A Relation Learning Hierarchical Framework for Multi-label Charge Prediction

Wei Duan¹, Lin Li^{1(✉)}, and Yi Yu²

¹ School of Computer Science and Technology, Wuhan University of Technology,
Wuhan, China

{dw8855, cathylinlin}@whut.edu.cn

² Digital Content and Media Sciences Research Division,
National Institute of Informatics, Tokyo, Japan
yiyu@nii.ac.jp

Abstract. In legal field, multi-label charge prediction is a popular and foundational task to predict charges (labels) by a case description (a fact). From perspectives of content analysis and label decision, there are two major difficulties. One is content confusion that the case descriptions of some charges are almost identical. The other is dynamic label number that the numbers of labels (label number) of different cases may be different. In this paper, we propose a relation learning hierarchical framework for multi-label charge prediction with two models, i.e., dynamic merging attention (DMA) and number learning network (NLN). Specially, DMA can improve the charge prediction performance by dynamically learning the similarity relation between a fact and external knowledge (provisions) and the difference relation between different provisions, which alleviates the phenomenon of content confusion. NLN mitigates the dynamic label number by learning the co-occurring relation between labels. Moreover, we put the two models into a unified framework to enhance their effects. Conducted on a public large real-world law dataset, experimental results demonstrate that our framework with DMA and NLN outperforms well-known baselines by more than 3%–23%.

Keywords: Multi-label charge prediction · External knowledge · Label number learning

1 Introduction

The legal charge prediction can be regarded as a multi-label text classification task that learns from a law case description (a fact) to predict its charges (labels). As shown in Table 1, someone has violated an arson and the other one has violated a fire sin based on the case descriptions. Therefore, the charge prediction task needs to focus on the description of illegal behaviors, recognizes the difference between two facts and finally classifies them into their corresponding charges. Through this example, there are two difficulties from two

different aspects. One is content confusion that there are quite similar descriptions between facts with different labels, such as arson and fire sin. The other is dynamic label number that the number of labels (label number) is not fixed for all the cases. For example, some facts have only one label, but others have multiple labels.

Table 1. The details of case samples

Charge (label)	Case description (fact)
Arson	...Then use the lighter carried with Ding to ignite the clothes and burn the air conditioners, TV sets, electric fans and other items in the east room...
Fire sin	...During the period, the defendant Wang did not ensure the safety of fire in the wild, caused a forest fire when burning paper money, and fled the scene after the fire...
<i>Charge (label)</i>	<i>Law provision (external knowledge)</i>
Arson	Deliberately set fire to public and private property and endanger public safety
Fire sin	A fire caused by the negligence of the actor, causing serious consequences and endangering public safety

Firstly, making use of external knowledge is an effective solution for content confusion [4, 12]. There is scope for improvement over previous approaches. Because they extract several most relevant knowledge (provisions) without considering impacts of the similarity relation between a fact and its provisions and the difference relation between different provisions. In this paper, we try to explore those impacts by our proposed dynamic merging attention (DMA). The DMA gains the self-attention scores of the fact and provisions, and furthermore because of different convolution kernels, it dynamically acquires various representations about the two relations from the scores.

Secondly, for dynamic label number, there are two widely used strategies to obtain labels from a probability, i.e., top-k based and threshold based. The two strategies need time and efforts to set parameters manually. In addition, they use a unified k or a threshold value, which causes an increase in errors especially when the prediction probability is not precise. There is a work which jointly learns a multi-label classification model and a threshold predictor to gain different fixed thresholds for the different labels [17]. However, it ignores the co-occurring relation between labels. If a charge is a label of a fact, then another related charge will be also its label with a high probability. Our proposed number learning network (NLN) can automatically learn the number of labels corresponding to a fact by incorporating the co-occurring relation to alleviate the dynamic label number. Specifically, our NLN extracts the relation to acquire the number of labels by a mapping rule and a convolution layer.

Moreover, the above two difficulties are mutually influential. The content confusion will generate the worse label probability, which makes it hard to get the correct number of labels. Similarly, even if we get better label probability, the dynamic label number will cause the number of labels to be wrong, resulting

in lower prediction accuracy. Therefore, this paper proposes a relation learning hierarchical framework for multi-label charge prediction with two models, i.e., DMA and NLN. Specifically, we collect the provisions corresponding to different charges as external knowledge and embed them and facts jointly. In our framework, there are two levels, i.e., label probability level and label number level. In the label probability level, we gain a knowledge-free representation by using some state-of-the-art deep models as feature extractors and get a knowledge-aware representation by DMA. Then, we construct a classifier to generate an output probability by these representations. In addition, in the label number level, we build NLN to predict the number of labels of each fact and obtain its charges according to first l ranking sorting by label probability (l is the number of labels). In short, we consider the impacts of the similarity relation, the difference relation and the co-occurring relation under a unified framework to alleviate these two difficulties at the same time.

To best of our knowledge, it is the first work to learn both the label probability and the number of labels in a unified framework for multi-label charge prediction. Overall, the major contributions of our work are as follows:

- We design DMA to extract the similarity relation and the difference relation from the fact and the provisions. Therefore, we can alleviate content confusion and improve the charge prediction performance.
- We propose NLN which learns the co-occurring relation from labels to effectively mitigate the dynamic label number in the multi-label charge prediction.
- We conduct experiments on the biggest published Chinese law dataset to verify the effectiveness of our proposed framework.

In this paper, there is an overview of related work about our research in Sect. 2. The details of our framework with dynamic merging attention (DMA) and number learning network (NLN) will be elaborated in Sect. 3. In Sect. 4, this paper reports the experimental results about our framework with two models. Finally, Sect. 5 makes a conclusion of our work and discuss our future work.

2 Related Work

2.1 Multi-label Charge Prediction

Because of canonical format of legal text and the large amount of data, the law field becomes a hot topic of natural language processing (NLP) and charge classification is a fundamental task in the legal field. In recent years, Hu et al. introduced multiple attributes as additional features to enhance the connection between facts and charges [4]. With the rise of joint learning, there are also attempts to combine legal article recommendations with charge prediction for multi-task learning [6, 12]; Some studies are based on reading comprehension and hierarchical multi-label classification [11, 16]. Inspired by the success of attention mechanism in NLP task, Wang et al. handled charge prediction task by incorporating an attention mechanism [17]. Different from them, our paper studies how to joint the impacts of the similarity relation, the difference relation and the co-occurring relation in a unified framework.

2.2 External Knowledge

In the recent study, external knowledge has shown to be effective in neural networks for NLP tasks, including word embedding [1], reading comprehension [5] and risk prediction [13]. Using external knowledge to enhance model capabilities through the attention mechanism is more common in current researches. Hu et al. handled the charge prediction task by the soft attention [4] and Kim et al. introduced the external knowledge into his deep learning model by the transformer [8]. Following the way of using external knowledge, our paper learns the similarity relation and the difference relation from law provisions to improve the charge prediction performance through DMA.

2.3 Label Number Learning

Label number learning is to learn the number of labels based on an output probability. There are some researches for using naïve top-k strategy [10] and threshold strategy [17]. Yang et al. summarized the whole process of multi-label classification and used a Multilayer Perceptron (MLP) to get thresholds for label output probabilities [19]. Lenc et al. referred to a simple neural network that can solve the difficulty by top-k strategy [10]. These works can get the fixed threshold for each label or fixed k for the labels, which still can not give different numbers of labels for different facts. Our proposed NLN can predict different numbers of labels of the facts by learning the co-occurring relation and it works jointly with DMA to further improve the accuracy of charge prediction in our framework.

3 Our Framework

3.1 Problem Definition

In this paper, the contents of the law provisions corresponding to the charges are used as external knowledge which is specified the details in the Table 1. We let $\hat{X} = \{\hat{X}_1, \hat{X}_2, \hat{X}_3, \dots, \hat{X}_M\}$ denotes a set of external knowledge, where M is the number of law provisions. We obtain word vectors $X = \{x_1, x_2, x_3, \dots, x_i, \dots, x_I\}^T$ for a fact and $\hat{X}_m = \{\hat{x}_{m1}, \hat{x}_{m2}, \hat{x}_{m3}, \dots, \hat{x}_{mn}, \dots, \hat{x}_{mN}\}^T \in \mathbb{R}^{N \times d}$ for the m -th provision, where d is the dimensionality of embedded vectors, I and N are the number of words of the fact and the m -th provision, respectively. We set the output of our framework as R and each R is a list, such as $R = [1, 55, 120]$. Each number in the list represents the index of the corresponding charge [18]. For example, the index of arson is 55. On the basis of this problem definition, we propose our relation learning hierarchical framework with DMA model and NLN model.

3.2 Overview of Our Framework

As shown in Fig. 1, our framework is divided in two levels, namely label probability level and label number level.

Fig. 1. An overview of the relation learning hierarchical framework

Label Probability Level. For a fact X , using a feature extractor can gain its knowledge-free representation h_1 . For X and all provisions $\hat{X}$, we get a knowledge-aware representation h_2 by DMA. Hence, the knowledge-mixed representation h_3 can be obtained by mixing these representations. Then two probabilities P_1 and P_2 can be obtained respectively from two representations h_1 and h_3 by a series of non-linear mappings. In addition, we get the label probability $P = \{p_1, p_2, p_3, \dots, p_J\}$ by weighted summation of them, where the number of charges is J and each $p_j (1 \leq j \leq J)$ is a probability of the corresponding label.

Label Number Level. By using NLN model, the label probability P is converted to the label number probability $\hat{P} = \{\hat{p}_1, \hat{p}_2, \hat{p}_3, \dots, \hat{p}_K\}$, where the maximal number of labels is K and each $\hat{p}_k (1 \leq k \leq K)$ is a probability of the corresponding number of labels. And then we get the index of maximal value of the label number probability $\hat{P}$ as the number of labels l . Finally, we sort the label probability P and extract the first l labels as its result R .

3.3 Label Probability Level with Dynamic Merging Attention

In this section, we set the input of label probability level to be $D_\ell = \{X, \hat{X}, Y_\ell\}$, where $Y_\ell = \{y_1, y_2, y_3, \dots, y_J\}$ is the one-hot encoding of charges. The knowledge-free h_1 can be obtained by most deep learning models as feature extractors. We design a DMA model to introduce the legal provisions as external knowledge into our framework for knowledge-aware h_2 . As shown in the top I of Fig. 1, we calculate the m -th self-attention score $A_m \in \mathbb{R}^{I \times N}$ to represent the similarity relation about the words of the fact and the words of m -th provision and it is defined in Eq. (1).

$$A_m = \frac{X \hat{X}_m^T}{\sqrt{d}} \quad (1)$$

By concatenating self-attention scores, the 3-dimensional self-attention vector is defined in Eq. (2),

$$A = [A_1; A_2; A_3; \dots; A_M] \quad (2)$$

where $[\cdot]$ denotes a concatenation operation. As shown in the top II of Fig. 1, using a convolution layer obtains the difference relation $D = \{D_1; D_2; D_3; \dots; D_O\}$, where each $D_o \in \mathbb{R}^I (1 \leq o \leq O)$ is a difference representation. Specially, the length of the convolution kernel is N and the number of output channels is O . Due to different convolution kernels, our DMA dynamically learns different kinds of difference representations from the similarity relation, where different difference representations contain the difference information about different provisions. As shown in the top III of Fig. 1, we take a pooling layer for each output channel D_o to obtain each dimension h_{2o} of knowledge-aware representation $h_2 = \{h_{21}, h_{22}, h_{23}, \dots, h_{2O}\}$ as shown in Eq. (3).

$$h_{2o} = \max_{1 \leq i \leq I} D_{o,i} \quad (3)$$

There is a knowledge-mixed representation $h_3 = [h_1, h_2]$ by concatenating the knowledge-free representation h_1 and the knowledge-aware representation h_2 . The result of using non-linear mappings for h_3 is not equal to concatenating the results of using non-linear mappings for h_1 and h_2 shown in Eq. (4),

$$f(h_3) \neq [f(h_1), f(h_2)] \quad (4)$$

where $f(\cdot)$ denotes a non-linear mapping. Since the information obtained by nonlinear mapping is different, we convert the knowledge-free representation h_1 and the knowledge-mixed representation h_3 to two output probabilities, namely $P_1 = \{p_{11}, p_{12}, p_{13}, \dots, p_{1J}\}$ and $P_2 = \{p_{21}, p_{22}, p_{23}, \dots, p_{2J}\}$, respectively by non-linear mappings. Equation (5) indicates the label probability P is obtained by weighted summation of output probabilities P_1 and P_2 . Among them, α denotes the trade-off parameter balancing the two terms. W and b are the weights and biases of the non-linear mappings. σ denotes a sigmoid function and ϕ denotes an activation function, usually the hyperbolic tangent function or the Rectified Linear Unit (ReLU).

$$\begin{aligned}
P &= \alpha * P_1 + (1 - \alpha) * P_2 \\
&= \alpha * (\sigma(W_1^{(2)}\phi(W_1^{(1)}h_1 + b_1^{(1)}) + b_1^{(2)})) \\
&\quad + (1 - \alpha) * (\sigma(W_2^{(2)}\phi(W_2^{(1)}h_2 + b_2^{(1)}) + b_2^{(2)}))
\end{aligned} \tag{5}$$

Since our label probability level with DMA gets different output probabilities P_1 and P_2 , we define loss $\mathcal{L}_{prob}$ of this level as shown in Eq. (6),

$$\begin{aligned}
\mathcal{L}_{prob}(\theta) &= -\alpha * \sum_{j=1}^J (y_j * \ln p_{1j} + (1 - y_j) * \ln(1 - p_{1j})) \\
&\quad - (1 - \alpha) * \sum_{j=1}^J (y_j * \ln p_{2j} + (1 - y_j) * \ln(1 - p_{2j})) + \lambda \|\theta\|_2
\end{aligned} \tag{6}$$

where λ is the hyperparameter, θ represents all learnable parameters in the label probability level, and right norm is for the regularization of the parameters. We conduct the sensitivity study for α and the experimental result shows that optimal weight is 0.2.

3.4 Label Number Level with Number Learning Network

It is arbitrary to directly sort a label probability P and then use the top-k strategy or threshold strategy to get a result R . In the label number level, we concentrate on the co-occurring relation to propose a network called number learning network (NLN) which can better mitigate the phenomenon of the dynamic label number. By using this model, the number of labels of each fact can be determined automatically.

Generally, the input of NLN is $\mathcal{D} = \{P, \hat{Y}\}$, where $Y = \{\hat{y}_1, \hat{y}_2, \hat{y}_3, \dots, \hat{y}_K\}$ is the one-hot encoding of the number of labels. As shown in bottom of Fig. 1, there is a specific mapping in the first hidden layer. The neurons of the input layer are in one-to-one correspondence with the neurons of the first hidden layer, which has a fixed weight. The formula for general neural networks is given in Eq. (7) and our specific expression is defined in Eqs. (8) and (9),

$$p_j^{(1)} = \phi(\sum_{j'} w_{jj'} p_{j'}^{(0)} + b_j) \tag{7}$$

$$w_{jj'} = \begin{cases} 1, & j = j' \\ 0, & j \neq j' \end{cases} \tag{8}$$

$$p_j^{(1)} = \phi(p_j^{(0)} + b_j) \tag{9}$$

where the output of first hidden layer is $P^1 = \{p_1^{(1)}, p_2^{(1)}, p_3^{(1)}, \dots, p_J^{(1)}\} \in \mathbb{R}^{1 \times J}$ and $p_j^{(0)}$ is p_j of P . And then we repeat P^1 L times to get $P^2 = \{P^1; P^1; \dots; P^1\} \in \mathbb{R}^{L \times J}$. It is easy to use a convolution layer for P^2 to get the co-occurring relation $C = \{C_1; C_2; C_3; \dots; C_{O'}\}$ and finally flat it, where O' is the number

of output channels of the convolution layer. Because of different convolution kernels, different co-occurring relations between different labels can be obtained. After these steps, the output probability $\hat{P}$ can be obtained by MLP. Besides, NLN is not complex to prevent gradient explosion or gradient vanishing.

In this level, there is a fine-tune and a screening work to learn the co-occurring relation through NLN, as shown in the bottom of Fig. 1. Each dimension of $\hat{P}$ represents the probability of the number of labels of a fact. For each fact, we select the index l of the maximum value from its label probability $\hat{P}$. Through first l ranking, the result R is obtained on the basis of the label probability P .

3.5 Loss Function

Since our framework consists of two levels, there are two losses, i.e., $\mathcal{L}_{prob}$ and $\mathcal{L}_{num}$. The $\mathcal{L}_{prob}$ is the loss of the label probability level and the $\mathcal{L}_{num}$ is the loss of the label number level. During the experiment, we first train the models of the label probability level, and then train the model of the label number level. The overall loss of our framework is defined in Eq. (10),

$$\begin{aligned}
 \mathcal{L} &= \min(\mathcal{L}_{prob}) + \min(\mathcal{L}_{num}) \\
 &= \min(-\alpha * \sum_{j=1}^J (y_j * \ln p_{1j} + (1 - y_j) * \ln(1 - p_{1j})) \\
 &\quad - (1 - \alpha) * \sum_{j=1}^J (y_j * \ln p_{2j} + (1 - y_j) * \ln(1 - p_{2j})) + \lambda \|\theta\|_2) \\
 &\quad + \min(\sum_{k=1}^K (\hat{y}_k * \ln \hat{p}_k + (1 - \hat{y}_k) * \ln(1 - \hat{p}_k)) + \beta \|\theta'\|_2)
 \end{aligned} \tag{10}$$

where β is the hyperparameter, θ' represents all learnable parameters in the label number level, and the rightest norm is for the regularization of the parameters.

4 Experiment

In this section, we introduce the dataset, the evaluation measure, the experimental configuration, and all the baselines. We compare our models with those baselines under our framework with the aim of answering the following research questions:

RQ1: How is the effectiveness of our DMA?

RQ2: How is the effectiveness of our NLN?

RQ3: How is the effectiveness of our framework with DMA and NLN?

4.1 Dataset and Evaluation Measures

The biggest published real-world Chinese law dataset is from Cail2018 [18] by the Supreme People’s Court of China. It consists of 154,592 samples for training and 32,508 samples for testing. Each sample contains a complex legal text description and three tasks. One of given task is multi-label charge prediction and the number of charge labels is 202. The details of dataset are illustrated in the Table 2. The ratio of the training dataset to the test dataset is about 5:1 given by Cail2018 [18]. Because this dataset is provided by Chinese ‘fa yan bei’ competition¹, we follow the evaluation measure in this competition, i.e., micro-F1 and macro-F1².

Table 2. The details of Cail2018 dataset

Dataset	Training	Valid	Test	Label	
Number	154592	17131	32508	202	
Label number	Num = 1	Num = 2	Num = 3	Num = 4	Num > 4
Number of training data	120475	30831	2914	288	96

4.2 Baselines

In this paper, we propose a relation learning hierarchical framework with DMA and NLN. Our framework adopts deep models as feature extractors, such as TextCNN [9], CRNN [14], DPCNN [7], CNN&Attention [15], Bi-GRU [2], Bi-LSTM [3], etc. For the RQ1, we compare our proposed DMA with the soft-attention and the transformer which are baselines [4, 8]; For the RQ2, we compare our proposed NLN with the threshold strategy and the top-k strategy which are baselines [10, 17]; For the RQ3, we compare our proposed DMA and NLN with DMA and the threshold strategy which is the label decision baseline in our framework.

4.3 Experiment Configuration

In the experiment, we adopt the word2vec directly after the legal text segmentation, so that the text is mapped into a 512-dimensional vector and the number of words of a fact and knowledge are 400 and 85, respectively. The width of a convolution kernel is generally 3, and the widths of convolution kernels of TextCNN are respectively 1, 2, 3, 4, 5. Loss functions of both levels are *Cross Entropy* during training. Because the number of samples which has the number of labels greater than 4 is too few in the Table 2, we simply set their number of labels as 4.

¹ <http://cail.cipsc.org.cn/>.

² https://en.wikipedia.org/wiki/Precision_and_recall.

Table 3. Results on RQ1 with DMA

	Type	micro-F1	macro-F1
TextCNN [9]	w/o attention	0.8459	0.7454
	w/ soft-attention [4]	0.8054	0.6923
	w/ transformer [8]	0.8473	0.7498
	w/ our DMA	0.8526	0.7657
CRNN [14]	w/o attention	0.7767	0.6119
	w/ soft-attention	0.7372	0.5823
	w/ transformer	0.8011	0.6732
	w/ our DMA	0.8391	0.7069
DPCNN [7]	w/o attention	0.8055	0.6503
	w/ soft-attention	0.7895	0.6258
	w/ transformer	0.8114	0.6883
	w/ our DMA	0.8228	0.7180
CNN+Attention [15]	w/o attention	0.8269	0.7136
	w/ soft-attention	0.7326	0.6587
	w/ transformer	0.8126	0.7047
	w/ our DMA	0.8237	0.7368
Bi-GRU [2]	w/o attention	0.7548	0.5816
	w/ soft-attention	0.7111	0.5925
	w/ transformer	0.7756	0.6291
	w/ our DMA	0.8229	0.7143
Bi-LSTM [3]	w/o attention	0.7754	0.6017
	w/ soft-attention	0.7098	0.5753
	w/ transformer	0.7726	0.5657
	w/ our DMA	0.7810	0.5768

4.4 Results on RQ1 by DMA

As shown in the Table 3, our DMA is better than other attention mechanisms in terms of micro-F1 and macro-F1. For example, using Bi-GRU as a feature extractor, our DMA performs best with 0.8229 micro-F1 score. In comparison, soft-attention gains 0.7111 and the more popular transformer gains 0.7756. As shown in the top I of Fig. 1, self-attention provides similarity information for a fact and external knowledge. As shown in the top II and II of Fig. 1, the convolutional layer and the pooling layer dynamically extract some differences from this information. Hence, our DMA can obtain the knowledge-aware representation which contains an understanding for a fact. Compared to our DMA, through external knowledge, soft-attention can only emphasize parts of a fact and it does not obtain the difference relation. Therefore, using soft-attention does not improve our framework and has a certain degree of negative impact. Compared to these attention mechanisms, we observe that our DMA significantly outperform others by dynamically learning the similarity relation and the difference

relation. In addition, it also indicates that it is significantly helpful for enhancing our framework with DMA to apply most deep learning models as feature extractors.

4.5 Results on RQ2 by NLN

In this section, we conduct experiments to show the performance of our NLN in terms of micro-F1 and macro-F1. As shown in the Table 4, since using the top-k strategy is worst on the both measures, we select the threshold strategy for comparison. For example, using TextCNN as a feature extractor, the score of our NLN model is 2.1% better than the threshold strategy on the macro-F1 metrics. For the label probability, our NLN contains a specific mapping like a fine-tune and a convolution layer to acquire the co-occurring relation. The experimental results also demonstrate the co-occurring relation is helpful for the label number learning.

Table 4. Results on RQ2 with NLN

	Type	micro-F1	macro-F1		Type	micro-F1	macro-F1
TextCNN	w/ top-k	0.7739	0.6252	CNN+Attention	w/ top-k	0.7039	0.5759
	w/ threshold	0.8459	0.7454		w/ threshold	0.8269	0.7136
	w/ NLN	0.8595	0.7618		w/ NLN	0.8344	0.7210
CRNN	w/ top-k	0.6390	0.5463	Bi-GRU	w/ top-k	0.6048	0.5054
	w/ threshold	0.7767	0.6119		w/ threshold	0.7548	0.5816
	w/ NLN	0.7804	0.6158		w/ NLN	0.7576	0.5844
DPCNN	w/ top-k	0.6490	0.5862	Bi-LSTM	w/ top-k	0.6314	0.4849
	w/ threshold	0.8055	0.6503		w/ threshold	0.7754	0.6017
	w/ NLN	0.809	0.6549		w/ NLN	0.7796	0.6020

4.6 Results on RQ3 by DMA and NLN

As shown in Fig. 2, our framework combined with DMA and NLN performs better than other baselines. For example, for the DPCNN as a feature extractor, our framework with the models outperforms the other by 4% and 6% on the micro-F1 and macro-F1, respectively. Since our framework is divided into two levels, each level uses a corresponding model to extract different relations, which causes better prediction results. The label probability level obtains a better label probability from the extracted relations, which provides a good basis for the label number prediction. The label number level improves the accuracy of the number of labels by the co-occurring relation and the better label probability. In this unified framework, combining these two levels finally yields better results. According Table 3 and Fig. 2, our framework with DMA and NLN outperforms these baselines by more than 3%–23%. For instance, using the Bi-GRU as a feature extractor, macro-F1 of our framework with DMA and NLN is 0.7163 which is 23% more than that of only using threshold strategy. It indicates our

DMA and NLN further enhance our framework. In the same way, it shows that we alleviate the above two difficulties by learning these relations together under a unified framework.

Fig. 2. Results on RQ3 with DMA and NLN

5 Conclusions and Future Work

In this paper, we propose a relation learning hierarchical framework with the two models, namely the dynamic merging attention (DMA) and the number learning network (NLN). Through learning the similarity relation, the difference relation and the co-occurring relation under a unified framework, it can effectively alleviate the content confusion and the dynamic label number difficulties on the multi-label charge prediction. By testing on real-world datasets, it verifies that our framework with the two models outperforms popular baselines significantly. In the future work, we plan to adopt this framework in the law article recommendation and term of penalty prediction task.

Acknowledgement. This work was partially supported by the Excellent Dissertation Cultivation Funds of Wuhan University of Technology (2018-YS-063), the National Natural Science Foundation of China (Grant No. 61602353) and Hubei Provincial Natural Science Foundation of China (Grant No. 2017CFA012).

References

1. Chen, Z., et al.: Revisiting word embedding for contrasting meaning. In: ACL, pp. 106–115 (2015)
2. Cho, K., et al.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: EMNLP, pp. 1724–1734 (2014)
3. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* **9**(8), 1735–1780 (1997)
4. Hu, Z., Li, X., Tu, C., Liu, Z., Sun, M.: Few-shot charge prediction with discriminative legal attributes. In: COLING, pp. 487–498 (2018)

5. Huang, Y., Yang, X., Zhuang, F., Zhang, L., Yu, S.: Automatic Chinese reading comprehension grading by LSTM with knowledge adaptation. In: Phung, D., Tseng, V.S., Webb, G.I., Ho, B., Ganji, M., Rashidi, L. (eds.) PAKDD 2018. LNCS (LNAI), vol. 10937, pp. 118–129. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-93034-3_10
6. Jiang, X., Ye, H., Luo, Z., Chao, W., Ma, W.: Interpretable rationale augmented charge prediction system. In: COLING, pp. 146–151 (2018)
7. Johnson, R., Zhang, T.: Deep pyramid convolutional neural networks for text categorization. In: ACL, pp. 562–570 (2017)
8. Kim, Y., Lee, H., Jung, K.: Attnconvnet at semeval-2018 task 1: attention-based convolutional neural networks for multi-label emotion classification. In: Proceedings of The 12th International Workshop on Semantic Evaluation, pp. 141–145 (2018)
9. Kim, Y.: Convolutional neural networks for sentence classification. In: EMNLP, pp. 1746–1751 (2014)
10. Lenc, L., Král, P.: Word embeddings for multi-label document classification. In: RANLP, pp. 431–437 (2017)
11. Long, S., Tu, C., Liu, Z., Sun, M.: Automatic judgment prediction via legal reading comprehension. In: Chinese Computational Linguistics, pp. 558–572 (2019)
12. Luo, B., Feng, Y., Xu, J., Zhang, X., Zhao, D.: Learning to predict charges for criminal cases with legal basis. In: EMNLP, pp. 2727–2736 (2017)
13. Ma, F., Gao, J., Suo, Q., You, Q., Zhou, J., Zhang, A.: Risk prediction on electronic health records with prior medical knowledge. In: KDD, pp. 1910–1919 (2018)
14. Shi, B., Bai, X., Yao, C.: An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(11), 2298–2304 (2017)
15. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 6000–6010 (2017)
16. Wang, P., Fan, Y., Niu, S., Yang, Z., Zhang, Y., Guo, J.: Hierarchical matching network for crime classification. In: SIGIR, pp. 325–334 (2019)
17. Wang, P., Yang, Z., Niu, S., Zhang, Y., Zhang, L., Niu, S.: Modeling dynamic pairwise attention for crime classification over legal articles. In: SIGIR, pp. 485–494 (2018)
18. Xiao, C., et al.: CAIL2018: a large-scale legal dataset for judgment prediction. arXiv preprint [arXiv:1807.02478](https://arxiv.org/abs/1807.02478) (2018)
19. Yang, Y., Gopal, S.: Multilabel classification with meta-level features in a learning-to-rank framework. *Mach. Learn.* **88**(1–2), 47–68 (2012)
20. Zhong, H., Guo, Z., Tu, C., Xiao, C., Liu, Z., Sun, M.: Legal judgment prediction via topological learning. In: EMNLP, pp. 3540–3549 (2018)