Few-Shot Charge Prediction with Multi-Grained Features and Mutual Information

Han Zhang[1], Zhicheng Dou[2,*],Yutao Zhu[3],Jirong Wen[4,5]
School of Information / Renmin University of China[1]
Gaoling School of Artificial Intelligence / Renmin University of China[2]
Université de Montréal
Beijing Key Laboratory of Big Data Management and Analysis Methods
Key Laboratory of Data Engineering and Knowledge Engineering, MOE
zhanghanjl@ruc.edu.cn, dou@ruc.edu.cn

Abstract

Charge prediction aims to predict the final charge for a case according to its fact description and plays an important role in legal assistance systems. With deep learning based methods, prediction on high-frequency charges has achieved promising results but that on few-shot charges is still challenging. In this work, we propose a framework with multi-grained features and mutual information for few-shot charge prediction. Specifically, we extract coarse- and fine-grained features to enhance the model's capability on representation, based on which the few-shot charges can be better distinguished. Furthermore, we propose a loss function based on mutual information. This loss function leverages the prior distribution of the charges to tune their weights, so the few-shot charges can contribute more on model optimization. Experimental results on several datasets demonstrate the effectiveness and robustness of our method. Besides, our method can work well on tiny datasets and has better efficiency in the training, which provides better applicability in real scenarios.

1 Introduction

Charge prediction aims to determine the final charge (e.g., *manslaughter*, *traffic offence*, or *theft*) for a case by analyzing the textual fact description of the defendants' behavior. As a subtask of legal judgment prediction (LJP), charge prediction plays an important role in legal assistance systems, thus has been widely applied in real scenarios. On the one hand, a charge prediction system can provide legal professionals, such as judges and lawyers, a quick and effective reference to improve their work efficiency. On the other hand, it can also provide non-legal professionals with some simple and useful legal advice.

Automatic charge prediction has been studied for decades. Early studies focused on applying mathematical or statistical methods, such as counting the specific attributes (e.g., crime time and place) of the cases (Kort, 1957; Keown, 1980). Later, researchers began to frame the charge prediction as a text classification problem and paid attention to designing manual features or extracting shallow features from fact descriptions to predict the charge (Liu et al., 2004; Liu and Hsieh, 2006; Katz et al., 2017). However, these features rely heavily on human expertise and are specific for different types of cases, which limits their application to a larger range of domains. Recently, deep learning based methods have also achieved promising results on charge prediction due to their superiority on automatic feature extraction and combination (Luo et al., 2017; Zhong et al., 2018; Yang et al., 2019; Xu et al., 2020).

However, the charge prediction is still a non-trivial problem. One of the challenges is **few-shot charges**, which is also the focus of this paper. In practice, the numbers of cases in different charges usually follow a long-tailed distribution, which means their case data are highly imbalanced. For example, in the real-world dataset Criminal (Hu et al., 2018), the most frequent ten charges (such as *theft* and *intentional injury*) cover around 78% cases, while the lowest frequent fifty charges (such as *scalping relics* and *tax-escaping*) cover less than 0.5% cases. Under this circumstance, deep learning based methods can hardly perform well because the training data are insufficient for these few-shot changes. Therefore, how to deal with the few-shot charges with limited cases is crucial for building a robust and effective charge prediction system.

Proceedings of the 20th China National Conference on Computational Linguistics, pages 1154-1166, Hohhot, China, August 13 - 15, 2021.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

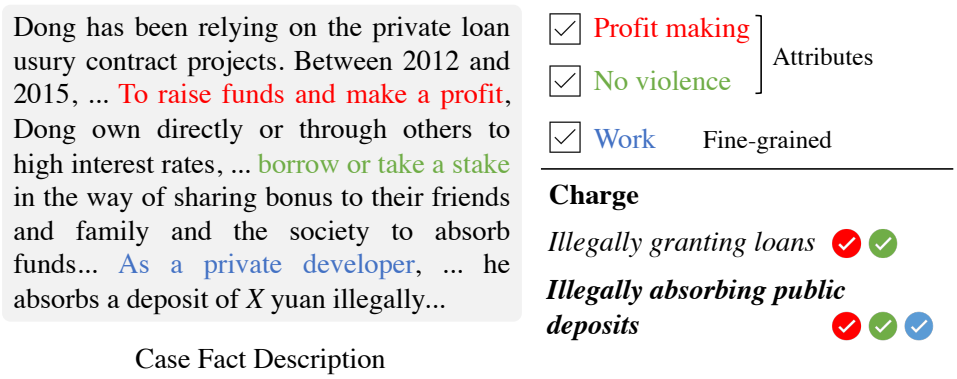

Figure 1: An example of charge prediction. The two few-shot charges have the same attributes, thus the fine-grained features are necessary for distinguishing them.

To alleviate this problem, Hu et al. (2018) introduced the attribute features of the law, which are shared by all charges, so as to transfer knowledge from high-frequency charges to low-frequency ones. He et al. (2019) proposed a sequence enhanced capsule model and leveraged the focal loss to alleviate the few-shot problem. However, we find that they still have some limitations: (1) The introduced attribute features are artificial, and they are usually very discrete on the cases of the few-shot charges and thus contribute less to distinguishing them from other charges. For example, the charge *illegally granting loans* and *illegally absorbing public deposits* have the same characteristics on profit-making purpose and nonviolent crime, and they can only be distinguished by more fine-grained characteristics such as the defendant's affiliation. (2) Existing works are all optimized by the cross entropy or its similar variants, and they do not consider that the cross-entropy is easily affected by the prior distribution, which makes it difficult to classify few-shot charges. For example, charge A has 1,000 cases and few-shot charge B has only 5 cases, charge A contributes more to the cross entropy loss as the loss of each sample is directly added up.

To tackle these problems, in this work, we **first** propose using multi-grained features. We introduce a convolutional network (CNN) with multiple kernels to extract coarse-grained features and a bilinear CNN (Lin et al., 2015) to extract fine-grained features from the case descriptions. These two kinds of features are then fused by a capsule network. The attribute features provided by Hu et al. (2018) are also leveraged as additional explicit knowledge. The fine-grained features, fused features, and attribute features are finally combined by a multi-layer perceptron for charge prediction. By this means, the representation of each description can be greatly enhanced and the classifier can obtain more information for predicting the charge. **Second**, we consider the prior probability distribution of different charges over the dataset, based on which we construct a mutual information loss function. The few-shot charges can thus be paid more attention during the training process. Finally, the whole model is optimized by both the charge prediction and attribute prediction tasks. Experimental results on a series of datasets with different sizes demonstrate the effectiveness and wide applicability of our method.

2 Related Work

2.1 Charge Predication

Early work on charge prediction focused on quantitative and statistical methods. For example, Kort (1957) counted various facts in the case to predict the crime. Keown (1980) introduced a linear model and a nearest neighbor method to predict crime. With the success of machine learning in some areas, researchers begin to model the charge prediction problem as a text classification problem. The basic idea is to extract features from the case description and make predictions by machine learning methods, such as linear models, logistic model trees (Lin et al., 2012), or SVMs (Sulea et al., 2017). However, these methods are built on manual features, which heavily rely on human expertise and are hard to apply on large datasets with various charges.

Recently, deep learning based methods have achieved promising results on charge prediction or legal

Proceedings of the 20th China National Conference on Computational Linguistics, pages 1154-1166, Hohhot, China, August 13 - 15, 2021.

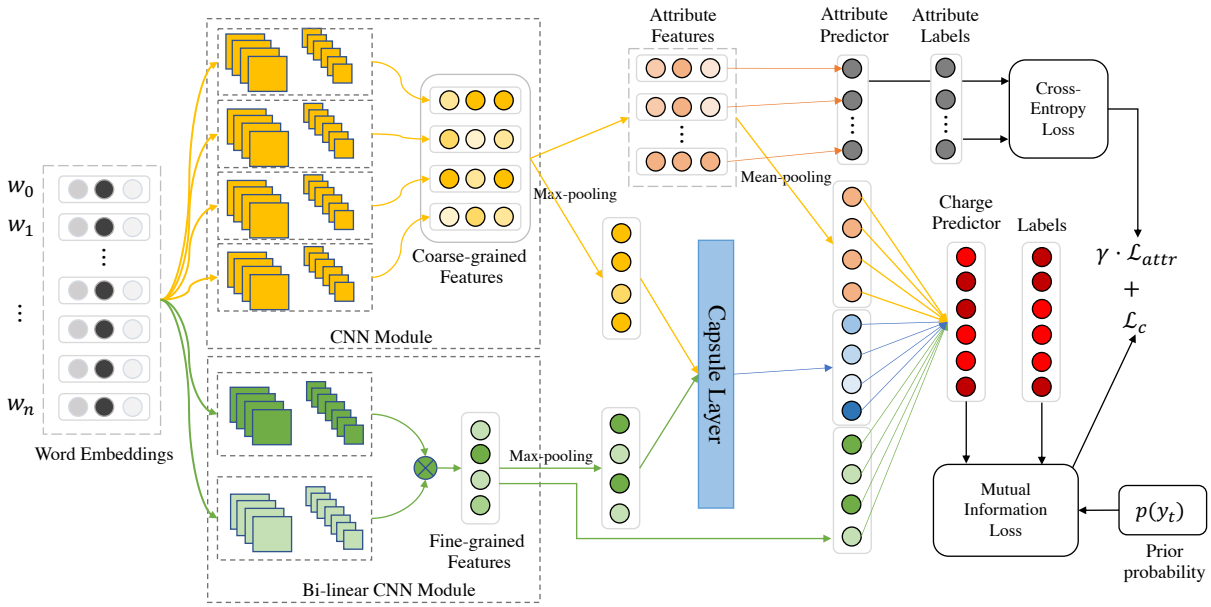

Figure 2: The framework of MFMI.

judgement prediction. For example, Wang et al. (2019) proposed a hierarchical text matching model to predict the cases. Xu et al. (2020) improved performance by introducing all laws into legal judge prediction task and building a relationship graph of all laws to distinguish the confusable cases. Liu et al. (2019) used a seq2seq model with attention to predict the cause of the decision. Chen et al. (2019) adopt gating mechanism to improve penalty prediction based on the charge. Pan et al. (2019) applied multi-scale attention to deal with the cases of multiple defendants. To deal with the problem of few-shot charge prediction, Hu et al. (2018) introduced the attribute characteristics (manually defined) as expertise knowledge into the charge prediction task, while He et al. (2019) applied a focal loss and designed a capsule network.

Different from the existing work, we propose to extract fine-grained features automatically from the case descriptions to enhance the representation and design a mutual information loss function to take the prior distribution of different charges into account. Both strategies are helpful in improving the performance of few-shot charges.

2.2 Few-shot Text Classification

In recent years, some studies have focused on text classification with few-shot samples. Gao et al. (2019) proposed a prototype network structure based on mixed attention. Considering the domain differences of the text, Xu et al. (2018) proposed the lifelong domain word embedding. Yu et al. (2018) proposed to integrate a variety of measures in cross-domain few-shot learning. For invisible classes and cold start problems, Xu et al. (2019) proposed an open world learning model to deal with invisible classes. Geng et al. (2019) proposed a dynamic routing induction method to encapsulate abstract class representations.

3 Methodology

We propose a multi-task framework based on Multi-grained Features and Mutual Information (MFMI) for few-shot charge prediction. The structure of our MFMI is shown in Figure 2. In general, MFMI considers three different kinds of features (namely coarse-grained features, fine-grained features, and attribute features) to represent a case and predict the charge based on the fused representations. To facilitate the few-shot charge prediction, we propose a loss function based on mutual information, with which the few-shot charges can contribute more on the model optimization, so the overall performance can be improved. Besides, similar to (Hu et al., 2018), we also add a supplementary task to predict whether the predefined attributes are contained in the fact, which is reported to be helpful in improving

Proceedings of the 20th China National Conference on Computational Linguistics, pages 1154-1166, Hohhot, China, August 13 - 15, 2021.

the performance. As a result, the whole model is trained with both our proposed mutual information loss on charge prediction and a cross-entropy loss on attribute prediction.

3.1 Task Formalization

Before introducing our model, we first give the formalization of the charge prediction task. Formally, assuming a fact description (text) containing n words as $X = (w_1, \cdots, w_n)$, where $w_i \in D$, and D is the fixed dictionary, the model is asked to predict (classify) the charge $y \in Y$ of the fact from the predefined charge set Y. Besides, as suggested by (Hu et al., 2018), we also predict the common attributes of the law articles as a supplementary task to boost the performance on few-shot charges prediction. It has the same input sequence X and aims at predicting the corresponding fact-findings of attributes $p = \{p_1, \cdots, p_k\}$ according to the fact. Here, k is the number of attributes, and $p_i \in \{0, 1\}$ is the label for a certain attribute. Generally, the charge prediction can be treated as a (multi-class) text classification task, while the attribute prediction can be regarded as a binary classification task.

3.2 Coarse-grained Features

As illustrated in Figure 2, the case description X is represented as a sequence of embeddings $\mathbf{X}$ by a looking-up operation on a pre-trained embedding table $\mathbf{E}$:

$$\mathbf{X} = [\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_n], \quad \mathbf{e}_i = \text{Look-Up}(\mathbf{E}(w_i)), \tag{1}$$

where $\mathbf{X} \in \mathbb{R}^{n \times d}$, and $\mathbf{e}_i \in \mathbb{R}^d$ is the embedding of the i-th word w_i in the fact description.

To further enhance the representation of the word, we apply 1D CNNs with different kernels over the embedding sequence and compute the coarse-grained features as:

$$\mathbf{X}_t^C = \text{1D-CNN}(\mathbf{X}, s_t), \quad t \in [1, T], \tag{2}$$

where s_t is the kernel size of the t-th CNN and T is the number of CNNs. With different kernels, the semantic information in the consecutive words can be integrated together. For example, if $s_t = 2$, we can obtain a sequence of bigram representations. As these features can only reflect shallow semantic information, we call them coarse-grained features. Note that we apply zero padding during the convolution, thus $\mathbf{X}_t^C$ has a dimension of $n \times m$, where m is the number of output channels. Other neural networks, such as RNN (Lai et al., 2015), can also be applied to represent the word sequence, here we choose CNN because it can be computed in parallel and provide better efficiency in a real system.

3.3 Fine-grained Features

In charge prediction, some charges usually have the same or similar fact descriptions. For example, the few-shot charge *illegally granting loans* and *illegally absorbing public deposits* are both nonviolent and have a profit-making purpose. It is hard to distinguish them based on only the coarse-grained features. Therefore, we propose to leverage a bilinear CNN module to extract fine-grained features for better distinguishing the charges. Bilinear CNN can integrate the information from two sub-CNNs by a bilinear transformation. In the original CNN structure, the features are usually refined through a max-pooling or mean-pooling operation, which can only take the first-order information into account. However, the pooling function of a bilinear CNN calculates the outer product of different feature channels. The outer product can capture the pairwise correlation between feature channels, providing a stronger characteristic representation than traditional CNNs.

Specifically, considering two 1D convolutional operations $f_A(\mathbf{X}, l)$ and $f_B(\mathbf{X}, l)$ computing at the position l of the embedding matrix $\mathbf{X}$, the bilinear transformation can be described as:

$$\text{Bilinear}(f_A, f_B) = f_A(\mathbf{X}, l)^\top f_B(\mathbf{X}, l). \tag{3}$$

As a result, if $f_A(\mathbf{X}, l)$ and $f_B(\mathbf{X}, l)$ have m_A and m_B channels respectively, the dimension of obtained features after bilinear transformation will be $m_A \times m_B$. Then, the features at all positions are aggregated

Proceedings of the 20th China National Conference on Computational Linguistics, pages 1154-1166, Hohhot, China, August 13 - 15, 2021.

by sum-pooling and produce an integrated feature representation $\Phi(\mathbf{X})$ as:

$$\Phi(\mathbf{X}) = \sum_{l \in L} \text{Bilinear}(f_A, f_B). \tag{4}$$

Finally, the output Φ of the bilinear transformation is flattened as a vector:

$$\mathbf{V} = \text{Flatten}(\Phi(\mathbf{X})). \tag{5}$$

To alleviate the overfitting problem (Lin et al., 2015), a signed square root and a L2 normalization are usually applied on $\mathbf{v}$ and output the fine-grained features $\mathbf{X}^F$:

$$\mathbf{V}' = \text{sign}(\mathbf{V})\sqrt{|\mathbf{V}|}, \quad \mathbf{X}^F = \mathbf{V}'/||\mathbf{V}'||_2. \tag{6}$$

3.4 Capsule Layer

A typical problem of CNNs is that they cannot obtain the relative position information among extracted features. To deal with this problem, we apply a capsule network (Sabour et al., 2017) to integrate the spatial information and fuse the features for better representation. Compared with traditional neural networks, the basic unit of capsule network is the capsule, which conducts a series of operations on input vectors rather than scalars. The connections between capsules are implemented by a dynamic routing mechanism, which contains several affine transformations and nonlinear functions[0]. Here, we use Capsule$(\cdot)$ to denote a capsule layer.

In particular, we use the capsule layer to aggregate the coarse-grained features $\mathbf{X}^C$ and fine-grained features $\mathbf{X}^F$ by:

$$\mathbf{X}^C = \text{Max-pooling}(\mathbf{X}_1^C, \cdots, \mathbf{X}_T^C), \quad \mathbf{X}^{\text{CAP}} = \text{Flatten}(\text{Capsule}(\mathbf{X}^C \oplus \mathbf{X}^F)), \tag{7}$$

where $\oplus$ is the concatenation operation. With such a capsule network, the spatial information of the features can be integrated into the refined representations.

3.5 Attribute Features

Inspired by (Hu et al., 2018), we also extract attribute features from the fact description. These attribute features are shared by various charges, thus can transfer knowledge from high-frequency charges to low-frequency ones.

Specifically, the attribute features are computed based on the coarse-grained features $\mathbf{X}_t^C$. First, we calculate attention weights $\mathbf{a} = \{\mathbf{a}_1, \cdots, \mathbf{a}_k\}$ for all attributes, where $a_i = [a_{i,1}, \cdots, a_{i,n}]$. For $\forall i \in [1, k]$ and $\forall j \in [1, T]$, $a_{i,j}$ is calculated by:

$$a_{i,j} = \frac{\exp(\tanh(\mathbf{W}^a \mathbf{X}_j^C)^\top \mathbf{u}_i)}{\sum_{t=1}^{T} \exp(\tanh(\mathbf{W}^a \mathbf{X}_t^C)^\top \mathbf{u}_i)}, \tag{8}$$

where $\mathbf{u}_i$ is the context vector of the i-th attribute (randomly initialized) to calculate how informative an element is to the attribute, and $\mathbf{W}^a$ is a weight matrix (parameter) shared by all attributes. Thereafter, we can obtain the fact-aware attribute features as:

$$\mathbf{X}_i^A = \sum_{t=1}^{T} a_{i,t} \mathbf{X}_t^C. \tag{9}$$

These features will be used for both charge prediction and attribute prediction (introduced in Section 3.7).

3.6 Aggregation and Prediction

To aggregate all obtained features, we concatenate them together and apply a multi-layer perceptron (MLP) to fuse them. Finally, the distribution Z_y of all crimes is calculated as:

$$Z_y = \text{MLP}(\mathbf{X}^F \oplus \mathbf{X}^{\text{CAP}} \oplus \mathbf{X}^A), \tag{10}$$

$$\mathbf{X}^A = \text{Mean-pooling}(\mathbf{X}_1^A, \cdots, \mathbf{X}_k^A). \tag{11}$$

[0]We suggest the reader to refer to the original paper (Sabour et al., 2017) for the details of capsule network.

Proceedings of the 20th China National Conference on Computational Linguistics, pages 1154-1166, Hohhot, China, August 13 - 15, 2021.

3.7 Optimization

In general, our model is a multi-task learning framework. Both a charge prediction loss and an attribute prediction loss are used for optimizing our model.

3.7.1 Charge Prediction Loss based on Mutual Information

As a multi-class classification problem, charge prediction is often optimized by a cross-entropy loss. Suppose that there are L categories, and the training data $(x, y) \sim \mathcal{T}$ have the distribution $p_\theta(y|x)$, the cross-entropy loss can be computed as:

$$\mathcal{L}_c = \mathbb{E}_{(x,y)\sim\mathcal{T}}[-\log p_\theta(y|x)], \tag{12}$$

$$p_\theta(y|x) = e^{Z_y} / \sum_{i=1}^{L} e^{Z_i}. \tag{13}$$

Directly applying cross-entropy loss into the charge prediction task is not suitable because the numbers of samples in different charge categories are extremely unbalanced. In this case, when we sample a batch of data during the training process, only a few low-frequency charges are contained, so they have less contribution for the optimization. As a result, the performance of the model on the few-shot charges is worse than that on others.

Essentially, whether a charge is few-shot or not is determined by its frequency in the training set, which is usually represented by the probability distribution. Therefore, the key point is how to integrate the prior probability distribution $p(y)$ of each charge into the loss function. When combining the cross entropy and the prior probability distribution, we find that the form fits the mutual information naturally (Menon et al., 2020):

$$\log \frac{p_\theta(y|x)}{p(y)} \sim Z_y, \tag{14}$$

which is equivalent to:

$$\log p_\theta(y|x) \sim Z_y + \log p(y). \tag{15}$$

In other words, we integrate the prior probability distribution into the loss function by adding a term $\log p(y)$ to Equation (12) as:

$$p_\theta(y|x) = \frac{e^{Z_y + \log p(y)}}{\sum_{i=1}^{L} e^{Z_i + \log p(i)}} = \left[1 + \sum_{i \neq y} \frac{p(i)}{p(y)} e^{Z_i - Z_y}\right]^{-1}. \tag{16}$$

The mutual information loss function is defined by taking Equation (16) into Equation (12).

3.7.2 Attribute Prediction Loss

As suggested by (Hu et al., 2018), we also add a supplementary task, namely, the attribute prediction, to improve the performance of our model on few-shot charge prediction. Specifically, we project the attribute features into the label space and use softmax function to get the final prediction result $\mathbf{p} = [p_1, \cdots, p_k]$ as:

$$\mathbf{z}_i = \text{softmax}(\text{MLP}(\mathbf{X}_i^A)), \quad p_i = \arg\max(\mathbf{z}_i), \tag{17}$$

where p_i is the prediction result of the i-th attribute, and $\mathbf{z}_i$ is the predicted binary probability distribution. As each attribute is equally important in the model, we can easily calculate the attribute prediction loss by summing up the cross-entropy of all attributes:

$$\mathcal{L}_{attr} = -\sum_{i=1}^{k} \sum_{j=1}^{2} \hat{z}_{ij} \log(z_{ij}), \tag{18}$$

where $\hat{z}_i$ is the ground-truth label, and z_i is the predicted probabilities distribution.

Proceedings of the 20th China National Conference on Computational Linguistics, pages 1154-1166, Hohhot, China, August 13 - 15, 2021.

Table 1: The statistics of all datasets.

Datasets	Criminal-S	Criminal-M	Criminal-L	Criminal-T
Train	61,589	153,521	306,900	6,860
Validation	7,755	19,250	38,429	2,730
Test	7,702	19,189	38,368	2,684

3.7.3 Overall Loss

Finally, we combine $\mathcal{L}_c$ and $\mathcal{L}_{attr}$ to get the overall loss function $\mathcal{L}$ as follows:

$$\mathcal{L} = \mathcal{L}_c + \gamma \cdot \mathcal{L}_{attr}, \tag{19}$$

where γ is a hyper-parameter, which is used to adjust the weights of these two loss functions.

4 Experiments

4.1 Datasets

We conduct experiments on two real datasets Criminal (Hu et al., 2018) and CAIL (Xiao et al., 2018).

Criminal: This dataset is for few-shot charges prediction. It contains three datasets with different sizes, denoted as Criminal-S (small), Criminal-M (medium), and Criminal-L (large), respectively. Each sample contains a fact description, a charge result, and attribute labels. The number of samples on these three datasets are as shown in Table 1. All datasets are divided into training, validation, and testing set with the ratio of 8:1:1.

In order to verify the performance of the model in terms of few-shot charges, all categories in the Criminal-S (small) dataset, are divided into three different classes according to their frequencies, where the charges with ≤ 10 cases are low-frequency charges and the charges with > 100 cases are high-frequency charges.

In practical situations, the number of cases for each charge is usually small. In order to further test the robustness and performance of the models in a real world scenario, on the basis of Criminal-S (small), the number of cases of all charges in the training set is limited to less than 100. That is, all high-frequency charges become medium-frequency. This dataset is named as Criminal-T (tiny), the categories of samples in the dataset is the same as the above three datasets. The statistics of all datasets are shonw in Table 1.

CAIL: This is a dataset for legal judgement prediction, which consists of three tasks (prediction of applicable law articles, charges, and term of penalty). The total number of samples is 101,619. The attribute information of each charge is not provided, thus we label it manually.

4.2 Baselines

On the Criminal dataset, we select two basic models and three state-of-the-art few-shot charge prediction models as baselines:

CNN and **LSTM**: These are two basic models for text classification. For CNN, we use $\{100, 200\}$ as the number of output channels; while for LSTM, we set the size of the hidden states as $\{100, 200\}$.

Fact-Law Attention (Luo et al., 2017): It improves the accuracy of charge prediction task by introducing additional legal articles and using attention mechanism to obtain most relevant legal article to enhance the representation capacity of fact description.

Secaps (He et al., 2019): It uses a capsule network to improve the model's representation ability and manually adjusts the category weights to improve the accuracy of few-shot charges.

Attribute-Attention (Hu et al., 2018): It constructs the features of artificial law attributes and predicts the charges and attributes simultaneously to improve performance.

On the CAIL dataset, we compare our model with three state-of-the-art models, which learn three legal judgement prediction tasks simultaneously, as introduced in Section 4.1:

Proceedings of the 20th China National Conference on Computational Linguistics, pages 1154-1166, Hohhot, China, August 13 - 15, 2021.

Table 2: Charge prediction results on the Criminal datasets. "†" indicates significant improvements ($p < 0.05$) compared with the best baseline.

	Criminal-S				Criminal-M				Criminal-L			
Model	Acc.	MP	MR	MF1	Acc.	MP	MR	MF1	Acc.	MP	MR	MF1
CNN-100	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-100	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
Secaps	**93.7**	67.8	66.3	65.8	94.7	**70.4**	68.3	68.2	**95.9**	77.2	73.5	73.7
Att. Attention	93.4	66.7	69.2	64.9	94.4	68.3	69.2	67.1	95.8	75.8	73.7	73.1
MFMI	**93.7**	**69.3**†	**70.5**†	**68.2**†	**94.9**	70.2	**75.0**†	**71.0**†	**95.9**	**78.7**†	**77.4**†	**76.4**†

Table 3: Macro F1 values of various charges on the Criminal-S dataset. "†" indicates significant improvements ($p < 0.05$) compared with the best baseline.

Charge Type	Low (<10)	Medium	High (>100)
# Charges	49	51	49
Secaps	52.4	59.2	85.9
Att. Attention	49.7	60.0	85.2
MFMI	**55.9**†	**63.5**†	85.7

Attribute-Attention-MTL: It has the same structure as the Attribute-Attention model, but is trained on three judgement prediction tasks.

TopJudge (Zhong et al., 2018): It leverages the dependencies of sub-tasks to improve the performance of legal judge prediction.

MPBFN (Yang et al., 2019): It designs a multi-view forward prediction and backward validation framework to utilize the dependencies among sub-tasks of legal judge prediction.

4.3 Evaluation Metrics

Following existing studies (Hu et al., 2018; He et al., 2019), we adopt Accuracy (Acc.), Macro precision (MP), Macro Recall (MR), and Macro F1 (MF1) as the evaluation metrics. Among them, MR and MF1 are the preferred evaluation metrics for multi-class classification problems, especially for those with imbalance categories.

4.4 Experiment Settings

We adopt THULAC[1] for word segmentation on all fact descriptions in the datasets. The maximum length of the fact description is set as 500. The pre-trained embedding table is obtained by Word2Vec (Mikolov et al., 2013) with the dimension of 100. For CNN module, we use four kernels with the sizes of {2,4,8,16}, and the number of output channels is set as 64. For bilinear CNN module, we set the kernel sizes as {8,12}, and the number of output channels is also 64. As for the capsule layer, we set the number of capsule as the number of categories, and the dimension of each capsule is 32. The number of routings is 3. Adam (Kingma and Ba, 2015) is applied as the optimizer with the learning rate of 1e-3. The settings of all baselines are consistent with their original papers.

[1] https://github.cosm/thunlp/THULAC-Python

Proceedings of the 20th China National Conference on Computational Linguistics, pages 1154-1166, Hohhot, China, August 13 - 15, 2021.

Table 4: Charge prediction results on the tiny Criminal-T dataset. "†" indicates significant improvements ($p < 0.05$) compared with the best baseline.

	Acc.	MP	MR	MF1
Secaps	81.4	63.8	67.4	63.8
Att. Attention	81.2	62.3	69.3	63.5
MFMI	**83.9**†	**63.7**	**70.7**†	**65.2**†

Table 5: Charge prediction results on the CAIL dataset.

	Acc.	MP	MR	MF1
TopJudge	82.10	**83.60**	78.42	79.05
MPBFN	82.14	82.28	80.72	80.72
Att. Attention-MTL	83.65	80.84	82.01	81.55
MFMI	**84.20**	81.34	**82.74**	**81.65**

4.5 Experimental Results

Experimental results on the Criminal datasets are shown in Table 2 to Table 4. We can observe:

(1) In general, our MFMI achieves consistently better performance on all three sizes of datasets. This indicates that our method has wide applicability over various application scenarios (sufficient or insufficient data).

(2) Compared with accuracy, all methods perform poorly in terms of the Macro F1 metric. This is mainly because of the imbalance of training samples among different charges, and indicates the shortage of prediction for few-shot charges. However, our model achieves promising improvements (3.3%, 3.9%, and 3.3% absolutely on three datasets respectively), which demonstrates the robustness and effectiveness of our model. To further validate the performance on few-shot charges, we compare the results of our model with the two best baseline models on different frequency of charges. As shown in Table 3, we divide all charges into three classes based on the case number. The charges with less than 10 samples are regarded as low-frequency, while those with more than 100 samples are high-frequency. The rest charges are treated as medium-frequency. By this means, we obtain 49 low-frequency charges, 51 medium-frequency charges, and 49 high-frequency charges. From the results, we can see that our model significantly improves the performance (6.68% and 5.83% compared with the baseline models) on both low- and medium-frequency charges. These results clearly demonstrate the effectiveness of our model in dealing with few-shot charges.

(3) Specifically, on Criminal-S, MFMI achieves the best results in terms of all evaluation metrics. The value of Macro F1 is significantly improved by 3.6% compared with the previous best method. This proves the superiority of our method on small datasets. In practice, the number of samples for each charge is usually very small. To mimic the application in such real scenario, we build a tiny dataset Criminal-T (tiny) based on Criminal-S (small) and test the performance of MFMI and the other two best baselines. Criminal-T contains only 12,274 samples from the original Criminal dataset. The results are shown in Table 4. We can observe that MFMI achieves new state-of-the-art results on Accuracy, Macro Recall, and Macro F1. The improvements are significant. This proves that our model is applicable in the real scenario and is capable of dealing with few-shot charges.

The experimental results on the CAIL dataset are shown in Table 5. It is worth noting that all baselines listed here are multi-task learning models, which are trained on three tasks (applicable law articles prediction, term of penalty prediction, and charge prediction). They use much more data and obtain more supervision signals. However, it is very interesting to see that our MFMI model achieves better results in terms of Accuracy, Macro Recall, and Macro F1. This implies that by leveraging fine-grained features and optimizing with the mutual information loss function, our method can make full use of the data and

Proceedings of the 20th China National Conference on Computational Linguistics, pages 1154-1166, Hohhot, China, August 13 - 15, 2021.

Table 6: Ablation results on the Criminal datasets. "†" indicates significant improvements ($p < 0.05$) compared with the best baseline.

	Criminal-S				Criminal-M				Criminal-L			
Model	Acc.	MP	MR	F1	Acc.	MP	MR	F1	Acc.	MP	MR	F1
MFMI	**93.7**	69.3	**70.5**†	**68.2**†	**94.9**	70.2	**75.0**†	**71.0**†	**95.9**	78.7	**77.4**	**76.4**†
w/o FF	92.1	63.1	68.8	63.6	93.6	66.4	73.7	68.1	94.0	70.1	76.9	72.3
w/ CE	**93.7**	**69.7**	67.0	67.1	94.7	**71.4**	69.8	69.4	95.8	**79.0**	73.1	74.8

Table 7: The time (seconds) taken per epoch.

Model	Criminal-T	Criminal-S
Att. Attention	200	2,150
Secaps	272	2,200
MFMI	**19**	**150**

provide better performance.

4.6 Ablation Study

To investigate the effectiveness of fine-grained features and our proposed mutual information loss, we conduct an ablation study by removing them from the full model, respectively. The two variants are denoted as "*w/o* FF" and "*w/* CE". The results are shown in Table 6. We can observe the performance degradation when removing either module. Specifically, when fine-grained features are not used, the performance of our model decreased significantly. The potential reason is that the fine-grained features can effectively improve the representation capability of our model. Thus the overall performance is improved. On the other hand, when replacing the mutual information loss with a normal cross-entropy (CE) loss, the accuracy has less change but the Macro F1 decreases significantly. By checking the results, we find that the performance on few-shot charges degrades more than the others. As the number of samples in few-shot charges is limited, the overall accuracy is less affected. This result demonstrates that the mutual information loss is effective on learning few-shot charges, which is consistent with our assumption.

4.7 Efficiency Analysis

We also compared the average time taken for each epoch during the training stage of our model and the two best few-shot charge prediction models. As shown in Table 7, our model spends less than a tenth of the time on both datasets compared to the baselines. This is because our model is based on several CNN modules, and the parallel training ability is far better than baselines with RNNs. In practice, our model with fast training speed is easier to deploy and apply.

5 Conclusion

In this work, we studied the problem of few-shot charge prediction. We proposed a multi-task learning framework, where both the charge prediction and attribute prediction are learned simultaneously. We extracted fine-grained and coarse-grained features to enhance the model's capability of representation, which are helpful for distinguishing the few-shot charges. Besides, we also proposed a loss function based on mutual information to enhance the learning on few-shot charges. Experimental results demonstrated the effectiveness of proposed methods on charge prediction, especially on the few-shot charges. Moreover, further experiments showed that our method has better scalability and efficiency. In the future, we will apply our method to other legal judgement prediction tasks and leverage the knowledge from other tasks to further improve the performance.

Proceedings of the 20th China National Conference on Computational Linguistics, pages 1154-1166, Hohhot, China, August 13 - 15, 2021.

6 Acknowledgements

We thank all the anonymous reviewers for their insightful comments. This work was supported by National Natural Science Foundation of China No. 61872370 and No. 61832017, and Beijing Outstanding Young Scientist Program NO. BJJWZYJH012019100020098.

References

Huajie Chen, Deng Cai, Wei Dai, Zehui Dai, and Yadong Ding. 2019. Charge-based prison term prediction with deep gating network. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 6361–6366. Association for Computational Linguistics.

Tianyu Gao, Xu Han, Zhiyuan Liu, and Maosong Sun. 2019. Hybrid attention-based prototypical networks for noisy few-shot relation classification. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6407–6414. AAAI Press.

Ruiying Geng, Binhua Li, Yongbin Li, Yuxiao Ye, Ping Jian, and Jian Sun. 2019. Few-shot text classification with induction network. *CoRR*, abs/1902.10482.

Congqing He, Li Peng, Yuquan Le, Jiawei He, and Xiangyu Zhu. 2019. Secaps: A sequence enhanced capsule model for charge prediction. In Igor V. Tetko, Vera Kurková, Pavel Karpov, and Fabian J. Theis, editors, *Artificial Neural Networks and Machine Learning - ICANN 2019: Text and Time Series - 28th International Conference on Artificial Neural Networks, Munich, Germany, September 17-19, 2019, Proceedings, Part IV*, volume 11730 of *Lecture Notes in Computer Science*, pages 227–239. Springer.

Zikun Hu, Xiang Li, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. 2018. Few-shot charge prediction with discriminative legal attributes. In Emily M. Bender, Leon Derczynski, and Pierre Isabelle, editors, *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 487–498. Association for Computational Linguistics.

Daniel Martin Katz, Michael James Bommarito, and Josh Blackman. 2017. A general approach for predicting the behavior of the supreme court of the united states. *PLOS ONE*, 12(4).

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

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Fred Kort. 1957. Predicting supreme court decisions mathematically: A quantitative analysis of the "right to counsel" cases. *American Political Science Review*, 51(01):1–12.

Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent convolutional neural networks for text classification. In Blai Bonet and Sven Koenig, editors, *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, pages 2267–2273. AAAI Press.

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. *Int. J. Comput. Linguistics Chin. Lang. Process.*, 17(4).

Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. 2015. Bilinear CNN models for fine-grained visual recognition. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 1449–1457. IEEE Computer Society.

Chao-Lin Liu and Chwen-Dar Hsieh. 2006. Exploring phrase-based classification of judicial documents for criminal charges in chinese. In Floriana Esposito, Zbigniew W. Ras, Donato Malerba, and Giovanni Semeraro, editors, *Foundations of Intelligent Systems, 16th International Symposium, ISMIS 2006, Bari, Italy, September 27-29, 2006, Proceedings*, volume 4203 of *Lecture Notes in Computer Science*, pages 681–690. Springer.

Chao-Lin Liu, Cheng-Tsung Chang, and Jim-How Ho. 2004. Case instance generation and refinement for case-based criminal summary judgments in chinese. *J. Inf. Sci. Eng.*, 20(4):783–800.

Zhiyuan Liu, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. 2019. Legal cause prediction with inner descriptions and outer hierarchies. In Maosong Sun, Xuanjing Huang, Heng Ji, Zhiyuan Liu, and Yang Liu, editors, *Chinese Computational Linguistics - 18th China National Conference, CCL 2019, Kunming, China, October 18-20, 2019, Proceedings*, volume 11856 of *Lecture Notes in Computer Science*, pages 573–586. Springer.

Bingfeng Luo, Yansong Feng, Jianbo Xu, Xiang Zhang, and Dongyan Zhao. 2017. Learning to predict charges for criminal cases with legal basis. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2727–2736. Association for Computational Linguistics.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. 2020. Long-tail learning via logit adjustment. *CoRR*, abs/2007.07314.

Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 3111–3119.

Sicheng Pan, Tun Lu, Ning Gu, Huajuan Zhang, and Chunlin Xu. 2019. Charge prediction for multi-defendant cases with multi-scale attention. In Yuqing Sun, Tun Lu, Zhengtao Yu, Hongfei Fan, and Liping Gao, editors, *Computer Supported Cooperative Work and Social Computing - 14th CCF Conference, ChineseCSCW 2019, Kunming, China, August 16-18, 2019, Revised Selected Papers*, volume 1042 of *Communications in Computer and Information Science*, pages 766–777. Springer.

Sara Sabour, Nicholas Frosst, and Geoffrey E. Hinton. 2017. Dynamic routing between capsules. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 3856–3866.

Octavia-Maria Sulea, Marcos Zampieri, Shervin Malmasi, Mihaela Vela, Liviu P. Dinu, and Josef van Genabith. 2017. Exploring the use of text classification in the legal domain. In Kevin D. Ashley, Katie Atkinson, Luther Karl Branting, Enrico Francesconi, Matthias Grabmair, Marc Lauritsen, Vern R. Walker, and Adam Zachary Wyner, editors, *Proceedings of the Second Workshop on Automated Semantic Analysis of Information in Legal Texts co-located with the 16th International Conference on Artificial Intelligence and Law (ICAIL 2017), London, UK, June 16, 2017*, volume 2143 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Pengfei Wang, Yu Fan, Shuzi Niu, Ze Yang, Yongfeng Zhang, and Jiafeng Guo. 2019. Hierarchical matching network for crime classification. In Benjamin Piwowarski, Max Chevalier, Éric Gaussier, Yoelle Maarek, Jian-Yun Nie, and Falk Scholer, editors, *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, pages 325–334. ACM.

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.

Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. 2018. Lifelong domain word embedding via meta-learning. In Jérôme Lang, editor, *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4510–4516. ijcai.org.

Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. 2019. Open-world learning and application to product classification. In Ling Liu, Ryen W. White, Amin Mantrach, Fabrizio Silvestri, Julian J. McAuley, Ricardo Baeza-Yates, and Leila Zia, editors, *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 3413–3419. ACM.

Nuo Xu, Pinghui Wang, Long Chen, Li Pan, Xiaoyan Wang, and Junzhou Zhao. 2020. Distinguish confusing law articles for legal judgment prediction. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 3086–3095. Association for Computational Linguistics.

Wenmian Yang, Weijia Jia, Xiaojie Zhou, and Yutao Luo. 2019. Legal judgment prediction via multi-perspective bi-feedback network. In Sarit Kraus, editor, *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 4085–4091. ijcai.org.

Proceedings of the 20th China National Conference on Computational Linguistics, pages 1154-1166, Hohhot, China, Augest 13 - 15, 2021.

Mo Yu, Xiaoxiao Guo, Jinfeng Yi, Shiyu Chang, Saloni Potdar, Yu Cheng, Gerald Tesauro, Haoyu Wang, and Bowen Zhou. 2018. Diverse few-shot text classification with multiple metrics. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1206–1215. Association for Computational Linguistics.

Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun. 2018. Legal judgment prediction via topological learning. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 3540–3549. Association for Computational Linguistics.

Proceedings of the 20th China National Conference on Computational Linguistics, pages 1154-1166, Hohhot, China, Augest 13 - 15, 2021.