et al. [2] focus on the interpretability of charge prediction task, and adopt reinforcement learning-based method to extract key information from input fact description. But they fail to consider the relevant articles which play a vital role in the civil law system. In this paper, we also ask the model to give corresponding explanations for the predicted results. For this purpose, we introduce a legal attention mechanism based on relevant articles to show which part does the model focus on.

3 Method

In this section, we propose an attentional neural network using relevant articles to assist charge prediction task. Similar to Luo et al. [9], we believe that the relevant articles of a specific case can help charge prediction. Moreover, we not only use fact description to extract relevant articles , but also use relevant articles to focus on the most important part of fact description. Compared with simply using multi-task learning to jointly model two tasks, our approach is more suitable for the charge prediction process in the civil law system. As show in Fig. 2, our model first takes fact description as input and outputs the fact representation sequence $\mathbf{d}^f$. $\mathbf{d}^f$ is then used to find the relevant articles. We then use an article document encoder to generate article representation sequence $\mathbf{d}^a$ for each relevant article. These article representation sequences are fed into the attention layer to calculate the attention-based fact representation $\mathbf{e}^f_{final}$. Finally, we use $\mathbf{e}^f_{final}$ to predict the appropriate charges for the input case.

3.1 Fact Document Encoder

Fact document encoder takes fact description as input and outputs fact representation sequence $\mathbf{d}^f = \left\{\mathbf{d}^f_1, \mathbf{d}^f_2, \ldots, \mathbf{d}^f_{T_f}\right\}$, where T_f is the length of the fact description. Zhong et al. [17] have shown the effectiveness of CNN model for text encoding in legal domain, we also adopt a CNN encoder based on previous work proposed by Kim [4].

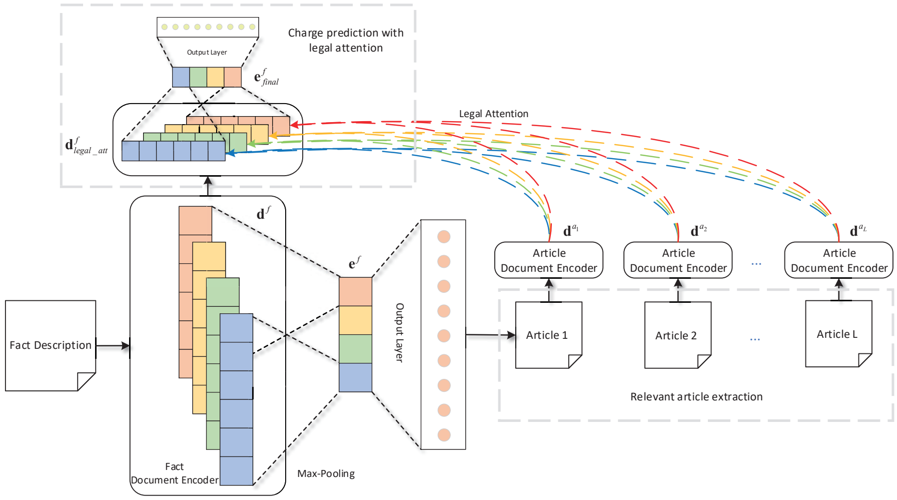

Fig. 2. Model overview.

We first use an embedding layer to convert the input fact description into embedding sequence $\mathbf{x}^f = \left\{\mathbf{x}^f_1, \mathbf{x}^f_2, \dots, \mathbf{x}^f_{T_f}\right\}$, where $\mathbf{x}^f_t \in \mathbb{R}^k$ and k is the dimension of word embedding.

Let $\mathbf{x}^f_{i:i+j}$ represent the concatenation of word embedding $\mathbf{x}^f_i, \mathbf{x}^f_{i+1}, \dots, \mathbf{x}^f_{i+j}$. We define a convolution operation with window size h as:

$$\mathbf{c}^f_{hi} = f\left(\mathbf{W}^f_h \cdot \mathbf{x}^f_{i:i+h-1} + \mathbf{b}^f_h\right) \tag{1}$$

where $\mathbf{W}^f_h$ and $\mathbf{b}^f_h$ are weight matrix and bias vector and $f(\cdot)$ is activation function. Specifically, we adopt multiple kernels with different window sizes. For each kernel $\mathbf{W}^f_*$, we apply convolution operation on the whole input sequence with padding at both ends of the sequence. The fact representation sequence is calculated by concatenating the results of convolution operations with different kernel:

$$\mathbf{d}^f = \left\{\mathbf{d}^f_1, \mathbf{d}^f_2, \dots, \mathbf{d}^f_{T_f}\right\}, \mathbf{d}^f_t = \text{concat}\left(\mathbf{c}^f_{*t}\right) \tag{2}$$

where $\mathbf{d}^f_t \in \mathbb{R}^m$ is the hidden state of word x^f_t and m is feature size.

3.2 Relevant Article Extractor

Training a classifier for each article is time consuming and hard to generalize due to the large number of articles. Therefore, we apply a simple affine transformation followed by sigmoid to calculate each article's score.

We first apply max pooling operation over $\mathbf{d}^f$ and obtain the fact representation $\mathbf{e}^f = \left[\mathbf{e}_1^f, \mathbf{e}_2^f, \ldots, \mathbf{e}_m^f\right]$ as:

$$\mathbf{e}_i^f = \max\left(\mathbf{d}_{1,i}^f, \mathbf{d}_{2,i}^f, \ldots, \mathbf{d}_{T_f,i}^f\right), \forall i \in [1, m] \tag{3}$$

then the article score is calculated by:

$$\mathbf{score}_{art} = \mathrm{sigmoid}\left(\mathbf{W}^s \mathbf{e}^f + \mathbf{b}^s\right) \tag{4}$$

where $\mathbf{W}^s$ and $\mathbf{b}^s$ are weight matrix and bias vector.

In order to prevent the misleading by irrelevant articles, we provide true relevant article labels in training step. In prediction step, we only chose articles with score higher than threshold τ as the truly relevant articles.

3.3 Article Document Encoder

We use the same framework described in Sec. 3.1 to encode the relevant articles as:

$$\mathbf{d}^{a_l} = \left\{\mathbf{d}_1^{a_l}, \mathbf{d}_2^{a_l}, \ldots, \mathbf{d}_{T_{a_l}}^{a_l}\right\}, \forall l \in [1, L] \tag{5}$$

where $\mathbf{d}_t^{a_l}$ is the hidden state of word $x_t^{a_l}$, L is the number of relevant articles and T_{a_l} is the length of l^{th} relevant article.

Since fact description and relevant articles usually have different emphases in description, we set different parameters for fact document encoder and article document encoder instead of sharing.

3.4 Attention-based Charge Prediction

Having fact representation sequence $\mathbf{d}^f$ and article representation sequence $\mathbf{d}^a$, we want to use $\mathbf{d}^a$ to assist the final charge prediction task. Therefore, we propose an attention mechanism based on relevant articles to focus on difference part of input fact description. Then the weighted sum over fact representation is used to make charge prediction.

Legal Attention. We share the same spirit with Vaswani et al. [14] that attention can be described as mapping a query and a set of key-value pairs to an output. Therefore, we use $\mathbf{d}^f$ and $\mathbf{d}^a$ to calculate the key vectors and query vectors as:

$$\begin{aligned} \mathbf{k}_i &= \tanh\left(\mathbf{W}^k \mathbf{d}_i^f\right), \forall i \in [1, T_f] \\ \mathbf{q}_i &= \tanh\left(\mathbf{W}^q \mathbf{d}_i^a\right), \forall i \in [1, T_a] \end{aligned} \tag{6}$$

where $\mathbf{W}^* \in \mathbb{R}^{d_{att} \times m}$ is weight matrix and d_{att} is the dimension of key vectors and query vectors.

Then legal attention matrix $\mathbf{A}$ is calculated by:

$$\mathbf{A} = \mathrm{softmax}\left(\left(\alpha_{ij}\right)_{T_a \times T_f}\right), \alpha_{i,j} = \mathbf{q}_i^T \mathbf{k}_j \tag{7}$$

We apply attention to $\mathbf{d}^f$ and get the fact description sequence with legal attention $\mathbf{d}_{legal\_att} = \left\{\mathbf{d}^f_{legal\_att_1}, \mathbf{d}^f_{legal\_att_2}, \ldots, \mathbf{d}^f_{legal\_att_{T_f}}\right\}$ as:

$$\mathbf{d}^f_{legal\_att_i} = \sum_{t=1}^{T_a} \alpha_{t,i} \mathbf{d}^f_i, \forall i \in [1, T_f] \tag{8}$$

We finally apply a max pooling over $\mathbf{d}^f_{legal\_att}$ to get the representation $\mathbf{e}_{legal\_att} = \left[e^f_{legal\_att_1}, e^f_{legal\_att_2}, \ldots, e^f_{legal\_att_{T_f}}\right]$ as:

$$\mathbf{e}^f_{legal\_att_i} = \max\left(\mathbf{d}^f_{legal\_att_{1,i}}, \mathbf{d}^f_{legal\_att_{2,i}}, \ldots, \mathbf{d}^f_{legal\_att_{T_f,i}}\right), \forall i \in [1, m] \tag{9}$$

Attention from Different Articles. Due to the multi-label property of our problem, we will get more than one relevant article by relevant article extractor. For each relevant article l, we obtain a fact representation with legal attention $\mathbf{e}^f_{legal\_att\_l}$, the final representation is then calculated by averaging all these vectors as:

$$\mathbf{e}^f_{final} = \text{mean}\left(\mathbf{e}^f_{legal\_att\_1}, \mathbf{e}^f_{legal\_att\_2}, \ldots, \mathbf{e}^f_{legal\_att\_L}\right) + \mathbf{e}^f \tag{10}$$

where we add a residual connection in order to reduce the impact of irrelevant articles and to simplify the training process.

Charge prediction. Given the final fact representation with legal attention $\mathbf{e}^f_{final}$, we feed it into a fully connected layer followed by sigmoid function to get the charge prediction result:

$$\hat{\mathbf{y}} = \text{sigmoid}\left(\mathbf{W}^p \mathbf{e}^f_{final} + \mathbf{b}^p\right) \tag{11}$$

where $\mathbf{W}^p$ and $\mathbf{b}^p$ are weight matrix and bias vector.

In prediction step, we use a threshold to select corresponding charge labels.

3.5 Training

The loss of our model contains two parts. In charge prediction part, we want to minimize the loss between $\hat{\mathbf{y}}$ and true distribution $\mathbf{y}_{charge}$. In relevant article extraction part, we want to minimize the loss between $\mathbf{score}_{art}$ and true distribution $\mathbf{y}_{art}$.

Due to the multi-label property of our problem, the loss is calculated by summing the cross-entropy loss over each label:

$$\begin{aligned} L_{charge} &= -\sum_{i=1}^{C} \mathbf{y}_{charge_i} \cdot \log\left(\hat{\mathbf{y}}_i\right) + \left(1 - \mathbf{y}_{charge_i}\right) \cdot \log\left(1 - \hat{\mathbf{y}}_i\right) \\ L_{art} &= -\sum_{i=1}^{N} \mathbf{y}_{art_i} \cdot \log\left(\mathbf{score}_{art_i}\right) + \left(1 - \mathbf{y}_{art_i}\right) \cdot \log\left(1 - \mathbf{score}_{art_i}\right) \end{aligned} \tag{12}$$

where C is the number of charges and N is the number of law articles.

Combining the two parts, our final loss is $L = L_{charge} + \alpha \cdot L_{art}$, where α is a weight factor of relevant law extraction task.

4 Experiments

4.1 Data Preparation

Our data is collected from the first large-scale Chinese legal dataset CAIL2018 [15]. CAIL2018 contains more than 2.6 million criminal cases with 202 criminal charges and 183 relevant articles, and there exist many low-frequency charges like smuggle and money laundering. In the following part, we only consider 100 charges with the highest frequency and 91 related articles. We randomly choose 203,823 cases for training, 20,000 for validation and 40,000 for testing. All the charges and articles have more than 100 training data. To model the multi-label property in real-world scenarios, we keep data with multiple charges or relevant articles, which account for 18.6%, 10.5%, 16.7% of training set, validation set and test set respectively.

Although there are some cases with more than one defendant in real-world, it's hard to deal with different parts of different defendants in one case. We therefore remove cases with multi-defendant and leave them for future work.

4.2 Baselines

We employ several text classification models and charge prediction models for comparison, and all the text classification models are trained with both task in multi-task framework:

CNN: CNN document encoder with multiple kernel sizes followed by max pooling [4].

Hierarchical Attention Network (HAN): A hierarchical network for document encoding in both word and sentence level proposed by Yang et al. [16].

Deep Pyramid CNN (DPCNN): Johnson and Zhang [3] propose a deep CNN model to capture global representation for document.

FactLaw: Luo et al. [9] propose an attention-based neural model jointly models charge prediction task and relevant article extraction task.

TopJudge: Zhong et al. [17] propose a neural model formalizing the dependencies among subtasks in legal judgment prediction.

4.3 Experimental Settings

Since our data is composed of Chinese and there are no delimiters in documents, we employ jieba[5] for Chinese segmentation. Word embeddings are trained using Skip-Gram model [11] on all fact descriptions with embedding size of 200.

We set maximum document length to 300. For HAN and FactLaw, we set maximum sentence length to 100, and one document contains no more than 20 sentences. For Recurrent Neural Network (RNN) based models, hidden size is set to 100. For CNN based models, filter size is set to 50 with window size in (2, 3, 4, 5). We set all threshold to 0.4 by validation. The parameter K in FactLaw is set to 10. The weight α of relevant article loss is set to 1.0.

[5] https://github.com/fxsjy/jieba

We employ Adam [5] as optimizer, and set learning rate to 0.001, dropout rate to 0.2 and batch size to 32. We evaluate our model using Micro-F1 and Macro-F1 in both charge prediction task and relevant article extraction task. Here Macro-F1 is calculated by averaging the F1 score of each category.

4.4 Experimental Results

Table 1. Relevant article extraction results and charge prediction results.

Model	Relevant Article Extraction		Charge Prediction	
	Micro-F1	Macro-F1	Micro-F1	Macro-F1
CNN	75.7	74.9	77.8	75.6
HAN	66.9	63.7	67.5	64.1
DPCNN	79.0	76.8	80.9	76.4
FactLaw	68.7	62.9	72.4	63.8
TopJudge	78.9	72.2	79.1	74.1
LegalAtt	**80.3**	**78.7**	**81.0**	**77.4**

As show in Table 1, our model outperforms other baselines on both relevant extraction task and charge prediction task.

In relevant article extraction task, our model is similar to traditional CNN model. But we use relevant articles to further assist the charge prediction task, which benefits both subtasks.

In charge prediction task, we share the similar spirit with FactLaw. Different from directly connecting the fact representation and article representation in FactLaw, we use an attention matrix to give a different weight to relevant and irrelevant information in fact description. This approach is like the real court scene in civil law system, where human judges use relevant articles to judge the details of fact descriptions. Moreover, FactLaw uses a fixed K to extract relevant articles, which affected by noise from irrelevant articles. In our model, we adopt a threshold τ to filter out irrelevant articles. Improved performance on relevant article extraction will further affects the charge prediction task.

4.5 Ablation Test

The performance of our model depends largely on the relevant articles, we therefore conduct some ablation tests to investigate the effectiveness of our model.

As show in Table 2, LegalAtt$-\tau$ refers to not use threshold τ but only fixed K, which is the same as FactLaw. Intuitively, LegalAtt$-\tau$ suffers from the noisy from irrelevant articles, as not all cases have K relevant articles. LegalAtt$-art$ means we do not provide relevant article labels for supervision in training step, and we set the parameter α to 0. All the parameters are learned by charge prediction task. The performance decrease significantly by 13.3% and 8.1% in

Table 2. Results of ablation test.

Model	Relevant Article Extraction		Charge Prediction	
	Micro-F1	Macro-F1	Micro-F1	Macro-F1
LegalAtt	**80.3**	**78.7**	**81.0**	**77.4**
LegalAtt$-\tau$	75.8	74.4	76.7	75.2
LegalAtt$-art$	70.6	65.4	72.0	69.3

Macro-F1 of relevant article extraction task and charge prediction task respectively. Therefore, relevant articles play a crucial role in overall model.

4.6 Case Study

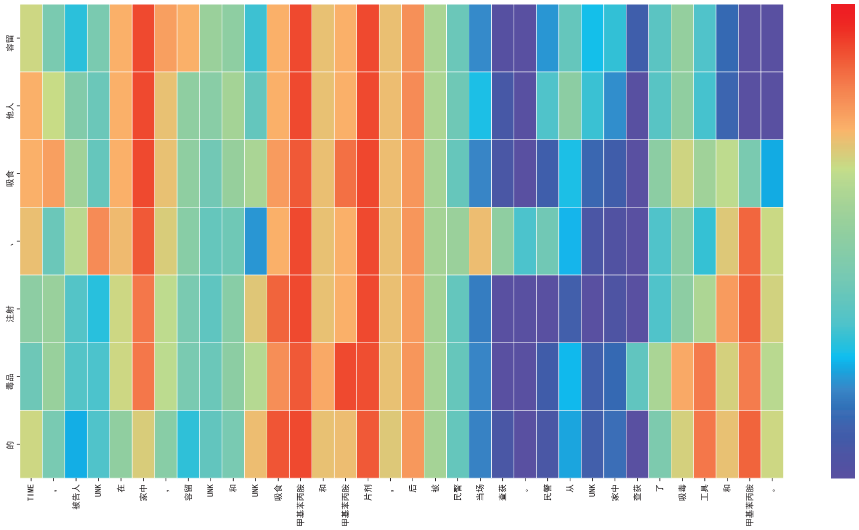

Fig. 3. Partial heat map of the attention matrix. The vertical axis is a fragment of legal text and the horizontal axis is a fragment of fact description.

In this part, we select a representative case to show how legal attention works in information filtering. In this case, the defendant violated the criminal law by illegally allowing others to take drugs at his home. Fig. 3 is a part of the overall heat map of attention matrix. Each cell represents the attention from word in relevant article to word in fact description. Cells with red color have higher weight, whereas cells with dark blue color have less weight. We can see that the relevant articl mainly focuses on three different parts. To facilitate the description, we remove all values less than 10^{-3} and obtain Fig. 4. As show in

Fig. 4, red part in relevant article focuses on the content about providing drugs for others(red part and green part in fact deacription), and green part in relevant article focuses on drugs(green part in fact deacription) and information about drugs(blue part in fact deacription). Specially, we notice that **drugs** in relevant article pay attention to **methamphetamine** which is kind of drugs.

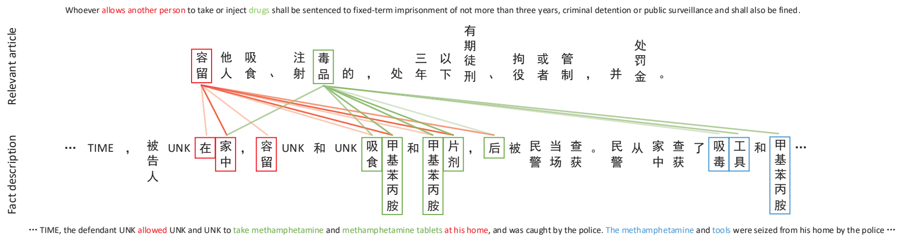
Whoever allows another person to take or inject drugs shall be sentenced to fixed-term imprisonment of not more than three years, criminal detention or public surveillance and shall also be fined.
Relevant article
容留 他人 吸食 、 注射 毒品 的 ， 处 三年 以下 有期徒刑 、 拘役 或者 管制 ， 并 处罚金 。
Fact description
… TIME ， 被告人 UNK 在 家中 ， 容留 UNK 和 UNK 吸食 甲基苯丙胺 和 甲基苯丙胺 片剂 ， 后 被 民警 当场 查获 。 民警 从 家中 查获 了 吸毒 工具 和 甲基苯丙胺 …
… TIME, the defendant UNK allowed UNK and UNK to take methamphetamine and methamphetamine tablets at his home, and was caught by the police. The methamphetamine and tools were seized from his home by the police …

Fig. 4. Visualization of heat map with threshold 10^{-3}. The text of different colors represents the translation of the content in the corresponding box.

5 Conclusion

In this paper, we focus on how to use relevant articles to assist the charge prediction task, and propose an attention-based neural model LegalAtt, which jointly models the relevant article extraction task and the charge prediction task. In this model, we use an attention matrix calculated by relevant articles to filter out irrelevant information in fact description. The attention mechanism can be regarded as an interpretable part of our model, which is crucial in legal domain. Experiments on real-world dataset show that our model can effectively use relevant articles to focus on different parts of the input fact description. As for future work, we will further explore the multi-defendant charges and cases in different law systems.

Acknowledgements

This work is supported by the National Social Science Fund of China (No.18ZDA315) and Big Data Application on lmproving Government Governance Capabilities National Engineering Laboratory Open Fund Project.

References

1. Hu, Z., Li, X., Tu, C., Liu, Z., Sun, M.: Few-shot charge prediction with discriminative legal attributes. In: Proceedings of the 27th International Conference on Computational Linguistics. pp. 487–498 (2018)

2. Jiang, X., Ye, H., Luo, Z., Chao, W., Ma, W.: Interpretable rationale augmented charge prediction system. In: Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations. pp. 146–151 (2018)
3. Johnson, R., Zhang, T.: Deep pyramid convolutional neural networks for text categorization. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 562–570 (2017)
4. Kim, Y.: Convolutional neural networks for sentence classification. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (2014)
5. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In Proceedings of ICLR (2014)
6. Lin, W.C., Kuo, T.T., Chang, T.J., Yen, C.A., Chen, C.J., Lin, S.d.: Exploiting machine learning models for chinese legal documents labeling, case classification, and sentencing prediction. Processdings of ROCLING p. 140 (2012)
7. Liu, C.L., Chang, C.T., Ho, J.H.: Case instance generation and refinement for case-based criminal summary judgments in chinese. Journal of Information Science and Engineering **20**(4), 783–800 (2004)
8. Liu, C.L., Hsieh, C.D.: Exploring phrase-based classification of judicial documents for criminal charges in chinese. In: International Symposium on Methodologies for Intelligent Systems. pp. 681–690. Springer (2006)
9. Luo, B., Feng, Y., Xu, J., Zhang, X., Zhao, D.: Learning to predict charges for criminal cases with legal basis. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 2727–2736 (2017)
10. Luong, T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 1412–1421 (2015)
11. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
12. Şulea, O.M., Zampieri, M., Vela, M., van Genabith, J.: Predicting the law area and decisions of french supreme court cases. In: Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017. pp. 716–722 (2017)
13. Tang, D., Qin, B., Liu, T.: Document modeling with gated recurrent neural network for sentiment classification. In: Proceedings of the 2015 conference on empirical methods in natural language processing. pp. 1422–1432 (2015)
14. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
15. Xiao, C., Zhong, H., Guo, Z., Tu, C., Liu, Z., Sun, M., Feng, Y., Han, X., Hu, Z., Wang, H., et al.: Cail2018: a large-scale legal dataset for judgment prediction. arXiv preprint arXiv:1807.02478 (2018)
16. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E.: Hierarchical attention networks for document classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 1480–1489 (2016)
17. Zhong, H., Zhipeng, G., Tu, C., Xiao, C., Liu, Z., Sun, M.: Legal judgment prediction via topological learning. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 3540–3549 (2018)