Few-Shot Charge Prediction with Discriminative Legal Attributes

Zikun Hu* **Xiang Li*** **Cunchao Tu** **Zhiyuan Liu†** **Maosong Sun**
Department of Computer Science and Technology, Tsinghua University
State Key Lab on Intelligent Technology and Systems, Tsinghua University
Beijing National Research Center for Information Science and Technology
{hzk14, x-l15}@mails.tsinghua.edu.cn
tucunchao@gmail.com, {liuzy, sms}@tsinghua.edu.cn

Abstract

Automatic charge prediction aims to predict the final charges according to the fact descriptions in criminal cases and plays a crucial role in legal assistant systems. Existing works on charge prediction perform adequately on those high-frequency charges but are not yet capable of predicting few-shot charges with limited cases. Moreover, these exist many confusing charge pairs, whose fact descriptions are fairly similar to each other. To address these issues, we introduce several discriminative attributes of charges as the internal mapping between fact descriptions and charges. These attributes provide additional information for few-shot charges, as well as effective signals for distinguishing confusing charges. More specifically, we propose an attribute-attentive charge prediction model to infer the attributes and charges simultaneously. Experimental results on real-work datasets demonstrate that our proposed model achieves significant and consistent improvements than other state-of-the-art baselines. Specifically, our model outperforms other baselines by more than 50% in the few-shot scenario. Our codes and datasets can be obtained from https://github.com/thunlp/attribute_charge.

1 Introduction

The task of automatic charge prediction aims to train a machine judge to determine the final charges (e.g., *theft*, *robbery* or *traffic offence*.) of the defendants in criminal cases. As a representative subtask of legal judgment prediction, charge prediction plays an important role in legal assistant systems and can benefit many real-world applications. For example, it can provide a handy reference for legal experts (e.g., lawyers and judges) and improve their working efficiency. Meanwhile, it can supply ordinary people who are unfamiliar with legal terminology and complex procedures with legal consulting.

As a typical task in legal intelligence, automatic charge prediction has been studied for decades and most existing works formalize this task under the text classification framework. At the early stage, researchers pay great efforts to extract efficient features from text or case profiles. For example, some works (Liu et al., 2004; Liu and Hsieh, 2006) utilize shallow textual features, including characters, words, and phrases, to predict charges. Katz et al. (2017) predict the US Supreme Court's decisions with efficient features extracted from case profiles (e.g., dates, locations, terms, and types). All these approaches require numerous human effort to design features and annotate training instances. Besides, these methods are hard to scale to other scenarios. Inspired by the successful usage of deep neural networks on natural language processing tasks (Kim, 2014; Baharudin et al., 2010; Tang et al., 2015), researchers propose to employ deep neural networks to model legal documents. For example, Luo et al. (2017) propose an attention-based neural network for charge prediction by incorporating the relevant law articles.

However, charge prediction is still confronted with two major challenges which make it non-trivial:

*Indicates equal contribution.
†Corresponding Author.

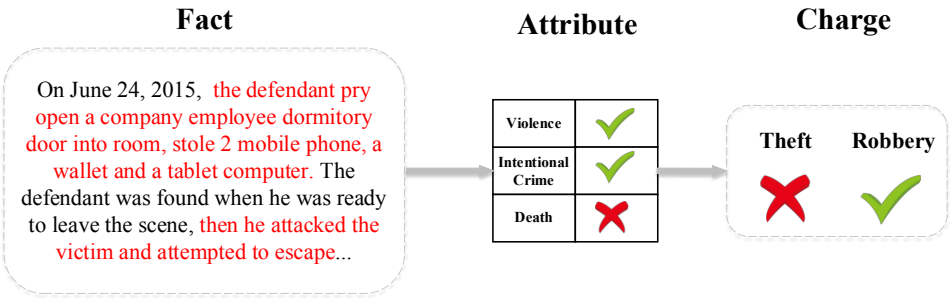

Figure 1: An illustration of the attribute-based charge prediction.

Few-Shot Charges. In practice, the case numbers of various charges are highly imbalanced. According to our statistics on a real-world dataset, the most frequent 10 charges (e.g., *theft*, *intentional injury*, and *traffic offence*.) cover 78.1% cases. On the contrary, the most low-frequency 50 (e.g., *scalping relics*, *disrupting the order of the court*, and *tax-escaping*.) charges only cover less than 0.5% cases and most of these charges own only around ten cases correspondingly. Previous works usually focus on these common charges and ignore the few-shot ones. Though deep neural models advance feature-engineering based charge prediction methods, they are unable to handle few-shot charges well due to the requirement of sufficient training data. Therefore, how to deal with these charges with limited cases is critical to a robust and effective charge prediction system.

Confusing Charges. Besides few-shot charges, there also exist many confusing charge pairs, such as (*theft*, *robbery*) and (*misappropriation of funds*, *embezzlement*). For each confusing pair, the definitions of two charges only differ in the verification of a specific act and the circumstances in corresponding cases are usually similar to each other. As illustrated in Fig. 1, many *robbery* case also contain the act of *theft*, and the existence of violence is the only key factor to distinguish these two charges. Thus, how to capture the crucial factors for distinguishing confusing charges is another challenge of charge prediction.

To address these issues, we propose to introduce discriminative legal attributes of charges into consideration and take these attributes as the internal mapping between facts and charge. More specifically, we select 10 representative attributes of charges, including *violence*, *profit purpose*, *buying and selling* and so on. Afterwards, we conduct a low-cost category-level annotation, i.e., for each charge, we annotate the value (including *yes*, *no*, or *not available*) of each attribute. This annotation indicates if an attribute is the essential condition of a charge.

With the attribute annotation of charges, we propose a novel multi-task learning framework to predict the attributes and charges of each case simultaneously. In this model, we employ attribute attention mechanism to capture the critical factual information relevant to a specific attribute. After that, we combine these attribute-aware representations with an attribute-free fact representation to predict the final charges. There are two reasons for introducing legal attributes into our charge prediction model. On one hand, these attributes can provide explicit knowledge about how to distinguish confusing charges. On the other hand, these attributes are shared by all charges, and the knowledge can transfer from high-frequency charges to low-frequency ones. Even for the few-shot charges, we can learn an efficient attribute-aware representation for prediction.

To investigate the advantage of our model on handling few-shot and confusing charges, we conduct experiments on three real-world datasets of Chinese criminal cases. Experimental results demonstrate that our model significantly and consistently outperforms other state-of-the-art models on all datasets and evaluation metrics. It is worth noting that, our model outperforms other baselines by more than 50% for the few-shot charges.

To summarize, we make three main contributions as follows:

(1) We are the first to focus on the few-shot and confusing problems in charge prediction. To address these issues, we introduce legal attributes of charges into charge prediction task for the first time.

(2) We propose a novel multi-task learning framework to infer the attributes and charges of a case jointly. To achieve it, we employ attribute attention mechanism to learn attribute-aware fact representa-

tions.

(3) We conduct efficient experiments on several real-world datasets, and our model significantly outperforms other baselines and achieves more than 50% improvements for few-shot charges.

2 Related Work

2.1 Zero-Shot Classification

Our work is relevant to zero-shot classification in computer vision. Many attribute-based models have been proposed under this task since attributes are shared among different classes and can offer an intermediate representation. Lampert et al. (2014) introduces direct attribute prediction(DAP) and indirect attribute prediction(IAP), and proposes attribute classifiers which can be pre-trained and dont need retraining when finding new suitable object class. Akata et al. (2013) proposes to transform the task of attribute-based classification to the label-embedding task. Jayaraman and Grauman (2014) introduces a random forest method stressing the unreliability of attribute prediction for unseen classes. They also extend it to the few-shot scenario.

Other than attributes, other external information can also be introduced to promote zero-shot classification. Elhoseiny et al. (2014) makes use of text description of the class label to transfer knowledge between text features and visual features. Zero-shot learning has also been used in applications besides object recognition, such as activity recognition (Zellers and Choi, 2017) and event recognition (Wu et al., 2014).

2.2 Charge Prediction

Researchers in the legal area have been working on automatically making the legal judgment for a long time. Kort (1957) applies quantitative methods to predict judgment by calculation numerical values for factual elements. Nagel (1963) makes use of correlation analysis to make predictions for reapportioning cases. Keown (1980) introduced mathematical models used for legal prediction such as linear models and the scheme of nearest neighbors. These methods are usually mathematical or quantitative, and they are restricted to a small dataset with few labels.

Since machine learning has been proven successful in many areas, researchers begin to formalize charge prediction as a text classification problem and make use of machine learning methods. Such work usually focuses on feature extraction from the case fact. Lin et al. (2012) fetches 21 legal factor labels for case classification. Mackaay and Robillard (1974) extracts N-grams and topics created by clustering semantically similar N-grams as features. Sulea et al. (2017) proposes a system based on SVM ensembles using the case description, ruling and time span of a case as input. However, these methods only extract shallow text features or manual labels which are hard to gather on a larger dataset. Whats more, the conventional models could not catch the subtle difference between similar crimes, thus they wouldnt perform well when the number of classes increases and more similar crimes appear.

With the successful usage of neural network methods on speech (Mikolov et al., 2011; Hinton et al., 2012; Dahl et al., 2012; Sainath et al., 2013), computer vision (CV) (Krizhevsky et al., 2012; Farabet et al., 2013; Tompson et al., 2014; Szegedy et al., 2015) and natural language processing (NLP) (Collobert et al., 2011; Kim, 2014; Bordes et al., 2014; Sutskever et al., 2014; Jean et al., 2015; Yang et al., 2016), researchers propose to employ neural models for legal tasks. Luo et al. (2017) proposes a hierarchical attentional network to predict charges and extract relevant articles jointly. However, this work only focuses on high-frequency charges, without paying attention to few-shot and confusing ones. To address these issues, we propose an attention-based neural model by incorporating several discriminative legal attributes.

3 Method

In this section, we propose a few-shot neural model which jointly models charge prediction task and legal attribute prediction task in a unified framework. In the following parts, we first introduce the discriminative charge attributes. Afterward, we give definitions of charge prediction and attribute prediction. Then

we describe the neural encoder of fact description and the attention-based attribute predictor. At last, we show the output layer and the loss function of our model.

3.1 Discriminative Charge Attributes

To distinguish confusing charges and provide additional knowledge for few-shot charges, we introduce 10 discriminative attributes for all the charges in Chinese criminal law. The detailed descriptions of these attributes are shown in Table 1. For each (*charge, attribute*) pair, it can be labeled as *Yes*, *No* or *NA*. For example, the charge of *manslaughter* should be labeled as *No* on *Intentional Crime*, *Yes* on *Death*, *NA* on *State Organ*. Note that, the fact-findings of a specific case can only be labeled as *Yes* or *No*. When convicting someone of a certain crime, the facts should conform to the description of the certain charge. Thus for a certain attribute, the label of a specific case and the label of the corresponding charge should be the same or not in conflict. In other words, for a certain attribute, the label of a case and the charge can only be (*Yes*, *Yes*), (*No*, *No*), (*Yes*, *NA*), or (*No*, *NA*). In practice, we conduct a low-cost annotation and annotate the attributes of 149 distinct charges manually. Then, we assign each case with the same attributes of its corresponding charge.

Attributes	Description
Profit Purpose	Whether the criminal commits a crime on the purpose of getting profit.
Buying and Selling	Whether the criminal has buying or selling behavior during the commission of the crime.
Death	Whether death is caused by the criminal.
Violence	Whether the criminal has the act of violence.
State Organ	Whether the case or the charge involves State organ or any functionary of a State organ.
Public Place	Whether the criminal commits a crime in a public place.
Illegal Possession	Whether the criminal commits a crime for the purpose of illegal possession.
Physical Injury	Whether a physical injury is caused by the criminal.
Intentional Crime	Whether the criminal commits an intentional crime.
Production	Whether the criminal commits a crime during the production.

Table 1: The descriptions of selected attributes.

3.2 Formalizations

3.2.1 Charge Prediction

The fact description of a case can be seen as a word sequence $\mathbf{x} = \{x_1, \ldots, x_n\}$, where n represents the sequence length, $x_i \in T$, and T is a fixed vocabulary. Given the fact description $\mathbf{x}$, the charge prediction task aims to predict a charge $y \in Y$ from a charge set Y.

3.2.2 Attributes Prediction

The attributes prediction task can be regarded as a binary classification task. It takes the same input sequence $\mathbf{x}$ as in the charge prediction task, and aims to predict the fact-findings of attributes $\mathbf{p} = \{p_1, \ldots, p_k\}$ according to the fact. Here, k is the number of selected attributes, and $p_i \in \{0, 1\}$ is the label for a certain attribute.

3.3 Fact Encoder

As illustrated in Fig. 2, fact encoder encodes the discrete input sequence into continuous hidden states. Here, we employ conventional Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) as fact encoder due to its ability to extract semantic meanings. LSTM is a variation of RNN and is capable of capturing long-term dependencies.

First, LSTM encoder converts each word $x_i \in \mathbf{x}$ into its word embedding $\mathbf{x}_i \in \mathbb{R}^d$, where d is the dimension of word embeddings. Then, we get the corresponding word embedding sequence as

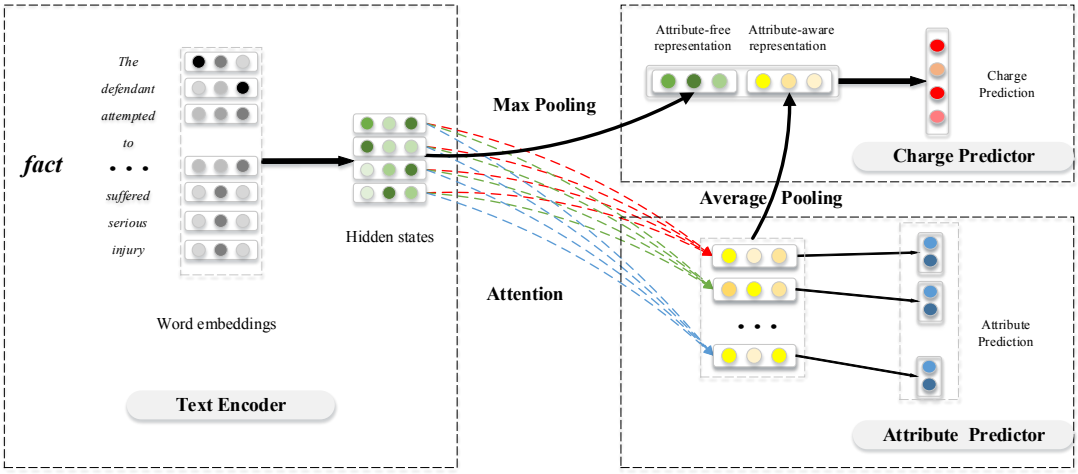

Figure 2: An illustration of the attribute-based charge prediction.

$\hat{\mathbf{x}} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$.

At each time step $t \in [1, n]$, the LSTM cell intakes $\mathbf{x}_t$, recalculates memory cell $\mathbf{c}_t$, and outputs new hidden state $\mathbf{h}_t$ as follows:

$$\begin{aligned}
\mathbf{f}_t &= \sigma\left(W_f \mathbf{x}_t + \mathbf{U}_f \mathbf{h}_{t-1} + \mathbf{b}_f\right), \\
\mathbf{i}_t &= \sigma\left(W_i \mathbf{x}_t + \mathbf{U}_i \mathbf{h}_{t-1} + \mathbf{b}_i\right), \\
\mathbf{o}_t &= \sigma\left(W_o \mathbf{x}_t + \mathbf{U}_o \mathbf{h}_{t-1} + \mathbf{b}_o\right), \\
\hat{\mathbf{c}}_t &= \tanh\left(W_c \mathbf{x}_t + \mathbf{U}_c \mathbf{h}_{t-1} + \mathbf{b}_c\right), \\
\mathbf{c}_t &= \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \hat{\mathbf{c}}_t, \\
\mathbf{h}_t &= \mathbf{o}_t \odot \tanh\left(\mathbf{c}_t\right).
\end{aligned} \tag{1}$$

Here, $\mathbf{f}_t$, $\mathbf{i}_t$ and $\mathbf{o}_t$ represent forget gate, input gate, and output gate respectively. $\odot$ means element-wise multiplication and σ is the sigmoid activation function. W, U, and b are weight matrices and bias vectors. After processing all time steps, we get a hidden state sequence $\mathbf{h} = \{\mathbf{h}_1, \ldots, \mathbf{h}_n\}$. At last, we feed it into a max-pooling layer to get the attribute-free representation $\mathbf{e} = [e_1, \ldots, e_s]$ as

$$e_i = \max(\mathbf{h}_{1,i}, \cdots, \mathbf{h}_{n,i}), \forall i \in [1, s]. \tag{2}$$

Here, s is the dimension of hidden states.

3.4 Attentive Attribute Predictor

Given the fact description $\mathbf{x}$, the attribute predictor aims to predict the label of every attribute. Inspired by (Yang et al., 2016), we employ an attention mechanism to select relevant information from facts and generate attribute-aware fact representations.

As shown in Fig. 2, attribute predictor takes the hidden state sequence $\mathbf{h} = \{\mathbf{h}_1, \ldots, \mathbf{h}_n\}$ as input. Our attentive attribute predictor then calculates attention weights $\mathbf{a} = \{\mathbf{a}_1, \ldots, \mathbf{a}_k\}$ for all attribute, where $\mathbf{a}_i = [a_{i,1}, \ldots, a_{i,n}]$. $\forall i \in [1, k]$ and $j \in [1, n]$, $a_{i,j}$ is calculated by:

$$a_{i,j} = \frac{\exp(\tanh(\mathbf{W}^a \mathbf{h}_j)^T \mathbf{u}_i)}{\sum_t \exp(\tanh(\mathbf{W}^a \mathbf{h}_t)^T \mathbf{u}_i)}. \tag{3}$$

Here, $\mathbf{u}_i$ is the context vector of the i-th attribute to calculate how informative an element is to the attribute i, and $\mathbf{W}^a$ is a weight matrix that all attributes share. Afterwards, we get attribute-aware representations of fact $\mathbf{g} = \{\mathbf{g}_1, \ldots, \mathbf{g}_k\}$, where $\mathbf{g}_i = \sum_t a_{i,t} \mathbf{h}_t$. At last, with representations g, we project it into the label space and use softmax function to get the final prediction results $\mathbf{p} = [p_1, \ldots, p_k]$, where p_i is the prediction result of attribute i and is calculated by:

$$\begin{aligned}
\mathbf{z}_i &= \text{softmax}(\mathbf{W}_i^p \mathbf{g}_i + \mathbf{b}_i^p), \\
p_i &= \arg\max(\mathbf{z}_i).
\end{aligned} \tag{4}$$

Here, $\mathbf{z}_i$ is the prediction probability distribution on *Yes* and *No*. $\mathbf{W}_i^p$ and $\mathbf{b}_i$ are weight matrix and bias vector of attribute i.

3.5 Output Layer

To integrate the fact descriptions and fact-findings of all attributes, we use both attribute-free and attribute-aware representations to predict the final charge of a case in the output layer. The predicted distribution y over all charges is calculated as follows:

$$\begin{aligned} \mathbf{r} &= \frac{\sum_i \mathbf{g_i}}{k}, \\ \mathbf{v} &= \mathbf{e} \oplus \mathbf{r}, \\ y &= \text{softmax}(\mathbf{W}^y \mathbf{v} + \mathbf{b}^y). \end{aligned} \tag{5}$$

Here, $\mathbf{r}$ is the average of attribute-aware representations. $\mathbf{r}$ and $\mathbf{e}$ are concatenated into the final fact representation $\mathbf{v}$. $\mathbf{W}^y$ and $\mathbf{b}^y$ are weight matrix and bias vector in the output layer.

3.6 Optimization

The training objective of our proposed model consists of two parts. The first one is to minimize the cross-entropy between predicted charge distribution y and the ground-truth distribution $\hat{y}$. The other one is to minimize the cross-entropy between predicted distribution and the ground-truth fact-finding of each attribute.

The charge prediction loss can be formalized as:

$$\mathcal{L}_{charge} = -\sum_{i=1}^{C} y_i \cdot log(\hat{y}_i), \tag{6}$$

where y_i is the ground-truth label, $\hat{y}_i$ is prediction probability, and C is the number of charges.

As each attribute is equally important in the model, we can easily calculate the attribution loss by sum up the cross-entropy of all attributes. However, when the attribute of a specific charge is *NA*, the label of the corresponding cases can be *Yes* or *No*. Therefore, we only add up the cross-entropy to the attribute loss when this attribute of the charge belongs to *Yes* or *No*. At last, we formulate the attribute loss as:

$$\mathcal{L}_{attr} = -\sum_{i=1}^{k} I_i \sum_{j=1}^{2} z_{ij} \cdot log(\hat{z}_{ij}), \tag{7}$$

where I_i is an indicator function. $I_i = 1$ if the i-th attribute of current charge is labeled as *Yes* or *No*, and $I_i = 0$ otherwise. Obviously, z_i is the ground-truth label, and $\hat{z}_i$ is predicted probabilities distribution on *Yes* and *No*.

Considering the two objectives, our final loss function $\mathcal{L}$ is obtained by adding $\mathcal{L}_{charge}$ and $\mathcal{L}_{attr}$ as follows:

$$\mathcal{L} = \mathcal{L}_{charge} + \alpha \cdot \mathcal{L}_{attr}, \tag{8}$$

where α is a hyper-parameter to balance the weight of the two parts in the loss function.

4 Experiments

In order to investigate the effectiveness of our model on criminal charges prediction, we conduct experiments on several real-world datasets and compare our model with several state-of-the-art baselines.

4.1 Dataset Construction

Since there are no publicly available datasets in previous works for charges prediction, we collect criminal cases published by the Chinese government from China Judgments Online[1]. As each case is well-structured and divided into several parts such as fact, court view, and penalty result, we select the fact

[1] `http://wenshu.court.gov.cn`.

part of each case as our input. Besides, we can easily extract the charges from the penalty result by regular expression. We have manually checked the extracted charges and there are few mistakes.

Although there are some cases that contain multiple defendants and multiple charges in real-world, considering the task would be too complex to solve if these cases contained, we removed the cases which have more than one charges in a verdict. Besides, in order to examine the performance of our method on few-shot charges, we keep 149 distinct charges (near 3 times as compared with (Luo et al., 2017)) with at least 10 cases.

After preprocessing, we randomly select about $400,000$ cases and construct three datasets with different scales, denoted as **Criminal-S(small)**, **Criminal-M(medium)** and **Criminal-L(large)**. The three different datasets contain the same number of charges but the different number of cases. The detailed statistics are shown in Table 2.

Datasets	**Criminal-S**	**Criminal-M**	**Criminal-L**
train	$61,589$	$153,521$	$306,900$
test	$7,702$	$19,189$	$38,368$
valid	$7,755$	$19,250$	$38,429$

Table 2: The statistics of different datasets.

4.2 Attribute Selection and Annotation

As mentioned in previous part, we propose to introduce discriminative attributes to enhance charge prediction. To select these attributes, we first train a LSTM based charge prediction model and obtain the confusion matrix of predicted charges on validation set. Then, we filter out the confusing charge pairs and provide them to three master students majoring in criminal. According to these confusing charge pairs, they define 10 representative attributes to distinguish these confusing pairs.

With the selected 10 attributes, we conduct a low-cost annotation over all charges. Here, the low-cost annotation means we only need to annotate 10 attributes for 149 charges manually, rather than all cases. As the selected attributes are discriminative and unambiguous, we asked these annotators to reach an agreement for each annotation. Totally, we spent less than 10 hours for annotation.

4.3 Baselines

We employ several typical text classification models and one charge predicting model as baselines:

TFIDF+SVM: We implement term-frequency inverse document frequency (TFIDF) (Salton and Buckley, 1988) to extract features of inputs, and employ SVM (Suykens and Vandewalle, 1999) as the classifier.

CNN: We implement the CNN with multiple filter widths (Kim, 2014) as text classifier.

LSTM: We implement a two-layer LSTM (Hochreiter and Schmidhuber, 1997) with a max-pooling layer as the fact encoder.

Fact-Law Attention Model: Luo et al. (2017) propose an attention-based neural charge prediction model by incorporating relevant law articles.

4.4 Experiment Settings and Evaluation Metrics

As all the case documents are written in Chinese without word cutting, we employ THULAC (Sun et al., 2016) for word segmentation and set the maximum document length to 500. For the TFIDF+SVM model, we set the feature size to $2,000$. For other neural models, we employ Skip-Gram model (Mikolov et al., 2013) to pre-train word embeddings with the embedding size of 100. We set the hidden state size of LSTM to 100.For the CNN based models, we set the filter widths to $(2, 3, 4, 5)$ with each filter size to 25 for consistency. The weight α of the attribute loss is set to 1.

Note that, the representation size of our model turns into 200 after concatenation. For a fair comparison, we add a 100×200 fully connected layer between after the pooling layer in CNN and LSTM,

denoted as CNN-200 and LSTM-200.

We use Adam (Kingma and Ba, 2015) as the optimizer, and set the learning rate to 0.001, the dropout rate (Srivastava et al., 2014) to 0.5 and the batch size to 64. We employ accuracy (Acc.), macro-precision (MP), macro-recall (MR) and macro-F1 as our evaluation metrics.

4.5 Results and Analysis

Datasets	Criminal-S				Criminal-M				Criminal-L			
Metrics	Acc.	MP	MR	F1	Acc.	MP	MR	F1	Acc.	MP	MR	F1
TFIDF+SVM	85.8	49.7	41.9	43.5	89.6	58.8	5 0.1	52.1	91.8	67.5	54.1	57.5
CNN	91.9	50.5	44.9	46.1	93.5	57.6	48.1	50.5	93.9	66.0	50.3	54.7
CNN-200	92.6	51.1	46.3	47.3	92.8	56.2	50.0	50.8	94.1	61.9	50.0	53.1
LSTM	93.5	59.4	58.6	57.3	94.7	65.8	63.0	62.6	95.5	69.8	67.0	66.8
LSTM-200	92.7	60.0	58.4	57.0	94.4	66.5	62.4	62.7	95.1	72.8	66.7	67.9
Fact-Law Att.	92.8	57.0	53.9	53.4	94.7	66.7	60.4	61.8	95.7	73.3	67.1	68.6
Our Model	**93.4**	**66.7**	**69.2**	**64.9**	**94.4**	**68.3**	**69.2**	**67.1**	**95.8**	**75.8**	**73.7**	**73.1**

Table 3: Charge prediction results of three datasets.

As shown in Table 3, we can observe that our model significantly and consistently outperforms all the baselines. Almost all existing methods perform poorly under the macro-F1 metric, which indicates their shortage of predicting few-shot charges. Conversely, our model achieves promising improvements (7.9%, 4.4%, and 5.2% absolutely on three datasets respectively), which demonstrates the robustness and effectiveness of our model.

To further verify the advance of our model on dealing with few-shot charges, we show the performance on charges with different frequencies. As shown in Table 4, we divide the charges into three parts according to their frequencies. Here, the charges with ≤ 10 cases are low-frequency, and the charges with > 100 cases are high-frequency. From this table, we find that our model achieves more than 50% improvements than baseline method for the low-frequency (i.e., few-shot) charges, which verifies the effectiveness of our model on handling few-shot issues.

Charge Type	Low frequency	Medium frequency	High frequency
Charge Number	49	51	49
LSTM-200	32.6	55.0	83.3
Our Model	**49.7 (↑ 17.1%)**	**60.0 (↑ 5.0%)**	**85.2 (↑ 1.9%)**

Table 4: Macro-F1 values of various charges on Criminal-S.

Datasets	Criminal-S				Criminal-M				Criminal-L			
Metrics	Acc.	MP	MR	F1	Acc.	MP	MR	F1	Acc.	MP	MR	F1
Our model	93.4	**66.7**	**69.2**	**64.9**	94.4	68.3	**69.2**	**67.1**	**95.8**	**75.8**	**73.7**	**73.1**
w/o attention	**93.5**	63.4	60.1	60.0	94.7	**68.8**	58.2	60.9	94.9	70.9	54.4	58.6
w/o concatenation	**93.5**	59.3	59.0	57.2	**95.0**	64.6	62.4	62.5	95.7	69.4	64.5	65.4

Table 5: Experimental results of ablation test.

4.6 Ablation Test

Our method is characterized by the incorporation of attention mechanism and attribute-aware representations. Thus, we design ablation test respectively to investigate the effectiveness of these modules. When taken off the attention mechanism, for each attribute we replace attention mechanism with a fully connected layer. When taken off the attribute-aware representations (i.e., without concatenating the averaged

Task	Charge Prediction	Attribute Prediction on *physical injury*
Ground Truth	Intentional Injury	Yes
Our model	Intentional Injury	Yes
LSTM-200	Affray	N/A

Table 6: Charge and attribute prediction result of the selected case.

attribute-aware representation), our method degrades into a typical multi-task learning based on LSTM for both charge and attribute prediction.

As shown in Table 5, we can observe that the performance drops obviously after removing the attention layer or the concatenation. The macro-F1 decreases at least 4%. Therefore, it can be seen that both attention mechanism and attribute-aware fact representation play irreplaceable roles in our model.

4.7 Case Study

In this part, we select a representative case to give an intuitive illustration of how the predicted attributes help to promote the performance of charge prediction. In this case, the defendant is convicted of intentional injury. It is often hard to decide whether to judge a case as affray or intentional injury since they are both related to violence. One important difference between them is that intentional injury has the feature of physical injury, while affray does not.

So we believe the attribute physical injury is essential in the charge prediction of this case. As shown in Table 6, our model correctly predicts the label of *physical injury* as *Yes*, and consequently predicts the charge as *intentional injury*. In contrary, the model LSTM-200 predicts it as *affray* incorrectly. In addition, we visual the heat map of this case when predicting the attribute $intentionalinjury$. Words with deeper background color have higher attention weights. From this figure, we observe that the attention mechanism can capture key patterns and semantics relevant to current attribute.

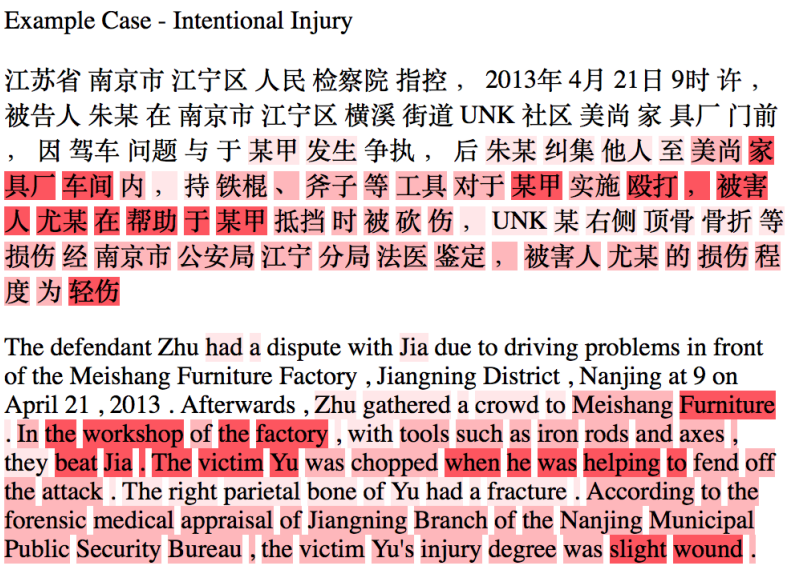

Figure 3: Visualization of attention mechanism.

.

5 Conclusion

In this work, we focus on the task of charge prediction according to the fact descriptions of criminal cases. To address the problem of prediction few-shot and confusing charges, we introduce discriminative legal

attributes into consideration and propose a novel attribute-based multi-task learning model for charge prediction. Specifically, our model learns attribute-free and attribute-aware fact representation jointly by utilizing attribute-based attention mechanism.

In future, we will explore the following directions:

(1) There are more complicated criminal cases, such as multiple defendants and charges. Thus, it is challenging to handle this general form of charge prediction.

(2) In this work, we only utilize several simple attributes of charges, while there exist more complex essential conditions of charges. How to take full usage of essential conditions of charges is expected to improve the interpretability of charge prediction models.

Acknowledgements

We thank all the anonymous reviewers for their insightful comments. This work is supported by the National Natural Science Foundation of China (NSFC No. 61661146007, 61572273) and Tsinghua University Initiative Scientific Research Program (20151080406). This research is part of the NExT++ project, supported by the National Research Foundation, Prime Ministers Office, Singapore under its IRC@Singapore Funding Initiative.

References

Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. 2013. Label-embedding for attribute-based classification. In *Proceedings of CVPR*, pages 819–826.

Baharum Baharudin, Lam Hong Lee, and Khairullah Khan. 2010. A review of machine learning algorithms for text-documents classification. *JAIT*, 1(1):4–20.

Antoine Bordes, Sumit Chopra, and Jason Weston. 2014. Question answering with subgraph embeddings. In *Proceedings of EMNLP*, pages 615–620.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *JMLR*, 12:2493–2537.

George E Dahl, Dong Yu, Li Deng, and Alex Acero. 2012. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE TASLP*, 20(1):30–42.

Mohamed Elhoseiny, Babak Saleh, and Ahmed Elgammal. 2014. Write a classifier: Zero-shot learning using purely textual descriptions. In *Proceedings of ICCV*, pages 2584–2591.

Clement Farabet, Camille Couprie, Laurent Najman, and Yann LeCun. 2013. Learning hierarchical features for scene labeling. *IEEE TPAMI*, 35(8):1915–1929.

Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Dinesh Jayaraman and Kristen Grauman. 2014. Zero-shot recognition with unreliable attributes. In *Proceedings of NIPS*, pages 3464–3472.

Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2015. On using very large target vocabulary for neural machine translation. In *Proceedings of ACL*, volume 1, pages 1–10.

Daniel Martin Katz, Michael J Bommarito II, and Josh Blackman. 2017. A general approach for predicting the behavior of the supreme court of the united states. *PloS one*, 12(4):e0174698.

R Keown. 1980. Mathematical models for legal prediction. *Computer/Law Journal*, 2:829.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of EMNLP*, pages 1746–1751.

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of ICLR*.

Fred Kort. 1957. Predicting supreme court decisions mathematically: A quantitative analysis of the ”right to counsel” cases. *American Political Science Review*, 51(1):1–12.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Proceedings of NIPS*, pages 1097–1105.

Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. 2014. Attribute-based classification for zero-shot visual object categorization. *IEEE TPAMI*, 36(3):453–465.

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 *Processdings of ROCLING*, pages 140–141.

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

Chao-Lin Liu, Cheng-Tsung Chang, and Jim-How Ho. 2004. Case instance generation and refinement for case-based criminal summary judgments in chinese. *JISE*, pages 783–800.

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 EMNLP*, pages 2727–2736.

Ejan Mackaay and Pierre Robillard. 1974. Predicting judicial decisions: The nearest neighbour rule and visual representation of case patterns. *Datenverarbeitung im Recht*, 3(3/4):302–331.

Tomáš Mikolov, Anoop Deoras, Daniel Povey, Lukáš Burget, and Jan Černockỳ. 2011. Strategies for training large scale neural network language models. In *Proceedings of ASRU workshop*, pages 196–201.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS*, pages 3111–3119.

Stuart S Nagel. 1963. Applying correlation analysis to case prediction. *Texas Law Review*, 42:1006.

Tara N Sainath, Abdel-rahman Mohamed, Brian Kingsbury, and Bhuvana Ramabhadran. 2013. Deep convolutional neural networks for lvcsr. In *Proceedings of ICASSP*, pages 8614–8618.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523.

Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *JMLR*, 15(1):1929–1958.

Octavia Maria Sulea, Marcos Zampieri, Mihaela Vela, and Josef Van Genabith. 2017. Exploring the use of text classi cation in the legal domain. In *Proceedings of ASAIL workshop*.

Maosong Sun, Xinxiong Chen, Kaixu Zhang, Zhipeng Guo, and Zhiyuan Liu. 2016. THULAC: An efficient lexical analyzer for chinese.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of NIPS*, pages 3104–3112.

Johan AK Suykens and Joos Vandewalle. 1999. Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich, Jen-Hao Rick Chang, et al. 2015. Going deeper with convolutions. In *Proceedings of CVPR*, pages 1–9.

Duyu Tang, Bing Qin, and Ting Liu. 2015. Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of EMNLP*, pages 1422–1432.

Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. 2014. Joint training of a convolutional network and a graphical model for human pose estimation. In *Proceedings of NIPS*, pages 1799–1807.

Shuang Wu, Sravanthi Bondugula, Florian Luisier, Xiaodan Zhuang, and Pradeep Natarajan. 2014. Zero-shot event detection using multi-modal fusion of weakly supervised concepts. In *Proceedings of CVPR*, pages 2665–2672.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J Smola, and Eduard H Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of NAACL*, pages 1480–1489.

Rowan Zellers and Yejin Choi. 2017. Zero-shot activity recognition with verb attribute induction. In *Proceedings of EMNLP*, pages 946–958.