Creating Auxiliary Representations from Charge Definitions for Criminal Charge Prediction

Liangyi Kang¹, Jie Liu¹, Lingqiao Liu², Qinfeng Shi², and Dan Ye¹

¹ Institute of Software, Chinese Academy of Sciences, Beijing, China
² School of Computer Science, The University of Adelaide, Australia
{kangliangyi15, ljie, yedan}@otcaix.iscas.ac.cn, {lingqiao.liu, javen.shi}@adelaide.edu.au

Abstract

Charge prediction, determining charges for criminal cases by analyzing the textual fact descriptions, is a promising technology in legal assistant systems. In practice, the fact descriptions could exhibit a significant intra-class variation due to factors like non-normative use of language, which makes the prediction task very challenging, especially for charge classes with too few samples to cover the expression variation. In this work, we explore to use the charge definitions from criminal law to alleviate this issue. The key idea is that the expressions in a fact description should have corresponding formal terms in charge definitions, and those terms are shared across classes and could account for the diversity in the fact descriptions. Thus, we propose to create auxiliary fact representations from charge definitions to augment fact description representation. The generated auxiliary representations are created through the interaction of fact description with the relevant charge definitions and terms in those definitions by integrated sentence- and word-level attention scheme. Experimental results on two datasets show that our model achieves significant improvement than baselines, especially for classes with few samples.

Introduction

The task of charge prediction is to determine appropriate charges, such as theft, seizing or robbery, for criminal cases by analyzing the textual fact descriptions. Automating charge prediction by using NLP technology could significantly reduce the human labor in organizing legal documents, and could be practically useful for an online legal assistant system.

Existing methods formulate the charge prediction task as a text classification problem, targeting at learning the representation of fact descriptions for prediction. Conventional methods [Liu and Liao 2005, Liu and Hsieh 2006, Lin et al. 2012, Sulea et al. 2017] design shallow text features to represent fact descriptions. Recently, deep learning provides end-to-end models to learn fact representations from fact descriptions [Luo et al. 2017, Hu et al. 2018, Zhong et al. 2018], which achieves the state-of-art result.

In practice, the fact description in a criminal case is written by prosecutors, lawyers, or defendants to state the detail of the criminal case. It comprises a substantial amount of diverse non-normative use of language. For example, the cases of robbery in Figure 1 all involve "theft", but the legal term "theft" may be implicitly expressed like "stole an electric vehicle" or "came forward to ride away Ke’s white Merida bicycle". Consequently, the representation of fact descriptions may exhibit considerable intra-class variation which may lead to prediction failure at the test stage. This could be more pronounced for charge classes with only a few examples since the samples are not sufficient for learning a predictive model robust to expression variation.

To address this issue, we introduce the charge definitions from criminal law to create more robust fact representations for charge prediction. We propose to create auxiliary fact representations from the charge definitions to augment the fact representation. Those auxiliary representations are essentially projections of the fact description in the semantic space of charge definitions. Our motivation is that the expressions in a fact description should have corresponding formal terms in charge definitions, and those formal terms can provide an alternative view of the expressions in fact description. Note that many of those formal terms are shared across charge classes and are less diverse. Thus, using elements in charge definitions to re-interpret fact description and generate auxiliary representations could have the potential to account for the diversity in the fact description.

Specifically, we design an integrated sentence- and word-level interaction model to generate two auxiliary fact representations. We identify the relevant charge definitions through sentence-level interaction between fact description and charge definitions, and then aggregate the holistic features of relevant charge definitions to create the first auxiliary representation, named as charge-related fact representation. The relevant charge definitions identified in the course of producing the first auxiliary representation will also serve for creating the second auxiliary representation. To create the second representation, we further consider finer-grained word-level interaction between the fact description and identified related charge definitions. Relevant words from relevant charge definitions are attended and aggregated through a recurrent neural network to generate the second auxiliary representation, named as charge-token-related fact represen-
Figure 1: Illustration of our method. The related charges are identified (indicated by the red arrow) via sentence-level attention and aggregated to create the first auxiliary representation, charge-related fact representation. Then key words in cases align to terms in identified charge definitions via word-level attention (aligned words are labeled by the same color), which are then used as the second auxiliary representation, charge-token-related fact representation.

tempt to extract shallow text features from fact descriptions or create hand-crafted features to represent fact descriptions, which are hard to generalize to large datasets due to the diverse expression of fact descriptions. Inspired by the success of deep learning, Luo et al. [2017], Ye et al. [2018], Hu et al. [2018], Zhong et al. [2018] employ neural models with external information to capture the high-level semantic information. Zhong et al. propose the LJP method, modeling multiple legal subtasks as a Directed Acyclic Graph (DAG) and using multi-task learning to assist prediction. Further, Luo et al. use a separate two-stage scheme to extract the related articles and then attend them attentively to fact representation for charge prediction. Ye et al. design 10 legal attributes to help the few-shot charges prediction. However, existing charge prediction models all need a large amount of feature engineering, either design features or build relations between subtasks. Instead, we augment fact representation to assist charge prediction by creating auxiliary representation from charge definitions in an end-to-end fashion.

Attention and Memory Our model is also related to attention and memory in deep learning (Bahdanau, Cho, and Bengio 2014; Vaswani et al. 2017; Sinha et al. 2018; Weston, Chopra, and Bordes 2014; Wang et al. 2018; Ebesu, Shen, and Fang 2018). Although researchers propose various neural architectures with memory and attention for NLP problems (Kumar et al. 2016; Wang et al. 2017; Gao et al. 2019), they either only consider sentence-level or only word-level alignment between sentences. In contrast, we combine them jointly to form auxiliary representation, where sentence-level interaction identifies relevant charges, and a fine-grained word-level interaction on the top of identified charge definitions makes the generated fact representation more robust.

The proposed model

Problem Formulation

Charge Prediction Charge prediction is to predict the corresponding charges $l$ for a given fact description $f$, where fact description $f$ consists of a sequence of words $\{w_1, w_2, \ldots, w_m\}$, and its label is a $C$ dimensional multi-hot vector – a fact description may correspond to multiple charges in $C$ charges. The charge definition for the $i$-th charge $l_i$ can be represented as a sequence of words $\{w_1^{l_i}, w_2^{l_i}, \ldots, w_n^{l_i}\}$.

Framework

To generate a robust fact representation for prediction, we propose an integrated sentence- and word-level interaction model. The architecture of our model is shown in Figure 2. As seen, the final fact representation is the concatenation of three representations.

- The original fact representation ($F_c$), derived from fact description only and obtained by the fact description encoder.
- The auxiliary representation $I$, charge-related fact representation ($F_\tau$), aggregated by the holistic representation of related charge definitions that are identified via
sentence-level interaction between fact and charge definitions.

- The auxiliary representation II, charge-token-related fact representation \( F_w \), on the top of the identified charge definitions, created by finer-grained word-level interaction between fact and identified charge definitions.

**Fact Description Encoder**

Given a fact description represented by a sequence of word embeddings \( f = \{w_1, w_2, \ldots, w_{|f|}\} \), we use Gated Recurrent Unite (Cho et al. 2014) to create a sequence of hidden states for encoding contextual information of each word.

\[
h_i = GRU(w_i, h_{i-1}),
\]

where \( h_i \) is the hidden state of the GRU at time step \( i \). The variable sequence is denoted as \( H = \{h_1, \ldots, h_i, \ldots, h_m\} \).

For a fact description, the words and consequently those hidden variables do not contribute equally to convey the semantic meaning of a fact, and long fact description will involve many less informative words. To suppressing the negative impact of the non-informative words, we use attention mechanism to assign each hidden state in \( H \) an importance weight \( \alpha_i \).

\[
a_i = W_2 \tanh(W_1 h_i^T),
\]

\[
\alpha_i = \frac{\exp(a_i)}{\sum_{k=1}^m \exp(a_k)},
\]

where \( W_1 \) and \( W_2 \) are trainable parameters. The holistic representation of original fact description \( F_c \) is computed as a weighted sum of \( H \):

\[
F_c = \sum_{i=1}^m \alpha_i h_i.
\]

**Charge Definitions Encoder**

Each charge class is associated with a charge definition, that is, \( l_i = \{w_{i1}, w_{i2}, \ldots, w_{im}\} \). We use a CNN to encode the sequence of \( n \) words into a sequence of vectors. Since we will deal with a large number of charge definitions, using CNNs (Kim 2014) gives us better training efficiency.

\[
e_j^i = \text{CNN}(w_{i1}, w_{i2}, \ldots, w_{im}),
\]

where the window size of CNN is \( s \). Then we sum up these vectors to create the holistic representation of each charge definition.

\[
L_i = \sum_{j=1}^{n} e_j^i.
\]

We also tried using GRUs to encode \( l_i \), but they require more computational resources and lead to worse performance. Thus we choose CNN as charge definitions encoder.

**Two Auxiliary Fact Representations from Charge Definitions**

The first auxiliary fact representation is created through the sentence-level interaction between the fact description and
charge definitions. Its creation process iterates between two steps: identifying related charges and attentively aggregating the holistic representation of charge definitions. After those iterations, relatedness weights of each charge will be obtained and they will also be used as the basis for creating the second auxiliary fact representation. The second auxiliary fact representation is generated from word-level interaction between fact description and identified charge definitions. It uses word-level attention to identify terms that align with the expression in the fact description, and aggregates those terms through a recurrent neural network.

We elaborate the creation of those two auxiliary representations as follows.

**Auxiliary Representation I: charge-related fact representation created via sentence-level interaction**

**Charges Identification** Identifying related charges is realized by calculating an attention weight for each charge to indicate the relatedness. Specifically, we exploit episodic memory attention mechanism (Xiong, Merity, and Socher 2016) to iteratively calculate the attention weight from the correlation between the charge definitions and fact description. Formally, we use following formulas to calculate the attention weight $g$ of each charge definition at the t-th iteration.

$$z_t = |L_t \odot F_c; L_t \circ m_t; |L_t - F_c|; |L_t - m_t|\rangle,$$

$$A_t(L_t, F_c, m_t) = W_2 \tanh(W_1^a z_t),$$

$$g_t(t) = \frac{\exp(A_t)}{\sum_{k=1}^C \exp(A_k)},$$

where $\odot$ is the element-wise product, $| |$ is the element-wise absolute value, and $\circ$ represents concatenation of the vectors. $W_1^a$ and $W_2^a$ are trainable weight matrices.

**Attentive Charges Aggregator** Once the attention weight of each charge is calculated, we update the memory by performing weighted summation over charge definition representations.

$$m_{t+1} = \sum_{i=1}^C g_t(t)L_i.$$

Finally, we concatenate original fact representation with the last memory and the previous memory, and feed them into a fully-connected layer to create the auxiliary charge-related fact representation by using the following equation:

$$F_k = f_c([F_c; m_T; m_{T-1}]),$$

where $f_c$ denotes the fully connected layer.

**Auxiliary Representation II: charge-token-related fact representation created via word-level interaction**

In the course of creating the above representation, both fact description and charge definitions are represented by holistic feature vectors. In other words, the interaction between fact and charge definitions is only at the sentence level. The second auxiliary representation steps further to introduce interaction at the word level. Specifically, for each hidden variable $h_k$ in the fact description, we first compute its matching score towards each $e_j^k$ in each charge definition $l_i$ by inner-product. Then $e_j^k$ is attentively aggregated to an intermediate representation $h_{k}^L$:

$$M_j = h_k \cdot e_j^T,$$

$$\beta_j = \frac{\exp(M_j)}{\sum_{p=1}^{m} \exp(M_p)},$$

$$h_{k}^L = \sum_{j=1}^{n_i} \beta_j e_j^k.$$

The above intermediate representation is defined w.r.t each charge definition $l_i$. In our method, we further perform a weighted summation over $h_{k}^L$ for different charge definition $l_i$. The weight is the attention weight $g(T)$ calculated at the last iteration $T$ in Eq. (9). Using this weight fits our intuition that the terms in the related charges are more relevant to the expressions in the fact description. Formally, we obtain

$$h_k^L = \sum_{i=1}^{c} g_i(T)h_{k}^L,$$

where $h_k^L$ can be viewed as a projection of $h_{k}^L$ in the space spanned by $e_j^k$.

After obtaining $h_k^L$ for each word in the fact description, we process the sequence $\{h_1^L, \ldots, h_k^L, \ldots, h_m^L\}$ by a new GRU and obtain the last hidden state $h_l$:

$$h_l = GRU(h_1^L, h_{l-1}).$$

We concatenate original and the projected fact representation, and feed them into a fully-connected layer to generate the auxiliary charge-token-related fact representation.

$$F_w = f_c([F_c; h_l]).$$

**The Output**

Finally, we concatenate all the generated representations and feed them into a fully-connected layer to generate the final fact representation $F$:

$$F = f_c([F_c; F_k; F_w]).$$

$F$ is then passed to the classifier layer to make charge prediction.

The loss function for training is as follows:

$$Loss = -\sum_{i=1}^{N} \sum_{j=1}^{c} y_{ij} \log(o_{ij}),$$

where $N$ is the number of training data, $C$ is the number of charges. $y_{ij} \in \{0, 1\}$ and $o_{ij}$ is the estimated likelihood of the $l$-th charge being true.
Table 1: Statistics of datasets.

| Datasets        | CAIL150K | CAIL30K |
|-----------------|----------|---------|
| Training samples| 154592   | 32506   |
| Test samples    | 32500    | 32500   |
| Charge classes  | 202      | 168     |

Experiments

In order to verify the effectiveness of our model on criminal charges prediction, we conduct experiments on two real-world datasets with different scales to compare our model against several baselines. Further analyses are also made to validate the significance of introducing charge definitions and various components of our model.

Setup

Dataset: We use publicly available datasets from Xiao et al. (2018) to conduct our experiments. There are two datasets with different scales: CAIL150K dataset and CAIL30K dataset. The criminal cases in these datasets are collected from the China Judgment Online with a single defendant. Table 1 shows the descriptive statistics of used datasets. It is worth noting that in these two datasets the distribution of charges is quite imbalanced. In CAIL150K, the top 30 most frequent charges cover 60% cases, and the 31% charges in the training set have less than 100 cases, taking up only 1.88% of the total number of cases. CAIL30K is a smaller dataset. In its training set, 42% charges have less than 10 cases, taking up only 0.89% of the total number of cases. The small number of samples makes it challenging to train a model that performs well on low-frequency classes.

As for charge definitions, they are extracted from articles in the Criminal Law of the People’s Republic of China. Specifically, in criminal law, except for articles irrelevant to specific charges, each article may include more than one charges, their corresponding charge definitions, and punishment. We use regular expressions to extract charge definitions, and merge charge definitions scattered in multiple articles. A snippet of cases and charge definitions is illustrated in Figure 1.

Training setup: As all the sentences in charge definitions and fact descriptions are written in Chinese without word segmenting, we apply jieba for word cut. We set the maximum length of fact description to 500, charge definitions to 110. We use pre-trained GloVe vectors as our initial word embeddings. In practice, we choose the 64 dimensional embedding vectors trained on our initial word embeddings. In practice, we choose the 64 dimensional embedding vectors trained on baidubaike. The iteration time $T$ in Eq. (9) is set as 3. Adam (Kingma and Ba 2014) is used as the optimizer and the learning rate is initialized as 0.005 and halved in every other epoch.

Baselines: We compare our model against several text classification models and charge prediction methods, which can be categorized into four categories:

1. **Not using charge definitions for classification.** We employ TFIDF (Salton and Buckley 1988) to extract text features from fact descriptions and use linear SVMs (Suykens and Vandewalle 1999) for charge prediction (TFIDF+SVM). We also implement deep learning models, such as multi-layers Convolution Neural Network (CNN) (Kim 2014), CNN (CNN_classify), Gated Recurrent Unit (GRU) (Cho et al. 2014), GRU (GRU_classify) and hierarchical LSTM (Sinha et al. 2018), LSTM (LSTM_classify) for fact descriptions encoding and classification.

2. **Using multi-task learning for classification.** Our method is somehow related to LJP (Zhong et al. 2018), which introduces related legal tasks and uses multi-task learning to train a better fact representation. We also reimplement it to compare with our method.

3. **Matching the fact description with charge definitions for classification.** We exploit TFIDF to extract text features from fact descriptions and charge definitions, then compare the fact description with each charge definition (TFIDF+match) to find the best matched charges. We also train a Siamese CNN (Koch, Zemel, and Salakhutdinov 2015) to match the representations of fact description and charge definitions.

4. **Augmenting fact description with charge definitions for classification.** We implement Fact-Law AN model that Luo et al propose to use relevant law articles, selected by SVMs, to serve as a legal basis for encoding the fact description. To demonstrate the advantage of our model in considering sentence- and word-level interaction jointly, we also implement improved memory network (Kumar et al. 2016), MemNet (Wang et al. 2017), and GA_Reader (Wang et al. 2017). These two methods are designed for question-answer task, which employ multi-iterative interaction between query and document at sentence- and word-level respectively for answer prediction. In the implementation, we replace query and document in GA_Reader and MemNet with the fact description and charge definitions.

Results

Evaluation Metrics: We employ accuracy (Acc.), macro-precision (MP), macro-recall (MR) and macro-F1 (MF1) as our evaluation metrics. The macro-precision/recall/F1 are calculated by averaging the precision, recall and F1 of each charge, which are metrics commonly used for multilabel classification task.

Overall Evaluation Results: Experimental results on two scale datasets are shown in Table 2. The observations are as follows:

- Generally speaking, models without incorporating charge definitions (TFIDF+SVM, CNN_classify, GRU_classify and LSTM_classify) perform inferior to their charge-definition-incorporated counterparts. This is evident by their lower MF1 scores (MF1 is a more comprehensive score for evaluating multi-label classification than Acc., MP, and MR). This observation clearly demonstrates the
Table 2: The experimental results [%] of baselines and our model on two datasets. Four different types of models are separated by lines and the best scores are highlight in bold font.

| Datasets   | Model        | CAIL150K | CAIL30K |
|------------|--------------|----------|---------|
|            | Acc. | MP | MR | MF1 | Acc. | MP | MR | MF1 |
| Not using charge definitions | TFIDF+SVM | 71.87 | 79.71 | 56.84 | 63.32 | 49.13 | 31.48 | 19.98 | 22.06 |
|            | CNN_classify | 79.23 | 70.80 | 62.27 | 64.97 | 52.75 | 23.64 | 21.95 | 20.59 |
|            | GRU_classify | 77.33 | 72.45 | 57.42 | 61.54 | 56.14 | 23.99 | 22.81 | 21.51 |
|            | LSTM_classify | 73.15 | 51.45 | 43.82 | 46.06 | 25.34 | 7.69 | 6.34 | 6.15 |
| Using multiple tasks | LJP | 25.26 | 25.78 | 24.32 | 25.55 | 15.29 | 15.45 | 15.68 | 15.56 |
| Match with charge definitions | TFIDF match | 13.03 | 31.21 | 40.29 | 26.52 | 12.19 | 37.37 | 35.41 | 27.60 |
|            | Siamese CNN | 72.98 | 74.52 | 64.64 | 66.55 | 50.66 | 32.74 | 33.74 | 29.28 |
| Augment with charge definitions | Fact-Law AN | 75.61 | 58.89 | 52.30 | 53.62 | 60.73 | 28.15 | 25.16 | 24.79 |
|            | GA Reader | 73.78 | 74.68 | 66.59 | 68.21 | 54.95 | 39.29 | 34.05 | 33.03 |
|            | MemNet | 80.18 | 80.09 | 67.13 | 70.78 | 62.40 | 32.62 | 27.54 | 27.64 |
|            | Ours | 81.05 | 82.06 | 68.33 | 72.43 | 67.99 | 46.13 | 36.00 | 37.62 |

Table 3: The experimental results of ablation test of our model on CAIL150K dataset.

| Models | Acc. | MP | MR | MF1 |
|--------|------|----|----|-----|
| Ours   | 81.05 | 82.06 | 68.33 | 72.43 |
| w/o Fc | 80.31 | 79.12 | 66.88 | 70.55 |
| w/o Fs,Fw | 77.33 | 72.45 | 57.42 | 61.54 |
| w/o Fw | 79.50 | 78.86 | 66.18 | 69.86 |
| w/o Fs | 80.62 | 80.54 | 66.97 | 71.28 |
| w/o Fs,gi | 80.54 | 76.90 | 64.34 | 67.98 |

benefit of introducing charge definitions to assist charge prediction.

• Incorporating charge definitions through matching based approaches (TFIDF match and Siamese CNN) works to some extent, although their performance is still worse than methods using more sophisticated interaction between fact description and charge definitions, i.e. GA_Reader, MemNet and Ours.

• Methods that Augment fact representation with charge definitions through end-to-end schema (GA_Reader, MemNet and Ours) attain better results than Fact-Law AN. The latter uses a separated two-stage framework to first identify the related charge definitions. This observation shows the importance of the end-to-end design. In addition, compared with GA_Reader and MemNet, which performs either sentence- or word-level interaction, our approach achieves better performance through considering sentence- and word-level interaction jointly.

• Our proposed model outperforms other baselines on two datasets. The improvement is especially significant on the CAIL30K dataset: our method surpasses the second best about 5% in MF1. Since the CAIL30K contains more classes with few training samples, the excellent performance of our approach suggests that our auxiliary representations may help to improve the generalization performance for classes with few samples.

• Finally, we compare our method against LJP. Like our method, LJP also uses external information for building the fact representation. Different from our method, they introduce multiple related tasks and adopt the multi-task learning for representation training. As shown in Table 2, we can see that our method achieves superior performance than LJP.

Further Analysis

Ablation Test We conduct ablation studies to verify the effectiveness of various components in our method. We consider several variations of our approach by removing some components of our model. The result is shown in Table 3. As seen, only using fact descriptions without any level auxiliary fact representations (w/o Fs,Fw) yields the worst performance, which proves the importance of the use of charge definitions. After adding either the sentence-level (w/o Fw) or the word-level auxiliary fact representation (w/o Fs), the performance can be significantly improved. It is observed that the performance of only adding charge-token-related fact representation (w/o Fs) is better than only adding charge-related fact representation (w/o Fw). We also created a variant of our method without using attention weight gi of each charge from Eq. (9) in the process of generating the attention weights.
Figure 4: Intra-class variance of different fact representations of the top-5 frequent classes in CAIL150K dataset. $F_c$ is fact representation only learned from fact description, $F_s$ is the $F_c$ augmented with charge-related fact representation, and $F$ is the $F_c$ augmented with all auxiliary fact representations.

charge-token-related fact representation (w/o $F_s,g_i$), which is implemented by setting the attention weight $g_i$ to $\frac{1}{i}$ instead of generated from charge identification part. It can be observed that the performance of w/o $F_s,g_i$ declines. This suggests that the two-level attention is necessary and using them jointly can get the best performance.

Impact of Exploiting charge definitions We analyze the effects of incorporating generated auxiliary fact representations for classes with few training data. As shown in Figure 3, we study the results of classes with less than 100 samples on CAIL150K dataset. We can find that the MF1 measure of many charges is zero if auxiliary representations are not used, and the results can be improved significantly if we add auxiliary representation from charge definitions. This observation highlights the benefit of introducing auxiliary representations for handling small sample cases.

Intra-class variance of different fact representations To investigate whether the fact representation of our method is more stable, we conduct the following experiment: we calculate the variance along each dimension of fact representations from five classes with the most amount of samples, and then use the average variance along all dimensions as an indicator of the intra-class variance of different fact representations. As shown in Figure 4, fact representation ($F_c$) only learned from fact description yields the largest intra-class variance. After augmenting fact representation from charge definitions through sentence-level interaction ($F_s$), the intra-class variance declines greatly. Specially, the final fact representation ($F$) with two auxiliary representations incorporated attains an even lower intra-class variance.

Case study Finally, we select a representative robbery case to give an intuitive illustration of the attention results on the sentence- and word-level interaction. As shown in Table 4, the case describes that the defendant is convicted of robbery due to stealing property and poking the victim to resist arrest. On the sentence-level interaction, with the increasing of iteration in Eq. (9), our model narrows down the candidate charges and finally identifies the correct related charges. We choose the iteration times as 3 since the performance cannot improve with more iterations.

On the word-level interaction, the attention mechanism makes the words in fact description align with the formal terms in charge definitions. To demonstrate this mechanism, Figure 5 shows for the words in fact description, which terms are focused on in the charge definition of robbery. The identified keywords in fact description are "electric vehicle", "resisting arrest" and "a knife", which correspond to key terms in robbery definition—"stolen goods", "resist arrest" and "use violence".

Table 4: Attention map of sentence-level attention of robbery case. $t_1$, $t_2$, and $t_3$ represent the iteration times in Eq. (9). The color darker means the charges are more related to the fact.

| Fact description: | The Defendant stole an electric vehicle, when blocked, resisted arrest to take out a knife to pock the victim to resist arrest... |
|-------------------|------------------------------------------------------------------------------------------------------------------|
| Charge: Robbery   | $t_1$ $t_2$ $t_3$                                                                                                   |
| Top5 Related Charges                                      | Robbery                                                                                                               |
|                   | Intentional injury                                                                                                  |
|                   | Theft                                                                                                                |
|                   | Negligent act causing severe injury                                                                                |
|                   | Endangering public security                                                                                        |

Figure 5: Attention map of word-level attention between robbery case and the charge definition of robbery. The dark color means a large value.

Conclusion In this work, we focus on the task of multilabel charge prediction for given fact descriptions of criminal cases. To address the problem of having a large expression variance in fact descriptions due to informal language use, we introduce charge definitions from criminal law to create auxiliary representations of the fact descriptions. The experimental results on two datasets show the effectiveness of our model on charge prediction. The significant improvement on the
classes with few training data validate that our method can benefit the small sample training scenario and the two-level auxiliary fact representations can help the model to generalize to the unseen description.

References

[Bahdanau, Cho, and Bengio 2014] Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

[Cho et al. 2014] Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.

[Dong et al. 2014] Dong, L.; Wei, F.; Tan, C.; Tang, D.; Zhou, M.; and Xu, K. 2014. Adaptive recursive neural network for target-dependent twitter sentiment classification. In Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 2: Short papers), volume 2, 49–54.

[Ebesu, Shen, and Fang 2018] Ebesu, T.; Shen, B.; and Fang, Y. 2018. Collaborative memory network for recommendation systems. In The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, 515–524. ACM.

[Gao et al. 2019] Gao, T.; Han, X.; Liu, Z.; and Sun, M. 2019. Hybrid attention-based prototypical networks for noisy few-shot relation classification.

[Hu et al. 2018] Hu, Z.; Li, X.; Tu, C.; Liu, Z.; and Sun, M. 2018. Few-shot charge prediction with discriminative legal attributes. In Proceedings of the 27th International Conference on Computational Linguistics, 487–498.

[Kim 2014] Kim, Y. 2014. Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882.

[Kingma and Ba 2014] Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

[Koch, Zemel, and Salakhutdinov 2015] Koch, G.; Zemel, R.; and Salakhutdinov, R. 2015. Siamese neural networks for one-shot image recognition. In ICML deep learning workshop, volume 2.

[Kumar et al. 2016] Kumar, A.; Irsoy, O.; Ondruska, P.; Iyyer, M.; Bradbury, J.; Gulrajani, I.; Zhong, V.; Paulus, R.; and Socher, R. 2016. Ask me anything: Dynamic memory networks for natural language processing. In International conference on machine learning, 1378–1387.

[Lin et al. 2012] Lin, W.-C.; Kuo, T.-T.; Chang, T.-J.; Yen, C.-A.; Chen, C.-J.; and Lin, S.-d. 2012. Exploiting machine learning models for chinese legal documents labeling, case classification, and sentencing prediction. Proceedings of ROCLING 140.

[Liu and Hsieh 2006] Liu, C.-L., and Hsieh, C.-D. 2006. Exploring phrase-based classification of judicial documents for criminal charges in chinese. In International Symposium on Methodologies for Intelligent Systems, 681–690. Springer.

[Liu and Liao 2005] Liu, C.-L., and Liao, T.-M. 2005. Classifying criminal charges in chinese for web-based legal services. In Asia-Paciﬁc Web Conference, 64–75. Springer.

[Luo et al. 2017] Luo, B.; Feng, Y.; Xu, J.; Zhang, X.; and Zhao, D. 2017. Learning to predict charges for criminal cases with legal basis. arXiv preprint arXiv:1707.09168.

[Salton and Buckley 1988] Salton, G., and Buckley, C. 1988. Term-weighting approaches in automatic text retrieval. Information processing & management 24(5):513–523.

[Sinha et al. 2018] Sinha, K.; Dong, Y.; Cheung, J. C. K.; and Ruths, D. 2018. A hierarchical neural attention-based text classiﬁer. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 817–823.

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

[Suykens and Vandewalle 1999] Suykens, J. A., and Vandewalle, J. 1999. Least squares support vector machine classiﬁers. Neural processing letters 9(3):293–300.

[Vaswani et al. 2017] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In Advances in neural information processing systems, 5998–6008.

[Wang et al. 2017] Wang, W.; Yang, N.; Wei, F.; Chang, B.; and Zhou, M. 2017. Gated self-matching networks for reading comprehension and question answering. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 189–198.

[Wang et al. 2018] Wang, S.; Mazumder, S.; Liu, B.; Zhou, M.; and Chang, Y. 2018. Target-sensitive memory networks for aspect sentiment classiﬁcation. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 957–967.

[Weston, Chopra, and Bordes 2014] Weston, J.; Chopra, S.; and Bordes, A. 2014. Memory networks. arXiv preprint arXiv:1410.3916.

[Xiao et al. 2018] Xiao, C.; Zhong, H.; Guo, Z.; Tu, C.; Liu, Z.; Sun, M.; Feng, Y.; Han, X.; Hu, Z.; Wang, H.; et al. 2018. Caiil2018: A large-scale legal dataset for judgment prediction. arXiv preprint arXiv:1807.02478.

[Xiong, Merity, and Socher 2016] Xiong, C.; Merity, S.; and Socher, R. 2016. Dynamic memory networks for visual and textual question answering. In International conference on machine learning, 2397–2406.

[Ye et al. 2018] Ye, H.; Jiang, X.; Luo, Z.; and Chao, W. 2018. Interpretable charge predictions for criminal cases: Learning to generate court views from fact descriptions. arXiv preprint arXiv:1802.08504.

[Zhong et al. 2018] Zhong, H.; Zipeng, G.; Tu, C.; Xiao, C.; Liu, Z.; and Sun, M. 2018. Legal judgment prediction via topological learning. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 3540–3549.