Interpretable Rationale Augmented Charge Prediction System

[1]Xin Jiang*, [1]Hai Ye*, [2]Zhunchen Luo*, [1]Wenhan Chao†, [1]Wenjia Ma
[1]School of Computer Science and Engineering, Beihang University
[2]Information Research Center of Military Science, PLA Academy of Military Science
[1,2]Beijing, China
[1]{xinjiang, yehai, chaowenhan, mawenjia}@buaa.edu.cn
[2]zhunchenluo@gmail.com

Abstract

This paper proposes a neural based system to solve the essential interpretability problem existing in text classification, especially in charge prediction task. First, we use a deep reinforcement learning method to extract rationales which mean short, readable and decisive snippets from input text. Then a rationale augmented classification model is proposed to elevate the prediction accuracy. Naturally, the extracted rationales serve as the introspection explanation for the prediction result of the model, enhancing the transparency of the model. Experimental results demonstrate that our system is able to extract readable rationales in a high consistency with manual annotation and is comparable with the attention model in prediction accuracy.

1 Introduction

Given a case's fact description, charge prediction aims to determine appropriate charge for the criminal suspect mentioned. Existing works generally treat charge prediction as a text classification problem, and have made a series of progress(Liu et al., 2004; Liu and Hsieh, 2006; Lin et al., 2012; Luo et al., 2017). However, in the field of justice, every decision may be a matter of life and death. It is necessary for judges and lawyers to understand the principles of the decisions, since people cannot completely trust the machine-generated judgement results without any interpretation provided.

Interpretability which means the ability of AI systems to explain their predictions, has attracted more and more attention. Hendricks et al. (2016) divide the concept of interpretation into *introspection explanation* which explains how a model determines its final output and *justification explanation* which produces sentences detailing how the evidence is compatible with the system output.

Works have been proposed to enhance the interpretability of AI&Law. From the justification aspect, Ye et al. (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. From the introspection aspect, Luo et al. (2017) propose to select supportive law articles and use the articles to enhance the charge prediction accuracy. The supportive law articles is treated as a kind of support for the predicted charge.

In this work, focusing on the introspection explanation of charge prediction, we learn to jointly extract rationales and make charge prediction. The task is not trivial: (1) The granularity of rationales is difficult to grasp – sentence level rationales are not concrete enough while word level rationales lose readability. (2) Corpus with rationale annotation is hard to obtain. (3) Methods of improving the prediction accuracy while having high interpretability are very essential, but have not been well studied. In order to overcome the difficulties above, we propose a hybrid neural framework to (1) extract readable and charge-decisive rationales in the form of key fact snippets from input fact description with the only supervision of charge labels, and (2) elevate charge prediction accuracy by a rationale augmentation mechanism.

* indicates equal contribution.
† Corresponding author.

This work is licenced under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/

Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations, pages 146–151
Santa Fe, New Mexico, USA, August 20-26, 2018.

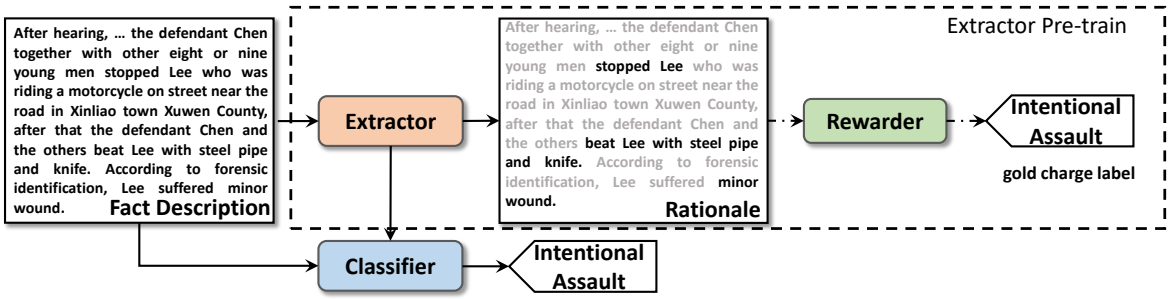

Figure 1: Architecture of Interpretable Rationale Augmented Charge Prediction System

2 Interpretable Rationale Augmented Charge Prediction System

In this section, we will first use mathematical language to define our task and then introduce the proposed Interpretable Rationale Augmented Charge Prediction System. We define the input fact description as word sequence $x = [x_1, x_2, \ldots, x_n]$, and the gold charge label y as a non-negative integer. Given x, we aim to extract rationales $r = \{x_i | z_i = 1, x_i \in x\}$ where $z_i \in \{0, 1\}$, and predict the charge based on x augmented by r. Figure 1 shows the overview of our system. The system takes the fact description in a case as input and outputs the predicted charge as well as the rationales. The rationales play an important role in the predicting process, so they can be seen as an explanation of the charge prediction. The system consists of two main components: **Extractor** and **Classifier**. We train these two components successively.

For the **Extractor** training phase, we apply a deep reinforcement method learning to extract rationales with the only supervision of charge labels. For the **Classifier** training phase, we freeze the parameters of **Extractor**, and the importance of each word is used to make a weighted sum over the RNN hidden states of all words. Then the weighted sum is used to make charge prediction.

2.1 Phrase-level Rationale Extraction

Considering the snippet-like rationales should be more integral in semantics, we propose to represent fact descriptions with phrases (as opposed to words). We split the fact description into phrases with a maximum length of 6. The phrase-level fact x^p is denoted as $[x_1^p, x_2^p, \ldots, x_m^p]$. x_i^p represents the i-th phrase in the fact description. x_i^p's representation is defined as the average word embedding in the phrase.

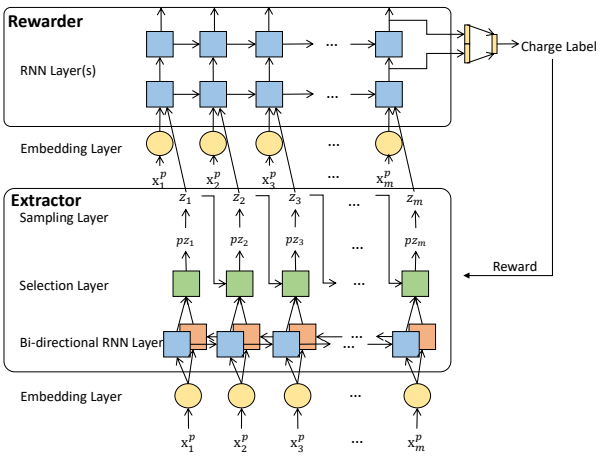

Figure 2: Architecture for Extractor Training

Figure 2 demonstrates the architecture for Extractor training. We introduce a latent variable z ($z \in \{0, 1\}^m$) to define the extraction of phrases. $z_t = 1$ represents the t-th word is chosen as an rationale phrase. The final goal of rationale extraction is to learn a distribution $p(z|x^p)$ over the phrase sequence. At time t, $p(z_t)$ is calculated as follows:

$$p(z_t|x^p, z_{<t}) = sigmoid(W^e[\overrightarrow{h_t}; \overleftarrow{h_t}; z_{t-1}] + b^e)$$
$$\overrightarrow{h_t} = \overrightarrow{f}(x_t^p, \overrightarrow{h_{t-1}}) \quad ; \quad \overleftarrow{h_t} = \overleftarrow{f}(x_t^p, \overleftarrow{h_{t+1}})$$

where $\overrightarrow{f}$ and $\overleftarrow{f}$ are Bi-RNN functions which read the input sequence forward and backward. Here we choose Bi-directional Gated Recurrent Units (Bi-GRU) as the recurrent units. z_t is sampled according to the probability $p(z_t)$. To model the distribution better, at time t, the information from current GRU outputs and the states of $z_{<t}$ are jointly considered to predict the label of x_t^p. The extracted rationales are $r = \{x_i^p | z_i = 1, x_i^p \in x^p\}$. The learning of rationale extraction needs a reward function to guide. Hence, we introduce **Rewarder**, a deep RNN model with 2 layers to model r, generate distribu-

tion over charge labels $\tilde{y}$ and then provide the reward. The final embedding of r is the concatenation of the last states of the two layers. $\tilde{y}$ is calculated as $\tilde{y} = sigmoid(W^r e_r + b^r)$.

To control the quantity of rationales, we introduce a novel penalty over z as $\Phi(z) = |\| z \| - \eta|$ where η is a constant to control $\| z \|$ around η in case of $\| z \|$ being too small or too large. We set η as 7 in this work. We define the loss function as $\mathcal{L}_\theta(r, y) = \| \tilde{y} - y \|_2^2 + \lambda\Phi(z)$. We use the gradient calculation in Lei et al. (2016). Sampling technique (Williams, 1992) is used to approximate the gradient.

2.2 Rationale Augmented Charge Prediction

We move to train **Classifier** utilizing the rationale information generated by **Extractor**. After the previous training, **Extractor** already has the ability to estimate the probabilities of the phrases being rationales. Though the phrase-level representation elevates the rationales' semantic integrality, it causes information loss in the averaging process.

In order to better utilize the information and make charge prediction more accurate, we adopt a RNN model with a rationale augment mechanism. Given the fact description word sequence $x = [x_1, x_2, \ldots, x_n]$, the hidden state at time t in the l-th layer is defined as follow:

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

where f is a unidirectional RNN function. The representation of fact description in layer l derived from the weighted sum of all the hidden states in layer l. Here, $p(z)$ is treated as the importance distribution on input fact description. And the weights a_t are calculated by a softmax layer based on $p(z_t|x)$, which is provided by the pre-trained **Extractor**. More precisely:

$$e_{doc}^{(l)} = \sum_1^n a_t h_t^{(l)}$$

$$a_t = \frac{\exp(p(z_t|x))}{\sum_{t=1}^n \exp(p(z_t|x))}$$

The final representation of a fact description is defined as the concatenation of the representation in each RNN layer: $e_{doc} = [e_{doc}^{(0)}; e_{doc}^{(1)}; \cdots; e_{doc}^{(L-1)}]$. Through an activation layer, e_{doc} generates the final distribution $\tilde{y}$ on the charges: $\tilde{y} = sigmoid(W^c e_{doc} + b^c)$. The loss function is defined as: $\mathcal{L}_\theta(x, y) = \| \tilde{y} - y \|_2^2$.

3 Experiments

3.1 Data Preparation

We construct the dataset from China Judgements Online[1]. 80k, 10k and 10k documents are randomly selected as training, validation and test set respectively. We extract the fact description and charge labels using regular expressions. We set up a length threshold of 256. Fact description longer than that will be stripped. We use HanLP[2] to tokenize the Chinese texts. We use CoreNLP (Manning et al., 2014) 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. We also use <name>, <num> and <date> to replace the names, numbers and dates in the corpus. Following Luo et al. (2017), we choose the same charge set involving 50 most common charges and leave the other charges as negative data. To evaluate the rationale extraction performance, we randomly select 1000 documents and ask three legal professionals to annotate the sentences mentioning illegal behaviors. Sentences chosen by at least two professionals are considered as gold rationale sentences. Kappa (Cohen, 1960) between the annotators is 0.773, proving the high consistency of the annotation.

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

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

MODEL	COMPARISON OF RATIONALE EXTRACTION PERFORMANCE
Bi-GRU$_{att}$	… 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 …
OURS$^-$	… 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 …
OURS	… 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 …
	MORE DEMONSTRATION OF OUR SYSTEM
CASE 1	$[Official\ Embezzlement]_{charge}$
	… PP 利用 其 担任 [公司 业务员 的 职务 便利]$_{key\ point}$， 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]$_{key\ point}$, PP took 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]$_{key\ point}$， 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]$_{key\ point}$, 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]$_{key\ point}$ 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 land, PP [inadvertently]$_{key\ point}$ ignited the mountain fire. 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 放火 烧 PP$_2$ 手机店 的 念头]$_{key\ point}$， 进入 PP$_2$ 的 店内 将 纸箱 点燃…
	…PP$_1$ hates PP$_2$ for business competition. After drinking, PP$_1$ [wanted to burn PP$_2$'s shop]$_{key\ point}$. PP$_1$ entered the shop and lighted the carton…
CASE 5	$[Negligent\ Homicide]_{charge}$
	… PP$_1$ 驾驶 货车 在 倒车 过程 中， [因 疏忽大意]$_{key\ point}$ 将 负责 指挥 倒车 的 PP$_2$ 挤 伤， 后 PP$_2$ 抢救 无效 死亡…
	…When reserving the truck, PP$_1$ [inadvertently]$_{key\ point}$ injured PP$_2$, who was in charge of commanding PP$_1$. PP$_2$ died later. …
CASE 6	$[Intentional\ Homicide]_{charge}$
	…PP$_1$ 从 家中 携带 匕首 出门 寻找 PP$_2$ [进行 报复]$_{key\ point}$， 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]$_{key\ point}$. He stabbed PP$_2$ and cut the neck back and forth, causing PP$_2$ to die on the spot…

Table 1: Examples of extracted rationales. The highlighted words are rationales extracted by models. Different colors are used to align Chinese original text and corresponding English translation. The cores which can directly influence the charges are artificially marked as "key point".

MODEL	CHARGE PREDICTION						RATIONALE EXTRACTION			
	MICRO			MACRO			MACRO			ACC
	P	R	F	P	R	F	P	R	F	
Bi-GRU	89.64	90.60	90.12	81.84	76.25	78.08	–	–	–	–
Bi-GRU$_{att}$	**90.22**	**91.16**	**90.68**	83.97	77.78	79.70	74.6	73.7	68.5	76.3
OURS$^-$	86.25	87.29	86.77	77.08	72.79	73.78	**78.5**	75.7	72.2	79.7
OURS	89.84	91.06	90.45	**84.28**	**77.99**	**80.34**†	70.5	**90.75**	**75.9**‡	**79.8**‡

Table 2: Charge prediction and rationale extraction results. "‡": significantly better than Bi-GRU$_{att}$ ($p<0.01$). "†": better than Bi-GRU$_{att}$ ($p<0.05$).

3.2 Baselines

We choose three types of baselines: Bi-GRU, Bi-GRU$_{att}$ and OURS$^-$. Bi-GRU reads the input sequence forward and backward. The final fact representation used for charge prediction is the average of the hidden states. Bi-GRU$_{att}$ is the base Bi-GRU model with an attention mechanism followed. We adopt similar attention calculation in Yang et al. (2016). OURS$^-$ consists of **Extractor** and the **Rewarder** used for training. That is, only the extracted rationales are used to make charge prediction. Additionally, it discards the concept of phrase. It can be seen as a modified version of Lei et al. (2016): simpler structure in $p(z|x)$ modeling, but almost the same classification performance.

3.3 Experimental Results and Case Study

Rationale Extraction We choose 20 most heavily weighted words in each document as extracted rationale words (almost equal to the rationale word count extracted by OURS). The result in Table 2 proves that our model significantly outperforms the attention model on rationale extraction. Table 1 presents the models' performance on rationale extraction. The first three same sentences are selected from a case with a charge of *negligent homicide* which is suitable for people causing one's death due to negligence. Only our model notices the fact that the shotgun fire was due to the slippery fall, which is a key point distinguishing the case from *intentional homicide*.

In addition, in the lower part of Table 1, we further present the rationale extraction performance of our system on several pairs of example with different but confusing charges. These examples demonstrate that our system can capture key points to distinguish the similar charges. In case 1, our system observes

the fact "his position as a company salesman" which is the key point of distinguishing *Official Embezzlemen* from *Larceny*. For the remaining cases, our system also seizes a series of key details such as "When $PP_2$was not at home", "inadvertently", and "for revenge", and correctly predicts the charges.

Charge Prediction We evaluate charge prediction performance using precision, recall and F1, in both micro and macro level. As shown in Table 2, Bi-GRU proves to be a strong baseline and the effect of attention mechanisms is obvious. Interestingly, though Bi-GRU$_{att}$ ranks first on all micro metrics, our model has better performance on macro metrics. This proves our method's competitive ability on subtle differences capturing, especially when making decision among infrequent but confusing charges. The huge gap between OURS$^-$ and OURS on charge prediction proves that our two-step rationale augmented base strategy fully utilizes the information contained in non-rationale text.

4 Conclusion

We propose a neural based system to jointly extract readable rationales and elevate charge prediction accuracy by a rationale augment mechanism. Sufficient experiments demonstrate that our model outperforms the attention based model on rationale capturing while having comparable classification accuracy.

Acknowledgements

We would like to appreciate the comments from anonymous reviewers and the data annotation from the the legal professionals. This work is supported by National Key Research and Development Program of China (Grant No. 2017YFB1402400) and National Natural Science Foundation of China (No. 61602490).

References

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. 2016. Generating visual explanations. In *ECCV (4)*, volume 9908 of *Lecture Notes in Computer Science*, pages 3–19. Springer.

Tao Lei, Regina Barzilay, and Tommi S. Jaakkola. 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*, pages 107–117.

Wan-Chen Lin, Tsung-Ting Kuo, Tung-Jia Chang, Chueh-An Yen, Chao-Ju Chen, and Shou-de Lin. 2012. Exploiting machine learning models for chinese legal documents labeling, case classification, and sentencing prediction. volume 17.

Chao-Lin Liu and Chwen-Dar Hsieh. 2006. Exploring phrase-based classification of judicial documents for criminal charges in chinese. In *ISMIS*, volume 4203 of *Lecture Notes in Computer Science*, pages 681–690. Springer.

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

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

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *ACL (System Demonstrations)*, pages 55–60. The Association for Computer Linguistics.

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

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J. Smola, and Eduard H. Hovy. 2016. Hierarchical attention networks for document classification. In *HLT-NAACL*, pages 1480–1489. The Association for Computational Linguistics.

Hai Ye, Xin Jiang, Zhunchen Luo, and Wenhan Chao. 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)*, pages 1854–1864.