A Recurrent Attention Network for Judgment Prediction

Ze Yang [§] Pengfei Wang[§⋆] Lei Zhang [§] Linjun Shou [†] Wenwen Xu [‡]
[§] School of Computer Science, Beijing University of Posts and Telecommunications, Beijing, China
[†] STCA NLP Group, Microsoft, Beijing, China
[‡] Information Science Academy, China Electronics Technology Group Corporation
{yangze01, wangpengfei, zlei}@bupt.edu.cn
lisho@microsoft.com
xuwenwenustb@163.com

Abstract. Judgment prediction is a critical technique in legal field. Judges usually scan both of the fact descriptions and articles repeatedly to select valuables information for a correct match(i.e., determine the correct articles for a given fact description). Previous works only analyze semantics to the corresponding articles, while the repeated semantic interactions between fact descriptions and articles are ignored, thus the performance may be limited. In this paper, we propose a novel Recurrent Attention Network(RAN for short) to address this issue. Specifically, RAN utilizes a LSTM to obtain both fact description and article representations, then a recurrent process is designed to model the iterative interactions between fact descriptions and articles to make a correct match. Experimental results on real-world datasets demonstrate that our proposed model achieves significant improvements over the state-of-the-art methods.

1 Introduction

Judgment prediction is a crucial and fundamental task in legal field. Given the fact, one attempts to automatically determine the correct law articles violated, which plays an important role in both professional and non-professional fields. For one hand, it can provide a reference for judges to improve work efficiency, on the other hand, it can provide legal advice to non-legal people.

Judgment prediction has been studied for decades [9,10,23], which is usually formalized as a multi-label classification problem. Previous works on this task usually exploit label correlations to improve the prediction performance. For example, Classifier Chain converts the multi-label task into a chain to model the correlation between labels [20]. Other methods such as BP-MLL [25], and kernel method [6] also model the label correlations, however, these methods can only be used to obtain low-dimensional relationships, and the high-order relationships are not taken into account.

⋆ Corresponding author

Table 1. An example of the judgment case, including a fact and two articles, where article 263 is the one that the fact violated.

fact:	*At 0:00 on October 9, 2011, the defendant Shi Jiliang, after a prior negotiation, was driven by Wei Mouyi to drive a BYD car and the rest of the people saw the victim Chen took the money at the teller machine...*
articles:	***Article263**: **Anyone who robs public or private property** in a large amount or who has been robbed several times shall be sentenced to fixed-term imprisonment of not more than three years....* ***Article264**: **Anyone who robs public or private property** by violence, coercion or other means shall be sentenced to fixed-term imprisonment of not less than three years and not more than ten years...*

Generally, article semantics(i.e, definition of articles) provide informative properties for judges to make a correct decision. We give an example in Table 1. Specifically, given the fact, a natural approach for judges is that they first browse all articles to select some candidates that are relevant with this fact (e.g, article 263 and article 264 are selected as both of two articles are relevant with robbery, similar information is marked in bold). Then a detail analysis of semantics between fact and candidates are applied to choose correct article. This process repeats several times for judges to make final decision.

Previous works, however, usually ignore label semantics for prediction. In addition, the repeated iterative information between fact and label semantics are ignored, thus the perforamnce may be limited.

In this paper, in order to address these issues, we propose a Recurrent Attention Network(RAN for short). Specially, the RAN utilizes LSTM and self-attention to embed both articles and facts into a low embedding space. After that, a recurrent block is designed to model the repeated interactions between facts and article semantics for a correct matching. To summarize, we make the following three main contributions:

- We formalize the judgment prediction task into a matching task to analyze the semantics matching between law articles and fact.
- We design a novel architecture of recurrent block to model the repeated semantic interactions between articles and facts.
- We conduct efficient experiments outperforms other baselines.Further analysis demonstrates the effectiveness of our proposed recurrent attention mechanism.

The rest of our paper is organized as follows. After a summary of related work in Section II, we describe the problem formalization of judgment prediction and our proposed model in section III. We provide experiments and evaluations in Section IV. Section V concludes this paper and discusses future directions.

2 Related Work

In this section, we briefly review two research areas related to our work: judgment prediction and attention mechanism.

2.1 Judgment Classification

Judgment prediction has been studied for a long time. At the early time, the researchers model legal predictions via statistical analysis [13, 19]. Recent attempts consider this task under text classification framework, the researchers usually extract efficient features from text and make use of machine learning methods [1, 9, 11] to learn a judgment prediction model. Inspired by the success of neural networks [3,12,17], researchers began to introduce neural network for modeling this task. Luo et al. proposed an attention-based neural network method to jointly model the judgment prediction task and the relevant article extraction task in a unified framework [15]. Hu et al. proposed an attribute-attentive charge prediction model to infer the articles and charges simultaneously [8].

As we can see, all of these works usually learn the mapping from fact to article, and ignore the semantic information of the article definition. Wang et al. introduced unified Dynamic Pairwise Attention Model for crime classification over articles [23]. In their work, a pairwise attention model based on article defnitions was incorporated into the classifcation model to help alleviate the label imbalance problem, however, their work ignore the interactive information between fact and article definition.

In our work, we try to fuse both the repeated interactive information and the aritcle semantic information into a unified model for judgment prediction.

2.2 Attention Mechanism

Attention mechanism is a technology widely used in neural networks. It is a method for automatically weighting a given input in order to extract important information. This mechanism was first used in the field of computer vision [18]. For instance, when we appreciate a painting, we first see the whole painting, then focus our attention on the part that attracts us and ignore the background information. The attention mechanism was first introduced into the field of NLP by machine translation [2], which method uses the attention mechanism between the source language and the target language to handle translation and alignment simultaneously. Luong et al. extended the previous work and proposed global and local attention [16]. Yin et al. performs the attention operation on the feature map for subsequent operations and achieved good results [24].

Many of the current works are based on a new attention mechanism called the self-attention mechanism [5,14]. In their works, the self-attention mechanism independently performed attention calculations on the original input and target input. Vaswani et al. replaced the RNN with the attention mechanism to build the entire model framework and proposed a multi-headed attention mechanism [22]. In their work, advanced results were achieved. Tan et al. proposed a deep attention neural network to model semantic role labeling with a self-attention sub-layer [21].

Our model is also related to the attention mechanism, but the main difference is that we utilize the article definition as external information, and we use

a recurrent attention block to capture multiple repeated interaction attention information between fact and article to support the judgment prediction.

3 OUR APPROACH

In this section, we first introduce the problem formalization of judgment prediction. After that, we then describe the proposed **RAN** model in detail. Finally, we present the learning and prediction procedure.

3.1 Formalization

We use $X = \{x^{(1)}, x^{(2)}, ..., x^{|X|}\}$ to denote all the facts, and $\mathcal{C} = \{y^{(1)}, y^{(2)}, ..., y^{(|C|)}\}$ for the set of all possible articles. $|X|$ and $|C|$ represent the total number of facts and labels. We use $\mathcal{L} = \{l^{(1)}, l^{(2)}, \ldots, l^{(|C|)}\}$ to represent the label description, $Y^{(i)}$ is a set of binary variables with $|C|$ elements, the j-th element is 1 or 0 to indicate whether the article is violated. In the following sections, we will use "label" instead of article for clarity.

Given all the facts X and label set $\mathcal{C}$, our task is to find an optimal $Y^{(i)}$ for a given fact $x^{(i)}$.

3.2 RAN

In this section, we present our **R**ecurrent **A**ttention **N**etwork(**RAN**) in details. Figure 1 shows the architecture of our model. Specifically, our model consists of three layers. Firstly, the encoder layer utilizes **LSTM** and self-attention to embed both label definitions and facts into a low-dimensional space. Secondly, the recurrent layer models the process by which judges repeatedly read the facts and label descriptions to obtain the repeated mutual information. Finally, the output layer gives the final prediction result of our model.

Encoder Layer In juridical field, each fact and label is described by a set of words. We take the one-hot word representation as the input, and we map each word to a vector in continuous space.

More formally, let $x = \{w_1, w_2, w_3, \ldots, w_m\}$ be a fact with m words, $l = \{w_1, w_2, w_3, \ldots, w_n\}$ be a label definition with n words, and w_i is the bag-of-word representation of i-th word. Let $\mathbf{V}^I = \{\boldsymbol{v}_t^I \in \mathbb{R}^{D_v} | t = 1, \ldots, N\}$ denote all the word vectors in a continuous space.

For each fact and label definition, we aggregate the word vectors to form the fact representation and label representation, a bidirectional **LSTM**(Bi-LSTM) is used to compute the hidden states for each word at step t as equation (1):

$$\begin{aligned} \overrightarrow{h_t} &= \overrightarrow{\text{LSTM}}(\overrightarrow{h}_{t-1}, w_t) \\ \overleftarrow{h_t} &= \overrightarrow{\text{LSTM}}(\overleftarrow{h}_{t-1}, w_t) \end{aligned} \tag{1}$$

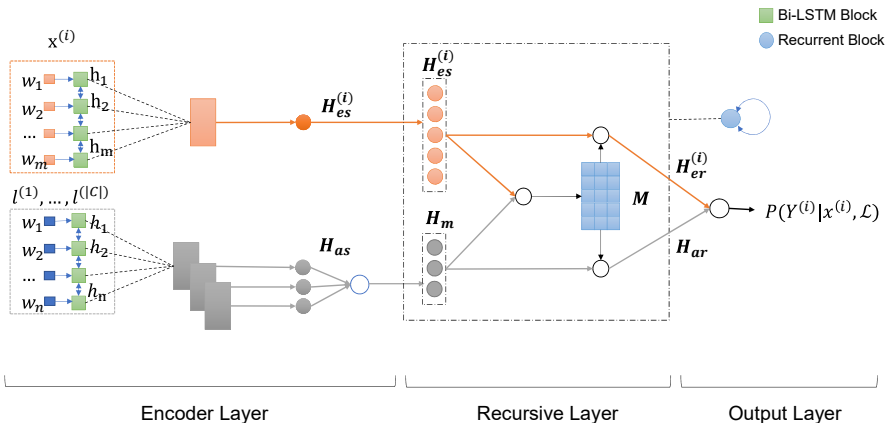

Fig. 1. The overall architecture of the proposed Recurrent Attention Model (RAN).

With the Bi-LSTM, we obtain the hidden representation of the i-th word by concatenating the hidden state of two directions, $h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}]$, then the fact x and the label l are mapped into continuous representations $H_e = [h_1, h_2, \ldots, h_m]$, $H_a = [h_1, h_2, \ldots, h_n]$, respectively.

As we all known, differnent words have different importance in one sentence. Inspired by the idea of slef-attention [22], we use self-attention to get the weighted fact and label representations H_{es} and H_{as}.

Recurrent Layer In the judicial judgment, a judge carefully reads the fact to obtain the important information, and select relevant articles as candidates, then a detail analysis of semantics between fact and the candidate articles are applied to decide final result, this process is often repeated several times to make the final determination. Different from Cui et al. [4], instead of using a simple interactive attention between source and target text. We design a recurrent attention block to model the judge's repeated reading behavior.

Through encoder layer, we get the word-level representations of fact and label, H_{es} and H_{as} respectively. We use aggregation operation to get sentence-level representations of all labels as follow:

$$H_m = [c(H_{as}^{(1)}), c(H_{as}^{(2)}), \ldots, c(H_{as}^{(|C|)})] \tag{2}$$

Where c is a aggregation operation, which average the word-level representations to form the sentence-level representations of each label.

Then, we calculate the matching score matrix M between label representation and fact's word-level representations as follows:

$$\mathbf{M}(j,k) = H_m(j) \cdot H_{es}(k) \tag{3}$$

where $M \in \mathbb{R}^{|C|*|x|}$, each value represent interactive value between fact's word and each label.

After getting the matching score matrix $\mathbf{M}$, we apply a column-wise softmax function to get probability distributions in each column, where each column represents an independent attention, and we use $\alpha(t)$ represent the label-level attention of each fact word at each step t, which can be seen as a fact word to label attention:

$$\begin{aligned} &\alpha(t) = softmax(\mathbf{M}(1,t), \mathbf{M}(2,t), \ldots, \mathbf{M}(|C|,t)) \\ &\alpha = [\alpha(1), \alpha(2), \ldots, \alpha(|x|)] \end{aligned} \tag{4}$$

Then we average all the $\alpha(t)$ to get an averaged label-level attention $\alpha^{'} \in \mathbb{R}^{|C|}$, where the averaging operation do not break the normalizing condition:

$$\alpha' = \frac{1}{|x|} \sum_{i=1}^{|x|} \alpha(t) \tag{5}$$

In the same way, we can use row-wise softmax to get label to fact word attention $\beta^{'} \in \mathbb{R}^{|x|}$. So far, we have obtained both sides attention $\alpha^{'}$ and $\beta^{'}$. Our motivation is to simulate the behavior of judges reading article and fact alternately. We propose a recurrent structure. Intuitively, this operation is continuously looped to learn the important mutual semantic information. The calculation process is shown as follows:

$$\begin{aligned} H_{es} &= H_{es} + H_{es}\mathbf{W}^{\alpha^{'}}\alpha' \\ H_m &= H_m + H_m\mathbf{W}^{\beta^{'}}\beta' \end{aligned} \tag{6}$$

Where $\mathbf{W}^{\alpha^{'}}$ and $\mathbf{W}^{\beta^{'}}$ is dimension transformation matrix. This process will be repeated several times as described above, after which we will get H_{er} and H_{ar}, representing H_{es} and H_m of the last circulation. They contain multi-level semantic interaction information for fact and label to support correct matching.

Output Layer To integrate the fact and global label information, we use both fact side and label side feature to predict the final result for a given instance in the output layer. The probability distribution over all labels is calculated as follows:

$$\begin{aligned} \boldsymbol{v}_{er} &= g(H_{er}) = \frac{1}{n}\sum_{t=1}^{n} \boldsymbol{v}_{er}(t) \\ \boldsymbol{v}_{ar} &= g(H_{ar}) = \frac{1}{n}\sum_{t=1}^{n} \boldsymbol{v}_{ar}(t) \\ \boldsymbol{v}_f &= \boldsymbol{v}_{er} \oplus \boldsymbol{v}_{ar} \\ \boldsymbol{v}_o &= \mathbf{W}^o\boldsymbol{v}_f + b^o \end{aligned}$$

Here, g is the operation of average, $\boldsymbol{v}_{er}$ represent the context representation of fact, $\boldsymbol{v}_{ar}$ is the averaged representation of all labels, which represent global label feature. $\oplus$ represent concatenate operation, $\mathbf{W}^o$ and $\mathbf{b}^o$ are learnable parameters.

3.3 Learning and Prediction

In order to learn parameters of **RAN** model, we use the stochastic gradient decent algorithm. We adopt binary cross-entropy loss in the training process as follows:

$$l = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{|\mathcal{C}|}[G^{(i)}(j)log(\sigma(\boldsymbol{v}_o^{(i)}(j))) + (1 - G^{(i)}(j))log(1 - \sigma(\boldsymbol{v}_o^{(i)}(j)))] \quad (7)$$

where σ is the sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$, $G^{(i)}(j) \in \{1, 0\}$ is a binary variable representing whether the label $y^{(j)}$ is violated by the instance $x^{(i)}$.

With the learned parameters, for each instance $x^{(i)}$, we can get the probability distribution of each label. With a threshold t, we can get the label set for the instance $x^{(i)}$, the calculation process is as follows:

$$Y^{(i)}(j) = \begin{cases} 1, & \textbf{if} \quad \mathbb{I}(\sigma(v_o^{(i)}(j))) > t \\ 0, & \textbf{else} \end{cases}$$

Where $\mathbb{I}$ is indication function, $Y^{(i)}(j)$ represent j-th element of $Y^{(i)}$, consider the label $y^{(j)}$ with output probability higher than t as the related label of $x^{(i)}$.

4 Experiment

In this section, we evaluate our proposed model on three real-world datasets. We first introduce the datasets, the experimental settings, then we compare our **RAN** with the baselines to demonstrate its effectiveness. Finally, we provide the analysis and discussion of experimental results.

4.1 Dataset

The experiments were evaluated on 3 real-world datasets:

- **CJO**: This dataset consists of 114576 samples, which dataset was collected by us from China Judge Online[1]. We removed the articles that appear less than 30 times because the data could not be used for training.
- **CAIL samll**: This dataset is a criminal case dataset for competition released by the Supreme People's Court of China [2], and the dataset consists of fact, as well as the article involved in each instance, the charges of the defendants, and the term of penalty.
- **CAIL2018** This dataset is the first large-scale Chinese legal dataset for judgment prediction [27]. CAIL2018 is a dataset, which is several times larger than other datasets in existing works.

The statistics of datasets is shown in Table 2. We split all the datasets into two non-overlapping parts, the training set and testing set, with a ratio 8:2.

[1] https://wenshu.court.gov.cn/
[2] http://cail.cipsc.org.cn/

Table 2. Basic statistics of the three datasets for experiments.

Dataset	number of samples	relevant articles	average fact length	average article size
CJO	114,576	137	825	1.17
CAIL small	204,231	183	263	1.27
CAIL 2018	1,710,856	183	279	1.04

4.2 Evaluation Metrics

Following the previous work, we use following evaluation metrics to evaluate the performance of models:

- **Jaccard similarity coefficients**: The Jaccard similarity coefficients is a widely used multi-label classification metric, it measures the similarity between two label sets, and it is defined as the size of the intersection divided by the size of the union of the label sets.
- **Macro-averaging**: macro-averaging is also a widely used metric in multi-label classification, which metric is calculated by counting the total true positives, false negatives, and false positives of each label, then calculate precision, recall, f1 for each label, and take their unweighted mean as macro-precision, macro-recall, macro-F_1.

4.3 Baselines

We adopt three types of baselines for comparison, including shollow model, nerual network based model, and attention based model.

- **KNN**: KNN [26] is a popular first-order multi-label method. Based on statistical information derived from the label sets of the neighboring instances of an unseen instance and use Bayesian inference to select assigned labels.
- **BR**: A first-order multi-label method [6]. In this model, transforms a multi-label classification with L labels into L single label classification, each classifier is a binary classifier by ignoring the correlations between labels, then unite all results of classifiers.
- **CC**: Classifier Chains [20] is a novel chaining method that can model label correlations while maintaining an acceptable computational complexity. this model train L classifiers with L labels, each next classifier is trained on the input space and all previous classifiers in the chain.
- **CNN**: A second-order multi-label method, which uses a convolution network for input representation [12], then inputted to linear layer followed by a sigmoid function to output the probabilities over the label space. The multi-label soft margin loss is optimized.
- **BiLSTM**: [7] which method is also a second-order method, and is a common way to model text and can get long-term associations.
- **DPAM**: DPAM [23] is a neural judgment prediction model by capture correlation between labels using a attention mechanism.

For all models, we set the maximum sentence length to 500 words. For shallow model, these model takes bag-of-words TF-IDF features as input, and uses chi-square to select top $5,000$ features [15]. For other models, we set the evidence representation and article representation size to 256. The size of the vocabulary is $10,000$ and out-of-vocabulary(OOV) words are replaced with unk. we use Adam optimization method to minimize the loss over the training data, we set the learning rate to be 0.001. Specially, Each LSTM in the **Bi-LSTM** is of size 128. For the CNN based models, we set the filter widths to $(3, 4, 5)$ with each filter size to 128 for consistency.

4.4 Comparison against Baselines

We compare our model **SAN** with the state-of-the-art baseline methods on judgement classification. The performance results on three datasets are shown in Table 3, MP, MR, MF and JS represent macro-precision, macro-recall, macro-F_1 and Jaccard similarity coefficients, respectively(the percentage numbers with omitted). The best performance in each case is underlined.

Table 3. Comparison between our method and all baselines on three datasets.

Dataset	metrics	Shallow Model				Neural Network Based Model		Attention Based Model	Our Model
		KNN	BR	CC	SVM	CNN	BiLSTM	DPAM	RSAN
CJO	MP	59.49	74.28	72.33	67.68	78.53	78.81	79.39	**81.52**
	MR	32.14	50.84	53.22	51.37	54.16	54.96	55.60	**55.75**
	MF	38.85	57.41	58.60	55.77	61.40	62.17	62.79	**63.34**
	JS	53.25	79.40	82.02	83.55	80.25	80.40	80.76	**80.96**
CAIL small	MP	31.75	41.59	42.12	43.07	78.32	79.93	80.35	**81.23**
	MR	20.11	30.23	32.49	39.66	54.73	57.77	62.03	**64.90**
	MF	22.93	33.57	35.58	40.14	61.35	63.98	67.42	**69.49**
	JS	38.85	59.74	62.59	71.98	74.12	75.09	76.00	**77.42**
CAIL 2018	MP	28.88	40.42	38.91	40.82	80.83	82.94	82.78	**84.01**
	MR	16.59	26.95	28.86	31.53	56.66	56.08	57.15	**57.52**
	MF	19.68	30.65	31.59	34.01	63.51	63.36	64.44	**64.92**
	JS	70.28	88.34	90.57	90.92	94.61	94.61	94.39	**94.68**

From the result, we have the following observations:

- It is not surprising that **KNN** and **BR** obtain the worst performance in terms of all evaluation metrics, these two methods covert multi-label to single label classification, and ignore the label relation.
- **CC** approach perform better than **KNN** and **BR**, which verify that modeling the correlation among multiple labels can improve the performance. But with the error propagation, the improvement is limited.
- The Neural network based models perform significantly better than shallow models. We take **CNN** as an example, comparing with the best shallow

model (CC), the improvement on CAIL small dataset over MP, MR, MF, and JS is around 36.2%, 22.24%, 25.77% and 11.53% respectively. it demonstrates that the neural network based model can effectively model text semantic information. This is also corresponding with the previous findings.
– Attention based Model(**DPAM**) achieve better performance than most text classification models(excluding **RAN**), which indicates that the article semantic information are integrated evidence.
– Our **RAN** obtains the best performance on all the evaluation metrics. Comparing with **DPAM**, **RAN** achieves significant improvement with the consideration of repeated interaction information between evidence and law articles. For example, compare with **DPAM** on CAIL Small, the relative performance improvement on MP, MR, MF, and JS is around 0.88%, 2.5%, 2.07%, 1.42%, respectively.

The experiments support our hypothesis, it's important to model the repeated mutual semantic information between evidence and article.

4.5 Analysis and Discussion

In this section, we first investigate the impact of the number of recurrent layer, then we utilize ablation test to explore the effectiveness of different layer.

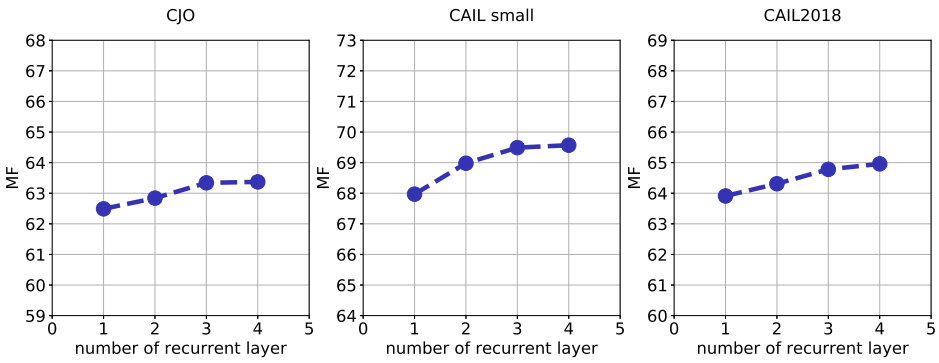

Fig. 2. The performance of different number of recurrent layers on three datasets.

The Impact of Recurrent Layer We perform the performance of our model on three datasets when the number of recurrent layers $n \in \{1, 2, 3, 4\}$, and the results are shown in the figure 2.

From the results, we can get the following observations:(1)As the number of recurrent layer n increases, the performance increase too. (2)As the number of n increases, the performance gain between two consecutive trials decreases. (3)It also indicates that after 3 layers, we have obtained stable information, and if we continue to increase the number of recurrent layer, there will be less performance improvement.

Ablation Test To further illustrate the significance of **RAN**, we evaluate the performance under difference scenario. We remove the recurrent layer(**R-r** for short), self-attention layer(**R-s** for short) to experiment separately. Result are shown in table 4.

Table 4. The ablation experiment on CAIL small.

Dataset	method	MP	MR	MF	JS
CJO	**DPAM**	79.39	55.60	62.79	80.76
	R-s	80.36	55.02	62.85	80.77
	R-r	78.78	54.56	61.97	80.35
	RAN	81.52	55.75	63.34	80.96
CAIL small	**DPAM**	80.35	62.03	67.42	76.00
	R-s	80.2	64.94	68.98	76.58
	R-r	79.37	62.08	66.75	76.93
	RAN	81.23	64.9	69.49	77.42
CAIL2018	**DPAM**	75.26	58.04	63.33	94.39
	R-s	75.3	58.06	63.9	94.64
	R-r	74.33	57.54	62.65	94.41
	RAN	77.69	58.69	64.88	94.68

We have the following findings:(1) Only the self-attention layer is retained, and the result is generally worse than **DPAM**. This is because the label that is confusing cannot be effectively distinguished due to the lack of associated information of the label. (2)Only keep the recurrent layer, the result is slightly better than **DPAM**, because the recurrent layer can effectively use the interactive attention mechanism. **DPAM** effectively utilizes the interaction information between the evidence and the label. (3)By combining the self-attention and recurrent attention mechanisms, **RAN** uses the important information obtained by self-attention for recurrent interaction, which enables the model to obtain more effective information.

This further testifies that the recurrent layer is capable of helping the model to acquire repeated semantic information for improving judgment prediction. f

5 Conclusion

In this paper, we propose an Recurrent Attention Network that can simulate the repeated reading behavior of judge, which method can utilize the semantic mutual information between evidence and article, Extensive experimental results show that the proposed model outperform the baselines. Further analysis demonstrates that our model not only obtain label correlation information, but also capture the multiple informative attention with the recurrent block.

In the future, we will seek to explore the following directions:(1) We will study how to improve the performance of the judgment prediction if more information is used as external knowledge. (2) Since judgment prediction has explicit logical

reasoning properties, we will seek the interpret ability of the model to better understand what the model does.

6 acknowledge

This research work was supported by the National Natural Science Foundation of China under Grant No.61802029, and the fundamental Research for the Central Universities under Grant No.500419741. We would like to thank the anonymous reviewers for their valuable comments.

References

1. Aletras, N., Tsarapatsanis, D., Preotiuc-Pietro, D., Lampos, V.: Predicting judicial decisions of the european court of human rights: a natural language processing perspective. PeerJ Computer Science **2**, e93 (2016)
2. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
3. Bordes, A., Glorot, X., Weston, J., Bengio, Y.: Joint learning of words and meaning representations for open-text semantic parsing. In: Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2012, La Palma, Canary Islands, Spain, April 21-23, 2012. pp. 127–135 (2012)
4. Cui, Y., Chen, Z., Wei, S., Wang, S., Liu, T., Hu, G.: Attention-over-attention neural networks for reading comprehension. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers. pp. 593–602 (2017)
5. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. CoRR **abs/1810.04805** (2018)
6. Elisseeff, A., Weston, J.: A kernel method for multi-labelled classification. In: International Conference on Neural Information Processing Systems: Natural and Synthetic. pp. 681–687 (2001)
7. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional lstm and other neural network architectures. Neural Networks **18**(5-6), 602–610 (2005)
8. Hu, Z., Li, X., Tu, C., Liu, Z., Sun, M.: Few-shot charge prediction with discriminative legal attributes. In: Proceedings of the 27th International Conference on Computational Linguistics. pp. 487–498. Association for Computational Linguistics (2018)
9. Katz, D.M., II, M.J.B., Blackman, J.: Predicting the behavior of the supreme court of the united states: A general approach. CoRR **abs/1407.6333** (2014)
10. Keown, R.: Mathematical models for legal prediction. The John Marshall Journal of Information Technology and Privacy Law **2**(1), 29 (1980)
11. Kim, M., Xu, Y., Goebel, R.: Legal question answering using ranking svm and syntactic/semantic similarity pp. 244–258 (2014)
12. Kim, Y.: Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL. pp. 1746–1751 (2014)

13. Kort, F.: Predicting supreme court decisions mathematically: A quantitative analysis of the right to counsel cases. American Political Science Review **51**(1), 1–12 (1957)
14. Lin, Z., Feng, M., dos Santos, C.N., Yu, M., Xiang, B., Zhou, B., Bengio, Y.: A structured self-attentive sentence embedding. CoRR **abs/1703.03130** (2017)
15. Luo, B., Feng, Y., Xu, J., Zhang, X., Zhao, D.: Learning to predict charges for criminal cases with legal basis. empirical methods in natural language processing pp. 2727–2736 (2017)
16. Luong, T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015. pp. 1412–1421 (2015)
17. Luong, T., Sutskever, I., Le, Q.V., Vinyals, O., Zaremba, W.: Addressing the rare word problem in neural machine translation. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers. pp. 11–19 (2015)
18. Mnih, V., Heess, N., Graves, A., Kavukcuoglu, K.: Recurrent models of visual attention. In: Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada. pp. 2204–2212 (2014)
19. Nagel, S.S.: Applying correlation analysis to case prediction. Tex. L. Rev. **42**, 1006 (1963)
20. Read, J., Pfahringer, B., Holmes, G., Frank, E.: Classifier chains for multi-label classification. Machine Learning **85**(3), 333–359 (2011)
21. Tan, Z., Wang, M., Xie, J., Chen, Y., Shi, X.: Deep semantic role labeling with self-attention. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018. pp. 4929–4936 (2018)
22. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA. pp. 6000–6010 (2017)
23. Wang, P., Yang, Z., Niu, S., Zhang, Y., Zhang, L., Niu, S.: Modeling dynamic pairwise attention for crime classification over legal articles. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018. pp. 485–494 (2018)
24. Yin, W., Schütze, H., Xiang, B., Zhou, B.: ABCNN: attention-based convolutional neural network for modeling sentence pairs. TACL **4**, 259–272 (2016)
25. Zhang, M.L., Zhou, Z.H.: Multilabel neural networks with applications to functional genomics and text categorization. IEEE Transactions on Knowledge and Data Engineering **18**(10), 1338–1351 (2006)
26. Zhang, M.L., Zhou, Z.H.: Ml-knn: A lazy learning approach to multi-label learning. Pattern recognition **40**(7), 2038–2048 (2007)
27. Zhong, H., Zhipeng, G., Tu, C., Xiao, C., Liu, Z., Sun, M.: Legal judgment prediction via topological learning. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 3540–3549 (2018)