Interpretable Charge Predictions for Criminal Cases: Learning to Generate Court Views from Fact Descriptions

[1]Hai Ye*, [1]Xin Jiang*, [2]Zhunchen Luo*, [1]Wenhan Chao†
[1] School of Computer Science and Engineering, Beihang University
[2] Information Research Center of Military Science, PLA Academy of Military Science
[1] Beijing 100191, China; [2] Beijing 100142, China
{yehai, xinjiang, chaowenhan}@buaa.edu.cn; zhunchenluo@gmail.com

Abstract

In this paper, we propose to study the problem of COURT VIEW GENeration from the fact description in a criminal case. The task aims to improve the interpretability of charge prediction systems and help automatic legal document generation. We formulate this task as a text-to-text natural language generation (NLG) problem. Sequence-to-sequence model has achieved cutting-edge performances in many NLG tasks. However, due to the non-distinctions of fact descriptions, it is hard for Seq2Seq model to generate charge-discriminative court views. In this work, we explore charge labels to tackle this issue. We propose a label-conditioned Seq2Seq model with attention for this problem, to decode court views conditioned on encoded charge labels. Experimental results show the effectiveness of our method. Dataset and codes of the paper can be obtained from https://github.com/oceanypt/Court-View-Gen.

1 Introduction

Previous work has brought up multiple legal assistant systems with various functions, such as finding relevant cases given the query (Chen et al., 2013), providing applicable law articles for a given case (Liu and Liao, 2005) and etc., which have substantially improved the working efficiency. As legal assistant systems, charge prediction systems aim to determine appropriate charges such as *homicide* and *assault* for varied criminal cases by analyzing textual fact descriptions from cases (Luo et al., 2017), but ignore to give out the interpretations for the charge determination.

Court view is the written explanation from judges to interprete the charge decision for certain criminal case and is also the core part in a legal document, which consists of *rationales* and a *charge* where the charge is supported by the rationales as shown in Fig. 1. In this work, we propose to study the problem of COURT VIEW GENeration from fact descriptions in cases, and we formulate it as a text-to-text natural language generation (NLG) problem (Gatt and Krahmer, 2017). The input is the fact description in a case and the output is the corresponding court view. We only focus on generating rationales because charges can be decided by judges or charge prediction systems by also analyzing the fact descriptions (Luo et al., 2017; Lin et al., 2012). COURT-VIEW-GEN has beneficial functions, in that: (1) improve the interpretability of charge prediction systems by generating rationales in court views to support the predicted charges. The justification for charge decision is as important as deciding the charge itself (Hendricks et al., 2016; Lei et al., 2016). (2) benefit the automatic legal document generation as legal assistant systems, by automatically generating court views from fact descriptions, to release much human labor especially for simple cases but in large amount, where fact descriptions can be obtained from legal professionals or techniques such as information extraction (Cowie and Lehnert, 1996).

COURT-VIEW-GEN is not a trivial task. High-quality rationales in court views should contain the important fact details such as the degree of injury for charge of *intentional injury*, as they are important basis for charge determination. Fact details are like the summary for the fact description similar to the task of DOCument SUMmarization (Yao et al., 2017). However, rationales are not the simple summary with only fact details, to support charges, they should be charge-discriminative with *deduced information* which does not appear in fact descriptions. The fact descriptions for charge of *negligent homicide* usually only describe someone being killed without direct statement about

* indicates equal contribution.
† Corresponding author.

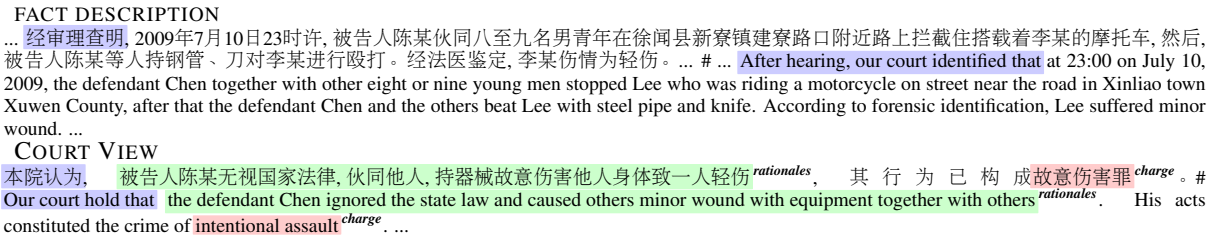

Figure 1: An example of fact description and court view from a legal document in a case.

the motive for killing, DOC-SUM will only summarize the fact of someone being killed, but rationales have to further contain the killing intention, aiming to be discriminative from those rationales for other charges like *intentional homicide*. However, it is hard to generate charge-discriminative rationales when input fact descriptions are not distinct among other facts with different charges. The fact descriptions for charge of *intentional homicide* are similar to those for charge of *negligent homicide* and also describe someone being killed but without clear motive, making it hard to generate charge-discriminative court views with accurate killing motives among the two charges.

Recently, sequence-to-sequence model with encoder-decoder paradigm (Sutskever et al., 2014) has achieved cutting-edge results in many NLG tasks, such as paraphrase (Mallinson et al., 2017), code generation (Ling et al., 2016) and question generation (Du et al., 2017). Seq2Seq model has also exhibited state-of-the-art performances on task of DOC-SUM (Chopra et al., 2016; Tan et al., 2017). However, non-distinctions of fact descriptions render Seq2Seq model hard to generate charge-discriminative rationales. In this paper, we explore charge labels of the corresponding fact descriptions, to benefit generating charge-discriminative rationales, where charge labels can be easily decided by human or charge prediction systems. Charge labels will provide with extra information to classify the non-discriminative fact descriptions. We propose a *label-conditioned* Seq2Seq model with attention for our task, in which fact descriptions are encoded into context vectors by an encoder and a decoder generates rationales with these vectors. We further encode charges as the labels and decode the rationales conditioned on the labels, to entail the decoder to learn to select gold-charge-related words to decode. Widely used attention mechanism (Luong et al., 2015) is fused into the Seq2Seq model, to learn to align target words to fact details in fact descriptions. Similar to Luo et al. (2017), we evaluate our model on Chinese criminal cases by constructing dataset from Chinese government website.

Our contributions in this paper can be summarized as follows:

- We propose the task of *court view generation* and release a real-world dataset for this task.
- We formulate the task as a text-to-text NLG problem. We utilize charge labels to benefit charge-discriminative court views generation, and propose a label-conditioned sequence-to-sequence model with attention for this task.
- Extensive experiments are conducted on a real-world dataset. The results show the efficiency of our model and exploiting charge labels for charge-discriminations improvement.

2 Related Work

Our work is firstly related to previous studies on legal assistant systems. Previous work considers the task of charge prediction as a text classification problem (Luo et al., 2017; Liu et al., 2004; Liu and Hsieh, 2006; Lin et al., 2012). Recently, Luo et al. (2017) investigate deep learning methods for this task. Besides, there are also works on identifying applicable articles for a given case (Liu and Liao, 2005; Liu and Hsieh, 2006; Liu et al., 2015), answering legal questions as a consulting system (Kim et al., 2014; Carvalho et al., 2015) and searching relevant cases for a given query (Raghav et al., 2016; Chen et al., 2013). As a legal assistant system, COURT-VIEW-GEN can benefit automatic legal document generation by generating court views from fact descriptions obtained from the last phase, through legal professionals or other technics like information extraction (Cowie and Lehnert, 1996) from raw documents in a case, if we generate legal documents step by step.

Our work is also related to recent studies on

model interpretation (Ribeiro et al., 2016; Lipton, 2016; Ling et al., 2017). Recently, much work has paid attention to giving textual explanations for classifications. Hendricks et al. (2016) generate visual explanations for image classification. Lei et al. (2016) propose to learn to select most supportive snippets from raw texts for text classification. COURT-VIEW-GEN can improve the interpretability of charge prediction systems by generating textual court views when predict the charges.

Our label-conditioned Seq2Seq model steams from widely used encoder-decoder paradigm (Sutskever et al., 2014) which has been widely used in machine translation (Bahdanau et al., 2014; Luong et al., 2015), summarization (Tan et al., 2017; Nallapati et al., 2016; Chopra et al., 2016; Cheng and Lapata, 2016), semantic parsing (Dong and Lapata, 2016) and paraphrase (Mallinson et al., 2017) or other NLG problems such as product review generation (Dong et al., 2017) and code generation (Yin and Neubig, 2017; Ling et al., 2016). Hendricks et al. (2016) propose to encode image labels for visual-language models to generate justification texts for image classification. We also introduce charge labels into Seq2Seq model to improve the charge-discriminations of generated rationales. Widely used attention mechanism (Luong et al., 2015; Xu et al., 2015) is applied to generate fact details more accurately.

3 COURT-VIEW-GEN Problem

Court View is the judicial explanation to interpret the reasons for the court making such charge for a case, consisting of the rationales and the charge supported by the rationales as shown in Fig. 1. In this work, we only focus on generating the part of rationales in court views. Charge prediction can be achieved by human or charge prediction systems (Luo et al., 2017). Final court views can be easily constructed by combining the generated rationales and the pre-decided charges.

Fact Description is the identified facts in a case (relevant events that have happened) such as the criminal acts (e.g. *degree of injury*).

The input of our model is the word sequential fact description in a case and the output is a word sequential court view (rationales part). We define the fact description as $\mathrm{x} = (x_1, x_2, \cdots, x_{|\mathrm{x}|})$ and the corresponding rationales as $\mathrm{y} = (y_1, y_2, \cdots, y_{|\mathrm{y}|})$. The charge for the case is denoted as v and will be exploited for COURT-VIEW-GEN. The task of COURT-VIEW-GEN is to find $\hat{\mathrm{y}}$ given x conditioned on the charge label v:

$$\hat{\mathrm{y}} = \arg\max_{\mathrm{y}} P(\mathrm{y}|\mathrm{x}, \mathrm{v}) \tag{1}$$

where $P(\mathrm{y}|\mathrm{x}, \mathrm{v})$ is the likelihood of the predicted rationales in the court view.

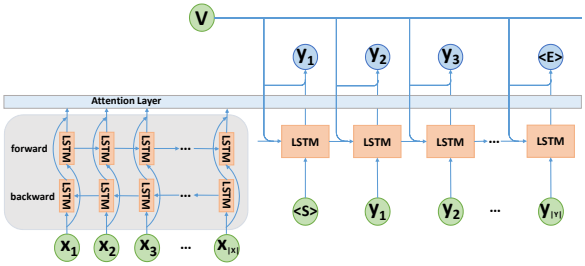

Figure 2: Label-conditioned Seq2Seq model with attention.

4 Our Model

4.1 Sequence-to-Sequence Model with Attention

Similar to Luong et al. (2015), our Seq2Seq model consists of an encoder and a decoder as shown in Fig. 2. Given the pair of fact description and rationales in court view (x, y), the encoder reads the word sequence of x and then the decoder will learn to predict the rationales in court view y. The probability of predicted y is given as follows:

$$P(\mathrm{y}) = \prod_{i=1}^{|\mathrm{y}|} P(y_i|y_{<i}, \mathrm{x}) \tag{2}$$

where $y_{<i} = y_1, y_2, \cdots, y_{i-1}$. We use a bidirectional LSTM (Hochreiter and Schmidhuber, 1997) as encoder and use another LSTM as decoder similar to Du et al. (2017).

Decoder. From the decoder side, at time t, the probability to predict y_t is computed as follows:

$$P(y_t|y_{<t}, \mathbf{c}_t) = \mathrm{softmax}(\mathbf{W}_1 \tanh(\mathbf{W}_0[\mathbf{s}_t; \mathbf{c}_t]))$$

where $\mathbf{W}_0$ and $\mathbf{W}_1$ are learnable parameters; $\mathbf{s}_t$ is the hidden state of decoder at time t; $\mathbf{c}_t$ is the context vector generated from the encoder side containing the information of x at time t; here the bias of model is omitted for simplification. The hidden state of $\mathbf{s}_t$ is computed as follows:

$$\mathbf{s}_t = \mathrm{LSTM}_d(y_{t-1}, \mathbf{s}_{t-1})$$

where y_{t-1} is the word embedding vector for pre-state target word at time $t-1$. The initial state for decoder is initialized by the last state of encoder.

Context vector of $\mathbf{c}_t$ is computed by summing up the hidden states of $\{\mathbf{h}_k\}_{k=1}^{|\mathrm{x}|}$ generated by the encoder with attention mechanism and we adopt global attention (Luong et al., 2015) in our work.

Encoder with Attention. We adopt a one-layer bidirectional LSTM to encoder the fact descriptions. The hidden state $\mathbf{h}_j$ at time j is computed as follows:

$$\mathbf{h}_j = [\overrightarrow{\mathbf{h}_j}; \overleftarrow{\mathbf{h}_j}]$$

where $\mathbf{h}_j$ is the concatenation of forward hidden state $\overrightarrow{\mathbf{h}_j}$ and backward hidden state $\overleftarrow{\mathbf{h}_j}$, specifically:

$$\overrightarrow{\mathbf{h}_j} = \overrightarrow{\mathrm{LSTM}}_e(x_j, \overrightarrow{\mathbf{h}}_{j-1})$$
$$\overleftarrow{\mathbf{h}_j} = \overleftarrow{\mathrm{LSTM}}_e(x_j, \overleftarrow{\mathbf{h}}_{j+1})$$

The hidden outputs $\{\mathbf{h}_k\}_{k=1}^{|\mathrm{x}|}$ will be used to compute the context vectors for decoder.

From the decoder side, by applying attention mechanism at time i, the context vector of $\mathbf{c}_i$ is generated as follows:

$$\mathbf{c}_i = \sum_{j=1}^{|\mathrm{x}|} \alpha_{ij} \mathbf{h}_j \tag{3}$$

where α_{ij} is the attention weight and is computed as follows:

$$\alpha_{ij} = \frac{\exp(\mathbf{s}_i^T \mathbf{W}_2 \mathbf{h}_j)}{\sum_{k=1}^{|\mathrm{x}|} \exp(\mathbf{s}_i^T \mathbf{W}_2 \mathbf{h}_k)} \tag{4}$$

where $\mathbf{s}_i$ is the hidden output state at time i in the decoder side.

4.2 Label-conditioned Sequence-to-Sequence Model with Attention

Given the tuple of fact description, rationales in court view and charge label $(\mathrm{x}, \mathrm{y}, \mathrm{v})$, the probability to predict y is computed as follows:

$$P(\mathrm{y}) = \prod_{i=1}^{|\mathrm{y}|} P(y_i | y_{<i}, \mathrm{x}, \mathrm{v}) \tag{5}$$

From this formula, encoding charge labels provides extra constrains comparing to Eq. (2), and restricts the target word searching space from the whole space to only gold-charge-related space for rationales generation, so model can generate more charge-distinct rationales. Charge labels are trainable parameters denoted by $\mathbf{E}^v$ where every charge will have a trainable vector from $\mathbf{E}^v$, which will be updated in the model training process.

As shown in Fig. 2, in the decoder side, at time t, y_t is predicted with the probability as follows:

$$P(y_t | y_{<t}, \mathbf{c}_t, \mathrm{v}) = \mathrm{softmax}(\mathbf{W}_1 \tanh(\mathbf{W}_0[\mathbf{s}_t; \mathbf{c}_t; \mathbf{E}^v_{[\mathrm{v}]}])) \tag{6}$$

where $\mathbf{E}^v_{[\mathrm{v}]}$ is the embedding vector of v obtained from $\mathbf{E}^v$. In this formula, we connect charge label v to $\mathbf{s}_t$ and $\mathbf{c}_t$ aiming to influence the word selection process. We hope that our model can learn the latent connections between the charge label v and the words of rationales in court views through this way, to decode out charge-discriminative words.

As shown in Fig. 2, we further embed the charge label v to highlight the computing of hidden state $\mathbf{s}_t$ at time t and $\mathbf{s}_t$ is merged as follows:

$$\begin{aligned} \mathbf{s}_t &= \mathrm{LSTM}_d(y_{t-1}, \mathbf{s}^v_{t-1}) \\ \mathbf{s}^v_{t-1} &= f_v(\mathbf{s}_{t-1}, \mathrm{v}) \\ f_v &= \tanh(\mathbf{W}^v[\mathbf{s}_{t-1}; \mathbf{E}^v_{[\mathrm{v}]}] + \mathbf{b}^v) \end{aligned} \tag{7}$$

where $\mathbf{W}^v$ and $\mathbf{b}^v$ are learnable parameters. In this way, the information of charge label can be embedded into $\mathbf{s}_t$. From Eq. (3) and Eq. (4), attention weights $\mathbf{c}_t$ are computed from $\mathbf{s}_t$, so encoding the charge label v to hidden states will make the model concentrate more on charge-related information from fact descriptions to help generate more accurate fact details.

4.3 Model Training and Inference

Suppose we are given the training data: $\{\mathrm{x}^{(i)}, \mathrm{y}^{(i)}, \mathrm{v}^{(i)}\}_{i=1}^N$, we aim to maximize the log-likelihood of generated rationales in court views given the fact descriptions and charge labels, so the loss function is computed as follows:

$$\begin{aligned} \mathcal{L}(\theta) &= -\sum_{i=1}^{N} \log P(\mathrm{y}^{(i)} | \mathrm{x}^{(i)}, \mathrm{v}^{(i)}; \theta) \\ &= -\sum_{i=1}^{N} \sum_{j=1}^{|\mathrm{y}^{(i)}|} \log P(y_j^{(i)} | y_{<j}^{(i)}, \mathrm{x}^{(i)}, \mathrm{v}^{(i)}; \theta) \end{aligned}$$

We split the training data into multiple batches with size of 64 and adopt adam learning (Kingma and Ba, 2014) to update the parameters in every batch data. At the inference time, we encode

# Training set	153706
# Dev set	9152
# Test set	9123
Avg. # tokens in fact desc.	219.9
Avg. # tokens in rationales	30.6
Num. of # charge labels	51
# Dict. size in fact desc.	222482
# Dict. size in rationales	21305

Table 1: Statistics of our dataset.

the fact descriptions and charge labels into vectors and use the decoder to generate rationales in court views based on Eq. (1). We adopt the algorithm of beam search to generate rationales. Beam search size is set to 5. To make generation process stoppable, an indicator tag "</s>" is added to the end of the rationales sequences, and when "</s>" is generated the inference process will be terminated. The generated word sequential paths will be ranked and the one with largest value is selected as the final rationales in court view.

5 Experiments

5.1 Data Preparation

Following Luo et al. (2017), we construct dataset from the published legal documents in China Judgements Online[1]. We extract the fact descriptions, rationales in court views and charge labels using regular expressions. The paragraph started with "经审理查明" ("our court identified that") is regarded as the fact description and the part between "本院认为" ("our court hold that") and the charge are regarded as the rationales. Nearly all the samples in dataset match this extraction pattern. Length threshold of 256 is set up, and fact description longer than that will be stripped, leaving too long facts for future study. We use the tokens of "<name>", "<num>" and "<date>" to replace the names, numbers and dates appearing in the corpus. We tokenize the Chinese texts with the open source tool of HanLP[2]. For charge labels, we select the top 50 charge labels ranked by occurrences and leave the left charges as others. Details about our dataset are shown in Table 1.

For cases with multiple charges and multiple defendants, we can separate the fact descriptions and the court views according to the charges or the defendants. In this work, we only focus on the cases with one defendant and one charge, leaving the complex cases for future study, so we can collect large enough data from the published legal documents without human to annotate the data.

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

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

5.2 Experimental Settings

Word embeddings are randomly initialized and updated in the training process, with the size of 512 tuned from $\{256, 512, 1024\}$. Charge label vectors are initialized randomly with size of 512. Maximal vocabulary size of encoder is set to 100K words and decoder is 50K by stripping words exceeding the bounds. Maximal source length is 256 and target is 50. The hidden size of LSTM is 1024 tuned from $\{256, 512, 1024\}$. We choose perplexity as the update metric. Early stopping mechanism is applied to train the model. The initial learning rate is set to 0.0003 and the reduce factor is 0.5. Model performance will be checked on the validation set after every 1000 batches training and keep the parameters with lowest perplexity. Training process will be terminated if model performance is not improved for successive 8 times.

5.3 Comparisons with Baselines

Evaluation Metrics. We adopt both automatic evaluation and human judgement for model evaluation. BLEU-4 score (Papineni et al., 2002) and variant Rouge scores (Lin, 2004) are adopted for automatic evaluation which have been widely used in many NLG tasks. We set up two evaluation dimensions for human judgement: 1) how *fluent* of the rationales in court view is; 2) how *accurate* of the rationales is, aiming to evaluate how many fact details have been accurately expressed in the generated rationales. We adopt 5 scales for both *fluent* and *accurate* evaluation (5 is for the best). We ask three annotators who knows well about our task to conduct the human judgement. We randomly select 100 generated rationales in court views for every evaluated method. The three raters are also asked to judge whether rationales can be adopted for use in comprehensive evaluation (*adoptable*) and record the number of adoptable rationales for every evaluated method.

Baselines.

• **Rand** is to randomly select rationales in court views from the training set (method of $\textbf{Rand}_{\textbf{all}}$). We also randomly choose rationales from pools with same charge labels ($\textbf{Rand}_{\textbf{charge}}$). Adopting Rand method is to indicate the low bound performance of COURT-VIEW-GEN.

• **BM25** is a retrieval baseline to index the fact description match to the input fact description

	AUTOMATIC EVALUATION			
MODEL (%)	B-4	R-1	R-2	R-L
Rand_{all}	6.4	26.5	6.2	25.1
$\text{Rand}_{\text{charge}}$	24.9	53.6	29.1	49.3
BM25_{f2f}	40.1	63.5	43.7	60.3
$\text{BM25}_{\text{f2f+charge}}$	42.8	67.1	47.4	63.8
MOSES+	6.2	39.8	20.8	18.6
NN-S2S	38.4	65.5	45.1	62.2
RAS†	44.1**	69.1**	50.3**	65.9**
Ours	**45.8**	**70.9**	**52.5**	**67.7**

	HUMAN JUDGEMENT		
MODEL	FLUENT	ACC.	ADOPT.(%)
BM25_{f2f}	4.95	3.66**	0.47**
$\text{BM25}_{\text{f2f+charge}}$	4.94	3.90**	0.50**
MOSES+	1.39**	1.31**	0**
NN-S2S	4.97	4.07**	0.62*
RAS†	4.96	4.25*	0.64*
Ours	4.93	**4.54**	**0.72**

Table 2: Results of automatic evaluation and human judgement with BLEU-4 and full length of F1 scores of variant Rouges. Best results are labeled as boldface. Statistical significance is indicated with $**(p < 0.01)$ and $*$ $(p < 0.05)$ comparing to our full model.

with highest BM25 score (Robertson and Walker, 1994) from the training set, and use its rationales as the result (**BM25_{f2f}**). Similar fact descriptions may have the similar rationales. Fact descriptions from pools with same charges are also retrieved (**$\text{BM25}_{\text{f2f+charge}}$**), to see how much improvement that adding charge labels can gender.

• **MOSES+** (Koehn et al., 2007) is a phrase-based statistical machine translation system mapping fact descriptions to rationales. KenLM (Heafield et al., 2013) is adopted to train a trigram language model on the target corpus of training set which is tuned on the validation set with MERT.

• **NN-S2S** is the basic Seq2Seq model without attention from Sutskever et al. (2014) for machine translation. We set one LSTM layer for encoding and another one LSTM layer for decoding. We adopt perplexity for training metric and select the model with lowest perplexity on validation set.

• **RAS**† is an attention based abstract summarization model from Chopra et al. (2016). To deal with the much longer fact descriptions, we exploit the more advanced bidirectional LSTM model for the encoder instead of the simple convolutional model. Another LSTM model is set as the decoder coherent to Chopra et al. (2016).

Experimental Results. In automatic evaluation from Table 2, the evaluation scores are relatively high even for method of $\text{Rand}_{\text{charge}}$, which indicates that the expressions of the rationales with same charge labels are similar with many overlapped n-grams, such that the rationales for *crime of theft* usually begin with "以非法占有为目的" ("in intention of illegal possession"). Accurately generating fact details like degree of injury or time of theft is more difficult. Retrieval method by adding charge labels is the strong baseline even better than basic Seq2Seq model. Adding attention mechanism will improve the performance indicated by the method of RAS† which is superior to retrieval methods. By exploiting charge labels, our full model achieves the best performance. The performances of statistical machine translation model are really poor, for it requiring the lengths of parallel corpus to be similar.

In human evaluation, we can see that retrieval methods can not accurately express fact details, for that it is hard to retrieve rationales containing details all matching the fact descriptions. However, our system can learn to generate fact details by analyzing fact descriptions. Dropping attention mechanism will have negative effects on model performance. RAS† has worse performance in ACC. whose main reason may lie in that RAS† can not generate charge-discriminative rationales with *deduced information*, which demonstrates that our task is not the simple DOC-SUM task. For the *fluent* evaluation, generation models are highly close to retrieval methods whose rationales are written by humans, which reflects that the generation models can generate highly natural rationales.

5.4 Further Analysis

Impact of Exploiting Charge Labels.

• **Charge2Charge Analysis.** We first analyze the effects of exploiting charge labels on model performance charge to charge, by dropping to encode charges based on our full model. From the results shown in Fig. 3, we can find that the results can be improved much by exploiting charge labels among nearly all charges. This result also indicates that the non-distinct fact descriptions are common among nearly all charges and reflects the difficulty of this task, but utilizing charge labels can release the seriousness of the problem.

• **Charge-discriminations Analysis.** We further evaluate the effects of charge labels for charge-discriminations improvement on specific charges with non-distinct fact descriptions: *intentional homicide*, *negligent homicide*, *duty embezzlement* and *corruption*. For every charge, two

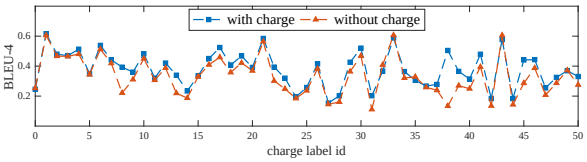

Figure 3: Results of impact of exploiting charge labels evaluated charge to charge in the metric of BLEU-4 (similar results can gender in other three metrics but are omitted for space saving).

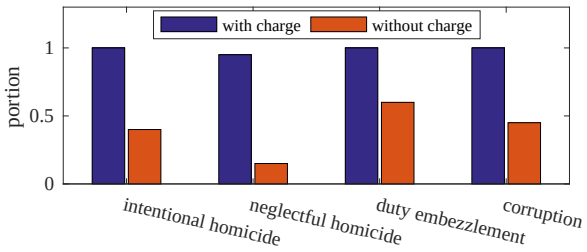

Figure 4: Portions of charge-discriminative rationales in court views for every charge with 20 candidates.

participants are asked to count the number of rationales that are relevant to the charge on 20 randomly selected candidates.

From Fig. 4, the number of charge discriminative rationales can be much improved among every charge by utilizing charge information, which demonstrates that charge labels can provide with much extra charge-related information to deal with latent information in fact descriptions. For crimes of *homicide*, the motives for killing are latent in the descriptions of killing without direct statement, but our system can learn to align the motives in rationales to the charge labels which are the strong distinct indicator for the two motives.

Ablation Study. We also ablate our full model to reveal different components of encoding charge labels for performance improvement. As shown in Table 3, " / softmax comp." is to remove the part in Eq. (6) and yields worse performance than our full model, but better than " / charge comp." that ignores to encode charge labels, which is same to the situation of " / hidden comp." that removes the part in Eq. (7). Our full model is still better than the ablated models. This finding shows that both of the methods of exploiting charge labels can improve model performance and stacking them will achieve better results.

Attention Mechanism Analysis. Heat map in Fig. 5 is used to illustrate the attention mechanism. The "slight injury" is aligned between the source and target. "responsibility" and "run" are well aligned to "away", which demonstrate the

	Ablation Study			
Model (%)	B-4	R-1	R-2	R-L
Our System	45.8**	70.9**	52.5**	67.7**
/ softmax comp.	45.7**	70.8**	52.3**	67.5**
/ hidden comp.	45.7**	70.2*	51.9*	67.0*
/ charge comp.	43.7	68.6	49.7	65.5

Table 3: Results of ablation study. Statistical significance is indicated with $**$ ($p < 0.01$) and $*$ ($p < 0.05$) comparing to the ablation of " / charge comp.".

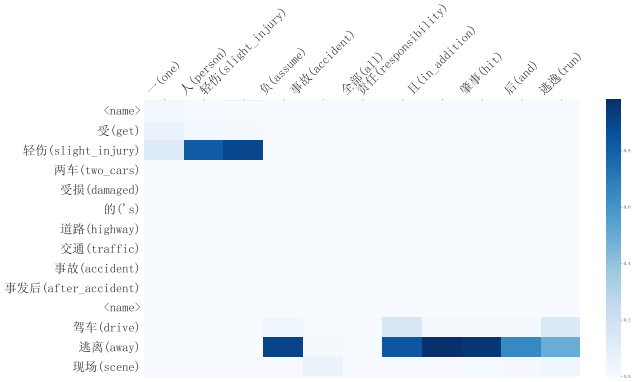

Figure 5: Heat map for attention mechanism analysis. The column is the source and the raw is the target.

efficiency of attention mechanism for generating fact details by forcing context vectors to focus more on fact details.

Performance by Reference Size. We further investigate the model performance by rationales length in court views. As shown in Fig. 6, not surprisingly the model performance drops when the length of reference rationales increases. Within the size of 30, BLEU-4 score can maintain around 0.4 and F1 score keeps around 0.5. Exceeding the length of 30, model performance decreases dramatically.

Human eval. vs. Automatic eval. *Are BLEU and Rouge suitable for* Court-View-Gen *evaluation*? Following the work of (Papineni et al., 2002; Liu et al., 2016), for the models evaluated in human judgemnet, we draw the linear regressions of their BLEU-4 and variant Rouge scores, as the function of Acc. and Adopt. from human judgement respectively as shown in Fig. 7. From

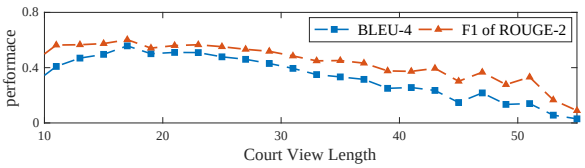

Figure 6: Model performance by rationales length with BLEU-4 and full length of F1 of Rouge-2.

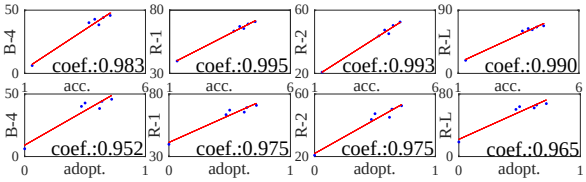

Figure 7: ACC. and ADOPT. of human judgement predict automatic evaluation scores.

the results, we can find that automatic evaluations track well with the human judgement with high correlation coefficients. This finding demonstrates that BLEU-4 and variant Rouges are adoptable for COURT-VIEW-GEN evaluation and provides the basis for future studies on this task.

Error Analysis. Our model has the drawback of *generating latent fact details*, which appear in rationales but are not clearly expressed in fact descriptions. For example, for the time of theft in charge of *larceny*, the term of "多次" ("several times") appears in rationales but may not be expressed in fact descriptions directly, only with descriptions of larceny but without exact term for this detail, so it will be hard for attention mechanism to learn to align "多次" in rationales to latent information in fact descriptions. In the generated rationales on test set, we find that only 42.4% samples can accurately extract out the term of "多次". It may need designed rules to deal with such details, like that count the time of theft from the descriptions, and if the time exceeds 1 then the term of "多次" can be generated in rationales.

5.5 Analysis through Cases

Fake Charge Label Conditioned Study. What generated rationales in court views will be if they are conditioned on fake charge labels? We select one fact description with gold charge of *intentional injury*, then generate rationales conditioned on fake charges of *defiance and affray crime*, *intentional homicide* and *neglectful homicide*.

From Fig. 8, the rationales conditioned on fake charges will be partly relevant to fake charge labels and also maintain fact details from the input fact description of gold charge. For the fake charge of *intentional homicide*, its fact details should be "caused someone dead", but instead express "causing someone slight injury" which is relevant to charge of *intentional injury*. For charge prediction systems, the discriminations between fact details and charges will help to remind people that the prediction results may be unreliable.

Case Study. Examples of generated rationales in court views are shown in Fig. 8. Generally speaking, our full label-conditioned model has high accuracy on generating fact details better than baseline models. For charges of *traffic accident crime* and *negligent homicide*, all fact details are generated. The extra information from charge labels helps the model to capture more important fact details, by forcing model to pay more attention to charge-related information in fact descriptions.

As for the charge-discrimination analysis, from the rationales of *negligent homicide*, we can infer that its fact description may relate to a traffic accident, which is non-distinct from that for *traffic accident crime*. Without encoding charge labels, Ours $_{/c}$ wrongly generates the rationales coherent to *traffic accident crime*, because traffic accidents are the strong indicator for traffic crimes, but the charge label will provide extra bias towards the *homicide crime*, so our full model can generate highly discriminative rationales. Utilizing charge labels, retrieval method can easily retrieve charge-related rationales, but hard to index rationales with accurate fact details. For charge of *larceny*, our full model extracts nearly all fact details but misses the fact of "多次"("several times"), reflecting the shortcoming of dealing with latent details.

6 Conclusion and Future Work

In this paper, we propose a novel task of court view generation and formulate it as a text-to-text NLG problem. We utilize charge labels to benefit the generation of charge-discriminative rationales in court views and propose a label-conditioned Seq2Seq model with attention for this task. Extensive experiments show the efficiency of our model and exploiting charge labels.

In the future: 1) More advanced technologies like reinforcement learning (Sutton and Barto, 1998) can be introduced to generate latent fact details such as the time of theft more accurately; 2) In this work, we only generate rationales in court views omitting charge prediction, it is interesting to see whether jointly generating the two parts will benefit both of the tasks; 3) Studying verification mechanism is meaningful to judge whether generated court views can really be adopted which is important for COURT-VIEW-GEN in practice; 4) More complex cases with multiple charges and multiple defendants will be considered in the future.

Model	[charge]	Generated Court Views Conditioned on Fake Charge Label
Gold	/故意伤害罪/ PP 故意 伤害 他人 身体 ， 致 一 人 轻伤 。# [intentional injury] PP intentionally injured others body , caused one people slight injury .	
Ours	/寻衅滋事罪/ PP 随意 殴打 他人 ， 致 一 人 轻伤 ， 情节 恶劣 。# [defiance and affray crime] PP beat others at will , caused one people slight injury . /故意杀人罪/ PP 故意 非法 剥夺 他人 生命 ， 致 一 人 轻伤 。# [intentional homicide] PP intentionally illegally deprived someone of life , caused one people slight injury . /过失致人死亡罪/ PP 过失 致 一 人 轻伤 。# [neglectful homicide] PP neglectfully caused one people slight injury .	

Model	[charge]	Generated Court Views
Gold	/交通肇事罪/ PP 违反 交通 运输 管理 法规 ， 造成 一 人 死亡 ， 二 人 受伤 的 交通 事故 ， 负 事故 的 全部 责任 。 # [traffic accident crime] PP violated traffic transportation management regulations , caused one people dead , two people injured , take accident's full responsibility . /过失致人死亡罪/ PP 在 驾驶 机动车 过程 中 ， 疏忽大意 ， 致使 他人 被 碾压 致死 。# [negligent homicide] PP when driving car , being neglectful , caused people dead by rolling . /盗窃罪/ PP 以 非法 占有 为 目的 ， 伙同 他人 多次 秘密 窃取 公民 财物 ， 数额 较 大 。# [larceny] PP in intention of illegal possession , ganged up with others and stole goods secretly in relatively large amount for several times .	
Ours	PP 违反 交通 运输 管理 法规 ， 发生 交通 事故 ， 致 一 人 死亡 ， 二 人 受伤 ， 负 事故 的 全部 责任 。 # PP violated traffic transportation management regulations , caused traffic accident , caused one people dead , two people injured , take accident's full responsibility . ✓ PP 因 疏忽大意 致 一 人 死亡 。 # PP neglectfully caused one people dead . ✓ PP 以 非法 占有 为 目的 ， 结伙 他人 秘密 窃取 他人 财物 ， 数额 较 大 。# PP in intention of illegal possession , ganged up with others and stole goods secretly in relatively large amount . ✗	
Ours / c	PP 违反 交通 运输 管理 法规 ， 发生 重大 交通 事故 ， 致 一 人 死亡 ， 负 事故 的 全部 责任 。 # PP violated traffic transportation management regulations , caused severe traffic accident , caused one people dead , took accident's full responsibility ✗ PP 违反 交通 运输 管理 法规 ， 发生 重大 交通 事故 ， 致 一 人 死亡 ， 负 事故 的 全部 责任 。# PP violated traffic transportation management regulations , caused severe traffic accident , caused one people dead , took accident's full responsibility . ✗ PP 以 非法 占有 为 目的 ， 秘密 窃取 他人 财物 ， 数额 较 大 。# PP in intention of illegal possession , stole goods secretly in relatively large amount . ✗	
BM25$_{f2f+c}$	PP 违反 道路 交通 运输 管理 法规 ， 致 一 人 死亡 且 负 事故 主要 责任 。# PP violated road traffic transportation management regulations , caused one people dead , took accident's main responsibility . ✗ PP 驾驶 车辆 过程 中 疏忽大意 ， 过失 致 一 人 死亡 。# PP when driving , neglectfully caused one people dead . ✓ PP 以 非法 占有 为 目的 ， 秘密 窃取 公民 财物 。# PP in intention of possession , stole goods secretly . ✗	

Figure 8: Fake charge label conditioned generated rationales in court views and examples of generated rationales.

Acknowledgments

Firstly, we would like to thank Yansong Feng, Yu Wu, Xiaojun Wan, Li Dong and Pengcheng Yin for their insightful comments and suggestions. We also very appreciate the comments from anonymous reviewers which will help further improve our work. This work is supported by National Natural Science Foundation of China (No. 61602490) and National Key R&D Plan (No. 2017YFB1402403). The work was done when Hai Ye interned in Beihang University from August, 2017 to January, 2018.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR* abs/1409.0473.

Danilo S. Carvalho, Minh-Tien Nguyen, Chien-Xuan Tran, and Minh-Le Nguyen. 2015. Lexical-morphological modeling for legal text analysis. In *New Frontiers in Artificial Intelligence - JSAI-isAI 2015 Workshops, LENLS, JURISIN, AAA, HAT-MASH, TSDAA, ASD-HR, and SKL, Kanagawa, Japan, November 16-18, 2015, Revised Selected Papers*. pages 295–311.

Yen-Liang Chen, Yi-Hung Liu, and Wu-Liang Ho. 2013. A text mining approach to assist the general public in the retrieval of legal documents. *JASIST* 64(2):280–290.

Jianpeng Cheng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers*.

Sumit Chopra, Michael Auli, and Alexander M. Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pages 93–98.

Jim Cowie and Wendy Lehnert. 1996. Information extraction. *Communications of the ACM* 39(1):80–91.

Li Dong, Shaohan Huang, Furu Wei, Mirella Lapata, Ming Zhou, and Ke Xu. 2017. Learning to generate product reviews from attributes. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Association for Computational Linguistics, pages 623–632.

Li Dong and Mirella Lapata. 2016. Language to logical form with neural attention. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 33–43.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers*. pages 1342–1352.

Albert Gatt and Emiel Krahmer. 2017. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *CoRR* abs/1703.09902.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified

kneser-ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9, Volume 2: Short Papers*. pages 690–696.

Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. 2016. Generating visual explanations. In *Computer Vision - ECCV 2016 - 14th European Conference, Proceedings, Part IV*. pages 3–19.

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

Mi-Young Kim, Ying Xu, and Randy Goebel. 2014. Legal question answering using ranking SVM and syntactic/semantic similarity. In *New Frontiers in Artificial Intelligence - JSAI-isAI 2014 Workshops, LENLS, JURISIN, and GABA, Kanagawa, Japan, October 27-28, 2014, Revised Selected Papers*. pages 244–258.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR* abs/1412.6980.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL 2007, Proceedings of the 45th Annual Meeting of the Association*.

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*. pages 107–117.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the ACL-04 Workshop. Association for Computational Linguistics*. pages 74–81.

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. *IJCLCLP* 17(4).

Wang Ling, Phil Blunsom, Edward Grefenstette, Karl Moritz Hermann, Tomás Kociský, Fumin Wang, and Andrew Senior. 2016. Latent predictor networks for code generation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers*.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 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, Volume 1: Long Papers*. pages 158–167.

Zachary Chase Lipton. 2016. The mythos of model interpretability. *CoRR* abs/1606.03490.

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.

Chao-Lin Liu and Chwen-Dar Hsieh. 2006. Exploring phrase-based classification of judicial documents for criminal charges in chinese. In *Foundations of Intelligent Systems, 16th International Symposium, ISMIS 2006, Bari, Italy, September 27-29, 2006, Proceedings*. pages 681–690.

Chao-Lin Liu and Ting-Ming Liao. 2005. Classifying criminal charges in chinese for web-based legal services. In *Web Technologies Research and Development - APWeb 2005, 7th Asia-Pacific Web Conference Proceedings*. pages 64–75.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016*. pages 2122–2132.

Yi-Hung Liu, Yen-Liang Chen, and Wu-Liang Ho. 2015. Predicting associated statutes for legal problems. *Inf. Process. Manage.* 51(1):194–211.

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*. pages 2717–2726.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. pages 1412–1421.

Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2017. Paraphrasing revisited with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Volume 1: Long Papers*. pages 881–893.

Ramesh Nallapati, Bowen Zhou, Cícero Nogueira dos Santos, Çaglar Gülçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, August 11-12, 2016*. pages 280–290.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. pages 311–318.

K. Raghav, P. K. Reddy, and V. B. Reddy. 2016. Analyzing the extraction of relevant legal judgments using paragraph-level and citation information. In *AI4JArtificial Intelligence for Justice*.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pages 1135–1144.

Stephen E. Robertson and Steve Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. (Special Issue of the SIGIR Forum)*. pages 232–241.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014*. pages 3104–3112.

Richard S Sutton and Andrew G Barto. 1998. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge.

Jiwei Tan, Xiaojun Wan, and Jianguo Xiao. 2017. Abstractive document summarization with a graph-based attentional neural model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers*. pages 1171–1181.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning, ICML*. pages 2048–2057.

Jin-ge Yao, Xiaojun Wan, and Jianguo Xiao. 2017. Recent advances in document summarization. *Knowl. Inf. Syst.* 53(2):297–336.

Pengcheng Yin and Graham Neubig. 2017. A syntactic neural model for general-purpose code generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers*. pages 440–450.