IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 29, 2021, accepted June 16, 2021, date of publication June 21, 2021, date of current version June 30, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3091323

ADAN: An Intelligent Approach Based on Attentive Neural Network and Relevant Law Articles for Charge Prediction

DAPENG LI[1,3], QIHUI ZHAO[2], JIAN CHEN[3], AND DAZHE ZHAO[1]

[1]School of Computer Science and Engineering, Northeastern University, Shenyang 110819, China
[2]Software College, Northeastern University, Shenyang 110819, China
[3]Neusoft Group Research, Northeastern University, Shenyang 110819, China

Corresponding author: Qihui Zhao (1910459@stu.neu.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFC0830601; in part by the National Natural Science Foundation of China under Grant 61772126, Grant 61972079, and Grant 61872073; and in part by the Fundamental Research Funds for the Central Universities under Grant N2016004 and Grant N2016002.

ABSTRACT The charge prediction task aims to predict appropriate charges for a given legal case automatically, which still confronts some challenging problems such as performance improvement and confusing charges issue. In this paper, inspired by the impressive success of deep neural networks in legal intelligence field, we present an end-to-end framework named law article deduplication attention neural network, ADAN, to address these problems. The incorporation of hierarchical sequence encoder and attention mechanism is employed to learn better semantic representations of fact description texts. To distinguish confusing charges, we use the relevant law articles of a given case as auxiliary information, and propose a novel difference aggregation mechanism among similar law articles for extracting effective distinguishable features. The experimental results on real-world datasets show that the performance of our proposed model is significantly better than existing methods on all evaluation metrics.

INDEX TERMS Charge prediction, hierarchical attention mechanism, bidirectional gated recurrent unit, text classification.

I. INTRODUCTION

In recent years, the application of artificial intelligence technology in the judicial field such as legal judgment prediction (LJP) has attracted increasing research attention [1]. As a representative subtask of LJP, charge prediction aims to predict appropriate charges by analyzing textual fact description of a legal case, which is useful in many real-world scenarios and has broad application prospects. For example, it can provide decision-making support for legal professionals (e.g., judges, prosecutors and lawyers) and improve their work efficiency. Furthermore, it can also help ordinary people who lack legal expertise to have a preliminary understanding of the legal case.

However, due to the complexity of judicial trial, it is not trivial to predict the appropriate charges by using artificial intelligence method: (1) there exists many confusing charge pairs in Chinese Criminal Law, such as (*robbery, snatching*) and (*bribery, bribery of non-official servant*). For each confusing charge pair, the circumstances of crime in corresponding cases usually have high similarity to each other and exist some subtle differences merely. For example, the difference between bribery and bribery of non-official servant lies in whether the subject of the crime is official servant or not. (2) The criminal facts in a specific case may involve multiple charges, which increases the difficulty of charge prediction. (3) There is a strong logical dependence between charge determination and relevant law articles. In the civil law system, including mainland China, the judges make decisions of criminal cases based on statutory laws, rather than decisions of precedent cases. As shown in Fig.1, a judgement document in China always includes case facts, relevant law articles (in the court view part) to support the judgement decision (the red text section in the given example). In the given case, the defendant was convicted of theft according to the criminal facts and relevant law articles. Therefore, how to use relevant law articles to improve the performance of charge prediction is another challenge.

The associate editor coordinating the review of this manuscript and approving it for publication was Michael Lyu.

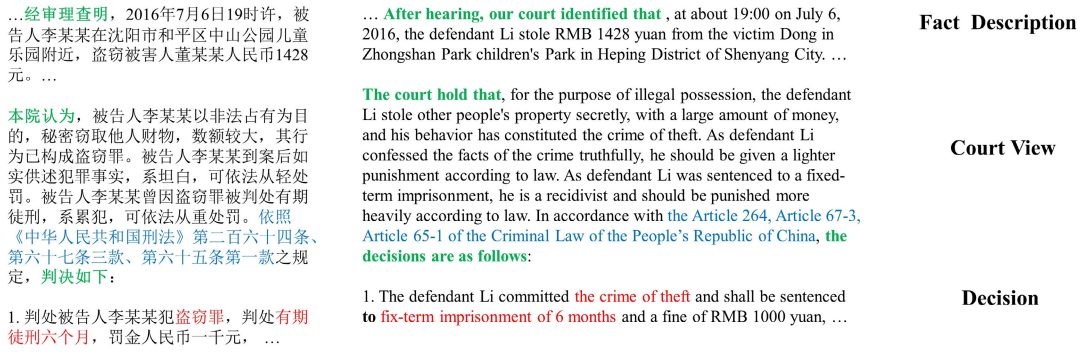


FIGURE 1. An example judgement document excerpt of a criminal case in our dataset.

Charge prediction has been studied for decades and the majority of existing works attempt to resolve the charge prediction task by formalizing it as a text classification problem. Early efforts either extracted shallow textual features [2], [3] such as key words, phrase, and term frequency, or adopted manually designing discriminative features [4], [5], to complete the charge prediction task. These methods have some disadvantages that cannot be solved, (1) legal expertise is needed for manual annotation or feature designing; (2) the generalization ability is poor; (3) a great quantity of time, manpower and materials are consumed on the processing of massive cases. With the improvement of computing power and the development of artificial intelligence, researchers tend to employ deep neural networks to extract semantic features from case documents automatically, and have achieved significant improvements in the legal intelligence field, such as semantic feature extraction [6], legal reading comprehension [7] and court view generation [8]. To resolve the issues of charge prediction mentioned above, several existing works attempt to incorporating law articles into the prediction model as auxiliary information. Luo *et al.* [9] attentively encoded relevant law articles extracted based on a case's fact description to an aggregated article embedding, then they concatenated article embedding with fact embedding and put them to a multi-label classifier. Nevertheless, its ability of distinguishing confusing charges still needs to be improved. To our knowledge, the phraseology of law articles is extremely refined, and the differences between law articles of confusing charges can be subtle. As show in Fig.2, bribery (article 385) and bribery of non-official servant (article 163) are a pair of confusing charges, and their articles have highly semantic similarity (i.e., the red text is the same sentences). Therefore, study [9] may result in similar article embeddings by encoding relevant articles directly, which are ineffective to distinguish confusing charges.

In order to address these problems, we propose an end-to-end framework for charge prediction, law article deduplication attention neural network, ADAN. We construct a two-tier sequence encoder by incorporating Bidirectional Gated Recurrent Units (Bi-GRU) [10] and hierarchical attention mechanism [11]. The encoder captures the importance of different words and sentences for charge prediction and

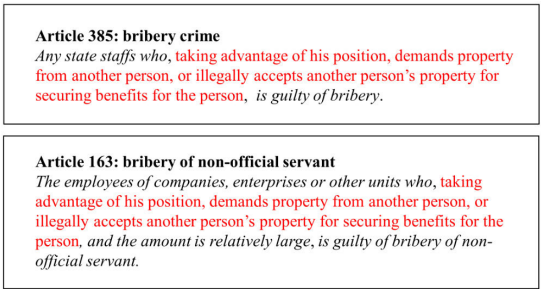


FIGURE 2. Examples of law articles of confusing charges.

generates a discriminative embedding of textual fact description. Considering the logical dependences between charge determination and relevant law articles, the relevant law articles of a given case is used as auxiliary information for charge prediction. Different from existing studies, we employ the SVM binary classifier to select top k candidate articles, and further adopt a difference aggregation mechanism and Bi-GRU to better understand the law articles and extract the differences among similar law articles. More importantly, we incorporate the article extraction task within our charge prediction framework, which not only provides another way to distinguish confusing charges more effectively, but also serves as legal basis to support the final decision.

The main contributions in this paper are summarized in the following points:

1) This study proposes an innovative framework, ADAN, to perform charge prediction task. ADAN employs a hierarchical attention neural network, which is beneficial for learning better representation from textual fact description.
2) We propose a novel difference aggregation mechanism among similar law articles to extract distinguishable features for distinguishing confusing charges effectively.
3) We carry out a series of experiments on real-world datasets. The experimental results show that the performance of our proposed model is significantly better than existing methods on all evaluation metrics.

The rest of this paper is organized as follows. Section2 gives the related research progress in recent years.

In section3, we describe all details of ADAN framework. Section4 provides the experimental results and analyses. Finally, conclusion and further work are presented in section5.

II. RELATED WORK

In early studies, researchers tend to rely on shallow textual features or manually designing tags to accomplish charge prediction task. Liu *et al.* [2] aimed to classify 12 common and frequently happened criminal crime based on a k-Nearest Neighbors algorithm with shallow textual features. Lin *et al.* [4] employed manually designing legal factors as the input of classifier to predict the judgement results. Boella *et al.* [12] exploited Support Vector Machine (SVM) classifier to classify legal cases and estimate which legal field they belong to. These early attempts are difficult to be applied because of the shortcomings of generalization ability and heavy manpower.

With the advent of development of natural language processing (NLP) and artificial intelligence technology, more and more deep learning methods had been widely used in the legal intelligence field. Wei *et al.* [13] applied convolution neural network (CNN) to legal text classification and their study illustrated that deep learning model performs much better than traditional machine learning methods in dealing with massive training dataset. Ye *et al.* [8] attempted to construct a Seq2Seq model, which expressed interpretable charge prediction as a court view generation problem. Guo *et al.* [14] incorporated Bidirectional Long Short-Term Memory (Bi-LSTM) with tensor decomposition to predict multiple accusation judgement in legal cases. Liu *et al.* [15] introduced charge keywords extracted by TF-IDF (term frequency-inverse document frequency) and TextRank for charge prediction. Zhong *et al.* [16] proposed a model based on reinforcement learning to visualize the LJP process and give interpretable judgments, which is also heavily dependent on expert knowledge.

Considering the correlation of different subtasks of LJP, Zhong *et al.* [17] proposed a topological multi-task framework that formalized the dependencies among different subtasks as a Directed Acyclic Graph for deep neural network learning. Li *et al.* [18] designed a multichannel neural network model framework with attention mechanism to complete entire LJP tasks. Xu *et al.* [19] proposed a new unified LJP model which capture the attention weights of different terms of penalty and the position of defendant. Yang *et al.* [20] presented a multi-perspective bi-feedback network with the word collocation attention mechanism for LJP task. They designed a multi-perspective forward prediction and backward verification framework to effectively utilize result dependencies among multiple LJP subtasks.

In summary, existing studies have achieved certain progress in legal judgement prediction field. Nevertheless, there are still some challenging problems for charge prediction, including confusing charge issue and performance improvement. This is why ADAN is introduced in this study.

III. OUR METHOD

A. PROBLEM FORMULATION

We introduce some terminologies and notations, and then formulate the charge prediction task.

Law Cases. As depicted in Fig.1, each law case consists of a fact description and several charges. The fact description is part of a judgement document, denoted by f. One or more charges are recorded in the decision part, denoted by $C = \{c_1, \ldots, c_m\}$, m is the number of charges.

Law Articles. There are two main types of statutory law in the Criminal Law of the People's Republic of China: basic law articles and auxiliary law articles. Basic articles define the common rules that must be complied with in the determination of charges and penalties. The purpose of our research is charge prediction, so we only use basic law articles, and remove the contents related to penalty in law articles which have little relevance and admittedly noisy for charge prediction. Formally, we denote the statutory law as a set of law articles $L = \{l_1, \ldots, l_n\}$ where n is the number of law articles. Each law article is represented as a text document too.

Charge Prediction. A completed law case can be represented by a tuple $[f, C]$. Based on a training dataset $D = \{(f, C)_i\}_{i=1}^{q}$ of size q, we aim to train a model $F(.)$ that can predict the charges for any test law case. Given a fact description of test law case f_{test}, $F(f_{test}, L) = \hat{C}$. Each charge is related to only one law article; therefore, charge prediction has been equivalent to article prediction.

B. AN OVERVIEW OF ADAN MODEL

As shown in Fig.3, in our framework ADAN, we convert each word in preprocessed fact description to word embedding. Then we input them into a hierarchical Bi-GRU encoder to generate the fact embedding v_f. The encoder has a two-tier structure: word-level and sentence-level. We introduce global word-level and sentence-level context vectors, i.e., c_{fw} and c_{fs}, to attentively capture informative words and sentences. Concurrently, fact description text is also used to extract top k relevant law articles by a module called Relevant Article Extractor (RAE). We employ the Article Difference Aggregator (ADA) to attentively aggregate the differences among these candidate articles and generate the article-side embedding v_a. Specially, context vectors for attention calculation in ADA are dynamically generated by v_f. Finally, v_f and v_a are concatenated and input to a softmax classifier to generate the predicted charge distribution P_c.

C. LEARNING FACT REPRESENTATION

1) TEXT PREPROCESSING

Because all judgement documents in our dataset are written in Chinese, we employ jieba[1] tool with a legal custom dictionary for word segmentation first. Next, some insignificant words such as modal particles are filtered by stop words list to avoid possible interference. Finally, we add sentence-end tags in

[1] https://github.com/fxsjy/jieba.

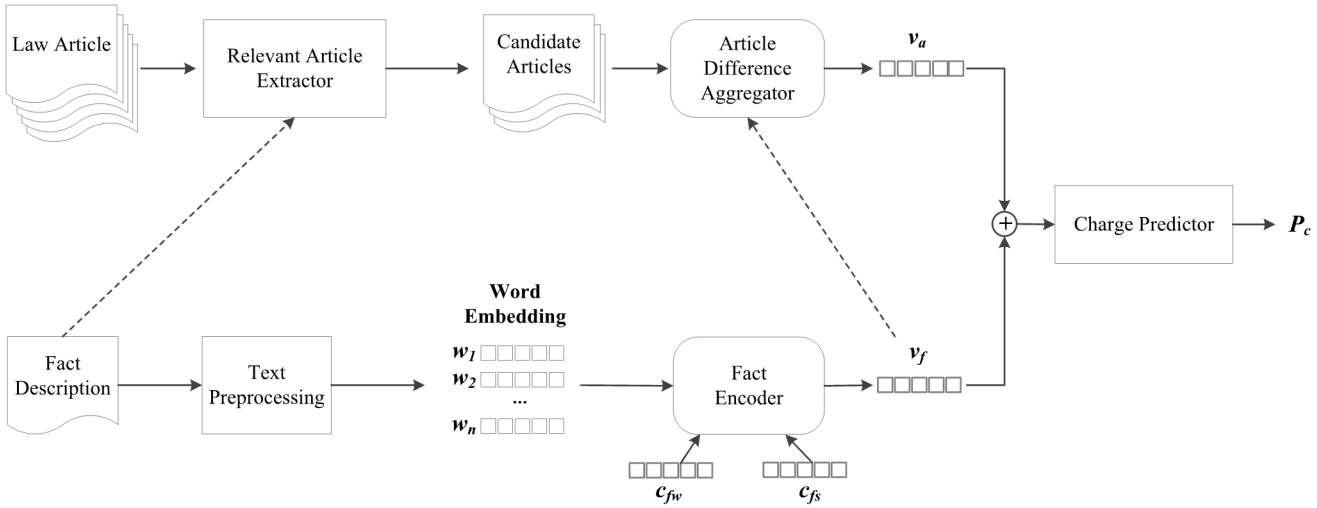


FIGURE 3. An overview of the ADAN framework.

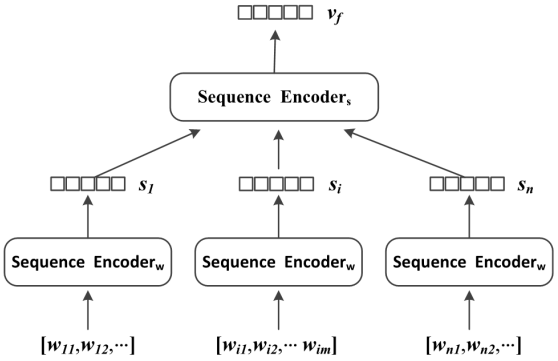


FIGURE 4. Hierarchical sequence encoder.

the cleaned fact descriptions. In this paper, word2vec and CBOW (Continuous Bag-of-Words) model are used to map each word in the text (including fact descriptions and law articles) into the same vector space [21], [22].

2) ATTENTIVE FACT ENCODER

The fact description of a law case has a hierarchical structure obviously, a sentence can be represented as a sequence of words, and the sequence of sentences constitutes a fact description. Inspired by [23], we adopt a two-tier structure to extract discriminative features with several word-level sequence encoders and a sentence-level sequence encoder, as show in Fig.4. Word embeddings of a sentence are fed into word-level encoders to generate sentence embeddings, and then sentence-level sequence encoder aggregate each sentence embedding to construct a fact embedding v_f.

Recurrent Neural Network (RNN) is a type of artificial neural network designed to recognize patterns in sequences of data. Typical RNNs include the traditional RNN, LSTM, GRU and their variants. A common LSTM unit consists of a memory cell and three gates. GRU unit is similar to LSTM but simpler, which removes the memory cell and one gate. Compared with LSTM, GRU has fewer parameters and faster convergence speed, which can save a lot of time in the case of large training data, so we employ Bi-GRU as the sequence encoder of both levels in this paper. Bi-GRU encodes the contexts of each element in two opposite directions by using a forward and a backward GRU [24], and then concatenates the states of both GRUs. Given a sequence $[x_1, x_2, \ldots, x_T]$where x_t is the embedding of element t, we can obtain the hidden state of Bi-GRU h_t at position t is:

$$h_t = [h_{ft}, h_{bt}] \tag{1}$$

where $\boldsymbol{h}_{ft}$ and h_{bt} are the forward and backward hidden states respectively.

However, directly using Bi-GRU outputs has the defect of treating informative elements equally with useless ones. The work of [23] inspired us to add an attention mechanism with hierarchical sequence encoder. As show in Fig.5, global context vectors c_w and c_s are introduced to calculate attention values in word-level and sentence-level respectively [25]. Note that they are global context vectors, where they are initialized randomly and learned during the training process.

Given the word-level Bi-GRU hidden state sequence $[h_{i1}, h_{i2}, \ldots, h_{iM}]$, a sequence of word-level attention values $[\alpha_{i1}, \alpha_{i2}, \ldots \alpha_{iM}]$, where $\alpha_{ij}\epsilon[0, 1]$ and $\sum_j \alpha_{ij} = 1$.The sentence-level vector s_i is calculated as:

$$\alpha_{ij} = \frac{exp(tanh((\mathbf{W}_w h_{ij})^T c_w))}{\sum_i exp(tanh((\mathbf{W}_w h_{ij})^T c_w))} \tag{2}$$

$$s_i = \sum_j \alpha_{ij} h_{ij} \tag{3}$$

Computational process of sentence-level attention values is similar to that of word-level:

$$\alpha_{ij} = \frac{exp(tanh((\mathbf{W}_s h_i)_i^T c_s))}{\sum_i exp(tanh((\mathbf{W}_s h_t)_i^T c_s))} \tag{4}$$

$$d = \sum_i \alpha_i h_i \tag{5}$$

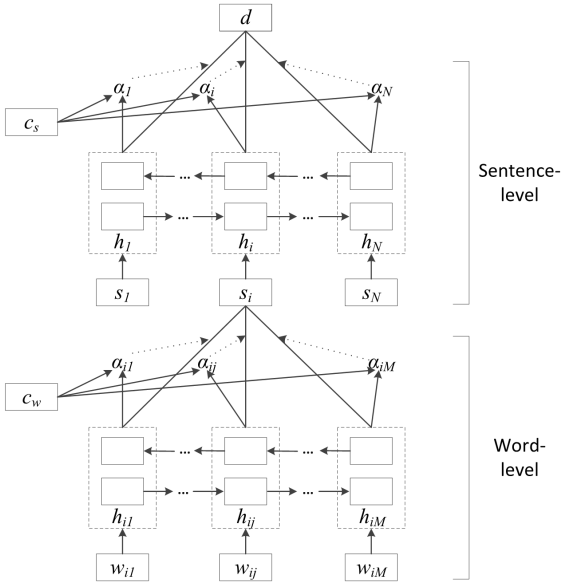


FIGURE 5. Hierarchical attention mechanism.

where $\mathbf{W}_w$ and $\mathbf{W}_s$ are word-level and sentence-level trainable weighted matrix respectively, and d is the document-level vector.

D. USING LAW ARTICLES

The method of using law articles to support charge prediction is designed based on the following observation: in general, dissimilar law articles are easy to distinguish due to the existence of sufficient distinctions, but it is difficult to distinguish similar law articles due to the lack of effective features. We adopt a two-step approach to deal with law articles. First, we should filter out a large number of irrelevant law articles based on fact description. Then, we extract distinguishable and crucial information from relevant law articles and generate article-side embedding attentively.

1) RELEVANT ARTICLE EXTRACTOR

As mentioned above, we need to extract the relevant law articles related to the given case first. Considering the large number of training law cases and law articles, a fast and easy-to-scale classifier should be introduced to obtain the correlation between the given case and law articles. So, we employed word-based SVM to build the binary classifier for each law article in our study with high efficiency and scalability, whereby the output of each classifier represents the relevance of this law article to the given case. Therefore, the number of binary classifiers is the same as the number of law articles in our dataset. When more articles are considered, we can simply add more binary classifiers accordingly, with the existing classifiers untouched. Specifically, word-level TF-IDF vectors, chi-square for feature selection and linear kernel are used for binary classification.

To evaluate the SVM relevant article extractor, a parameter k is used to control the number of relevant law articles. Obviously, the value of k directly affects the performance of relevant article extraction.

2) ARTICLE DIFFERENCE AGGREGATOR

Previous work [9] directly encoded extracting relevant articles and produce an aggregated article embedding attentively. As mentioned above, this method cannot extract the distinguishable features from similar law articles. For this reason, we proposed an innovative framework to aggregate the differences among similar law articles. The design concept is based on a simple fact: if the texts of two confusing law articles have high similarity, we should reduce similarity by removing duplicated texts between them. The idea is simple, but in practice, we must elaborately design the principle of text deduplication, otherwise the leftover texts may also have similarity and generate misleading information.

Before article deduplication, all punctuation marks are deleted in the law article texts and form a short text collection made up of law articles, denoted by L. Next, we extract TF-IDF vectors of all law article text in L, and compute the cosine similarity between each pair of TF-IDF vectors. Then, for each law article, we rank other law articles according to their cosine similarity, and perform text deduplication operation based on the textual contents of top i articles. It is worth pointing out that we do not use the deduplicated law article for comparison. Because excessive text deduplication may result in too little information remained, which is not conducive to the subsequent effective feature extraction, we only delete three or more consecutive identical words (i.e., Chinese characters) when comparing a pair of law articles. After completing the deduplication operations one by one, a new short text collection composed of deduplicated law articles for article differences aggregation is generated, denoted by L'. Finally, each word in deduplicated law article text is converted to word embedding.

Since we have transformed the law articles into short texts without punctuations, as show in Fig.6, different from the fact decoder, both article encoder and article aggregator are a bi-directional GRU network. After extracting top k relevant articles by RAE module, we have obtained the serial numbers of these articles. Instead of the original text, the deduplicated text with the same serial number from L' are input into the article encoder respectively. Specially, article encoders attentively produce article embeddings a_i by using context vector c_{ae}, which is dynamically generate for each case according to its corresponding fact embedding v_f:

$$c_{ae} = W_e V_f + b_e \tag{6}$$

where W_e and b_e are the weight matrix and the bias. By using dynamic context vector, our model can pay more attention to the informative words with respect to the specific fact description, rather than just selecting generally informative ones. The produced article embedding a_i contains rich distinguishable features that can distinguish from other confusing articles.

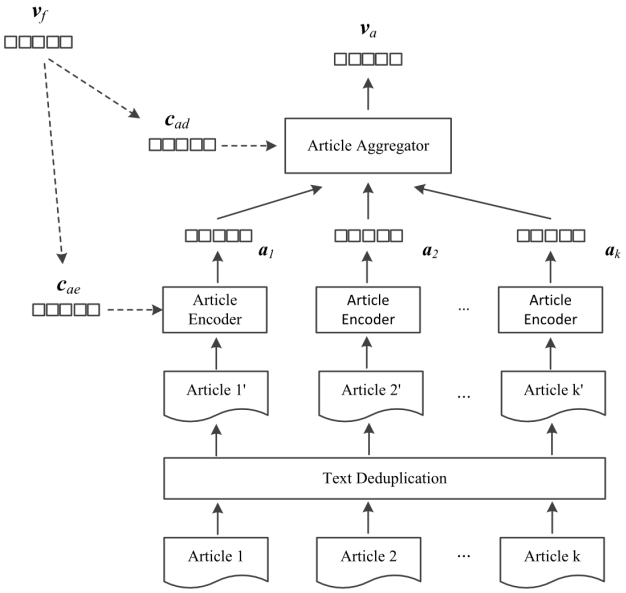


FIGURE 6. Article difference aggregator.

The article aggregator aims to obtain a more refined embedding v_a at the article-level by attentively selecting highly relevant law articles for charge prediction. We also dynamically generate the context vector c_{ad} by:

$$c_{ad} = W_d v_f + b_d \tag{7}$$

where W_d and b_d are the weight matrix and the bias. The attention values can be seen the relevance of each articles to the given case.

E. THE OUTPUT

We concatenate the fact-side embedding v_f and the aggregated article-side embedding v_a, and pass them to a two full connection layers to generate a new vector v', which is input to a softmax classifier to generate the predicted charge distribution $P_c = [p_{c1}, p_{c2}, \ldots p_{cG}]$, each $p_{cg} \epsilon [0, 1]$ represents the predicted probability of charge g being applicable for a given case, and G is the number of charges in our dataset.

$$P_c = softmax(W_c v' + b_c) \tag{8}$$

where W_c and b_c are the trainable weight matrix and bias respectively. We determine a threshold τ by using validation dataset, and the charges with predicted probability higher than are regarded as positive predictions. The loss function for measuring the prediction loss is cross-entropy:

$$Loss = -\sum_{1}^{G} r_{cg} log(p_{cg}) \tag{9}$$

where p_{cg} and r_{cg} are the predicted probability and ground-truth of charge g respectively. The ground-truth distribution r_c is produced by setting $r_{cg} = \frac{1}{m}$ for positive labels and $r_{cg} = 0$ for negative labels, where m is the number of positive labels.

IV. EXPERIMENTS

A. EXPERIMENTAL DATASETS

We construct our experimental datasets on the basis of CAIL2018 [26]. Totally, CAIL2018 contains 2,676,075 criminal cases, which are annotated with 183 criminal law articles and 202 criminal charges. Each case in CAIL2018 consists of fact description and corresponding judgement results including relevant law articles (basic law articles only), charges, and prison terms. Before the experiment was conducted, we do some preprocess on the datasets. We filter out the cases with fewer than 10 meaningful words first. Next, the cases with charges appeared less than 100 times are removed. Thirdly, we delete some law article labels which are not related to specific charges such as article 383 and 386. Finally, we obtain two refined CAIL2018 datasets called RCAIL-L and RCAIL-S. RCAIL-S dataset contains a large number of cases with easily confused charges such as *bribery* and *bribery of non-official servant*, which is mainly used to evaluate the ability of the models to distinguish confusing charges. RCAIL-L contains more charges and cases to evaluate the overall prediction performance of the models. The details of experimental datasets are shown in Table 1.

TABLE 1. Statistics of experimental dataset.

Dataset	RCAIL-L	RCAIL-S
Training Set	1,161,619	80,000
Testing Set	126,731	10,000
Validation Set	110,818	10,000
Charges	103	20
Law articles	103	20

B. BASELINES AND EXPERIMENTAL SETTINGS

To evaluate the performance of ADAN framework, several baselines are used to compare with our method.

- **TF-IDF+SVM**: a word-based SVM text classification model with TF-IDF features [27].
- **RCNN**: a convolutional recurrent neural network for document classification [28].
- **HAN**: a hierarchical attention recurrent network for document classification [23].
- **TOPJUDGE**: a topological multitask framework by using a directed acyclic graph structure to capture topological dependencies among LPJ subtasks [17].
- **FLA**: an attentive neural network for charge prediction by incorporating the applicable law articles [9].

We train the word embeddings on all fact descriptions and deduplicated law article texts and set word embedding size as 200. The hidden state size of Bi-GRU is set to 100, the two full connection layers are of size 200 and 150. The maximum sentence length and the maximum document length are set to 100 words and 15 sentences respectively. The prediction threshold τ is 0.4.

For the model training, we use Adam [29] as the optimizer, learning rate, dropout rate, and batch size are set as 0.001, 0.5, and 128. We terminate the training process of every model if there is no performance improvement.

C. ARTICLE EXTRACTION AND DEDUPLICATION EXPERIMENTS

In order to achieve the optimal performance of ADAN framework, we first make experiments to determine the values of two important parameters: the number of candidate articles extracted by the relevant article extractor k and the number of articles used for comparison when doing text deduplication i.

To evaluate the SVM relevant article extractor, we select 20,000 labeled law cases from our dataset randomly as reference. The recall rate of top k extraction is calculated as 0.768, 0.875, 0.945, 0.957 regarding k as 5, 10, 20 and 30, respectively. The SVM extractor can achieve over 94 recall for top 20 article extraction, which is good enough for subsequent operation. Therefore, k is set to 20. Furthermore, we calculate the micro-F1 of the extractor is only 0.71. If we use the extractor's results directly, it will seriously affect the performance of charge prediction, which is the reason for using the attention mechanism in ADA module.

In order to find the optimal value of i, we set i from 1 to 8 to perform charge prediction on the validation dataset. The prediction results are shown in Fig.7, ADAN achieve the best performance when i is set to 3 and 4. We speculate on the reason for this result is that when the value of i is too small, the similarity among similar law articles is not reduced enough, and when the value of i is greater than 4, information loss is increased and the performance of charge prediction decreased as well. We also observe that the F1-socre remain stable if i is set to greater than 7. The possible reason is that most of similar law articles have been deduplicated, and increasing the value of i cannot affect the performance of charge prediction. Therefore, the value of variable i is set to 4, it means that we perform text deduplication operation based on the textual contents of top 4 articles.

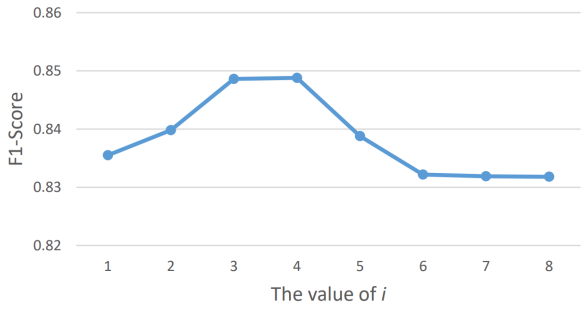


FIGURE 7. The charge prediction F1-score based on different i value on the validation dataset.

D. CHARGE PREDICTION RESULTS

To compare the performance of the baselines and our ADAN model, we employ metrics including accuracy (Acc.), macro-Precision (MP), macro-Recall (MR), and macro F1-score (F1), which are widely used for evaluating text classification task. Table 2 and Table 3 show the experimental results on RCAIL-L and RCAIL-S respectively. We can observe that ADAN model significantly outperforms all the baselines. Compared with the state-of-the-art FLA model, ADAN improved the F1-score of charge prediction by 1.82% and 3.52% on RCAIL-L and RCAIL-S datasets respectively. Specially, F1-score has a higher improvement on RCAIL-S dataset, which shows that ADAN performs better than other baseline models in the discrimination of confusing charges. The advantage of ADAN relative to FLA lies in that, instead of extracting features of relevant law articles directly, our approach uses difference aggregation mechanism to generate law article representation containing rich distinguishable features, which is more effective for charge prediction. In addition, the experimental results also demonstrate that the performance of all baseline models on RCAIL-L is better than that on RCAIL-S, since training data in RCAIL-L are more sufficient.

TABLE 2. Charge prediction results on RCAIL-L dataset.

Metrics	Acc	MP	MR	F1
TF-IDF+SVM	86.55	75.36	71.55	73.41
RCNN	88.41	81.67	76.63	79.07
HAN	90.79	82.63	82.68	82.65
TOPJUDGE	91.68	84.35	83.31	83.83
FLA	91.58	85.48	84.07	84.77
ADAN	**93.66**	**86.98**	**85.66**	**86.31**

TABLE 3. Charge prediction results on RCAIL-S dataset.

Metrics	Acc	MP	MR	F1
TF-IDF+SVM	78.78	72.26	68.38	70.27
RCNN	81.63	74.51	71.56	73.01
HAN	83.55	76.55	77.38	76.96
TOPJUDGE	83.31	78.48	76.60	77.53
FLA	85.86	78.51	78.16	78.33
ADAN	**87.53**	**81.61**	**80.57**	**81.09**

The other baseline models, i.e., TF-IDF+SVM, RCNN, and HAN perform worse, since they just use fact description, without considering the correlation between charge and law articles. This indicates the significance of incorporating relevant law articles as auxiliary information. The extracted relevant law articles inevitably contain noise, which should be properly handled. That RNN-based models achieve better performance than CNN model also shows the advantage of RNNs in processing sequential textual data. The HAN, FLA and ADAN model stand out from the rest due to the hierarchical structure representation of judgment documents. Although TF-IDF+SVM is used for relevant law articles extraction in our study, but we can observe that neural network models have better performance of charge prediction, which indicates that neural network is a better way to extract latent semantic features from law case texts.

E. ABLATION STUDY

Our approach is characterized by the incorporation of attention mechanism of fact encoder, law article duplication

mechanism and attentive article difference aggregation. In order to deeply analyze the effects of these modules, we designed ablation experiments for comparison. As show in Table 3, removing any of the above modules will affect the prediction performance of ADAN model.

TABLE 4. Ablation study on RCAIL-S dataset.

Metrics	Acc	MP	MR	F1
ADAN	**85.66**	**86.61**	**83.57**	**85.06**
W/O attention for fact	80.41	79.25	78.56	78.89
W/O duplication	82.58	83.23	82.15	82.69
W/O attention for article	83.58	84.15	82.83	83.48

''W/O attention for fact'' means removing the attention mechanism of fact encoder from ADAN framework. It signifies that encoder treats the importance of each word or sentence as the same. The experimental result indicates that the prediction performance decline significantly without attention mechanism, which shows its importance for generating semantic representation of judgement documents. Furthermore, ''W/O duplication'' means that we encode the relevant law articles directly without text duplication. The prediction performance of the model also declined because of the lack of discriminative features. ''W/O attention for article'' denotes that attention mechanism is not used in the process of article difference aggregation. It weakens the ability to find effective law articles for charge prediction from the top k extractions and certainly causes performance degradation.

V. CONCLUSION

In this paper, we propose an end-to-end framework named ADAN for charge prediction. The incorporation of hierarchical sequence encoder and attention mechanism is employed to learn better semantic representations of fact descriptions. In addition, a novel difference aggregation mechanism among similar law articles is proposed for extracting distinguishable features to improve the prediction performance. The experimental results on real-world datasets show that the performance of our proposed model is significantly better than existing methods on all evaluation metrics. The future work mainly consists of two aspects: we need to consider more complicated situations such as law cases involving multiple defendants on the one hand; on the other hand, the usage of law articles to improve the performance of prison term prediction is also worth studying.

REFERENCES

[1] Q. Huang and X. Luo, ''State-of-the-art and development trend of artificial intelligence combined with law,'' *Comput. Sci.*, vol. 45, no. 12, pp. 1–11, 2018.

[2] C. Liu, C. Chang, and J. Ho, ''Case instance generation and refinement for case-based criminal summary judgments in Chinese,'' *J. Inf. Sci. Eng.*, vol. 20, no. 4, pp. 783–800, 2004.

[3] C. Liu and C. Hsieh, ''Exploring phrase-based classification of judicial documents for criminal charges in Chinese,'' in *Proc. 16th Int. Symp. Found. Intell. Syst. (ISMIS)*, Bari, Italy, Sep. 2006, pp. 681–690.

[4] W. Lin, T. Kuo, T. Chang, C. Yen, C. Chen, and S. Lin, ''Exploiting machine learning models for Chinese legal documents labeling case classification, and sentencing prediction,'' in *Proc. 24th Conf. Comput. Linguistics Speech Process.*, Chung-Li, Taiwan, Sep. 2012, pp. 140–141.

[5] Z. Hu, X. Li, C. Tu, Z. Liu, and M. Sun, ''Few-shot charge prediction with discriminative legal attributes,'' in *Proc. 24th Conf. Comput. Linguistics Speech Process.*, Santa Fe, NM, USA, Aug. 2018, pp. 487–498.

[6] Z. Lin, H. Chi, and B. Xu, ''Research of criminal case semantic feature extraction method based on the convolutional neural network,'' *Math. Pract. Theory*, vol. 46, no. 17, pp. 129–142, 2017.

[7] S. Long, C. Tu, Z. Liu, and M. Sun, ''Automatic judgment prediction via legal reading comprehension,'' 2018, *arXiv:1809.06537*. [Online]. Available: http://arxiv.org/abs/1809.06537

[8] H. Ye, X. Jiang, Z. Luo, and W. Chao, ''Interpretable charge predictions for criminal cases: Learning to generate court views from fact descriptions,'' in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, New Orleans, LA, USA, Jun. 2018, pp. 1854–1864.

[9] B. Luo, Y. Feng, J. Xu, X. Zhang, and D. Zhao, ''Learning to predict charges for criminal cases with legal basis,'' in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Copenhagen, Denmark, Sep. 2017, pp. 2727–2736.

[10] T. Bansal, D. Belanger, and A. McCallum, ''Ask the GRU: Multi-task learning for deep text recommendations,'' in *Proc. 10th ACM Conf. Rec. Syst.*, Boston, MA, USA, Sep. 2016, pp. 107–114.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, ''Attention is all you need,'' in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, New York, NY, USA, Dec. 2017, pp. 6000–6010.

[12] G. Boella, L. D. Caro, and L. Humphreys, ''Using classification to support legal knowledge engineers in the Eunomos legal document management system,'' in *Proc. 5th Int. Workshop Juris-Inform.*, 2011, pp. 1–12.

[13] F. Wei, H. Qin, S. Ye, and H. Zhao, ''Empirical study of deep learning for text classification in legal document review,'' in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Seattle, WA, USA, Dec. 2018, pp. 3317–3320.

[14] X. Guo, H. Zhang, L. Ye, and S. Li, ''RnRTD: Intelligent approach based on the relationship-driven neural network and restricted tensor decomposition for multiple accusation judgment in legal cases,'' *Comput. Intell. Neurosci.*, vol. 2019, pp. 1–18, Jul. 2019.

[15] Z. Liu, M. Zhang, R. Zhen, Z. Gong, N. Yu, and G. Fu, ''Multi-task learning model for legal judgment predictions with charge keywords,'' *J. Tsinghua Univ., Sci. Technol.*, vol. 59, no. 7, pp. 497–504, 2019.

[16] H. Zhong, Y. Wang, C. Tu, T. Zhang, Z. Liu, and M. Sun, ''Iteratively questioning and answering for interpretable legal judgment prediction,'' in *Proc. AAAI Conf. Artif. Intell.*, York, PA, USA, Feb. 2020, pp. 1250–1257.

[17] H. Zhong, Z. Guo, C. Tu, C. Xiao, Z. Liu, and M. Sun, ''Legal judgment prediction via topological learning,'' in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Brussels, Belgium, Oct./Nov. 2018, pp. 3540–3549.

[18] S. Li, H. Zhang, L. Ye, X. Guo, and B. Fang, ''MANN: A multichannel attentive neural network for legal judgment prediction,'' *IEEE Access*, vol. 7, pp. 151144–151155, 2019.

[19] Z. Xu, X. Li, Y. Li, Z. Wang, and X. Lai, ''Multi-task legal judgement prediction combining a subtask of the seriousness of charges,'' in *Proc. 19th China Nat. Conf. (CCL)*, Hainan, China, Oct./Nov. 2020, pp. 415–429.

[20] W. Yang, W. Jia, X. Zhou, and Y. Luo, ''Legal judgment prediction via multi-perspective bi-feedback network,'' in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Macau, China, Aug. 2019, pp. 4085–4091.

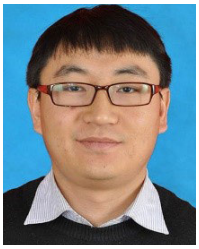
[21] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, ''Distributed representations of words and phrases and their compositionality,'' in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, vol. 26, no. 1, pp. 3111–3119.

[22] F. Xiong, Y. Deng, and X. Tang, ''Word2vec parameter learning explained,'' *J. Nanjing Normal Univ., Eng. Technol. Ed.*, vol. 2015, no. 2, pp. 43–48, 2015.

[23] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, ''Hierarchical attention networks for document classification,'' in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, San Diego, CA, USA, Feb. 2016, pp. 1480–1489.

[24] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, ''Learning phrase representations using RNN encoder–decoder for statistical machine translation,'' in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Doha, Qatar, 2014, pp. 1724–1734.

[25] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. Int. Conf. Learn. Represent.*, San Diego, CA, USA, Feb. 2015, pp. 1–15.
[26] H. Zhong, C. Xiao, Z. Guo, C. Tu, Z. Liu, M. Sun, Y. Feng, X. Han, Z. Hu, H. Wang, and J. Xu, "Overview of CAIL2018: Legal judgment prediction competition," 2018, *arXiv:1810.05851*. [Online]. Available: http://arxiv.org/abs/1810.05851
[27] J. A. K. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Process. Lett.*, vol. 9, no. 3, pp. 293–300, Jun. 1999.
[28] R. Wang, Z. Li, J. Cao, T. Chen, and L. Wang, "Convolutional recurrent neural networks for text classification," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Austin, TX, USA, Jul. 2019, pp. 2267–2273.
[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, San Diego, CA, USA, May 2015, pp. 1–41.



DAPENG LI received the M.S. degree in computer science from Northeastern University, Shenyang, China, in 2008, where he is currently pursuing the Ph.D. degree in computer science. His research interests include legal intelligence and natural language processing.

QIHUI ZHAO received the B.E. degree from the Software College, Shenyang University of Technology, in 2014, and the M.E. degree from the Software College, Northeastern University, in 2017, where he is currently pursuing the degree. His research interests include natural language processing including, legal intelligence, and deep learning.

JIAN CHEN received the Ph.D. degree in computer science and technology from Northeastern University, Shenyang, China, in 2010. In 2016, he was a Visiting Research Associate with the King's College London, London, U.K. He is currently an Associate Professor with Northeastern University and a Senior Software Engineer with Neusoft Corporation, Shenyang. His research interests include resource allocation, D2D communication, location technology, network management, and signal and image processing.

DAZHE ZHAO received the B.S. degree in mathematics from Liaoning University, in 1982, the M.S. degree in computer science from the Shenyang Institute of Computing Technology, Chinese Academy of Sciences, in 1990, and the Ph.D. degree in computer science from Berlin Technical University, Berlin, Germany, in 1996. She is currently a Professor with the Research Institute, Northeastern University, Shenyang, China. Her research interests include software engineering, medical image processing and analysis, and knowledge discovery and data mining.