Legal Summarization for Multi-role Debate Dialogue via Controversy Focus Mining and Multi-task Learning

Xinyu Duan*[2], Yating Zhang*[1], Lin Yuan[2], Xin Zhou[1], Xiaozhong Liu[3]
Tianyi Wang[1], Ruocheng Wang[2], Qiong Zhang[1], Changlong Sun[1], Fei Wu[2]
Alibaba Group[1],Hangzhou, Zhejiang, China
Zhejiang University[2], Hangzhou, Zhejiang, China
Indiana University Bloomington[3], Bloomington, Indiana, USA
duanxinyu@zju.edu.cn, ranran.zyt@alibaba-inc.com

ABSTRACT

Multi-role court debate is a critical component in a civil trial where parties from different camps (plaintiff, defendant, witness, judge, etc.) actively involved. Unlike other types of dialogue, court debate can be lengthy, and important information, with respect to the controversy focus(es), often hides within the redundant and colloquial dialogue data. Summarizing court debate can be a novel but significant task to assist judge to effectively make the legal decision for the target trial. In this work, we propose an innovative end-to-end model to address this problem. Unlike prior summarization efforts, the proposed model projects the multi-role debate into the controversy focus space, which enables high-quality essential utterance(s) extraction in terms of legal knowledge and judicial factors. An extensive set of experiments with a large civil trial dataset shows that the proposed model can provide more accurate and readable summarization against several alternatives in the multi-role court debate scene.

CCS CONCEPTS

• **Information systems**; • **Computing methodologies** → **Kn owledge representation and reasoning**; **Multi-task learning**; • **Applied computing** → **Law**;

KEYWORDS

legal summarization, multi-role debate dialogue, controversy focus, multi-task learning

ACM Reference Format:

Xinyu Duan, Yating Zhang, Lin Yuan, Xin Zhou, Xiaozhong Liu , Tianyi Wang, Ruocheng Wang, Qiong Zhang, Changlong Sun, Fei Wu. 2019. Legal Summarization for Multi-role Debate Dialogue via Controversy Focus Mining and Multi-task Learning. In *Proceedings of The 28th ACM International Conference on Information and Knowledge Management, (CIKM '19), November 3–7, 2019, Beijing, China* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3357384.3357940

* Both authors contributed equally to this research.

CIKM '19, November 3–7, 2019, Beijing, China

ACM ISBN 978-1-4503-6976-3/19/11…$15.00
https://doi.org/10.1145/3357384.3357940

1 INTRODUCTION

"Trial judges are suffering from 'daunting workload"'[1] is becoming an increasingly critical issue, which challenges the efficiency of legal justice ecosystem in different nations. According to the report of statistics, the typical active federal district court judge closed around 250 cases in a year, and as workloads rise in Federal courts, judge counts however remain flat [4, 5]. Therefore, applying novel artificial legal intelligence techniques to facilitate the lawsuit process so as to alleviate the overwhelmed workload of judges is of great significance [24].

Multi-role court debate is a special dialogue scene that commonly occurs in civil trials where parties from different camps (plaintiff, defendant, witness etc.) debate with each other based on the essential controversy focuses summarized by the presiding judge according to the plaintiff's complaint[2] and the defendant's answer[3]. After the trial, the judge needs to investigate the debate recording to locate the important information and to find clues to respond to each controversy focus until a verdict is accomplished[4]. However, a civil trial can last a few hours, even days, depending on the issues being litigated[5], and statistically, an 1-2 hour's trial can host 10-30 thousand words, which is the key reason that most courts are overwhelmed and under-resourced. In such context, enabling auto-legal summarization for trial debate can be a novel but critical task.

Document summarization has been well-studied [6–9, 19, 28], among which abstractive approach [6, 8, 28] is gaining increasing popularity recently because of its capability to generate new content. However, such approach can be inappropriate for trial debate scenario for two reasons. First, the judge needs to collect original evidences/facts from debate [14] and the employed method should be not "creative". Second, the summarized debate synopsis, as training decoding data, can be unavailable in most cases.

Prior extractive algorithms, however, should be not directly applied to address this problem. Unlike most existing summarization tasks, legal knowledge can play an important role for trial debate investigation. In another word, debate text data should not be used per se for summarization, and the case associated controversy focuses and legal knowledge need be employed for utterance representation learning. Fig. 1 depicts an example, where the important utterances

[1] Reported by the New York Times http://tiny.cc/tbo95y
[2] The first document filed with the court by a plaintiff claiming legal rights against another. https://dictionary.law.com/Default.aspx?selected=261
[3] A written pleading filed by a defendant to respond to a complaint in a lawsuit. https://dictionary.law.com/Default.aspx?selected=2407
[4] "The Civil Litigation Process" http://www.torhoermanlaw.com/civil-lawsuit-process
[5] "Why Does A Lawsuit Take So Long?" https://millerlawpc.com/lawsuit-take-long/

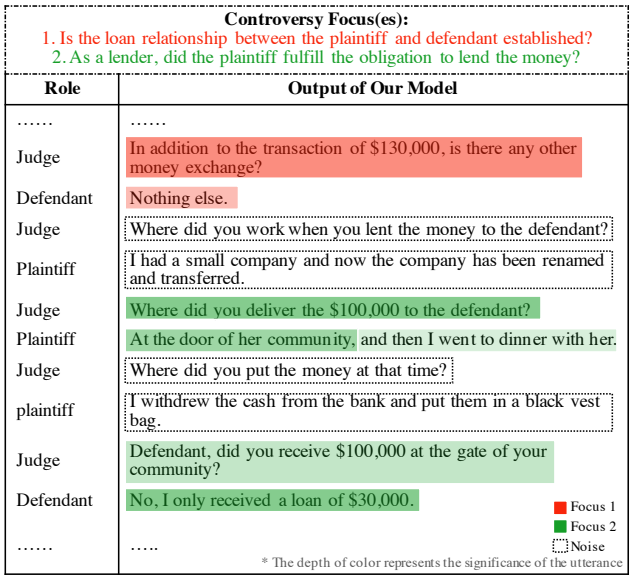

Figure 1: The visualization of a toy example of the system output

center around the case associated controversy focuses. Unfortunately, the utterances sharing the same controversy focus do not necessarily share the similar text content. In addition, each court debate associates uncertain number of parties (e.g., it may contain multiple plaintiffs, defendants and witnesses) and undetermined controversy focuses (e.g., the number and content of the controversy focuses may vary for different cases), which makes it even harder to automatically learn/represent the assignment between debates and the controversy focuses. Moreover, since the controversy focuses are drafted by the case judge, different debate dialogues may have different controversy focuses, and each controversy focus is concluded by the presiding judge in natural language. As a result, the assignment of the utterance to the controversy focuses cannot be solved as a classical classification task.

Motivated by such observation, in this paper, we propose a neural controversy focus centered extractive debate summarization model which jointly learns the assignment of the utterances in the court debate to the corresponding controversy focuses and simultaneously locates the essential utterances for each controversy focus for summarization. Fig. 2 depicts the systematic structure of the proposed model. In a joint learning process, various kinds of information, e.g., legal knowledge graph, role of the litigant, and semantic information of the debates, are encapsulated to regularize the summarization of a civil trial through the fine-tuning of the court debate representation.

To the best of our knowledge, this work is the first pioneer investigation of summarization for the multi-role court debate scene, which can be important and essential to make the judges' work more efficient. To sum up, our contributions are as follows:

(1) We propose an innovative and effective method to quantify a court debate by leveraging multi-view utterance representation.
(2) We propose an end-to-end model in a manner of multi-task learning process for a novel legal intelligence problem to address multi-role and multi-focus court debate summarization.

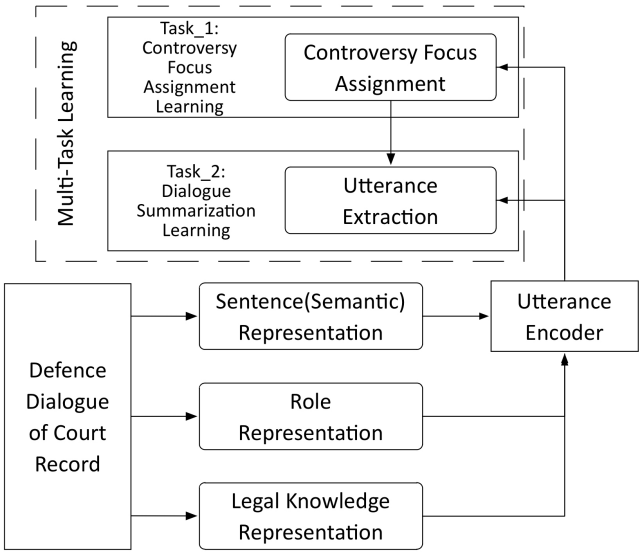

Figure 2: The structure of our proposed model: Controversy Focus-based Debate Summarizatio (CFDS)

(3) The proposed model is capable of digging the legal concepts behind the debate to make the alignment between the controversy focuses and the debate itself by leveraging the legal knowledge graph.
(4) We evaluate the proposed model via 5,477 court records of civil trials trained with more than 200 thousand debates of court recordings. The experimental results demonstrate the proposed approach significantly improves the performance for both tasks.
(5) To motivate other scholars to investigate this novel but important problem, we make the experiment dataset publicly available (while removing the sensitive information). To the best of our knowledge, this is the first civil trial court debate dataset.

2 PROBLEM FORMULATION

Legal summarization for multi-role debate aims to extract informative utterances for each controversy focus to represent the critical information in a debate dialogue. Given a debate dialogue $D = (u_1, u_2, ..., u_L)$, containing L utterances where each utterance u_i is composed of a sequence of l words $s_i = (w_{i1}, w_{i2}, ..., w_{il})$ and the role of its speaker r_i, along with a set of M controversy focuses $F = (f_1, f_2, ...f_M)$ w.r.t the debate dialogue D, an extractive summary of such debate should commit two missions: first, assigning the utterances to either one of the controversy focuses (or the category of *Noise*)[6]; second, extracting the important utterances in terms of each controversy focuses and generating multiple summarization.

To be clarified, the definition of important notations in the following sections are illustrated as follows:

- D: a debate dialogue containing L utterances;
- u_i: the i_{th} utterance in D;
- s_i: the text content of u_i;

[6]Statistically, 71.2% of utterances in the experiment data are regarded as noises, i.e., independent to the case controversy focuses. More details will be available in the dataset section.

- r_i: the role of the speaker in u_i (i.e. presiding judge, plaintiff, defendant and witness);
- F: a set of controversy focuses w.r.t the debate dialogue D
- f_m: a controversy focus in F;
- G: the Legal Knowledge Graph (LKG) containing Q nodes;
- v_q: a node in G and its semantics is represented by a sequence of words;
- p_{iz}: a path containing n nodes on G w.r.t utterance u_i;
- k_i: the legal knowledge addressed in utterance u_i, which aggregates all its candidate path $p_{iz}, z \in [1, Z]$ where Z is a pre-defined number of candidate paths for each utterance.

Note that $\boldsymbol{u_i}$, $\boldsymbol{s_i}$, $\boldsymbol{r_i}$, $\boldsymbol{f_m}$, $\boldsymbol{v_q}$, $\boldsymbol{p_{iz}}$, $\boldsymbol{k_i}$ represent the embedding representations of the corresponding variables introduced above.

3 MODEL

In this section, we introduce the proposed summarization framework, Controversy Focus-based Debate Summarization (CFDS). As Fig. 3 depicts, CFDS is a multi-task model with multi-view utterance encoder. CFDS has three major components: (1) the sentence (semantic) representation learning, (2) the role representation learning and (3) the legal knowledge augmentation mechanism (to locate the relevant paths for each debate utterance on the legal knowledge graph), as well as two decoders for specific tasks, i.e., controversy focus assignment and utterance extraction. Unlike prior efforts in dialogue mining and representation learning, each debate utterance needs to be projected to the paths on the legal knowledge graph (legal knowledge augmentation), and the multi-view utterance encoder characterizing three kinds of debate legal information - sentence (semantic) embedding, role embedding and legal knowledge embedding. Meanwhile, the decoder for controversy focus assignment interprets the candidate controversy focus for the target utterance, which can be critical to tell the utterance importance for summarization (utterance extraction). CFDS is a highly specialized model designed for legal summarization and representation learning.

3.1 Multi-view Utterance Encoder

Characterizing utterance information in a debate context can be essential for summarization. Unlike existing studies in dialogue representation learning, it is necessary to estimate utterance semantic, role and legal information in a debate context. Detailed method can be found in this section.

3.1.1 Sentence (Semantic) Representation. At the sentence (semantic) level, we employ a bidirectional LSTM[7], namely *Bi-LSTM*^{S}, to encode the semantic meaning of the sentence while maintaining its syntactics:

$$\begin{aligned} \overrightarrow{h_{it}^s} &= \overrightarrow{\text{Bi-LSTM}^S}(w_{it}),\ t \in [1, l] \\ \overleftarrow{h_{it}^s} &= \overleftarrow{\text{Bi-LSTM}^S}(w_{it}),\ t \in [l, 1] \\ h_{it}^s &= [\overrightarrow{h_{it}^s}, \overleftarrow{h_{it}^s}] \end{aligned} \tag{1}$$

where l is the number of words in s_i. h_{it}^p is the concatenation of the forward hidden state $\overrightarrow{h_{it}^p}$ and the backward one $\overleftarrow{h_{it}^p}$. In this way, h_{it}^p summarizes the information of adjacent words on both sides via the tendency of RNNs to better represent recent inputs, and h_{it}^p still focuses on the current word w_{it}. The sentence (semantic) embedding $\boldsymbol{s_i}$ is the average of all h_{it}^s, where $t \in [1, l]$.

3.1.2 Role Representation. In terms of role embedding, we use dense vectors to represent members from different roles (e.g., presiding judge, plaintiff, defendant and witness) in the debate dialogue. The role information can be critical for debate legal summarization, and different roles may show different attitude towards the same controversy focus. In most cases, the plaintiffs tend to certify the effectiveness of each controversy focus while the defendants try to deny it. The witnesses provide evidences and depositions that are beneficial to the plaintiffs or the defendant and the presiding judges attempt to verify the legality, the veracity and the relevancy of these clues by a series of questions. Characterizing role embedding can help summarizer better capture different aspects of the same controversy focus, which can be quite different from classical summarization tasks. The role embedding $\boldsymbol{r_i}$ is randomly initialized and jointly learnt during the training process.

3.1.3 Legal Knowledge Representation. The Legal Knowledge Graph (LKG) used in this work is generated by legal professionals[8]. Fig. 4 visualizes an exemplar portion of the LKG. Each node represents a judicial factor. The LKG is more like an ontology, depicting the legal judge requirements for the target type of cases. For instance, for private loan disputes (PLD) cases, Fig. 4 presents an exemplar case that can occur in various scenarios, e.g., loan consent, loan payment, or partial principal payoff. Each scenario may also contain several sub-scenarios. In this work, the proposed method is able to locate the relevant legal knowledge (via LKG) for each utterance in the debate, which is utilized as an important input of the CFDS multi-task learning model.

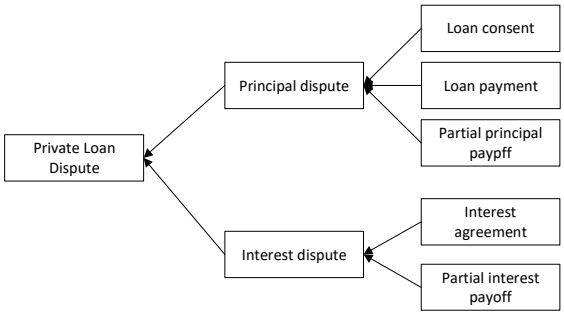

Figure 4: A portion of the Legal Knowledge Graph

Legal Knowledge Augmentation: As an auxiliary, LKG provides important legal information which endows our model with an enhanced understanding of the debate dialogue. How to obtain the legal knowledge from LKG is an essential issue in the proposed model. Inspired by the experience from presiding judge, in a court trial, the controversy focuses are the main problems to be solved and the debate dialogue (as Fig. 1 shows) is centered around these controversy focuses. In this study, given the target debate and the

[7]We have also tested LSTM and Transformer as the encoder in our model, and the comparison can be found in Sec. 5.3.

[8]In this study, because of dataset restriction, we focus on private loan disputes (PLD) cases. For LKG generation, seven judges from three civil courts, who have experience with PLD, contributed to the graph generation. More detailed information can be found in Sec. 4.1.

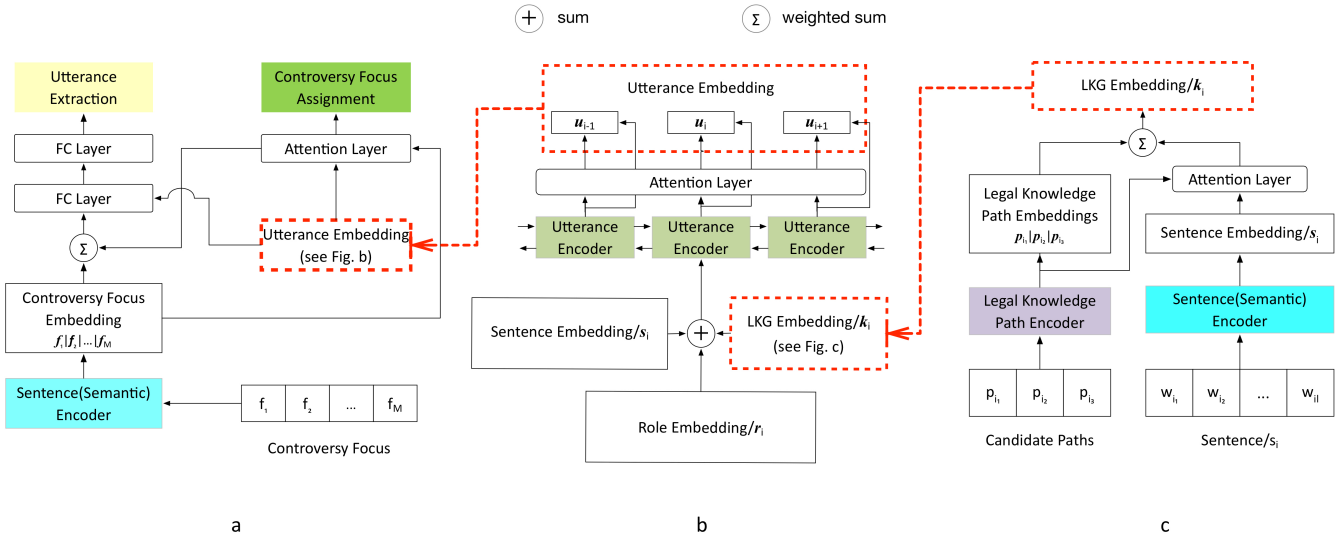

Figure 3: Network architecture of the proposed CFDS model.

case associated controversy focuses, the proposed method can first activate LKG by detecting relevant nodes w.r.t the current debate dialogue, and then mine the candidate paths on the activated-LKG for each utterance in the debate.

LKG Activation. In this step, we employ controversy focuses F to restrict the searched nodes w.r.t the whole debate. For each f in F, we calculate its semantic relevance score with each candidate node v on LKG (see Eq. 2). Then, we select the top ranked nodes[9] by leveraging the semantic relevance score and also remove the node whose score is lower than a threshold[10]. In such context, we obtain an activated-LKG G^* for debate dialogue D:

$$E(f, v) = \frac{1}{n_f n_v} \sum_{i=1}^{n_f} \sum_{j=1}^{n_v} cos(\mathbf{w_{fi}}, \mathbf{w_{vj}}) \tag{2}$$

where we regard each piece of text as a cluster of words and then compute the average similarity between the two clusters.

Candidate Path Mining. Within the activated-LKG G^*, we then seek for the relevant knowledge w.r.t each utterance u in the debate dialogue D. To compute the relevance score between the semantics of sentence s and v on G^*, Eq. 2 is applied and top relevant nodes[11] on activated-LKG G^* are selected. Finally, we set the shortest path from each selected node to the root node to form a set of candidate paths for each utterance on G^*, represented as P.

Fig. 5 offers a toy example depicting the process of candidate paths mining on LKG. f_1, f_2, f_3 are the controversy focuses of a debate dialogue, and they activate nodes F, C, I and J on LKG respectively (see Fig. 5a). Then, by using the aforementioned semantic similarity algorithm, the semantics of sentence s_i can be further projected to nodes C, I. Thus, we obtain 2 potential paths (i.e., $C \rightarrow A$, $I \rightarrow D \rightarrow A$) on LKG (see Fig. 5b).

[9]We experimentally set top 3 nodes in this work.
[10]The threshold is practically set to 0.7 in this study.
[11]We experimentally set top 3 nodes in this work.

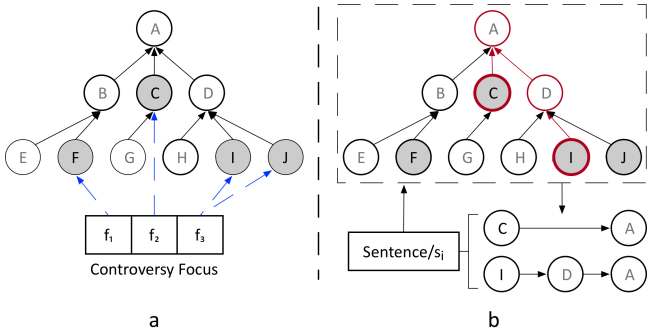

Figure 5: A toy example depicting the process of candidate paths mining on LKG

Legal Knowledge Embedding The obtained candidate paths P in the previous step are transmitted to the legal knowledge representation module as shown in Fig. 3c. For each path in P, we analogize the node on the path as the word in a sentence so that the representation of a path can be calculated in a way of sentence representation. As for the path encoder, we utilize a bidirectional LSTM[7], namely $Bi\text{-}LSTM^P$ to capture the semantics and dependencies between nodes by summarizing the information from adjacent nodes. Note that the legal knowledge encoder $Bi\text{-}LSTM^P$ and sentence encoder $Bi\text{-}LSTM^S$ are two different bidirectional LSTM models.

In the debate dialogue, one sentence may involve multiple legal scenes, which results in different legal knowledge paths. It is reasonable that judges may recall related knowledge from LKG in different aspects in terms of the semantics of the sentence. To integrate various legal knowledge paths behind sentence s_i, we introduce an attention mechanism to learn the importance of each legal knowledge path w.r.t the sentence and aggregate the representation of these paths to form the legal knowledge embedding $\boldsymbol{k_i}$. We calculate the normalized attention score α_{iz}^{sp} between $\boldsymbol{s_i}$ and

each legal knowledge path $\boldsymbol{p_{iz}}$:

$$\alpha_{iz}^{sp} = \frac{exp(\boldsymbol{s_i}^T \cdot W^{sp} \cdot \boldsymbol{p_{iz}})}{\sum_{z=1}^{Z} exp(\boldsymbol{s_i}^T \cdot W^{sp} \cdot \boldsymbol{p_{iz}})} \tag{3}$$

Then, we obtain the legal knowledge embedding $\boldsymbol{k_i}$ by computing the weighted sum over all legal knowledge path embedding considering the attention score α^{sp}:

$$\boldsymbol{k_i} = \sum_{z=1}^{Z} \alpha_{iz}^{sp} * \boldsymbol{p_{iz}} \tag{4}$$

3.1.4 Utterance Representation. With all the information calculated in the previous steps, the multi-view utterance encoder takes three kinds of embedding as input, i.e., sentence embedding $\boldsymbol{s_i}$, role embedding $\boldsymbol{r_i}$, and legal knowledge embedding k_i (see Fig. 3b). The statistics of the experimental dataset shows that 83% debate dialogue contains more than 200 utterances. In general, an utterance has relevance with adjacent and distant utterances to some extent. Inspired by this observation, we model the local context of adjacent utterances using an utterance encoder while strengthen the relevance between the distant utterances to access global context in the dialogue via an attention layer. The utterance encoder and the dialogue attention layer are two essential components in utterance representation. Within utterance encoder, we first aggregate the information of sentence, role and the legal knowledge by summing up their embeddings[12] and form a unified vector x_i. Then we apply another bidirectional LSTM[7], namely $Bi\text{-}LSTM^U$ which takes x_i as input and generates the hidden output h_i^u (see Eq. 1).

To strength the relevance between the current utterance and the distant utterances, we employ the attention mechanism to acquire c_i as follows:

$$\begin{aligned} \alpha_{ij}^{u} &= \frac{exp(h_i^{uT} \cdot W^u \cdot h_j^u)}{\sum_j exp(h_i^{uT} \cdot W^u \cdot h_j^u)} \quad , j \in [1, L] \,\&\, j \neq i \\ c_i &= \sum_j \alpha_{ij}^u * h_i^u \end{aligned} \tag{5}$$

Finally, we concatenate[13] h_i^u and c_i and get the utterance embedding $\boldsymbol{u_i}$:

$$\boldsymbol{u_i} = [h_i^u, c_i] \tag{6}$$

3.2 Task-specific Decoders

3.2.1 Task 1: Controversy Focus Assignment. The first task in the proposed multi-task architecture is to assign a controversy focus to each utterance (shown in Fig. 3a). Different debate dialogue may have different controversy focuses, and each controversy focus is concluded by the presiding judge according to the content of debate dialogue D. The controversy focus is expressed by natural language (see Fig. 1). Since the number of controversy focus is varied in different debate dialogue and each controversy focus is differed from semantics and syntactic, we can hardly cope with this task by using text classification. In this study, we calculate the relevance between utterance u_i and each controversy focus f_m in F w.r.t the debate dialogue D.

To do so, we need compute the embedding of each controversy focus. As both controversy focuses and sentences in the debate are natural language, we utilize the sentence (semantic) encoder $Bi\text{-}LSTM^S$, thus obtaining the controversy focus embedding $\boldsymbol{f_m}$ (see Eq. 1). In addition, not every utterance u_i is assigned with a controversy focus. Some utterances don't belong to any controversy focus and they can be regarded as the irrelevant content, namely *Noise*, in the debate dialogue. Thus, a category *Noise* is created for every debate dialogue and use a dense vector to represent it. Then we calculate the attention score α_{ij}^f of utterance u_i with $\boldsymbol{f_j}$:

$$\alpha_{im}^{f} = \frac{exp(u_i^T \cdot W^f \cdot \boldsymbol{f_m})}{\sum_{m=1}^{M+1} exp(u_i^T \cdot W^f \cdot \boldsymbol{f_m})} \tag{7}$$

Controversy focus with the highest normalized score α_{ij}^f is the assigned controversy focus to the utterance u_i.

3.2.2 Task 2: Utterance Extraction. The second task aims at extracting the important utterances from the debate dialogue in terms of different controversy focuses and forming multiple summarization. The utterance extractor takes two aspects into consideration: the utterance content and controversy focuses. In order to enhance the utterance representation learning considering the global legal information, we employ the normalized controversy focus distribution as the input to this task (see Eq. 8).

$$\boldsymbol{F_i} = \sum_{m=1}^{M+1} \alpha_{im}^{f} * \boldsymbol{f_m} \tag{8}$$

Then $\boldsymbol{F_i}$ and $\boldsymbol{u_i}$ are concatenated and fed into the fully connected layers as follows:

$$o_i = sigmoid(W_2^{fc} \cdot ReLU(W_1^{fc} \cdot [\boldsymbol{F_i}, \boldsymbol{u_i}])) \tag{9}$$

where W_1^{fc} and W_2^{fc} are two weight matrix and $o_i \in [0, 1]$ is the output of utterance extractor which indicates the probability of extracting utterance u_i.

3.3 Parameter Optimization

In controversy focus assignment learning, we use cross-entropy to formulate the problem as follows:

$$loss_1 = -\frac{1}{|D|} \sum_{|D|} \sum_{i=1}^{L} \sum_{m=1}^{M+1} y_{im}^f log \alpha_{im}^f \tag{10}$$

where y_{im}^f denotes the ground truth and $|D|$ is the number of debate dialogue.

As for utterance extraction learning, we use mean square error to compute the loss between the ground truth y_i^e and the prediction:

$$loss_2 = \frac{1}{|D|} \sum_{|D|} \sum_{i=1}^{L} \frac{1}{2} (o_i - y_i^e)^2 \tag{11}$$

Denoting all the parameters in our model as θ. Therefore, the optimization objective function in our learning process is given by

[12] We also tried embedding concatenation method, but the performance is worse than embedding summation.

[13] We have tried to sum up these two embedding, but the performance is worse than embedding concatenation.

$$\min_{\theta} loss = loss_1 + loss_2 + \lambda \|\theta\|_2^2 \quad (12)$$

To minimize the objective function, we use stochastic gradient decent (SGD) with the diagonal variant of AdaGrad in [33]. At time step t, the parameter θ is updated as follows:

$$\theta_t \leftarrow \theta_{t-1} - \frac{\rho}{\sqrt{\sum_{i=1}^{t} g_i^2}} g_t \quad (13)$$

where ρ is the initial learning rate and g_t is the sub-gradient at time t.

According to the evaluation results on the development set, all the hyperparameters are optimized on the training set.

4 EXPERIMENTAL SETTINGS

4.1 Datasets

For experiment, we collected 279,494 court debate records of civil Private Loan Disputes (PLD) cases, among which 5,477 cases have more than 2 controversy focuses (pre-generated by the case judge). Legal experts labeled the most important utterances for each controversy focus in those 5,477 cases (for model training and testing). After three rounds of training (lead by PLD judges), experts achieved Kappa coefficient = 0.8 (substantial agreement). In addition, for PLD cases, legal experts constructed a LKG for PLD legal knowledge characterization. The court record is a multi-role debate dialogue associating four roles, i.e., judge, plaintiff, defendant and witness. The statistics of the dataset is shown in Table 1. To the best of our knowledge, this is the very first large debate dataset enables text mining and NLP studies. We release all the experiment data to motivate other scholars to further investigate this problem[14]. To break the language barrier as well as the privacy issue, we convert all the words and labels into indices and we also provide the word embedding pre-trained on 279,494 court debate records. Note that the experiment is based on the PLD court record data mentioned above, but the method can be generalized to any other types of cases as long as the target LKG and training data are available.

Table 1: Statistics of our proposed dataset

utterances/record	maximum	2107
	minimum	20
	average	210.2
controversy focuses/record	maximum	6
	minimum	2
	average	2.4
utterances/controversy focus	average	30.5

4.2 Training Details

Word2Vec is used to pre-train the word embeddings, which are then used to initialize the embeddings in the model. The word embeddings are trained via 279,494 PLD court debate dialogues[15]. The dimensionality of word embedding, controversy focus embedding, sentence embedding, role embedding, legal knowledge embedding are set to 300. The LSTM dimension is set to 150. In this case a combination of forward and backward LSTM gives us 300 dimensions. The dropout is set to 0.7. Based on these settings, we optimize the objective function with a weight decay 0.001 and a learning rate of 0.0005. We perform the mini-batch gradient descent with a batch size of 32 for 20 epochs in total.

[14]https://github.com/zhouxinhit/Legal_Dialogue_Summarization

[15]The minimum of term frequency is set to 5, and we use "CBOW" to train the model.

5 RESULT DISCUSSIONS

5.1 Evaluation Metrics

- **Task 1: Controversy Focus Assignment:** This task can be formulated as a classification task. We evaluate the performance of each model based on two popular classification metrics: Micro F1 and Macro F1.
- **Task 2: Utterance Extraction**: To automatically assess the quality of summaries irregardless of the controversy focus assignment, we used ROUGE score [15] to compare different models. We report ROUGE-1 and ROUGE-2 as the means of assessing informativeness and the ROUGE-L for assessing fluency.
- **Task 1+ Task 2:** The multi-task learning aims at summarizing different utterances from a debate dialogue according to the controversy focuses. Thus, utterance extraction and controversy focus assignment should be assessed simultaneously. To better validate different models, we evaluate the extraction results using ROUGE in terms of each controversy focus in a debate dialogue. Thus, a debate dialogue has multiple ROUGE score $ROUGE_i, i \in [1, M]$ corresponding to each controversy focus f_i to the dialogue. We define a macro ROUGE score as:

$$^*ROUGE = \frac{1}{M} \sum_{i=1}^{M} ROUGE_i \quad (14)$$

5.2 Baselines

To demonstrate the effectiveness of the proposed model from different angles, three groups of baselines are employed targeting on task 1 only (i.e., controversy focus assignment), task 2 only (i.e., utterance extraction) and addressing task 1 and 2 simultaneously.

- *Task 1:Controversy Focus Assignment*
  - **TextCNN**: A convolutional neural networks trained on top of pre-trained word vectors for sentence-level classification tasks proposed by [13]. To fit for the controversy focus assignment, we use the same mechanism of task 1 which is shown in Fig. 3a.
  - **TextRNN**: We replace the CNN to RNN in TextRNN.
  - **FastText**: A simple neural network that averages the word feature to represent a sentence [12]. We add the necessary components to the model as what we do for TextCNN.
- *Task 2:Utterance extraction*
  - **LEAD3**: The commonly used baseline by selecting the first three sentences as the summary.
  - **TextRank**: An unsupervised algorithm based on weighted-graphs proposed by [27].
  - **SummaRunner**: A RNN based sequence model for extractive summarization by extra features such as information content, salience and novelty proposed by [19].

Table 2: Main Results of All Tested Methods.

	Method	Task1		Task2			Task1 + Task2		
		Micro F1	Macro F1	ROUGE-1	ROUGE-2	ROUGE-L	*ROUGE-1	*ROUGE-2	*ROUGE-L
Baselines	TextCNN	0.649	0.351	–	–	–	–	–	–
	TextRNN	0.72	0.419	–	–	–	–	–	–
	fastText	0.691	0.364	–	–	–	–	–	–
	LEAD3	–	–	0.2	0	0.4	–	–	–
	TextRank	–	–	32.2	10.4	38.6	–	–	–
	SummaRunner	–	–	29.9	21.0	36.3	–	–	–
	NEUSUM	–	–	29.8	22.1	36.9	–	–	–
	Xnet	–	–	30.3	22.8	37.1	–	–	–
	TextRNN+Summa.	–	–	–	–	–	15.4	5.8	18.5
	TextRNN+NEUSUM	–	–	–	–	–	16.5	6.3	20.1
	TextRNN+Xnet	–	–	–	–	–	18.1	7.0	21.5
	Transformer$^+$	0.654	0.445	54.2	34.6	54.2	**34.8**	12.4	34.8
	LSTM$^+$	0.716	0.421	52.4	37.6	58.0	25.1	13.9	29.7
Ours	Ours (Loss1)	0.709	0.436	–	–	–	–	–	–
	Ours (Loss2)	–	–	53.2	38.9	60.7	–	–	–
	Ours	**0.721**	**0.482**	**58.8**	**45.2**	**64.0**	31.6	**19.9**	**36.8**

  - **Xnet**: [21] is composed of a hierarchical document encoder and an attention-based extractor with attention over external information.
  - **NEUSUM**: A neural network framework proposed by [35] for extractive document summarization by jointly learning to score and select sentences.
- *Task 1+Task 2*
  - **TextRNN+Summa.**, **TextRNN+NEUSUM** and **TextRNN+Xnet** are three composite models[16].
  - **Transformer$^+$**: *Transformer* is proposed on [31] which is based solely on attention mechanisms. We also add LKG and role features to *Transformer* to prove the general effectiveness of the proposed features on other framework[17].
  - **LSTM$^+$**: We replace all bidirectional LSTM with LSTM in our proposed model.

5.3 Overall Performance

To comprehensively validate the proposed model for each individual of the two tasks as well as the joint of two tasks, we report results from three perspectives: (1) comparison against baselines, (2) the usefulness of multi-task framework, and (3) the effectiveness of each feature component (ablation test).

Comparison against baselines. The performance of all tested methods is reported in Table 2. We have the following observations from the results: (1) **Ours (Loss1)** performs significantly better than the first three single-task baselines (**TextCNN**, **TextRNN** and **fastText**) over $macro\_F_1$ score. Note that the method **Ours(Loss1)** is regarded as our sub-model but in single-task framework by training with only $loss_1$ (see Eq.10). As for the task 2, our method **Ours (Loss2)** training with only $loss_2$ (see 11) outperforms all the single-task baselines (row 4-8 in Table 2) by a big margin. We notice the baseline **LEAD3** is quite low in the court debate scenario because the opening three utterances of a trial are usually court disciplines. The overwhelming performance of our models when tackling with single tasks proves the advantage of the proposed utterance encoder and its generality for different tasks. (2) The two end-to-end models, **Transformer$^+$** and **LSTM$^+$**, beat the other three composite models over all the metrics when evaluating task 1 and task 2 together which indicates the effectiveness of end-to-end models for training the multi-task framework. (3) **Ours** is our main model designed to jointly learn the two tasks. It outperforms the best multi-task model (**Transformer$^+$**) over almost all the metrics (p-value< 0.001) except macro ROUGE-1 score, which indicates the effectiveness of the selection of encoder among Bi-LSTM, transformer and LSTM. **Transformer$^+$** is regarded as the best baseline for further comparison with the proposed method.

The usefulness of multi-task framework. To compare the results of our single-task approaches (**Ours(Loss1)** and **Ours(Loss2)**) and the proposed multi-task framework (**Ours**) (see Table 2), it shows that training in multi-task framework helps to significantly improve the two tasks simultaneously (**Ours** v.s. **Ours(Loss1)** at task 1 and (Ours) v.s. **Ours(Loss2)** at task 2) (p-value< 0.001).

The effectiveness of each feature component. To assess the contribution of different components in the proposed method, we conduct ablation tests for best baseline and best proposed model respectively. Table 3 reports the macro ROUGE-1, ROUGE-2 and ROUGE-L scores when training on all features and when training on all features except the particular one. Note that we didn't add attention layer for **Transformer$^+$** since it contains self-attention mechanism already, so we mainly test the significance of the other features for the best baseline. According to the results shown in Table 3, all the features contribute positively to the results. To be specific, the *global contextual* feature, namely the attention layers in Fig. 3b, has largest impact - their removal causes 22% increase in the error (RIE). As for *LKG*, it shows significant impact for both best baseline (29.8% in RIE) and our method (16.5% in RIE). Furthermore, the *role* information also has great influence on the summarization

[16] **TextRNN** shows best performance in Task1 and **SummaRunner**, **TextRNN+NEUSUM** and **Xnet** are the best three baselines for Task2.

[17] Since *Transformer* contains self-attention mechanism which has implicitly taken the context of distant utterances into consideration, so we didn't add attention layer in this baseline.

task over dialogues. This result proofs our initial hypotheses that legal knowledge and role information can be critical for debate summarization.

Table 3: Ablation Test.

	Method	*ROUGE-1	*ROUGE-2	*ROUGE-L
Best Baseline	All	**34.8**	**12.4**	**34.8**
	-Role	20	6	24.4
	-LKG	12.1	2.9	15.4
Ours	All	**31.6**	**19.9**	**36.8**
	-Global	19	10.6	22.8
	-Role	30	18.5	34.4
	-LKG	22.1	12.8	26.4

5.4 Case Study

To help readers better consume the algorithm outcomes and compare different algorithms, in Fig. 6, we show two case studies (case A and B) to compare the performance of the proposed method against two baseline methods. Note that **Transformer**$^+$ performs best under multi-task scenario among those single-model baselines, while **TextRNN+Xnet** shows the best performance as a composite model over all the other baselines.

Comparing with the baselines, the proposed method, as cases depicted, can understand the multi-role debate more effectively with respect to the target controversy focus(es). For example, the two baselines failed to assign a number of utterances (ID A-1,2,5,6 and B-8,9) to the correct focus because of incapacitated legal semantic representation and controversy focus projection. Unlike the baselines, the proposed method encodes the debate contextual information by leveraging both focus assignment task and the utterance extraction task[18], which can provide important auxiliary features for legal context characterization and summary generation.

As another critical finding, Legal Knowledge Graph can be significant for debate representation learning and legal summarization. In case B, for instance, both baselines are unable to distinguish the utterances (ID B-4,5,10,11) from the noises. It mainly because these utterances are semantically ambiguous to the controversy focus(es), however, LKG provides essential information to build connections between the text content and the controversy focus by projecting them into the same hierarchical legal conceptual space. By using LKG, both debate semantics and the latent relationships (among the legal concepts) are explored. For instance, on the LKG, the most possible candidate path of the utterance B-5 is *Actual beneficiary → Borrowing on behalf of the others → Agreement content → Loan agreement*. This path bridges the target utterance and the first controversy focus (about the loan agreement), which can be critical for legal reasoning and summary generation.

5.5 Error Analysis

For the bad cases[19], we conclude three major occasions that cause the misprediction of our proposed model: (1) 30.5% errors occur when the utterances contain some relevant semantics but not the cause-effect relationship with a certain controversy focus, our model wrongly connect them together.(2) 27.1% errors come from the controversy focus assignment for the utterances near the consecutive switches of controversy focuses in debate dialogue. The sudden change of controversy focus w.r.t successive utterances makes it difficult to assign a proper controversy focus to each utterance. (3) We also find that if one or two utterances are interspersed in a series of noises, they are prone to be mispredicted. To boost the model in the future, enhancing the legal knowledge characterization mechanism to improve the utterance representation can be a promising approach.

[18]**TextRNN+Xnet** does not consider context information for focus assignment. **Transformer**$^+$ and the associated self-attention mechanism is able to cope with the context, so it outperforms **TextRNN+Xnet**.

[19]Since the legal summarization is a multi-task, if the focus assignment and the utterance extraction are not comprehensively predicted, it is defined as a bad case.

6 RELATED WORK

6.1 Legal Intelligence

LegalTech (Legal Technology) is not a novel idea which has been forecast decades ago that the law would be transformed by IT [29, 30], while some of the predictions have already come to pass. A number of researchers from both legal field and computer science area have been exploring the potentials and methodologies to predict or anticipate the judicial decision, aiming at helping lawyers and lower court judges. In the recent work, several work focus on legal judgment prediction for the criminal cases [32, 34]. Oard et al. studied the information retrieval for e-discovery [22, 23]. Zhou et al. leveraged multi-view dispute representation for predicting e-commerce judge result[37]. Besides the prediction task, the task of legal text summarization have been regarded as another important mission for improving the understanding of the lengthy legal document by providing a brief summary. For instance, Megala et al. [16] and Polsley et al. [26] proposed several statistics-based methods (e.g., TFIDF) and some manually defined features to extract important part of the document.

Compared to all the prior efforts, this study can be novel in the following aspects. First, we work on the legal summarization for multi-role debate dialogue during the trial procedure of a case. Different from the other legal documents, the debate are much complicated but extremely important due to the vague/informal expressions conveyed through dialogues and the frequent switch among the speakers, which undoubtedly makes the understanding and the learning process much challenging. Second, the keyword or simple statistics-based methods are not suitable to address this problem because the diverse expressions of spoken language compared to written language. In this study, we pioneer this problem by quantifying a court debate through the novel multi-role and multi-focus deep debate representation learning along with sophisticated multi-task learning.

6.2 Deep learning and representation learning

The success of machine learning algorithms generally depends on data representation. Although specific domain knowledge can be used to help design representations, learning with generic priors can also be used, and the quest for AI is motivating the design of more powerful representation-learning algorithms implementing such priors[1]. In the field of NLP, the representation learning has penetrated to almost every granularity of text from character to the document representation. For instance, word2vec [17] and Glove

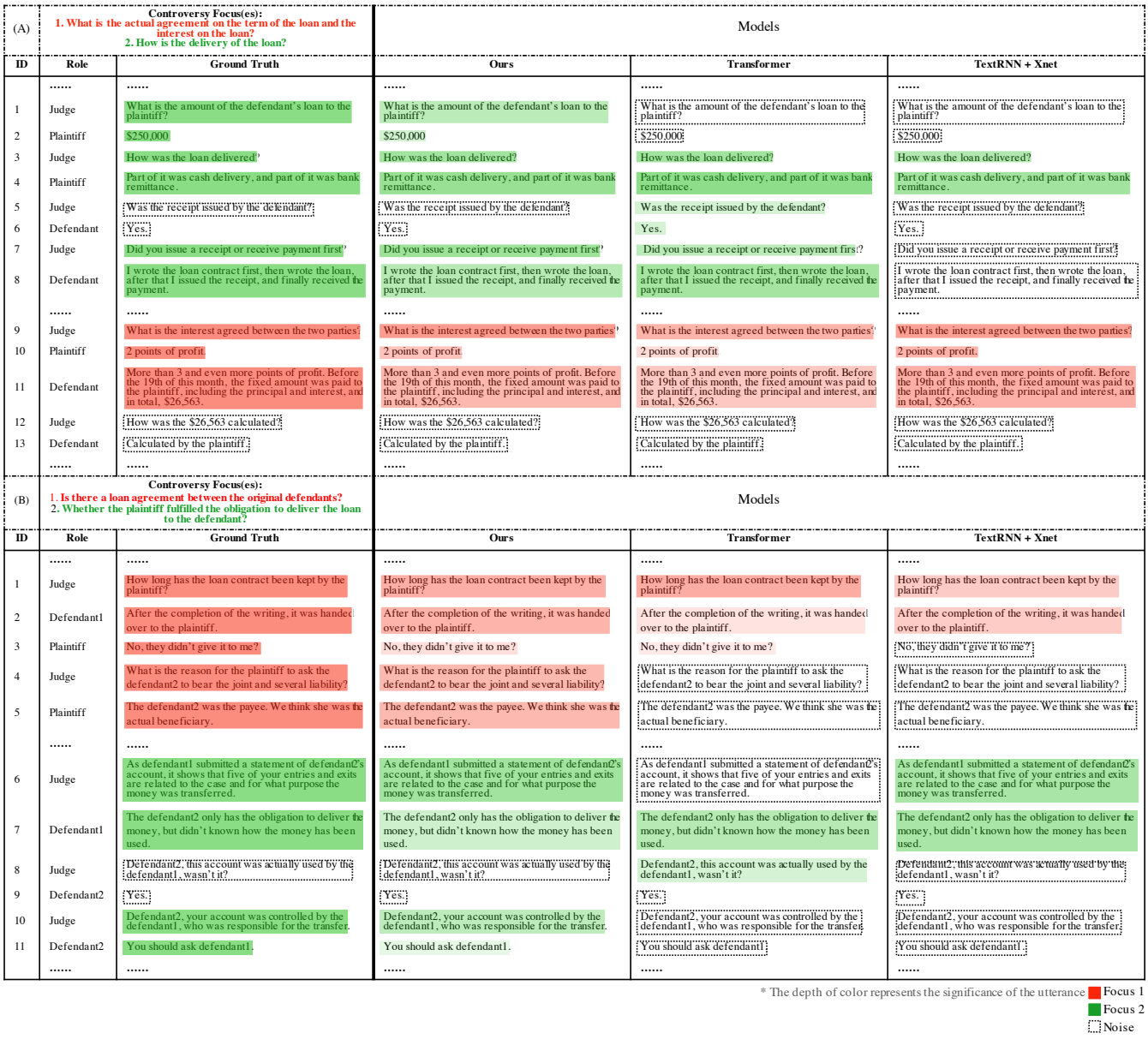

Figure 6: Case Study. Note that the depth of color corresponds to the score computed by Eq. 9

[25] are the two widely adopted *word embedding* techniques which can be traced back to the *distributed representations* introduced by Hinton [10], and developed in the context of statistical language modeling [2]. Above the word, the sentence can be also represented as a low-dimensional vector through convolutional network [13] or RNN-based network [11] by considering the sequential information within the sentences.

In this work, we introduce a novel multi-role and multi-focus debate utterance representation by not only effectively coping with sentence-level text representation but also leveraging multiple types of domain-specific features (e.g., Legal Knowledge Graph (LKG) and discrete role features).

6.3 Document summarization

Document summarization has been extensively studied for years, among which abstractive approach [6, 8, 20, 28, 36] is appealing due to its capability of generating new content but not appropriate for trail debate scenario as we mentioned in the beginning of this paper. As an effective approach, extractive methods can be used when massive training dataset is not readily available. For instance, Murray et al. applied several unsupervised methods, such as Latent Semantic Analysis (LAS) [9] and Maximal Marginal Relevance (MMR) [3], for automatic speech summarization over a meeting corpus [18]. Recently, deep neural networks based approaches have become popular for extractive document summarization [7, 19, 21]. In this work, we propose an end-to-end model in a manner of

multi-task learning process to address multi-role court debate summarization. Different from traditional summarization output, our proposed framework enables to generate multi-focus summaries for various controversy focuses respectively which can assist judges to consume and adjudicate cases. To prove the effectiveness of our model, we also set several of the previous work as baselines to be compared with our proposed framework under the court debate scenario.

7 CONCLUSION

As an interdisciplinary study, performing legal summarization over court debate dialogues can be practically useful to assist the judges to adjudicate cases. In this work, we introduce a delicately designed multi-role and multi-focus utterance representation technique and provide an end-to-end solution which is highly specialized for controversy focus-based debate summarization via jointly leaning. The empirical findings validate our hypothesis that learning various tasks jointly can improve the performance over state-of-the-art approaches. Additionally, legal knowledge graph proves essential information to enhance the performance of multi-task learning. Through results and error analysis, we proof the significance of each facet in debate summarization and also anticipate to discover a better legal knowledge characterization mechanism for future work. Different from classical NLP problems, the cost of legal text mining can be much higher since the legal NLP application needs very high precision in scenarios such like legal dispute prediction, legal knowledge generation, controversy focus generation, etc. These applications can be domain dependent which leads to the necessity of leveraging information from various aspects.

8 ACKNOWLEDGMENTS

This work is supported by National Key R&D Program of China (2018YFC0830200;2018YFC0830206).

REFERENCES

[1] Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence* 35, 8 (2013), 1798–1828.

[2] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research* 3, Feb (2003), 1137–1155.

[3] Jaime G Carbonell and Jade Goldstein. 1998. The Use of MMR and Diversity-Based Reranking for Reodering Documents and Producing Summaries. (1998).

[4] Judge Information Center. [n. d.]. As Workloads Rise in Federal Courts, Judge Counts Remain Flat. available from https://trac.syr.edu/tracreports/judge/364/. Accessed: 2019-05-06.

[5] Judge Information Center. [n. d.]. Some Federal Judges Handle Inordinate Caseloads. available from https://trac.syr.edu/tracreports/judge/501/. Accessed: 2019-05-06.

[6] Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. *arXiv preprint arXiv:1805.11080* (2018).

[7] Jianpeng Cheng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words. *arXiv preprint arXiv:1603.07252* (2016).

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

[9] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science* 41, 6 (1990), 391–407.

[10] Geoffrey E Hinton et al. 1986. Learning distributed representations of concepts. In *Proceedings of the eighth annual conference of the cognitive science society*, Vol. 1. Amherst, MA, 12.

[11] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[12] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759* (2016).

[13] Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882* (2014).

[14] US Legal. [n. d.]. The Trial Stage. available from https://civilprocedure.uslegal.com/the-course-of-a-civil-lawsuit/the-trial-stage/. Accessed: 2019-05-06.

[15] Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*.

[16] S Santhana Megala, A Kavitha, and A Marimuthu. 2014. Feature Extraction Based Legal Document Summarization. *International Journal of Advance Research in Computer Science and Management Studies* 2, 12 (2014), 346–352.

[17] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[18] Gabriel Murray, Steve Renals, and Jean Carletta. 2005. Extractive summarization of meeting recordings. (2005).

[19] Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Thirty-First AAAI Conference on Artificial Intelligence*.

[20] Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023* (2016).

[21] Shashi Narayan, Ronald Cardenas, Nikos Papasarantopoulos, Shay B Cohen, Mirella Lapata, Jiangsheng Yu, and Yi Chang. 2018. Document modeling with external attention for sentence extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2020–2030.

[22] Douglas W Oard, Jason R Baron, Bruce Hedin, David D Lewis, and Stephen Tomlinson. 2010. Evaluation of information retrieval for E-discovery. *Artificial Intelligence and Law* 18, 4 (2010), 347–386.

[23] Douglas W Oard, William Webber, et al. 2013. Information retrieval for e-discovery. *Foundations and Trends® in Information Retrieval* 7, 2–3 (2013), 99–237.

[24] OECD. 2013. What makes civil justice effective? *OECD Economics Department Policy Notes* 18 (2013).

[25] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[26] Seth Polsley, Pooja Jhunjhunwala, and Ruihong Huang. 2016. CaseSummarizer: A System for Automated Summarization of Legal Texts. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*. 258–262.

[27] Radim Rehurek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Citeseer.

[28] Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368* (2017).

[29] Richard Susskind. 2000. *Transforming the law: Essays on technology, justice and the legal marketplace*. Oxford University Press, Inc.

[30] Richard E Susskind and Richard E Susskind. 1996. *The future of law: facing the challenges of information technology*. Number s 279. Clarendon Press Oxford.

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[32] Pengfei Wang, Ze Yang, Shuzi Niu, Yongfeng Zhang, Lei Zhang, and ShaoZhang Niu. 2018. Modeling Dynamic Pairwise Attention for Crime Classification over Legal Articles. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 485–494.

[33] Matthew D Zeiler. 2012. ADADELTA: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701* (2012).

[34] Haoxi Zhong, Guo Zhipeng, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun. 2018. Legal Judgment Prediction via Topological Learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 3540–3549.

[35] Qingyu Zhou, Nan Yang, Furu Wei, Shaohan Huang, Ming Zhou, and Tiejun Zhao. 2018. Neural Document Summarization by Jointly Learning to Score and Select Sentences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 654–663.

[36] Qingyu Zhou, Nan Yang, Furu Wei, and Ming Zhou. 2017. Selective encoding for abstractive sentence summarization. *arXiv preprint arXiv:1704.07073* (2017).

[37] Xin Zhou, Yating Zhang, Xiaozhong Liu, Changlong Sun, and Luo Si. 2019. Legal Intelligence for E-commerce: Multi-task Learning by Leveraging Multiview Dispute Representation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 315–324.