A Multiple Attention Label-Enhanced Model for Legal Article Prediction

1st Xuanqing Zhang
School of Computer Science and Technology
Henan Polytechnic University
Jiaozuo, China e-mail: xqzhang_ch@163.com

2nd Junyi Chen*
School of Computer and Artificial Intelligence
Zhengzhou University
Zhengzhou, China email: junyichen_ch@sina.com

3rd Yuke Wang
School of Computer and Artificial Intelligence
Zhengzhou University
Zhengzhou, China email: yukeewang@foxmail.com

2022 4th International Conference on Frontiers Technology of Information and Computer (ICFTIC) | 979-8-3503-2195-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICFTIC57696.2022.10075319

***Abstract*—Legal Article Prediction (LAP) aims to automatically identify the relevant articles according to the content of the prosecution document of each case and the established law articles, which is of great value in legal assistance systems and in improving the work efficiency of legal practitioners. In this paper, we formulate LAP as a multi-label learning problem and present a Multiple Attention Label-Enhanced Model (MALE), which first utilizes two encoders to obtain the base representation of fact description and established law articles, second employs a labelenhanced method fully fusing the information of the two from different perspectives to get the co-dependent representation of the fact description and each law article finally. Experimental results on the CAIL2018 public dataset demonstrate that MALE achieves significant performance improvements over existing neural models in predicting relevant law articles.**

***Keywords*—law article prediction; multi-label classification; deep neural network**

I. Introduction

Legal article prediction(LAP) is one of the subtasks of legal judgment prediction(LJP), which aims to automatically identify the relevant law articles based on cases' fact descriptions and established law articles. In many jurisdictions, the excessive workload of courts often leads to substantial delays. While suitable AI predictive models can effectively assist legal professionals in their work, thereby speeding up their work. At present, this direction is attracting increasing attention.

In practice, the judging process often starts with identifying the relevant law articles, then making a charge prediction and deciding the terms of penalty, which means that LAP often has an impact on the other two subtasks. Therefore, it is crucial and necessary to develop a model that can accurately predict the relevant articles. Fig. 1 shows the factual description and relevant law articles for an anonymous legal case. The defendant violated two law articles at the same time, leading to two charges. The defendants' different behaviors are reflected by different keywords in the fact description. As highlighted in the figure, there is a strong correlation between the fact description and the content of relevant law articles. For example, words related to law article "264" in the fact description are highlighted in red, and those related to law article "266" is in blue. Therefore, it is desirable to develop a

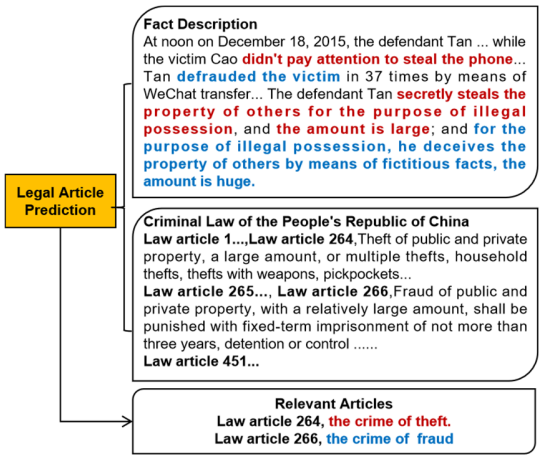

Fig. 1. An example of LAP scenario. The words in red and blue indicates those highly related to the two law articles.

model that can capture this type of correspondence and use the content of law articles to guide the judging process, as human judges do. Also in this way, it is possible to further improve the accuracy of the LJP and provide a degree of interpretability for legal professionals by highlighting the corresponding relevant text fragments, as in Fig. 1.

However, existing methods [1]–[3] usually treat LAP as a multi-class classification task, assuming that each case is only associated with a single law article, which ignores the multi-labeling nature of LAP. However, it is common for a case to be associated with multiple law articles in realworld scenarios. Some recent works [4], [5] have also noticed this issue and proposed some excellent models. Inspired by them, we take the content of the established law articles as label-enhanced information and integrate it into the model to assist in predicting the relevant articles. At the same time, we observe that the keywords in the fact description of each case span a wide range, and the content of established articles is quite different from that of factual descriptions. It is difficult to make their information sufficiently interactive with a simple attention mechanism.

To address the above issues, we propose a novel model called Multiple Attention Label-Enhanced (MALE), which first utilizes a multi-scale convolutional network to encode fact descriptions and a Bi-GRU to encode established articles. Next,

* is corresponding author.

it employs a multiple attention mechanism that contains internal and external attention modules so that the established articles and factual descriptions can interact sufficiently. The internal attention based on coattention [6] focuses on locating the key elements in the fact description and established articles, while the external attention module aims to fuse established articles with different weights for each fact description.

II. RELATED WORK

Deep neural networks have achieved great success in LAP. [7] proposed an attention-based neural network to predict the relevant article and charge. [8] proposed a LAP model based on a deep pyramid convolutional neural network to acquire long-distance dependencies of legal texts. Some studies exploit the relationship among the three subtasks of LJP (i.e., law article prediction, charge prediction, and the term of penalty prediction.) to help improve the performance of LAP. [2] proposed a topological structure to perform the multi-task prediction. Subsequently, [9] added backward verification to the topology and proposed a multi-perspective bi-feedback network with the word collocation.

Inspired by how human judges use the established law articles, [1] proposed a novel graph neural network to automatically learn subtle differences between confusing law articles. [10] proposed a transfer learning model to solve the data imbalanced problem of LAP by transferring prior knowledge from frequent cases to low-frequency ones. [5] proposed a law article element-aware multi-representation model which generates multiple representations of a fact for classification. [4] improved the encoding of fact description to capture longdistance dependencies of legal texts and guide the prediction with the encoded knowledge of the established articles.

MALE differs from those models by adapting advanced multiple-label learning techniques [11]–[14] to explore the rich contextual information contained in the fact description, and then adopting a multiple-attention mechanism that includes internal attention and external attention modules to keep the established articles and fact descriptions interacting sufficiently.

III. METHOD

A. Problem Statement

Giving a training dataset $D=\{\mathcal{X}_i, \mathcal{A}, \mathcal{Y}_i\}_{i=1}^{s}$, where s denotes the size of training dataset. $A = \{A_1,A_2,...,A_k\}$ is the set of established law articles, k is the number of law articles. The goal of LAP is to produce a mapping function $F : X \rightarrow Y$ that can predict relevant articles for legal cases with fact descriptions in the test set $\mathcal{X}^{test} \notin \{\mathcal{X}_i\}_{i=1}^{s}$, i.e.,

$F(X^{test},A) = Y_b$. Since each case may have multiple relevant law articles, so $Y_b = \{y_1,y_2,...,y_i\}$,where $y_i \in A$.

B. Embedding

The fact description of a case is a word sequence $X_i = \{x_1,x_2,...,x_n\}$,where n denotes the document length. We use an embedding layer to convert X into embedding sequence $\mathbf{E}_i = \{\mathbf{e}_1,\mathbf{e}_2,...,\mathbf{e}_n\}$, where $\mathbf{e}_i \in R^d$. The pretrained word2vec [15] embeddings are used. Similarly, law articles are some word sequences. After embedding, article k can be denoted as $A_k = \{a_{k1},a_{k2},...,a_{kl}\}$,where l denotes the law article length.

C. Fact Description Encoder

[1], [3]–[5], [17] have shown that the article content can benefit LAP. Different from them, MALE adopts a multiresidual convolution neural network [14] to encode the case description in order to capture rich contextual information via various n-grams. It mainly consists of the following two parts.

1) Multi-filter convolution layer: Multi-filter convolution layer [16] consists of m filters with different kernel sizes, $s_1,s_2,...,s_m$. It is formulated as:

$$C_i = \bigwedge_{j=1}^{n} tanh(\boldsymbol{W}_{si}\boldsymbol{E}^{j:j+s_i-1} + \boldsymbol{b}_{si}), \quad (1)$$

where $\bigwedge_{j=1}^{n}$ represents a convolution operation from left to right, $\mathbf{E}^{j:j+s_i-1} \in R^{s_i \times d}$ is the concatenation of word embedding and is the sub-matrix of $\mathbf{E}$. $\mathbf{W}_{s_i} \in R^{f\times(s_i\times d)}$ and $\mathbf{b}_{s_i} \in R^f$ are the learnable weight matrix and the bias vector. The padding number is set to $\lfloor \frac{s_i}{2} \rfloor$, the stride is 1 for each filter, and each filter has the same out-channel size f.

2) Residual block: Residual block can further enlarge the receptive field [14] and capture even rich information from longer context, which consists of three convolution filters, formulated as:

$$F_{i,1} = \bigwedge_{j=1}^{n} tanh(\boldsymbol{W}_{i,1}\boldsymbol{C}_{\boldsymbol{i}}^{j:j+s_i-1}), \quad (2)$$

$$F_{i,2} = \bigwedge_{j=1}^{n} tanh\boldsymbol{W}_{i,2}\boldsymbol{F}_{\boldsymbol{i,1}}^{j:j+s_i-1}, \quad (3)$$

$$F_{i,3} = \bigwedge_{j=1}^{n} tanh\boldsymbol{W}_{i,3}\boldsymbol{F}_{\boldsymbol{i,1}}^{j:j}, \quad (4)$$

$$F = concat(F_{i,2} + F_{i,3}), \forall i \in [1,m] \quad (5)$$

where $\boldsymbol{W_{i1}} \in \boldsymbol{R^{O\times(s_i\times f)}}$, $\boldsymbol{W_{i2}} \in \boldsymbol{R^{O\times(s_i\times O)}}$ and $\boldsymbol{W_{i3}} \in \boldsymbol{R^{O\times(1\times f)}}$ are weight matrices, $\boldsymbol{C_i^{j:j+s_i-1}} \in \boldsymbol{R^{s_i\times f}}$ is the submatrix of the output of multi-filter convolution layer $\mathbf{C}_i$. The settings of padding and stride are the same as the convolution layer. Since there are m filters, and each has one residual block, the final output of residual convolution layer is a concatenation of $\mathbf{F_i}$. Finally we get the representation of fact description $\mathbf{F} \in R^{n\times(m\times o)}$, where $m \times o$ is the hidden dimension of $\mathbf{F}$, we will use h to represent next.

D. Internal Attention Module

MALE makes improvements on the basis of coattention [6] to enable fact description to fully interact with each law article.

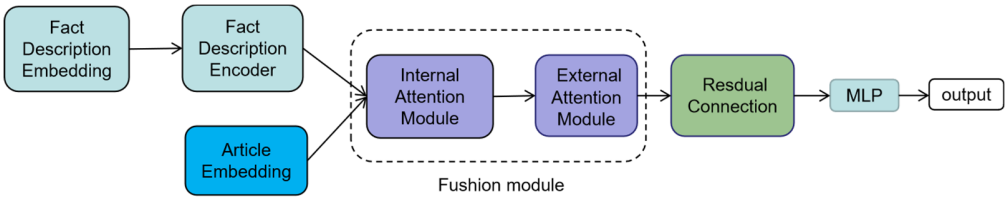

Fig. 2. Overview of MALE.

For each law article, there are the following calculations. Firstly, calculate the affinity matrix $\mathbf{P} = A_k^t \ \mathbf{F} \in \mathrm{R}^{n \times l}$,which contains affinity scores corresponding to fact description and law article words. Then normalize the row-wise and columnwise of $\mathbf{P}$ to get the attention weights α and β:

$$\alpha = softmax(P) \in R^{l \times n} \tag{6}$$

$$\beta = softmax(P^T) \in R^{n \times l}, \tag{7}$$

Secondly, compute the attention contexts of each law article in light of the represent of fact description $\mathbf{F}$. Then modify the fact encoding by the way of concatenation, and put the law article encoding into the space of fact encoding. Lastly,we get the co-dependent representation $\mathbf{M}_k$ of each law article and fact description by internal attention fusion.

$$\mathbf{D} = \beta \mathrm{A_k} \in R^{n \times h}, \tag{8}$$

$$\mathbf{F`} = [\mathbf{F}, \mathbf{D}] \in R^{n \times 2h}, \tag{9}$$

$$M_k = \alpha F` \in R^{l \times 2h} \tag{10}$$

E. External Attention Module

After each article fully interacts with the fact description, we get $\mathbf{M} = [\mathbf{M}_1,...,\mathbf{M}_k] \in \mathrm{R}^{k \times l \times 2h}$, followed by the sumpooling operation, as $\mathbf{M}^* = pooling(\mathbf{M}) \in \mathrm{R}^{k \times 2h}$, to get the representation of fact description for each article. Due to it has a different correlation with the fact description,we calculate the weight of each article based on the fact description $\mathbf{F}$ before fusion. It is formulated as:

$$\mathbf{t} = pooling(\mathbf{F})\mathbf{W}_t + \mathbf{b}_t, \tag{11}$$

$$\gamma = softmax(\mathbf{t}) \in \mathrm{R}^k, \tag{12}$$

$$\mathbf{a} = \gamma \bullet pooling(\mathbf{M}^*) \in \mathrm{R}^k. \tag{13}$$

where $W_t \in \mathrm{R}^{h \times k}, b_t \in \mathrm{R}^k$.

Next, we add a residual connection on the output of the external attention module to avoid forgetting information. Then, we transform it with a linear layer to get the final output $\mathbf{y}_f = MLP(\mathbf{a} + \mathbf{t})$.

F. Prediction and Training

The final prediction is computed as $y\hat{} = \mathrm{sigmoid}(\mathbf{y}_f)$. The training objective is to minimize the binary cross-entropy loss between the prediction $y\hat{}$ and the target y.

$$\mathcal{L} = -\frac{1}{s}\sum_{i=1}^{s} y_i log(\hat{y_i}) + (1 - y_i) log(1 - \hat{y_i}). \tag{14}$$

IV. EXPERIMENTS

A. Dataset

We use CAIL2018[1] released by the Supreme People's Court of China, and the personal privacy information in the dataset has been anonymized when released publicly. It contains 183 relevant articles. The details of CAIL2018 can be found in [18]. Following [4], [5], we used CAIL-small and its filtered version CAIL-split, where relevant articles with a frequency of less than 300 were removed.

B. Results and Analysis

We implemented models based on the NeuralClassifier [22] toolkit. The experiments results are shown in Table I. Due to the long-tailed distribution of the dataset, we mainly use the F1 score for comparison. By conparison we have the following findings. (1)The three models designed for LJP (i.e., TopJudge, LEMM, and Mulan) and label-enhanced models (i.e., CAML and MultiResCNN) almost always outperform the generic text classifiers, except that CAML only outperforms CNN. Since CAML is composed of CNN and label attention modules, this result is as expected, which just verifies that label attention is effective for document classification. (2)Comparing the models designed for LJP, TopJudge performs poorly. It may be because TopJudge is designed for single-label multi-task scenarios, which joins the topological relationship of three subtasks without injecting label information. (3)The horizontal comparison shows that the evaluation scores of all models have decreased in CAIL-small, especially Macro- scores. It is obviously caused by the addition of small sample cases. (4) Among the above models, the overall performance of MALE is well, exceeding Mulan on CAIL-split/small by 1.05%/0.46% in Micro-F1 score and 0.62%/0.57% in Macro-F1 score. It shows that the multiple attention module we proposed is effective.

V. CONCLUSION

In this paper, we propose a label-enhanced model with multiple attention for LAP. It encodes the fact description and established law articles at the same time, which uses multiple attention modules from different perspectives to fully integrate the two to improve the performance of LAP. Experiments show that our model achieves certain performance improvements. Finally, we must emphasize that models are only auxiliaries and cannot replace humans in making decisions. Furthermore, due to the rigor and particularity of the legal field, the decisionmaking process must be visible, legitimate, and

[1] https://github.com/thunlp/CAIL

reasonable. In the future, we will further explore improving the interpretability of LAP.

TABLE I THE RESULTS OF LAP ON CAIL-SPLIT AND CAIL-SMALL.

	CAIL-split						CAIL-small					
	Precision		Recall		F1		Precision		Recall		F1	
	Micro-	Macro-	Micro-	Macro-	Micro-	Macro-	Micro-	Macro-	Micro-	Macro-	Micro-	Macro-
TextCNN [16]	79.94	79.69	75.87	72.21	77.85	74.64	76.53	60.11	73.43	52.42	74.95	53.42
RCNN [19]	79.88	80.22	83.50	80.70	81.65	79.60	77.29	65.83	81.51	56.64	79.34	58.30
Transformer [20]	78.34	79.88	81.49	77.91	79.88	77.60	77.03	66.34	78.01	54.69	77.52	56.97
DPCNN [21]	79.99	80.27	78.32	73.17	79.14	75.12	77.58	60.60	75.27	49.84	76.41	51.51
CAML [11]	76.33	78.42	81.15	77.25	78.67	76.55	75.62	62.01	78.51	55.49	77.04	56.36
MultiResCNN [14]	79.98	81.37	84.80	82.87	82.32	81.13	79.72	71.43	82.35	63.07	81.01	64.36
TopJudge [2]	80.31	81.20	79.40	73.88	79.85	76.04	78.13	61.11	72.47	46.65	75.19	50.17
LEMM [5]	83.73	83.91	84.55	82.11	84.13	82.46	81.47	73.53	81.65	61.21	81.56	64.69
Mulan [4]	81.72	83.68	84.98	81.82	83.32	81.91	80.75	74.44	83.21	64.75	81.96	66.80
MALE	84.59	84.88	84.15	81.22	84.37	82.37	82.65	76.48	82.51	64.34	82.58	67.37

REFERENCES

[1] N. Xu, P. Wang, L. Chen, L. Pan, X. Wang, and J. Zhao, "Distinguish confusing law articles for legal judgment prediction," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, Jul. 2020, pp. 3086–3095.

[2] H. Zhong, Z. Guo, C. Tu, C. Xiao, Z. Liu, and M. Sun, "Legal judgment prediction via topological learning," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, 2018, pp. 3540–3549.

[3] Y. Feng, C. Li, and V. Ng, "Legal judgment prediction via event extraction with constraints," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, Dublin, Ireland, May 2022, pp. 648–664.

[4] J. Chen, L. Du, M. Liu, and X. Zhou, "Mulan: A multiple residual article-wise attention network for legal judgment prediction," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 21, no. 4, pp. 1–15, apr 2022.

[5] H. Zhong, J. Zhou, W. Qu, Y. Long, and Y. Gu, "An element-aware multi-representation model for law article prediction," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, Nov. 2020, pp. 6663–6668.

[6] C. Xiong, V. Zhong, and R. Socher, "Dynamic coattention networks for question answering," in *International Conference on Learning Representations*, 2017, pp. 1–14. [Online]. Available: https://openreview.net/forum?id=rJeKjwvclx

[7] B. Luo, Y. Feng, J. Xu, X. Zhang, and D. Zhao, "Learning to predict charges for criminal cases with legal basis," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, Sep. 2017, pp. 2727–2736.

[8] H. Zhang, X. Wang, H. Tan, and R. Li, "Applying data discretization to dpcnn for law article prediction," in *Natural Language Processing and Chinese Computing*, 2019, pp. 459–470.

[9] W. Yang, W. Jia, X. Zhou, and Y. Luo, "Legal judgment prediction via multi-perspective bi-feedback network," in *Proceedings of the TwentyEighth International Joint Conference on Artificial Intelligence, IJCAI19*. International Joint Conferences on Artificial Intelligence Organization, July 2019, pp. 4085–4091.

[10] Y.-S. Chen, S.-W. Chiang, and M.-L. Wu, "A few-shot transfer learning approach using text-label embedding with legal attributes for law article prediction," *Applied Intelligence*, vol. 52, no. 3, pp. 2884–2902, 2022.

[11] J. Mullenbach, S. Wiegreffe, J. Duke, J. Sun, and J. Eisenstein, "Explainable prediction of medical codes from clinical text," in *NAACL-HLT*, 2018.

[12] A. Rios and R. Kavuluru, "Few-shot and zero-shot multi-label learning for structured label spaces," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 3132–3142.

[13] J. Lu, L. Du, M. Liu, and J. Dipnall, "Multi-label few/zero-shot learning with knowledge aggregated from multiple label graphs," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 2935–2943.

[14] F. Li and H. Yu, "Icd coding from clinical text using multi-filter residual convolutional neural network," in *AAAI*, 2020.

[15] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems*, vol. 26, pp. 3111–3119, 2013.

[16] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, Oct. 2014, pp. 1746–1751.

[17] L. Yue, Q. Liu, B. Jin, H. Wu, K. Zhang, Y. An, M. Cheng, B. Yin, and D. Wu, "NeurJudge : A circumstance-aware neural framework for legal judgment prediction," in *SIGIR2021*, pp. 973–982.

[18] C. Xiao, H. Zhong, Z. Guo, C. Tu, Z. Liu, M. Sun, Y. Feng, X. Han, Z. Hu, H. Wang, and J. Xu, "CAIL2018: A large-scale legal dataset for judgment prediction," *arXiv preprint axXiv:1807.02478*, 2018.

[19] S. Lai, L. Xu, K. Liu, and J. Zhao, "Recurrent convolutional neural networks for text classification," in *AAAI*, 2015, pp. 2267–2273.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[21] R. Johnson and T. Zhang, "Deep pyramid convolutional neural networks for text categorization," in *ACL*, 2017, pp. 562–570.

[22] L. Liu, F. Mu, P. Li, X. Mu, J. Tang, X. Ai, R. Fu, L. Wang, and X. Zhou, "Neuralclassifier: An open-source neural hierarchical multilabel text classification toolkit," in *ACL*, 2019, pp. 87–92.

[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.