Journal of Artificial Intelligence Research 66 (2019) 743-764 Submitted 03/2019; published 11/2019

Interpretable Charge Prediction for Criminal Cases with Dynamic Rationale Attention

Wenhan Chao CHAOWENHAN@BUAA.EDU.CN
Xin Jiang XINJIANG@BUAA.EDU.CN
School of Computer Science and Engineering, Beihang University, Beijing, China

Zhunchen Luo ZHUNCHENLUO@GMAIL.COM
Information Research Center of Military Science, PLA Academy of Military Science, Beijing, China

Yakun Hu HUYAKUN@BUAA.EDU.CN
Wenjia Ma MAWENJIA@BUAA.EDU.CN
School of Computer Science and Engineering, Beihang University, Beijing, China

Abstract

Charge prediction which aims to determine appropriate charges for criminal cases based on textual fact descriptions, is an important technology in the field of AI&Law. Previous works focus on improving prediction accuracy, ignoring the interpretability, which limits the methods' applicability. In this work, we propose a deep neural framework to extract short but charge-decisive text snippets – rationales – from input fact description, as the interpretation of charge prediction. To solve the scarcity problem of rationale annotated corpus, rationales are extracted in a reinforcement style with the only supervision in the form of charge labels. We further propose a dynamic rationale attention mechanism to better utilize the information in extracted rationales and predict the charges. Experimental results show that besides providing charge prediction interpretation, our approach can also capture subtle details to help charge prediction.

1. Introduction

Given the fact descriptions of criminal cases, charge prediction aims to determine appropriate charges (e.g. larceny, intentional homicide or robbery) for the criminals suspect mentioned. The technology is important for legal assistant systems which can assist the judges and lawyers during the trials. It is also applied to legal advice systems which help non-legal professionals by providing professional legal advice for them.

Existing works generally treat charge prediction as a text classification problem, and have made a series of progress (Liu, Chang, & Ho, 2004; Liu & Hsieh, 2006; Lin, Kuo, Chang, Yen, Chen, & Lin, 2012; Luo, Feng, Xu, Zhang, & Zhao, 2017). However, just like in other rigorous fields such as medicine and finance, what the users urgently demand is not only the prediction results but also the principles of the decisions. Therefore, it is necessary to improve the interpretability of the charge prediction and other prediction systems in rigorous fields.

Interpretability refers to the ability of AI systems to interpret their predictions and has been studied for decades. For the first time, Hendricks, Akata, Rohrbach, Donahue, Schiele, and Darrell (2016) divide the concept of interpretation into *introspection explanation* and *justification explana-*

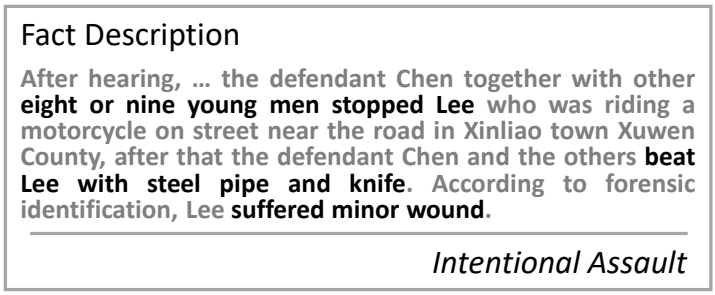

Figure 1: An example of a case charged *Intentional Assault*. We provide the charge prediction as well as a rationale explanation.

tion. Introspection explanation details how a model determines its final output, while justification explanation produces sentences detailing how the evidence compatible with the system output.

In this work, we are committed to making charge prediction interpretable by extracting rationales which refer short but charge-decisive text snippets from fact description. The rationales are extracted in the process of charge prediction and have a decisive effect on the prediction results. Therefore, they can be regarded as an introspection explanation of charge prediction, elevating the charge prediction interpretability. We hold that good rationales must meet three criteria: 1) the total amount should be small; 2) the rationales must be charge-decisive; 3) the rationales must express complete semantics. As shown in Figure 1, the model's input is the fact description of a criminal case and the output consists of the charge prediction result and the supportive rationales.

The task is not trivial. First, to meet the previous criteria, the granularity of rationales is difficult to grasp. Given sentence level rationales, we may still not know where the key point is, while word level rationales cannot clearly describe what exactly happened. A form (e.g. phrase) between word and sentence would be more appropriate. Besides, considering the practicability of the approaches, it is better to complete the rationale extraction task without rationale supervision. Since corpus with rationale annotation is hard to obtain. Zhang, Marshall, and Wallace (2016) propose an interpretable text classification approach utilizing human-annotated rationale sentences as supervised attention. However, due to the difficulty of obtaining rationale annotated corpus, the practicability of their method is limited. Finally, methods of improving the prediction accuracy while having high interpretability are very essential, but have not been well studied. Lei, Barzilay, and Jaakkola (2016) propose a Generator-Encoder architecture to select small key snippets of input text as the rationales for the sentiment prediction. Their model has the advantage of training without any rationale annotations. However, the information loss in the selection process leads to a reduction in prediction accuracy.

In order to overcome the three difficulties above, we first propose to extract rationales at phrase level, which can guarantee information concentration and semantic integrity at the same time. Besides, we adopt a reinforcement style neural method (Jiang, Ye, Luo, Chao, & Ma, 2018) which stems from Lei et al. (2016) to extract rationales using the only supervision of charge label. Finally, we further propose a dynamic rationale attention mechanism to utilize the information in extracted rationales, to improve the charge prediction performance.

We train and evaluate our model on real Chinese criminal cases by collecting legal documents from China Judgements Online[1]. Experimental results demonstrate that besides providing charge prediction interpretation, our approach can also capture subtle details to help with charge prediction. On the evaluation indicators of rationale extraction and charge prediction, our approach outperforms the attention based model. This paper's contributions can be summarized as follows:

• We emphasize the importance of interpretability to charge prediction task, and put forward a feasible solution by extracting rationales.

• We introduce a neural framework to solve the problem by not only extracting rationales but also utilizing the rationales to improve the accuracy of charge prediction. The proposed dynamic rationale attention mechanism can inspire other text classification tasks.

• We build and release a real dataset of Chinese criminal judgement documents, which can be used to study charge prediction and other related issues in AI&Law.

2. Related Work

Our work is first related to the field of AI&Law. A lot of works have been presented to promote judicial informatization and intelligentialize in the last few decades. Among them, charge prediction is a particularly important task. Previous works (Liu et al., 2004; Liu & Hsieh, 2006; Lin et al., 2012; Luo et al., 2017; Hu, Li, Tu, Liu, & Sun, 2018) usually consider charge prediction as a text classification problem and make efforts to mining word-based or phrase-based features from input text. Among them, Liu and Hsieh (2006) and Liu et al. (2004) adopt K-Nearest Neighbors (KNN) as the classifier with words or phrases extracted as shallow textual features. Besides, Lin et al. (2012) and Hu et al. (2018) improve charge prediction performance by creating the concept of key attributes between facts and charges. Recently, Luo et al. (2017) first propose a neural attention framework to jointly extract related articles and elevate charge prediction accuracy. Further more, works in AI&Law also focus on identifying appropriate law articles for given cases (Aletras, Tsarapatsanis, Preotiuc-Pietro, & Lampos, 2016), retrieving similar historical cases (Raghav, Reddy, & Reddy, 2016; Chen, Liu, & Ho, 2013), and predicting the overall outcome of cases (Aletras et al., 2016; Katz, Bommarito II, & Blackman, 2017), etc.

Our work is also related to machine learning interpretability (Simonyan, Vedaldi, & Zisserman, 2013; Park, Hendricks, Akata, Schiele, Darrell, & Rohrbach, 2016; Hendricks et al., 2016; Lipton, 2016; Ribeiro, Singh, & Guestrin, 2016; Ling, Yogatama, Dyer, & Blunsom, 2017), which has been considered to be increasingly crucial in various fields including computing vision (CV) and natural language processing (NLP). According to the definition of interpretability by Hendricks et al. (2016), these works may focus on introspection explanation (Simonyan et al., 2013; Zhang et al., 2016), justification explanation (Hendricks et al., 2016; Ling et al., 2017), or both (Park et al., 2016). The application scenarios for these works include image captioning, medical diagnosis, algebraic problem solving, etc. More similar to our work, there are some works proposed to enhance the interpretability of AI&Law. Ye, Jiang, Luo, and Chao (2018) consider court views as the explanation for the pre-decided charges. They use a charge-conditioned Seq2Seq model to generate court views based on criminal cases' fact descriptions and the given charge labels. Luo et al. (2017) propose to select supportive law articles and use the articles to enhance the charge prediction accuracy. The supportive law articles are treated as a kind of support for the predicted charge.

1. http://wenshu.court.gov.cn

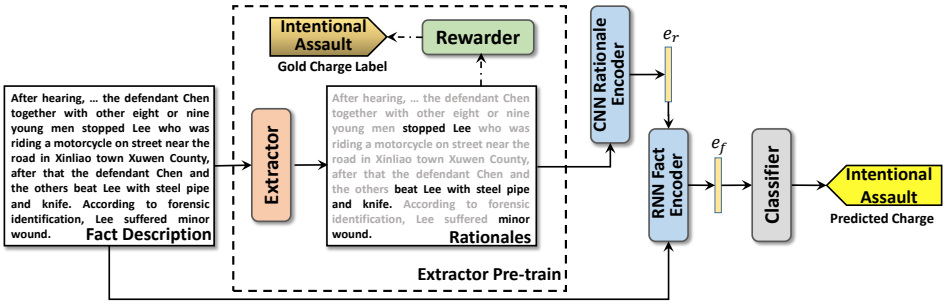

Figure 2: Overview of Our Interpretable Rationale Integrated Charge Prediction Model

As an important part of our approach, attention mechanism is highly relevant to our work. It has made a series of progress in various tasks, including text classification (Yang et al., 2016), speech recognition (Chorowski, Bahdanau, Serdyuk, Cho, & Bengio, 2015), image captioning (Xu et al., 2015), text summarization (Rush, Chopra, & Weston, 2015), etc. As a text classification task, our work shares more similarities with Yang et al. (2016) who propose a hierarchical attention based recurrent neural network (RNN) on document classification to attentively distinguish informative words and sentences. The main difference is that, the *context vector* (Yang et al., 2016) is no longer static but dynamically generated from the extracted rationales. This means that we can utilize the information in the extracted rationales to generate a more reasonable weight distribution. It turns out that our rationale attention indeed captures important information more precisely.

Essentially, our work is a text classification issue. In recent years, remarkable results have been achieved in the field of text classification (Reis & Culotta, 2018; Epshteyn & Dejong, 2006; Yang & Nenkova, 2017; Denoyer & Gallinari, 2004; Stamatatos, 2008; Kim, 2014; Joulin, Grave, Bojanowski, & Mikolov, 2016; Yang et al., 2016). Joulin et al. (2016) propose a simple but effective deep learning baseline for text classification, which only represents documents by averaging the embeddings of the appearing words. Kim (2014) applies a convolutional neural network (CNN) to text classification and achieved the state-of-art results on sentence classification at the time. In the modeling process of rationales, we adopt the same single-layer CNN.

3. Approach

Considering charge prediction as a text classification task, we aim to provide the rationales while predicting the charge labels. We define the input fact description as word sequence $x = [x_1, \ldots, x_n]$, and the gold charge label y as a non-negative integer. The rationales refer to the most charge-decisive text snippets in x. Given x, we aim to extract rationales $r = \{x_i | z_i = 1, x_i \in x\}$ ($z_i \in \{0, 1\}$), and predict $\tilde{y}$ based on x and r.

Figure 2 shows the overview of our approach. Firstly, rationales are extracted from the input fact description by EXTRACTOR. Then, a two-stack attention based neural network separately models the rationales and the fact description as two vectors e_r and e_f. e_r serving as a rationale attention context vector, participates in e_f's building process. Finally, e_f is fed into CLASSIFIER and utilized to predict the charge. The training process thus consists of two phases. First, EXTRACTOR is pre-trained using the only supervision of charge label. REWARDER is proposed to provide the reward and be jointly trained with EXTRACTOR. Then, freezing the parameters of EXTRACTOR and utilizing the rationales extracted, the overall system is trained to predict the charge.

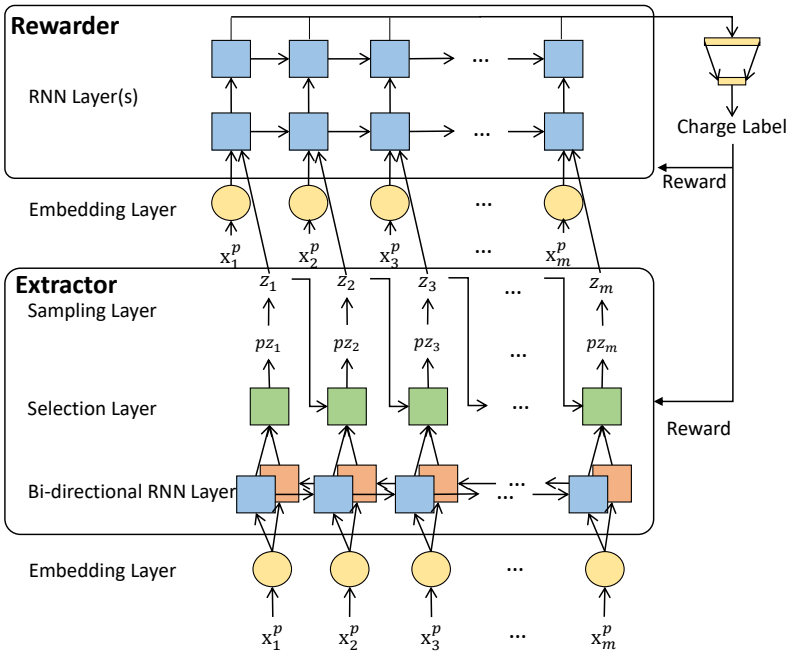

Figure 3: Architecture for EXTRACTOR Training

3.1 Phrase-Level Rationale Extraction

Here, we will present how we conduct EXTRACTOR training using the only supervision of charge label. Figure 3 shows the architecture for EXTRACTOR training. The architecture consists of two components, namely EXTRACTOR and REWARDER. EXTRACTOR is responsible for extracting an appropriate number of charge decisive text snippets from the input fact description as the rationales. REWARDER models the rationales, predicts the charge distribution $\tilde{y}$, and then calculates the discrepancy between $\tilde{y}$ and the gold charge y.

3.1.1 EXTRACTOR

Inspired by Lei et al. (2016), we introduce the rationales as latent variables and train the model to extract them from input fact description in a reinforcement learning style, without rationale annotation. Considering the snippet-like rationales should be more integral in semantics, we parse the fact description and split the sentences into phrases with a maximum length of 6 words. The phrase-level fact x^p is denoted as $[x_1^p, x_2^p, \cdots, x_m^p]$. x_i^p represents the i-th phrase in the fact description. x_i^p's distributed representation $e(x_i^p)$ is defined as the average word embedding in the phrase.

A latent variable z ($z \in \{0,1\}^m$) is introduced to define the extraction of phrases. The final goal of rationale extraction is to learn a distribution $p(z|x^p)$ over the phrases. The distribution can be represented through a recurrent neural model. At time t, $p(z_t)$ is calculated as follows:

$$
\begin{aligned}
p(z_t|x^p, z_{<t}) &= \text{sigmoid}(W_0[\overrightarrow{h_t}; \overleftarrow{h_t}; z_{t-1}] + b_0) \\
\overrightarrow{h_t} &= \overrightarrow{f}\,(e(x_t^p), \overrightarrow{h_{t-1}}) \\
\overleftarrow{h_t} &= \overleftarrow{f}\,(e(x_t^p), \overleftarrow{h_{t+1}})
\end{aligned}
$$

where W_0 and b_0 are trainable parameters; $\overrightarrow{f}$ and $\overleftarrow{f}$ are Bi-RNN functions which read the input sequence forward and backward. In this paper, we choose Bi-directional Gated Recurrent Units (Bi-GRU) (Cho et al., 2014) for the recurrent units. After the calculation of $p(z_t)$, a binary selection will be processed according to the probability $p(z_t)$ to generate z_t. From the above formulas, we can see that at time t, the information from current GRU output and states of $z_{<t}$ are jointly considered to predict the label of x_t^p. The extracted rationales are $r = \{x_i^p | z_i = 1, x_i^p \in x^p\}$.

3.1.2 REWARDER

The learning of rationale extraction needs a reward function to guide. Here, we introduce a deep RNN model with L layers to model r, generate distribution over charge labels $\tilde{y}$ and then calculate discrepancy between $\tilde{y}$ and the target charge y. Given the extracted rationale sequence $r = [r_1, r_2, \ldots, r_m]$, the hidden state at time t in the l-th layer is defined as follow:

$$h_t^{(l)} = \begin{cases} f(h_t^{(l-1)}, h_{t-1}^{(l)}) & l > 0 \\ f(e(r_t), h_{t-1}^{(0)}) & l = 0 \end{cases}$$

where $e(r_t)$ represents the pre-trained embedding of r_t and f is a unidirectional RNN function. Without loss of generality, the RNN unit here is also set to GRU. The final embedding of r is the average of all the hidden states in the last layer. In this paper, we set L as 2. $\tilde{y}$ is calculated as $\tilde{y} = sigmoid(W_1 e_r + b_1)$ where W_1 and b_1 are trainable parameters.

3.1.3 JOINT TRAINING

We take r as a latent variable and joint train EXTRACTOR and REWARDER. To control the number of phrases extracted from facts, we introduce a penalty over z as $\Phi(z) = |\| z \| - \eta|$ where $\| z \| = \sum_{i=1}^{m} z_i$ and η is a constant to control $\| z \|$ around η in case of $\| z \|$ being too small or too large. Square error is used to define the loss, and the final cost function is $\mathcal{L}_{\theta_e \theta_r}(x, z, y) = \| \tilde{y} - y \|_2^2 + \lambda \Phi(z)$, where θ_e and θ_r represent the trainable parameters of EXTRACTOR and REWARDER. Considering that z is never given during the training process, what we actually optimize is the expectation of the loss:

$$\min_{\theta_e \theta_r} \mathbb{E}_{z \sim Extrator(x)} [\mathcal{L}_{\theta_e \theta_r}(x, z, y)]$$

Since value space of z is exponential, we cannot traverse all the cases. We use sampling technique (Williams, 1992) to minimize the expected loss.

We jointly train EXTRACTOR and REWARDER, but adopt different gradients for them. Specifically, for the parameters of EXTRACTOR:

$$\begin{aligned} & \nabla_{\theta_e} \mathbb{E}_{z \sim Extrator(x)} [\mathcal{L}_{\theta_e \theta_r}(x, z, y)] \\ = & \nabla_{\theta_e} \sum_z \mathcal{L}_{\theta_e \theta_r}(x, z, y) \cdot p(z) \\ = & \sum_z \mathcal{L}_{\theta_e \theta_r}(x, z, y) \cdot \frac{\nabla_{\theta_e} p(z)}{p(z)} \cdot p(z) \end{aligned}$$

since $(lnf(x))' = f'(x)/f(x)$, we get

$$
\begin{aligned}
&= \sum_{z} \mathcal{L}_{\theta_e\theta_r}(x,z,y) \cdot \nabla_{\theta_e} \ln p(z) \cdot p(z) \\
&= \mathbb{E}_{z\sim Extrator(x)} \left[\mathcal{L}_{\theta_e\theta_r}(x,z,y) \cdot \nabla_{\theta_e} \ln p(z)\right]
\end{aligned}
$$

At this point, we can adopt sampling technique (Williams, 1992) to sample z and then approximate the gradient of this function. For the parameters of REWARDER, gradient can be calculated as below and approximated by the same sampling technique:

$$
\begin{aligned}
&\nabla_{\theta_r} \mathbb{E}_{z\sim Extrator(x)} \left[\mathcal{L}_{\theta_r\theta_r}(x,z,y)\right] \\
&= \nabla_{\theta_r} \sum_{z} \mathcal{L}_{\theta_e\theta_r}(x,z,y) \cdot p(z) \\
&= \sum_{z} \nabla_{\theta_r} \mathcal{L}_{\theta_e\theta_r}(x,z,y) \cdot p(z) \\
&= \mathbb{E}_{z\sim Extrator(x)} \left[\nabla_{\theta_r} \mathcal{L}_{\theta_e\theta_r}(x,z,y)\right]
\end{aligned}
$$

3.2 Interpretable Charge Prediction with Dynamic Rationale Attention

After the rationale extraction training, EXTRACTOR can already extract an appropriate number of charge decisive phrases as the rationales. Evaluation results on rationale annotated corpus also indicate that our model can extract rationales more accurately than attention based model (Yang et al., 2016). However, evaluation results on charge prediction show that it is not enough to use them alone to make accurate predictions. This is because that words not selected as rationales still contain useful information for charge prediction. We hold that since our model can extract rationales accurately, a re-evaluation of word weights by modeling the extracted rationales as context vector may be a better approach than attention model with static context vector. Therefore, we design the follow-up complete model.

3.2.1 CNN RATIONALE ENCODER

Considering that the amount of rationales we extract in one case is quite small, we simply regard them as a sentence and adopt the simple single-layer convolutional neural network (CNN) (Kim, 2014) which has been widely used in feature extraction (Ye, Yan, Luo, & Chao, 2017; Zeng, Liu, Lai, Zhou, & Zhao, 2014; Ye, Chao, Luo, & Li, 2017), to model them. Given a rationale word sequence $r = [r_1, r_2, \ldots, r_n]$, we build a matrix $M \in \mathbb{R}^{n\times d}$ with the d-dimensional pre-trained embedding of each word. And then the matrix is fed to a CNN with a series of convolving filters. The filters with different heights h, are responsible for gripping different h-gram features. After applied to M, each filter $f_i \in \mathbb{R}^{h_i\times d}$ induces a *feature map* $fm_i \in \mathbb{R}^{n-h_i+1}$. We then apply a Rectified Linear Unit (ReLU) (Krizhevsky, Sutskever, & Hinton, 2012) activation to the feature maps. Finally, a max-overtime pooling operation (Collobert et al., 2011) is applied to the feature maps. The maximum value of each feature map is selected and concatenated to form e_r, the vector representation of the rationale snippets.

3.2.2 RNN FACT ENCODER

We simply regard the fact description as a single word sequence $x = [x_1, x_2, \ldots, x_n]$, and use a Bi-directional Gated Recurrent Unit (Bi-GRU) to encode it as follows:

$$
\begin{aligned}
h_t &= [\overrightarrow{h_t}; \overleftarrow{h_t}] \\
\overrightarrow{h_t} &= \overrightarrow{GRU}(e(x_t), \overrightarrow{h_{t-1}}) \\
\overleftarrow{h_t} &= \overleftarrow{GRU}(e(x_t), \overleftarrow{h_{t+1}})
\end{aligned}
$$

In general, the average of the hidden states of all words is used as the final embedding vector. However, this method can not stand out the important information from useless information. We solve this problem by introducing the attention mechanism. Given the Bi-GRU hidden state sequence $h = [h_1, h_2, \cdots, h_T]$, the final embedding vector is calculated as follows:

$$
\begin{aligned}
e_f &= \sum_{t}^{n} a_t h_t \\
a_t &= \frac{\exp(tanh(W_2 h_t)^T c_r)}{\sum_t \exp(tanh(W_2 h_t)^T c_r)}
\end{aligned}
$$

Different from Yang et al. (2016), we use a dynamically generated context vector c_r from our CNN RATIONALE ENCODER, instead of using a static global context vector. c_r is dynamically calculated from e_r: $c_r = tanh(W^3 e_r + b^3)$. We believe that in a text classification task, we should vary our focus with the topic of the document, rather than using the same criteria to allocate attention. Treating e_r as a representation of the document's topic, we intend to increase the weights of words that are related to e_r.

3.2.3 CLASSIFIER

Our CLASSIFIER consists of a linear full connected layer and an activation layer. The fact embedding e_f are fed into the linear full connected layer to produce another fact embedding e. Then, through the activation layer, e generates the final distribution on the charges. In this paper, we choose sigmoid as the activation function, and correspondingly, mean squared error is used to measure the training loss:

$$
\begin{aligned}
\tilde{y} &= sigmoid(W_4 e_f + b_4) \\
\mathcal{L} &= \sum_{i=1}^{N} \sum_{j=1}^{C} (y_{ij} - \tilde{y}_{ij})^2
\end{aligned}
$$

where N is the training set size, and C is the charge label set size. y_{ij} is set to 1 if the i-th case's charge is j. Otherwise, it is set to 0.

4. Experiments

In this section, we describe the corpus used to train and evaluate the proposed model, the experimental setup, the baseline methods, the metrics, and the experiment results.

经审理查明，... [犯罪嫌疑人陈某伙同八九个年轻人拦下了在徐闻县新寮镇道路上骑摩托的李某。被告陈某和其他人用钢管和刀子殴打李某。经法医鉴定，李某受轻伤。] fact description

.....

本院认为，李某伙同他人，持械伤害被害人，造成被害人轻伤，其行为已构成 [*故意伤害罪*] charge。

After hearing, ... [the defendant Chen together with other eight or nine young men stopped Lee who was riding a motorcycle on street near the road in Xinliao town Xuwen County, after that the defendant Chen and the others beat Lee with steel pipe and knife. According to forensic identification, Lee suffered minor wound.] fact description

.....

Our court hold that, Li, together with others, injured victim with weapons, causing minor injuries. His behavior has constituted the crime of [*Intentional Assault*] charge.

Figure 4: An fragment of legal document

4.1 Dataset and Data Preparation

We construct a dataset from China Judgements Online, which contains a large number of documents on various cases throughout the country. 80,000, 10,000 and 10,000 identically distributed documents are randomly selected as training, validation and test set respectively. As shown in Figure 4, paragraphs that begin with “”(“After hearing, the court identified”) are extracted as fact descriptions. Fact descriptions longer than 256 will be stripped. We collected 400 most common charges from PRC Criminal Law. The ones appear between “”(“His/Her behavior has constituted”) and the next period are counted as the charges of the case. Cases with multiple suspects are filtered out and left for future work. For the cases with multiple charges, we regard the combination of charges as a new independent charge. We choose the 50 most common charges in the dataset and the cases with other charges are counted as negative data. Figure 5 shows the proportions of data for each charge in our dataset. The top ten charges account for 70 % of the total dataset, reflecting the imbalance of our dataset.

We use HanLP[2] to tokenize the Chinese texts. CoreNLP (Manning, Surdeanu, Bauer, Finkel, Bethard, & McClosky, 2014) is used to parse the syntax tree, and words in a subtree with a max length of 6 make up a phrase. There are 2.8 words in each phrase on average. In order to eliminate interference as much as possible, we use regular expressions to match names, numbers and dates in the corpus, and use “<name>”, “<num>” and “<date>” to replace them respectively. Table 1 shows the statistics of our dataset in detail.

4.2 Experimental Setup

The hidden size is set to 200 for all GRUs. For CNN RATIONALE ENCODER, we set different filter heights of 3, 4, 5, with 128 feature maps each. 200 dimensional word embeddings are pre-trained using the continuous bag-of-words architecture (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) on around 1M legal documents. The word embedding contains 313,729 words. Words that are not in the set of pre-trained words are initialized randomly.

We choose a batch size of 64 and adopt Adam stochastic optimization method (Kingma & Ba, 2014) to learn the trainable parameters. We apply a dropout regularization layer (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014) after the RNN and CNN outputs, and L2 regularization to all trainable parameters.

2. https://github.com/hankcs/HanLP

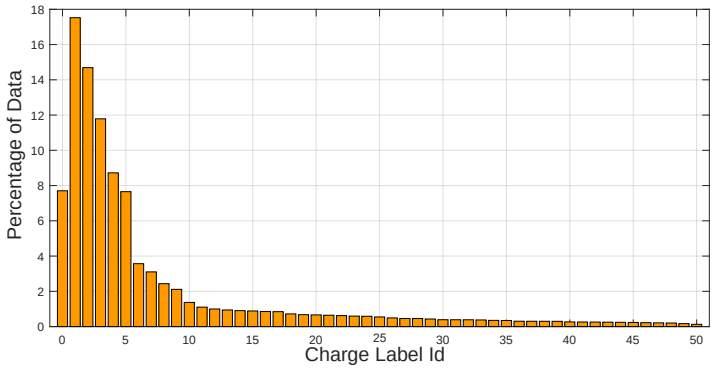

Figure 5: Percentages of data for each charge in our dataset. 0 for all the other rare charges.

# Charges	50
# Training Set	80,000
# Validation Set	10,000
# Test Set	10,000
Avg. # words in Fact Description	219.9
# Annotated Docs	934
# Sentences in Annotated Docs	6,644
# Sentences Annotated as Rationale	2,279

Table 1: Statistics of the legal document dataset

4.3 Metrics and Baselines

Rationale Extraction We evaluate rationale extraction performance using *precision*, *recall*, and *F1* at a document level and *accuracy* at a sentence level. That is, rationales are extracted from the documents, and the sentences hit by at least one rationale snippet are regarded as model-extracted rationale sentences. For comparison, we implement an attention based recurrent neural network (GRU_ATT_F) which takes fact description as input. Discarding the concept of sentence, it is a simplified version of Yang et al. (2016), the state-of-art general document classification model. We use GRU as the recurrent unit. η is tuned to control the number of rationale extracted by our EXTRACTOR. We select the corresponding number of words according to the weights given by the attention based model.

Charge Prediction We evaluate charge prediction performance using *precision*, *recall* and *F1*. Considering that both the *precision* and *recall* are calculated according to certain one category which only represents a local effect, we average them at both micro and macro level to evaluate the performance of our model in a global aspect. The calculation is based on four indicators: False Positive (FP) stands for the number of instances which are labeled as positive while they are negative; Accordingly, False Negative (FN) is the number of instances which are labeled as negative while they are positive. True Positive (TP) and True Negative (TN) represent the number of positive and negative instances that are correctly labeled, respectively. *Precision* and *recall* are defined as

follow:

$$Precision = \frac{TP}{TP + FP}$$
$$Recall = \frac{TP}{TP + FN}$$

and *F1* is the harmonic mean of *Precision* and *recall*:

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

Micro average means calculating metrics globally by counting the total TP, TN, FP, FN. Macro average means calculating metrics for each label, and finding their unweighted mean.

In order to demonstrate the superiority of our method in charge prediction accuracy, six baselines are set as follows:

• $\mathbf{SVM_R}$ and $\mathbf{SVM_F}$ are SVM-based methods which respectively take extracted rationales and fact descriptions as input. We choose linear kernel for kernel function and Bag-of-words(BOW) for feature representation. 10,000 feature words are selected by TF-IDF scores.

• $\mathbf{GRU_R}$ and $\mathbf{GRU_F}$ are both bi-directional and single layer RNN-based methods, which also respectively take extracted rationales and fact descriptions as input. GRU is chosen as the recurrent units. The average of all the hidden states is used to represent the input.

• $\mathbf{GRU\_ATT_R}$ and $\mathbf{GRU\_ATT_F}$ are attention versions of $\mathbf{GRU_R}$ and $\mathbf{GRU_F}$. The weighted average of all the hidden states is used to represent the input.

4.4 Experimental Results

We evaluate our model from two aspects: rationale extraction and charge prediction, reflecting the interpretation and prediction performance of the model respectively.

4.4.1 RATIONALE EXTRACTION

Figure 6 presents the rationale extraction evaluation results of our model (RAT_ATT) and the attention based RNN model (GRU_ATT_F). As shown in the figure, as the quantity of extracted rationales increases, the *precision* and *accuracy* of RAT_ATT gradually drop and *recall* increases correspondingly. *F1* value ranges from 68% to 72% and reaches the peak when the rationale word count is 20. The rationale extraction performance of GRU_ATT_F has similar trend of change, but it is always weaker than our model. Since our model can be more accurate in rationale extraction than the attention based RNN model, we should further take this advantage to improve the charge prediction performance. This is exactly why we introduce the follow-up CNN RATIONALE ENCODER and RNN FACT ENCODER.

Case Study Table 2 presents several cases with the rationales highlighted by our model. Taking the phrase as the basic unit, the rationales show good readability and make it easy to understand. These confusing charges usually share superficially similar fact descriptions. However, the key factors that truly distinguish between the confusing charges are not the direct crime, but the subjective and objective state in which the suspects committed the crimes, including the suspects' psychological states, crime manifestations, and social roles, etc. It is usually hard for models to capture. For example, in PRC Criminal Law, an important factor that distinguishes *Official Embezzlement* from

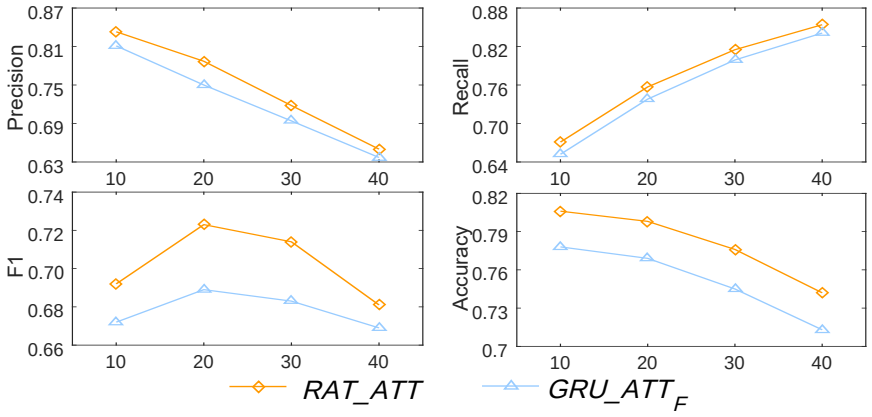

Figure 6: Evaluation results on rationale extraction. *Accuracy* is the number of all correctly predicted sentences divided by the total number of sentences. For *precision*, *recall* and *F1*, we calculate the metrics of each document, and then average them over all articles.

DEMONSTRATION OF OUR RATIONALE EXTRACTION
CASE 1 $[Official\ Embezzlement]_{charge}$
… PP 利用 其 担任 公司 业务员 的 职务 便利， 从 公司 仓库 提 走 多部 手机， 后 将 手机 卖掉， 货款 挥霍 …
…Using his position as a company salesman, PP took several phones from the company's warehouse, sold the phones, and squandered the money…
CASE 2 $[Larceny]_{charge}$
…PP_1 趁 PP_2 家中 无 人 之 机， 进入 到 PP_2 家 卧室 内 伺机 盗窃 。 被 PP_2 回家 后 发现， PP_1 翻 墙 逃跑…
…When PP_2 was not at home, PP_1 went to PP_2's bedroom to steal. When PP_2 came home, PP_1 fled the wall and ran…
CASE 3 $[Negligently\ Causing\ Fire]_{charge}$
…在 焚烧 耕地 上的 杂草 时， 不慎 引发 山林 火灾 。案发 后， PP 积极 救火， 主动 向 上级 说明 失火 情况…
… When burning weeds on cultivated land, PP inadvertently ignited the mountain fire. Then, PP actively doused the fire and reported the fire situation …
CASE 4 $[Arson]_{charge}$
…PP_1 因 生意 竞争 与 PP_2 产生 积怨 。PP_1 酒后 萌生 火烧 PP_2 手机店 的 念头， 进入 店内 将 纸箱 点燃…
…PP_1 hates PP_2 for his business competition. After drinking, PP_1 produced the idea of burning PP_2's shop. He entered the shop and lights the carton…
CASE 5 $[Negligent\ Homicide]_{charge}$
… 在 狩猎 过程 中， PP 因 地 滑 摔跤， 导致 其 所 持 鸟 铳 击 发 走火， 将 走 在前面 的 PP 打 伤 致死 …
…In the process of hunting, PP fell down due to the slippery ground , leading to the shotgun fire, killing PP who was walking in front …
CASE 6 $[Intentional\ Homicide]_{charge}$
…PP_1 从 家中 携带 匕首 出门 寻找 PP_2 进行 报复， 将 PP_2 捅倒 后， 在 颈 部 来回 割， 致 PP_2 当场 死亡…
…PP_1 took the dagger and looked for PP_2 for revenge. He stabbed PP_2 and then cut him back and forth on the neck, causing PP_2 to die on the spot…

Table 2: Examples of extracted rationales. The highlighted words are rationales extracted by our EXTRACTOR. Different colors are used to align Chinese original text and corresponding English translation.

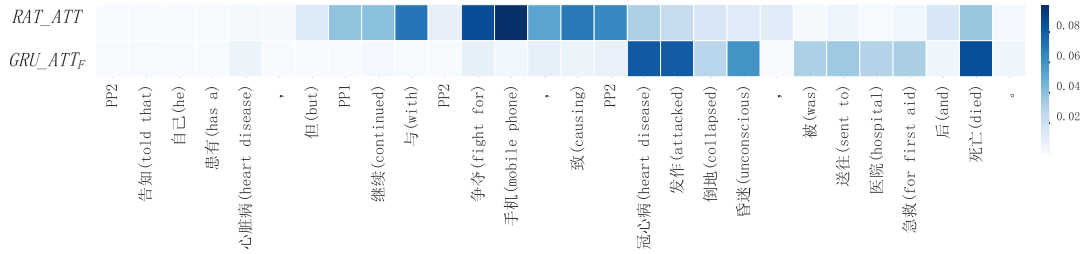

Figure 7: Heat map for rationale-attention mechanism analysis.

Models	Micro			Macro		
	P	R	F1	P	R	F1
SVM_R	86.33	86.26	86.34	78.60	68.05	71.72
SVM_F	87.51	86.73	87.25	81.53	67.62	72.48
GRU_R	87.36	88.67	88.00	77.35	74.04	75.03
GRU_F	89.64	90.60	90.12	81.84	76.25	78.08
GRU_ATT_R	88.09	89.21	88.65	79.54	75.83	77.15
GRU_ATT_F	**90.19**	91.08	90.53	**83.65**	77.93	80.06
RAT_ATT	90.17	**91.35**	**90.77**	83.11	**80.02**‡	**81.02**‡

Table 3: Charge prediction results.The amount of rationale phrases is 13(around 36 Chinese words). When η takes other values, we get similar performance. "‡": significantly better than GRU_ATT_F (p<0.01)

Larceny is "*utilizing the convenience of duties*". In case 1, our model successfully captures the key fact that "*using his position as a company salesman*". It is exactly the key factor in convicting this case as *Official Embezzlement*. In case 3, the fact "*inadvertently ignited*" is also extracted by our model as a rationale. Distinguishing from "subjective intentional arson", it is the key to convicting the suspect in the case as a charge of *Negligently Causing Fire* rather than *Arson*. In case 5 and 6, the difference between the shoot killing caused by a slippery fall and the revenge killing is the essential factor distinguishing these two homicide cases which are also extracted by our EXTRACTOR.

4.4.2 CHARGE PREDICTION

We set η between values $\{9, 11, 13, 15\}$, and train the series of the models respectively. As shown in Table 3, SVM model is a strong baseline and the advantage of neural based model is obvious. As we expected, although the rationales contain the core information of the documents, the performance of classification based on rationales only is still discounted. Thanks to the availability of all information, GRU_ATT_F achieves the best results in addition to RAT_ATT. Though the improvement of our RAT_ATT on *micro-F1* is not significant, the significant gap of *macro-F1* between RAT_ATT and GRU_ATT_F proves that our framework has more advantages on rare charges.

Figure 8 further details the *F1* gaps between RAT_ATT and GRU_ATT_F in charge prediction for all the charges. Both taking the fact description as a word sequence and build representation by weighting the GRU hidden states, the performance gap is entirely due to the introduction of rationale attention. As shown in the figure, although the two models have almost the same performance on the small-id charges (the more frequent the charge appears, the smaller the id), our RAT_ATT has

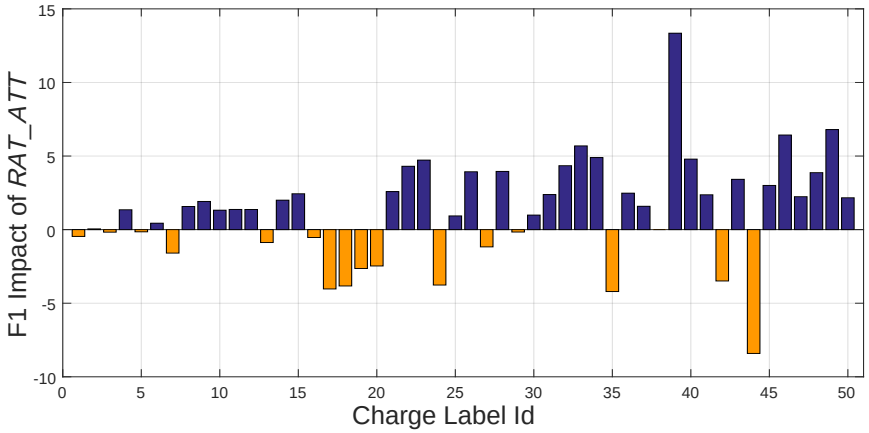

Figure 8: *F1* gaps between RAT_ATT and GRU_ATT_F in charge prediction for all the charges.

more obvious advantages on the rare and indistinguishable charges. To some degree, this result also shows that our model is more accurate in subtle detail grasping.

Case Study We select a fragment of a case with a charge of *Negligent Homicide*, and visualize the attention map generated by RAT_ATT and GRU_ATT_F in Figure 7. The attention distribution generated by RAT_ATT notices the fact "fight for mobile phone", "heart disease attacked", and "died". Because the model notices the unintentional character of the suspect, it successfully predicts the case as a charge of *Negligent Homicide*. However, GRU_ATT_F ignores the most important fact "in the fight for mobile phones", which directly determines the conviction of the case. Therefore, GRU_ATT_F mistakenly judges the case as *Intentional Homicide*.

4.4.3 FURTHER ANALYSIS

Impact of Rationale Amount We evaluate the impact of different amount of extracted rationales on charge prediction. By adjusting the parameter η, we control our EXTRACTOR to extract different amounts of fact description as rationales, and then utilize the rationales to make charge prediction through the proposed dynamic rationale attention mechanism. Figure 9 presents the *F1* scores of charge prediction with different amounts of rationale extracted. As the number of extracted rationale snippets increases from 5 to 13, the *F1* score of charge prediction gradually increases. However, when the number continues to rise, *F1* score remains stable. Therefore, we can draw the conclusion that a certain number (here is 13) of rationales is sufficient to optimize the charge prediction performance while ensuring focus.

Dynamic Attention in Text Classification Through the experiments above, we have already demonstrated the effectiveness of dynamic attention mechanism in improving charge prediction accuracy. We further design and implement experiments to demonstrate the superiority of dynamic attention mechanism over other attention mechanisms in text classification tasks.

We evaluate the performance of the proposed dynamic attention model on four multi-classification datasets (20Newsgroups[3](Hingmire, Chougule, Palshikar, & Chakraborti, 2013), TREC[4](Li & Roth,

3. http://web.ist.utl.pt/acardoso/datasets/
4. http://cogcomp.cs.illinois.edu/Data/QA/QC/

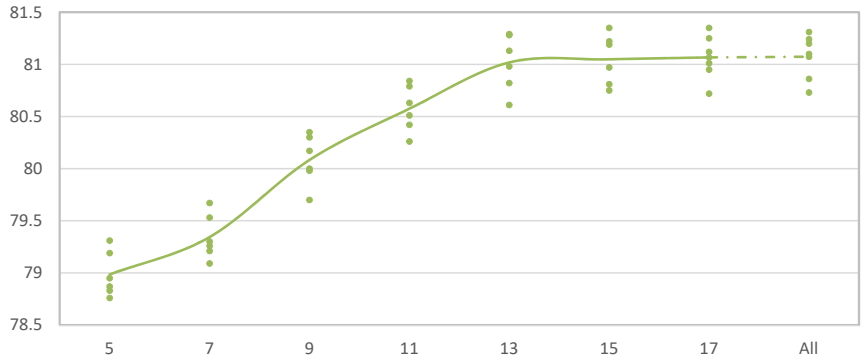

Figure 9: *Macro-F1*(y-axis) score of charge prediction when various amounts of phrase are extracted fact description as rationales (x-axis). "All" means all the text in fact descriptions is extracted. The points on the line are the average of the six values after the maximum and minimum values removed from 8 independent experiments.

Dataset	**#C**	**#Train**	**#Valid**	**#Test**	**Avg.#Len**	**#V**	**#V_{pre}**
20Newsgroups	4	7520	836	450	276	67088	41887
TREC	6	5452	-	500	10	9685	5812
AG	4	120000	-	7600	35	62975	55029
SST-1	5	8544	1101	450	19	19126	17479
MR	2	10662	-	**10-CV**	20	21425	17690
Subj	2	10000	-	**10-CV**	23	21322	19983

Table 4: Details of the text classification datasets. **#C**: number of classes, **#V**: vocabulary size, **#V_{pre}**: number of words present in the set of pre-trained word embeddings, **10-CV**: 10-fold cross validation.

2002), AG[5](Zhang, Zhao, & LeCun, 2015), SST-1[6](Socher, Perelygin, Wu, Chuang, Manning, Ng, & Potts, 2013)),and two binary classification datasets (MR[7](Pang & Lee, 2005), Subj[8](Pang & Lee, 2004)). Table 4 shows the details of these datasets.

Baselines are selected and divided into four types: traditional machine learning methods based on feature engineering, deep learning models based on convolutional neural network, deep learning models based on recurrent neural network for automatic feature extraction, and models distilling important information by self-attention mechanism.

5. http://www.di.unipi.it/ gulli/AG_corpus_of_news_articles.html
6. http://nlp.stanford.edu/sentiment/
7. https://www.cs.cornell.edu/people/pabo/movie-review-data/
8. https://www.cs.cornell.edu/people/pabo/movie-review-data/

Model	20News	SST-1	Subj	TREC	AG	MR
2DCNN	96.5	**52.4***	94.0	96.1	-	82.3
CNN-non-static+UNI	-	50.8	93.7	94.4	-	82.1
MVCNN	-	49.6	93.9	-	-	-
Tree-LSTM	-	51.0	93.2	-	91.8	80.7
ID-LSTM	-	50.0	93.5	-	92.2	81.6
HS-LSTM	-	49.8	93.7	-	92.5	82.1
CRAN_prand	-	50.0	94.1	-	-	82.8
Self-Attentive	-	47.2	92.5	-	91.1	80.1
Att-BLSTM	94.6	49.8	93.5	93.8	91.5	81.0
Dyn-Att+Bi-LSTM	**97.1***	50.4	**94.8***	**98.4**‡	**92.8***	**83.3***
Adasent	-	-	**95.5***	92.4	-	**83.1***
Combine-skip	-	-	93.6	92.2	-	76.5

Table 5: Results of Dyn-Att+Bi-LSTM (Dynamic attention mechanism with bidirectional LSTM) and the baselines on the datasets. The " ‡" indicates result with the best performance among all models listed in the table. And the "*" indicates result better than all the neural (attention included) based models.

- **Combine-skip**: It is a general unsupervised sentence representation model that draws lessons from word2vec's skip-gram model[9] to predict the last sentence and the next sentence (Kiros et al., 2015).
- **Adasent**: Adasent effectively forms a hierarchy of representations from words to phrases and then to sentences through recursive gated local composition of adjacent segments (Zhao, Lu, & Poupart, 2015).
- **2DCNN**: A combined framework that utilizes Bi-LSTM to capture long-term sentence dependencies, and extracts features by 2D convolution and 2D max pooling operation for sequence modeling tasks (Zhou et al., 2016a).
- **CNN-non-static+UNI**: Instead of randomly initializing the convolutional filters, this model (Li, Zhao, Liu, Hu, & Du, 2017) encode semantic features into them, which helps the model focus on learning useful features at the beginning of the training.
- **MVCNN**: A Multichannel variable-size convolution for sentence classification (Yin & Schütze, 2015).
- **ID-LSTM/HS-LSTM**:The former selects only import, task-relevant words, and the latter discovers phase structures in a sentence (Zhang, Huang, & Zhao, 2018).
- **Tree-LSTM**: A generalization of LSTM to tree-structured network topologies (Tai, Socher, & Manning, 2015).

9. http://code.google.com/p/word2vec/

- **CRAN**: A network Architecture combines recurrent neural network with CNN-based attention model (Du, Gui, Xu, & He, 2017). This method also uses CNN to generate the weight of a word. Unlike our model, it utilizes the convolution result at the position of specific word to align with the hidden state of RNN and calcute the weight.

- **Self-Attentive**: A self-attention based model using a 2-D matrix to represent the embedding, with each row of the matrix attending on a different part of the sentence (Lin, Feng, dos Santos, Yu, Xiang, Zhou, & Bengio, 2017).

- **Att-BLSTM**: Attention-based Bidirectional LSTM to capture the most important semantic information in a sentence (Zhou et al., 2016b).

Table 5 presents the performance of the proposed dynamic attention mechanism and other state-of-the-art models on six classification datasets. In order to maximize the superiority of the proposed dynamic attention mechanism, we treat all the texts as rationales and feed them to our model, eliminating the calculation of rationale extraction. The experimental results show that our model gets the best performance on 20Newsgroups, TREC and AG. Each of them is a dataset with more than two categories, of which 20Newsgroups and AG have four, and TREC has six. And even on datasets with two categories, our model still outperforms all the models based on deep neural network. It should be emphasized that on the six-category dataset TREC, our model outperforms the second place (2DCNN) by 2.3 percentage points and outperforms the static attention based model (Att-BLSTM) by 4.6 percentage points. The huge improvement proves the effectiveness of the original intention of dynamic attention mechanism – different concerns for different categories.

5. Conclusion

In this paper, we propose a neural framework to jointly extract charge decisive rationales and utilize them with a dynamic rationale attention mechanism to make charge prediction. The extracted rationales reflect the operating mechanism of the model and serve as an introspection explanation of the machine-generated decision, elevating the interpretability of charge prediction. Sufficient experiments demonstrate that our model outperforms the static attention based model on both rationale extraction and charge prediction. The case study on the indistinguishable cases also proves the superiority of our model in subtle details capturing and attention weights distributing. Furthermore, experimental results on text classification tasks also demonstrate strong competitiveness of the proposed dynamic attention mechanism in multi-category text classification tasks. As for the utilization methods of the rationale information, this paper only proposes one of them. There can be various other ways to take advantage of rationale information to improve charge prediction accuracy.

Acknowledgements

This work was supported by National Key Research and Development Program of China (Grant No. 2017YFB1402400), National Natural Science Foundation of China (No. 61602490), and State Key Laboratory of Software Development Environment (SKLSDE-2013ZX-15). Sharing equal contribution, Wenhan Chao and Xin Jiang are the co-first authors of the paper. Zhunchen Luo is the corresponding author.

References

Aletras, N., Tsarapatsanis, D., Preotiuc-Pietro, D., & Lampos, V. (2016). Predicting judicial decisions of the european court of human rights: a natural language processing perspective. *PeerJ Computer Science*, *2*, e93.

Chen, Y., Liu, Y., & Ho, W. (2013). A text mining approach to assist the general public in the retrieval of legal documents. *JASIST*, *64*(2), 280–290.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pp. 1724–1734.

Chorowski, J., Bahdanau, D., Serdyuk, D., Cho, K., & Bengio, Y. (2015). Attention-based models for speech recognition. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pp. 577–585.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, *12*, 2493–2537.

Denoyer, L., & Gallinari, P. (2004). Bayesian network model for semi-structured document classification. *Inf. Process. Manage.*, *40*(5), 807–827.

Du, J., Gui, L., Xu, R., & He, Y. (2017). A convolutional attention model for text classification. In *Natural Language Processing and Chinese Computing - 6th CCF International Conference, NLPCC 2017, Dalian, China, November 8-12, 2017, Proceedings*, pp. 183–195.

Epshteyn, A., & Dejong, G. (2006). Generative prior knowledge for discriminative classification. *Journal of Artificial Intelligence Research*, *27*(1), 25–53.

Hendricks, L. A., Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., & Darrell, T. (2016). Generating visual explanations. In *ECCV (4)*, Vol. 9908 of *Lecture Notes in Computer Science*, pp. 3–19. Springer.

Hingmire, S., Chougule, S., Palshikar, G. K., & Chakraborti, S. (2013). Document classification by topic labeling. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pp. 877–880. ACM.

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

Jiang, X., Ye, H., Luo, Z., Chao, W., & Ma, W. (2018). Interpretable rationale augmented charge prediction system. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pp. 146–151, Santa Fe, New Mexico. Association for Computational Linguistics.

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of tricks for efficient text classification. *CoRR*, *abs/1607.01759*.

Katz, D. M., Bommarito II, M. J., & Blackman, J. (2017). A general approach for predicting the behavior of the supreme court of the united states. *PloS one*, *12*(4), e0174698.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pp. 1746–1751.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *CoRR*, *abs/1412.6980*.

Kiros, R., Zhu, Y., Salakhutdinov, R., Zemel, R. S., Urtasun, R., Torralba, A., & Fidler, S. (2015). Skip-thought vectors. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pp. 3294–3302.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *NIPS*, pp. 1106–1114.

Lei, T., Barzilay, R., & Jaakkola, T. S. (2016). Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pp. 107–117.

Li, S., Zhao, Z., Liu, T., Hu, R., & Du, X. (2017). Initializing convolutional filters with semantic features for text classification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1884–1889.

Li, X., & Roth, D. (2002). Learning question classifiers. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pp. 1–7. Association for Computational Linguistics.

Lin, W., Kuo, T., Chang, T., Yen, C., Chen, C., & Lin, S. (2012). Exploiting machine learning models for chinese legal documents labeling, case classification, and sentencing prediction. *IJCLCLP*, *17*(4).

Lin, Z., Feng, M., dos Santos, C. N., Yu, M., Xiang, B., Zhou, B., & Bengio, Y. (2017). A structured self-attentive sentence embedding. *CoRR*, *abs/1703.03130*.

Ling, W., Yogatama, D., Dyer, C., & Blunsom, P. (2017). Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pp. 158–167.

Lipton, Z. C. (2016). The mythos of model interpretability. *CoRR*, *abs/1606.03490*.

Liu, C., Chang, C., & Ho, J. (2004). Case instance generation and refinement for case-based criminal summary judgments in chinese. *J. Inf. Sci. Eng.*, *20*(4), 783–800.

Liu, C., & Hsieh, C. (2006). Exploring phrase-based classification of judicial documents for criminal charges in chinese. In *ISMIS*, Vol. 4203 of *Lecture Notes in Computer Science*, pp. 681–690. Springer.

Luo, B., Feng, Y., Xu, J., Zhang, X., & Zhao, D. (2017). Learning to predict charges for criminal cases with legal basis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pp. 2727–2736.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014). The stanford corenlp natural language processing toolkit. In *ACL (System Demonstrations)*, pp. 55–60. The Association for Computer Linguistics.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *NIPS*, pp. 3111–3119.

Pang, B., & Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, p. 271. Association for Computational Linguistics.

Pang, B., & Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pp. 115–124. Association for Computational Linguistics.

Park, D. H., Hendricks, L. A., Akata, Z., Schiele, B., Darrell, T., & Rohrbach, M. (2016). Attentive explanations: Justifying decisions and pointing to the evidence. *CoRR*, *abs/1612.04757*.

Raghav, K., Reddy, P. K., & Reddy, V. B. (2016). Analyzing the extraction of relevant legal judgments using paragraph-level and citation information. *AI4J–Artificial Intelligence for Justice*, 30.

Reis, V. L. D., & Culotta, A. (2018). Robust text classification under confounding shift. *Journal of Artificial Intelligence Research*, *63*, 391–419.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "why should I trust you?": Explaining the predictions of any classifier. In *KDD*, pp. 1135–1144. ACM.

Rush, A. M., Chopra, S., & Weston, J. (2015). A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pp. 379–389.

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631–1642.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*(1), 1929–1958.

Stamatatos, E. (2008). Author identification: Using text sampling to handle the class imbalance problem. *Inf. Process. Manage.*, *44*(2), 790–799.

Tai, K. S., Socher, R., & Manning, C. D. (2015). Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pp. 1556–1566.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, *8*, 229–256.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A. C., Salakhutdinov, R., Zemel, R. S., & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pp. 2048–2057.

Yang, Y., & Nenkova, A. (2017). Combining lexical and syntactic features for detecting content-dense texts in news. *Journal of Artificial Intelligence Research*, *60*.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A. J., & Hovy, E. H. (2016). Hierarchical attention networks for document classification. In *HLT-NAACL*, pp. 1480–1489. The Association for Computational Linguistics.

Ye, H., Chao, W., Luo, Z., & Li, Z. (2017). Jointly extracting relations with class ties via effective deep ranking. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pp. 1810–1820.

Ye, H., Jiang, X., Luo, Z., & Chao, W. (2018). Interpretable charge predictions for criminal cases: Learning to generate court views from fact descriptions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pp. 1854–1864.

Ye, H., Yan, Z., Luo, Z., & Chao, W. (2017). Dependency-tree based convolutional neural networks for aspect term extraction. In *Advances in Knowledge Discovery and Data Mining - 21st Pacific-Asia Conference, PAKDD 2017, Jeju, South Korea, May 23-26, 2017, Proceedings, Part II*, pp. 350–362.

Yin, W., & Schütze, H. (2015). Multichannel variable-size convolution for sentence classification. In *Proceedings of the 19th Conference on Computational Natural Language Learning, CoNLL 2015, Beijing, China, July 30-31, 2015*, pp. 204–214.

Zeng, D., Liu, K., Lai, S., Zhou, G., & Zhao, J. (2014). Relation classification via convolutional deep neural network. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*, pp. 2335–2344.

Zhang, T., Huang, M., & Zhao, L. (2018). Learning structured representation for text classification via reinforcement learning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pp. 6053–6060.

Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pp. 649–657.

Zhang, Y., Marshall, I. J., & Wallace, B. C. (2016). Rationale-augmented convolutional neural networks for text classification. In *EMNLP*, pp. 795–804. The Association for Computational Linguistics.

Zhao, H., Lu, Z., & Poupart, P. (2015). Self-adaptive hierarchical sentence model. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, pp. 4069–4076.

Zhou, P., Qi, Z., Zheng, S., Xu, J., Bao, H., & Xu, B. (2016a). Text classification improved by integrating bidirectional LSTM with two-dimensional max pooling. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pp. 3485–3495.

Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., & Xu, B. (2016b). Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers*.