AI Open 3 (2022) 172–181

Contents lists available at ScienceDirect

AI Open

journal homepage: www.keaipublishing.com/en/journals/ai-open

BCA: Bilinear Convolutional Neural Networks and Attention Networks for legal question answering

Haiguang Zhang, Tongyue Zhang, Faxin Cao, Zhizheng Wang, Yuanyu Zhang, Yuanyuan Sun *, Mark Anthony Vicente

Department of Computer Science and Technology, Dalian University of Technology, China

ARTICLE INFO

Keywords:
Deep learning
Text classification
Attention mechanism
Convolutional Neural Networks
Judicial Examination
Bilinear model

ABSTRACT

The National Judicial Examination of China is an essential examination for selecting legal practitioners. In recent years, people have tried to use machine learning algorithms to answer examination questions. With the proposal of JEC-QA (Zhong et al. 2020), the judicial examination becomes a particular legal task. The data of judicial examination contains two types, i.e., Knowledge-Driven questions and Case-Analysis questions. Both require complex reasoning and text comprehension, thus challenging computers to answer judicial examination questions. We propose **B**ilinear **C**onvolutional Neural Networks and **A**ttention Networks (**BCA**) in this paper, which is an improved version based on the model proposed by our team on the Challenge of AI in Law 2021 judicial examination task. It has two essential modules, **K**nowledge-**D**riven **M**odule (**KDM**) for local features extraction and **C**ase-**A**nalysis **M**odule (**CAM**) for the semantic difference clarification between the question stem and the options. We also add a post-processing module to correct the results in the final stage. The experimental results show that our system achieves state-of-the-art in the offline test of the judicial examination task.

1. Introduction

The judicial examination is a legal professional qualification examination that follows the relevant regulations and is the most crucial selection examination for legal talents in China. It is known as one of the most challenging examinations in China because of the tremendous legal knowledge, and the annual-pass rate only reaches 10%. Accurately answering judicial questions requires the respondent to possess complex reasoning and text comprehension, which are difficult for humans and computer algorithms. Using algorithms to answer judicial questions is conducive to promoting the development of legal question answering and making NLP technology more widely used in the legal field. In the future, the judicial examination will highlight the ability of students to use artificial intelligence to deal with and solve legal problems (Zheng, 2021).

The judicial examination task was set up in the judicial examination track of Challenge of AI in Law (CAIL2020) to open to both the research community and the legal industry. It is similar to Multi-choice Machine Reading Comprehension (Multi-choice MRC) in that it selects the appropriate one from a set of candidate options. However, the difference is that Multi-choice MRC uses a short passage to help answer questions, while this task does not. The dataset for this task is derived from JEC-QA (Zhong et al., 2020) with 21072 indefinite multiple-choice questions, and each question has four options. The questions can be divided into Knowledge-Driven questions (KD-questions) and Case-Analysis questions (CA-questions). KD-questions focus on understanding specific legal concepts, while CA-questions focus on analysing real legal cases and involve different people, places, times, and events. Answering questions accurately requires five types of reasoning skills: Word Matching, Concept Understanding, Numerical Analysis, Multi-Paragraph Reading, and Multi-Hop reasoning. Based on the statistics of Zhong et al. (2020), 65.9% of KD-questions require word matching, indicating that some options have a direct or indirect implicit semantic connection to the question stem. 66.2% of CA-questions require multi-hop reasoning, i.e., multiple logical reasoning steps to get the answer. These questions have implicit semantic relations between options and the question stem. Also, their options possess a particular semantic association. That is, if there are two conflicting descriptions of options, at least one of them is wrong. As shown in Tables 1 and 2, KD-questions focus on understanding the legal concept of the respondents, while CA-questions focus on inferring legal-related information from situational cases. Therefore, we can see a gap between KD-questions and CA-questions. However, previous methods in other areas using information

* Corresponding author.
E-mail addresses: haiguang@mail.dlut.edu.cn (H. Zhang), zty9818@mail.dlut.edu.cn (T. Zhang), 32009168@mail.dlut.edu.cn (F. Cao), wzz_dllg@mail.dlut.edu.cn (Z. Wang), knockknock404@mail.dlut.edu.cn (Y. Zhang), syuan@dlut.edu.cn (Y. Sun).

https://doi.org/10.1016/j.aiopen.2022.11.002
Received 21 June 2022; Received in revised form 15 October 2022; Accepted 12 November 2022
Available online 16 November 2022

Table 1
KD-questions analysis.

Statement	Candidate options
房屋所有权人以营利为目的, 将以划拨方式取得使用权的国有土地之上建成的房屋出租。其租金中所含土地收益依法应如何处理? *To make a profit, the house owner rents out the house on the state-owned land that has obtained the right to use it. How should the law deal with land income included in the rent?*	✗ A 租金中所含土地收益归承租人 *The land income included in the rent goes to the lessee.* ✗ B 租金中所含土地收益归出租人 *The land income included in the rent goes to the lessor.* ✓C 应当将租金中所含土地收益上缴国家 *The land income included in the rent should be returned to the state.* ✗ D 应当将租金中扣除所含土地收益 *The land income should be deducted from the rent.*

Table 2
CA-questions analysis.

Statement	Candidate options
以下事实中, 不能在甲、乙之间产生民事法律关系的是? *Which facts cannot create a civil legal relationship between Party A and Party B?*	✓A 甲、乙书面约定2012年8月8日领取结婚证, 海枯石烂, 永不变心. *Party A and Party B agreed to receive the marriage certificate on August 8, 2012. The sea is dry, the stone is rotten, and the heart will never change.* ✓B 甲赌博时输给乙2万元并当场给付 *Party A loses 20,000 yuan to Party B when gambling and pays it on the spot.* ✗ C 甲、乙约定某日在白玫瑰大酒店签订合作开发房地产合同 *Party A and Party B agree to sign a cooperative real estate development contract at White Rose Hotel on a particular day.* ✓D 甲向乙问路, 乙欣然指路, 但因乙疏忽而指错方向 *Party A asks Party B for directions, and Party B readily points the way, but Party B is negligent and points in the wrong direction.*

retrieval (Verma et al., 2020) and some reading comprehension (Seo et al., 2016; Yin et al., 2016; Zhu et al., 2018; Wang et al., 2018) strategies failed to tackle the gap between different types of questions.

To make up for this deficiency, we propose Bilinear Convolutional Neural Networks and Attention Networks (BCA) in this paper, in which we transform the task into a binary classification task of judging whether the question stem and each option are matched. Considering the relationship between the question stem and the options, the input of the model is set to the form of "$[CLS]$+*question stem*+$[SEP]$+*answer*+$[SEP]$" (hereinafter called the "question–answer pair"). Here $[CLS]$ is placed at the beginning of the sentence and represents the semantics of the entire sentence, and $[SEP]$ is a sign indicating the separation of two sentences.

Furthermore, we design two structures to adapt to their different characteristics according to the analysis of two different types of questions. In Table 1, the red fonts represent irrelevant or opposite to the question stem, and the green fonts represent the contents related to the question stem. In Table 2, the scenarios described by the four options are quite different, so answering is to mine the semantic relationship between several question–answer pairs, pay attention to more important information, and reduce attention to other information. To capture semantic relations from question–answer pair, we use convolutional neural networks (CNN) (Krizhevsky et al., 2012) to get local features in the input and fuse the original features. Therefore, we represent the tensors encoded by the pre-trained models of the respective question–answer pairs of the two types of questions in two different networks to assign weights to different options and better focus on the correct ones. In order to better express the relationship between the models and the questions, we name modules the **K**nowledge **D**riven **M**odule(**KDM**) and **C**ase-**A**nalysis **M**odule(**CAM**), respectively.

2. Related work

The judicial examination task is a special legal task of knowledge question and answer, similar to multiple-choice reading comprehension. In this summary, we briefly introduce the work in related fields.

Machine Reading Comprehension (MRC). MRC is a fundamental Natural Language Processing (NLP) task that tests how well a machine understands natural language by asking it to answer questions based on a given context. Many datasets have been published in the field of reading comprehension (Hermann et al., 2015; Rajpurkar et al., 2016; He et al., 2017; Lai et al., 2017). These have extensively promoted the development of reading comprehension (Wang and Jiang, 2016; Seo et al., 2016; Wang et al., 2017a; Dhingra et al., 2016; Richardson et al., 2013). Multi-choice MRC has also been studied in recent years (Zhu et al., 2018; Zhang et al., 2019; Jin et al., 2020; Zhu et al., 2021b). In addition, some scholars use retrieval methods to answer questions (Lin et al., 2018; Wang et al., 2017b; Clark and Gardner, 2017; Zhu et al., 2021a).

Yes/No Answers on Legal Question Answering. This approach is a method to judge whether a question is associated with an answer. If the answer is related to the question, the result is true; otherwise, it is false. In the past decade, much effort has been devoted to the study of legal questions. For example, Martinez-Gil (2021) systematically summarizes the research achievements of Legal Question Answering in recent years and their advantages and disadvantages. Kim et al. (2013) have put different methods to answer yes/no questions in legal bar exams. After that, Kim et al. (2014) also adopted several unsupervised methods (TF-IDF and Latent Dirichlet Allocation (LDA)) and one supervision method (Ranking SVM). Additionally, a final improvement includes the use of paraphrases (Kim et al., 2016). Then, other researchers (Taniguchi and Kano, 2016; Taniguchi et al., 2018) use case-role analysis and FrameNet, respectively, to solve the problem. Finally, Kano et al. (2017) studied linguistic structures for the first time.

Prompt-learning. Recently, the fine-tuned Pre-trained Language Models have achieved tremendous success in various NLP tasks, such as question answering (Yang et al., 2019; Adiwardana et al., 2020), text classification (Minaee et al., 2021; Ding et al., 2021). PLMs can learn syntactic (Goldberg, 2019), semantic (Ma et al., 2019) and structural (Jawahar et al., 2019) information about language. To this end, inspired by GPT-3 (Brown et al., 2020), prompt-learning has been proposed to transfer downstream tasks as some cloze-style objectives and achieved superior performance, especially in few-shot learning (Liu et al., 2021a). Along this line, many hand-crafted prompts have been made for various tasks. Furthermore, to avoid time-consuming and labour-intensive prompt design, a series of automatic prompt generation, methods have been explored recently (Li and Liang, 2021; Lester et al., 2021; Shin et al., 2020).

Judicial Examination Tasks. Some teams also proposed a few solutions in previous judicial examination tasks. The team at Soochow

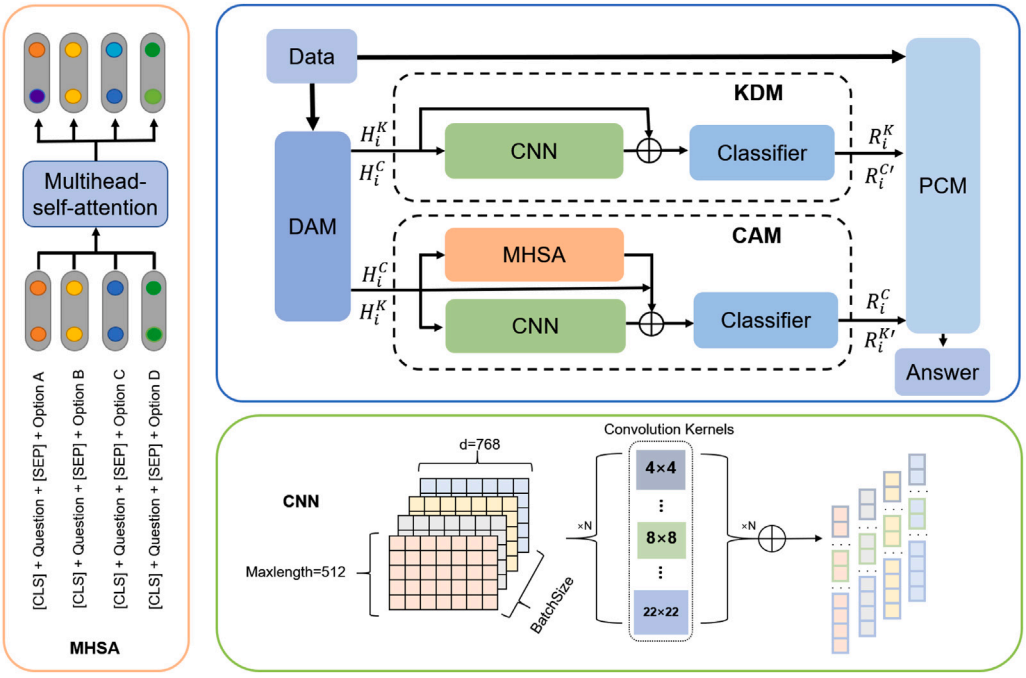

Fig. 1. Architecture of BCA.

University used a method of data enhancement. The team of Nanjing Qingdun Information Technology Company Limited has expanded the label of the dataset. Wu and Luo (2021) employed an aligned graph network. Traditional reading comprehension models used to solve multiple-choice questions strongly depend on reading fragments and a weak ability to exploit the dependencies between options. Most of them only apply to datasets in the public domain and are not suitable for solving some problems in the legal domain. Therefore, they are not suitable for the judicial examination task. In addition, the previous solutions dealt with the two types of question types together and did not compare and distinguish the differences between the question types, so we have carried out a systematic study to address these challenges in this study.

3. Method

3.1. Problem definition

Tables 1 and 2 shows that given questions related to law, each question has four options. The respondent needs to select the correct options from the candidate options, and the answer is considered correct only when the predicted answer list and the standard answer list are the same.

3.2. Model architecture

In this section, we introduce each module of BCA: Bilinear Model (BM), including Knowledge-Driven Module (KDM) and Case-Analysis Module (CAM), Data Augmentation Module (DAM), and Prompt and Choose Module (PCM), as shown in Fig. 1. Firstly, we augment the training set with DAM to improve the model's ability to generalize on the test set and find relationships between options. Secondly, due to the different problem-solving properties of KD-questions and CA-questions, we design KDM and CAM for them, respectively. After DAM processes the data, it is sent to KDM and CAM for analysis and processing. Finally, to avoid the model-generated answer list being empty and to correct errors in which the answer list violates certain features of the question stem, we use a post-processing module PCM for final correction.

3.3. Data Augment Module (DAM)

Data Augment. To improve the generalization ability of the system on the test set, we augment the training set with data augmentation. Our method is to splice some of the correct options to get new options and finally splice with the question to get a new question–answer pair.

Our data augmentation falls into three patterns. For questions with three correct and incorrect options, we choose two of three options to splice after the question stem, regardless of the order, and splicing all three options with the question stem will generate four new question–answer pairs. We concatenate two correct options after the question stem for questions with two correct and two incorrect options. For questions with four correct options, we choose two of four options to be spliced after the question stem, regardless of order, resulting in six new question–answer pairs. Experiments show that data augmentation has a particularly positive effect on the model's accuracy.

Encode. We represent the data as $\mathrm{Q}^{\mathrm{T}}, \mathrm{O}^{\mathrm{T}}$:

$$\{Q_i^T = (q_1, \ldots, q_n); O_i^{T,l} = (o_1, \ldots, o_m) \mid T \in \{\mathrm{K}, \mathrm{C}\}, l \in \{1,2,3,4\}\} \tag{1}$$

where Q represents the question stem, O represents the option, q and o represent tokens, n and m represent the number of question stem and option tokens, T represents the type of the question, K represents KD-questions, C represents CA-questions, l represents the order of the options, i represents the ith question sample.

After the data enhancement module, we get $O_i'^T$. Then for the two question types, the input question–answer pairs are $I_{i,j}^T$. We use RoBERTa (Liu et al., 2019) to encode them and get the high-dimensional semantic representation $H_{i,j}^T$. The formula is as follows:

$$O'^T_{i,j} = \mathrm{Augmentation}\left(O_i^{T,l}\right) \tag{2}$$

$$I_{i,j}^T = [CLS] + Q_i^T + [SEP] + O'^T_{i,j} + [SEP] \tag{3}$$

$$H_{i,j}^T = \mathrm{RoBERTa}\left(I_{i,j}^T\right) \tag{4}$$

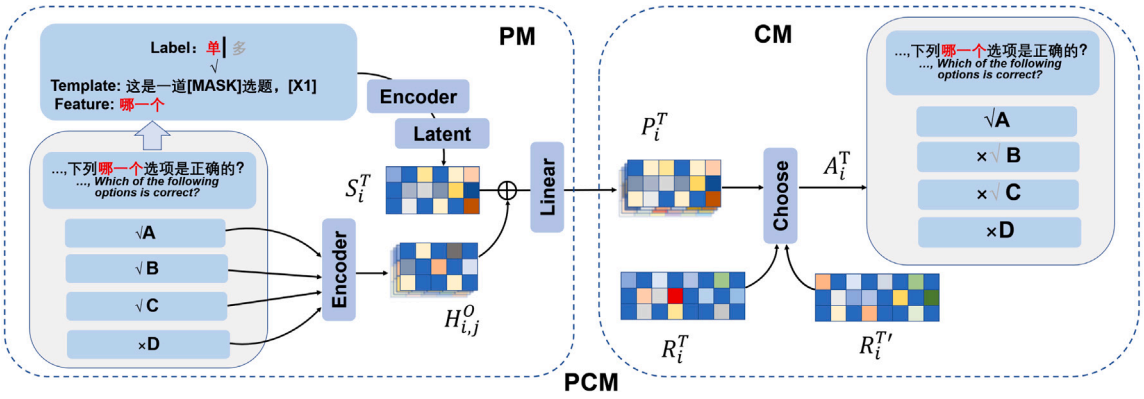

Fig. 2. Architecture of PCM.

where j represents the jth question–answer pair of the ith question. $H_{i,j}^{T} \in \mathrm{R}^{(n+\mathrm{more}(m))\times d}$, $more(.)$ calculates the total length after splicing multiple options and d is the dimension of the token embedding layer. For better representation and processing, we denote all question–answer pairs belonging to the same question as H_i^T.

3.4. Knowledge-Driven Module (KDM)

We found that the focus of answering KD-questions is to mine whether some expressions in the answer violate the legal concept in the question, which requires the model to have a solid ability to extract local features to mine the semantics of local fields in the text. Here we adapt CNN to answer legal questions. One of the primary motivations for using deeper models, such as neural networks with many layers, is the potential for significantly improved representational efficiency compared to shallower neural network models. We extract linguistic features between two sentences and compare those features to determine textual entailment. Not all language features are directly related to this task, so we will capture relevant features and connect them locally. It limits the network architecture using local connections known as receptive fields. In addition, the CNN possesses a strong generalization ability and ample room for improvement (Rakhlin, 2016), and we apply it in our model to extract local semantic information implied by legal sentences. Inspired by CNN text classification (Krizhevsky et al., 2012), we feed H_i^K into multiple two-dimensional convolutional layers with different convolution kernel sizes and stack the outputs of all convolutional layers to obtain the final output of the convolutional neural network. The formula is as follows:

$$Conv_i^K = [Conv2d_1(H_i^K); \cdots ; Conv2d_N(H_i^K)] \tag{5}$$

where $Conv2d_N$ represents the two-dimensional convolutional layer, N represents the number of convolutional layers, and $[;...;]$ stands for concatenating vectors.

Finally, we superimpose the $Pooled_i^K$ obtained by H_i^K after removing redundant information through the pooling layer and the convolutional neural network output to fuse global semantic information and local semantic information. Then it is sent to the classifier for binary classification, and R_i^K is obtained. Here we use the $softmax$ function to normalize the label dimension, and the same is true of $softmax$ mentioned later. The formula is as follows:

$$R_i^K = \mathrm{softmax}([Conv_i^K; Pooled_i^K]W_K + b_K) \tag{6}$$

where W_K and b_K are trainable parameters.

$$Pooled_i^K = \mathrm{Tanh}(H_i^K W_{pool} + b_{pool}) \tag{7}$$

where Tanh(.) is the activation function, W_{pool} and b_{pool} are trainable parameters.

3.5. Case-Analysis Module (CAM)

We found that compared to KD-questions, the situational differences expressed between the four options in CA-questions were more significant. To make the model learn the semantic difference between different options, we input the pooling layer outputs $Pooled_i^C$ of multiple question–answer pairs belonging to the same question into a multi-head self-attention module MHSA (Yang et al., 2016) with the corresponding pooling layer outputs, as follows:

$$Att_i^C = \mathrm{MultiHead}\left(Pooled_i^C\right) \tag{8}$$

$$Pooled_i^C = \mathrm{Tanh}(H_i^C W_{pool} + b_{pool}) \tag{9}$$

where W_{pool} and b_{pool} are trainable parameters.

$$\mathrm{MultiHead}(Q,K,V) = [head_1; \ldots ; head_h]W^O \tag{10}$$

$$head_i = \mathrm{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{11}$$

$$Attention(Q,K,V) = \mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{12}$$

In CAM, question–answer pairs of the same question are not shuffled, allowing the attention module to function between question–answer pairs of the same question. Eq. (12) uses the scaled point cumulative attention mechanism, where d_k represents the word embedding dimension of the pre-trained model.

Finally, we add the output of the attention module to RoBERTa output vector through the residual network and then concatenate it with the output of CNN to combine the local semantic information and the global semantic association between different options. The output is finally sent to the classifier for binary classification to obtain R_i^C:

$$\mathrm{R_i^C} = \mathrm{softmax}([Conv_i^C; Pooled_i^C + Att_i^C]W_C + b_C) \tag{13}$$

where W_C and b_C are trainable parameters.

$$Conv_i^C = [Conv2d_1(H_i^C); \ldots ; Conv2d_N(H_i^C)] \tag{14}$$

To better use cross information, it is worth noting that when predicting KD-questions, H_i^K is input into CAM, $Conv_i^{K'}$ is obtained after passing through the convolutional layer, and then

$$R_i^{K'} = \mathrm{softmax}([\mathrm{Conv_i^{K'}}; \mathrm{Pooled_i^{K'}} + \mathrm{Att_i^{K'}}]\mathrm{W_C} + b_C) \tag{15}$$

is used as supplementary information to R_i^K. The same is true for predicting CA-questions, and

$$R_i^{C'} = \mathrm{softmax}([\mathrm{Conv}_i^{C'}; Pooled_i^{C'}]W_K + b_K) \tag{16}$$

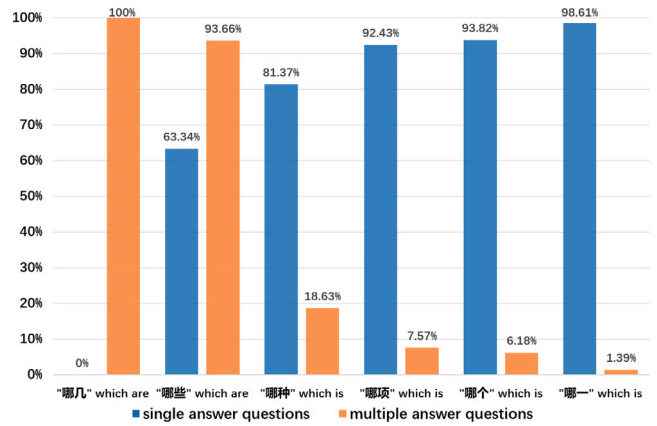

Fig. 3. The single-multiple choice ratio of questions with unique feature words in question stems. The horizontal axis is the feature word related to the number of answer lists in the question stem, and the vertical axis represents the proportion of single-choice or multiple-choice questions in all the questions containing a certain feature word.

is obtained through KDM as supplementary information to $R_i^{C'}$.

In addition, KDM and CAM both train the model end-to-end by minimizing the cross-entropy loss L.

$$L = \sum_i \text{CrossEntropy}(y_i, p_i) \tag{17}$$

where y_i represents the label of the ith sample, the positive label is 1, the negative label is 0, and p_i represents the probability that the ith sample is predicted to be a positive class.

3.6. Prompt and Choose Module (PCM)

The disadvantage of using binary classification for this task is that the predicted results greatly exceed or are far below the number of correct options. Facts have proved that if the number of correct options is not well controlled, the answer to the question must be wrong. So, in our system, to avoid the anomalous phenomenon of the predicted answer list of the test set, we design a post-processing module to judge and select the correct options, which includes the following three aspects.

Firstly, the prediction results of some questions are illogical. It is obvious that there are high-frequency feature words (Li et al., 2021) in the stem, and the result is contrary to the stem (e.g., In the Chinese context, "which are" indicates that this question is likely to be a multiple-choice question, but the length of answer list is zero or just one). For this reason, we conduct statistical analysis on the information of the question stem, as shown in Fig. 3. We can see that the characteristic words have a positive significance for classifying single-multiple choice. Defined as follows: Feature word information $F = (f_1, \ldots, f_w)$, where f_i represents a feature word, and w represents the number of feature words.

On the other hand, the data with feature words only occupies 40%, but the contribution rate to classification is limited, and some of the remaining nearly 60% cannot be excluded from other important information, so we assume that these samples also have a positive effect on the results. Besides, some language models are self-supervised tasks with a gap with downstream tasks. Generally speaking, the larger the gap, the greater the impact on the task. Bert (Devlin et al., 2018)'s self-supervised tasks are masked language mode (MLM) and next sentence prediction (NSP), and the downstream tasks are not necessarily just the two. Therefore, the model of downstream tasks is closer to the form of self-supervised tasks, which will give full play to the advantages of the pre-training model. In addition, the pre-training model can be seen as a form of knowledge base, which contains much hidden semantics. To reduce the loss of the pre-training model and the downstream tasks, we design the single/multiple choice task to be close to MLM. Coincidentally, the main goal of prompt-learning (Han et al., 2021) is to reduce the gap between the pre-training target and the downstream fine-tuning target as much as possible. So we use this advantage to extract the relationship between samples and labels and strengthen it in prompt-learning. We designed the PM module to pre-judgment the number of the right option, as shown in Fig. 2. The manual template is as follows: Template: 这是一道$[MASK]$选题 , $[X1]$

This is a $[MASK]$ choice question , $[X1]$ where $[X1]$ is a symbol of the question's stem, $[MASK]$ represents the masked word in a sentence, and the purpose is to let the model predict the word through contextual understanding. Q_i^T and $O_{i,j}^T$ are coded with BERT, including $[CLS]$ and $[SEP]$:

$$H_i^P = \text{BERT}\left(Q_i^T\right) \tag{18}$$

$$H_{i,j}^O = \text{BERT}\left(O_{i,j}^T\right) \tag{19}$$

where j is the jth option of the ith question.

Finally, we use the feature word information F and the question stem information H_i^P to calculate their similarity through the pre-training model ALBERT and combine the partial weights of the prompt-learning to obtain the similarity matrix S_i^T:

$$S_i^T = \text{similar}_{\text{ALBERT}}\left(\text{Prompt}\left(H_i^P\right), F, H_i^P\right) \tag{20}$$

Then S_i^T combines the encoding information $H_{i,j}^O$ of the option to obtain P_i^T, which is sent to CM.

$$P_{i,j}^T = [S_i^T; H_{i,j}^O]W_P + b_P \tag{21}$$

CM combines $P_{i,j}^T$ and R_i^T and uses $R_i^{T'}$ as an aid, and then applies the similarity module to predict the maximum similarity of the selected options to obtain A_i^T. The purpose of Choose(.) is to re-predict illogical answers, and we sort by option similarity, supplemented by $R_i^{T'}$ to get the final answer.

$$A_i^T = \text{Choose}\left(P_i^T, R_i^T, R_i^{T'}\right) \tag{22}$$

where P_i^T represents all $P_{i,j}^T$ belonging to the ith question.

The loss function of PCM is:

$$L_{PCM} = \sum_i \text{CrossEntropy}(y_i^*, p_i^*) \tag{23}$$

* represents the masked position during training, p_i^* represents the probability value of y_i^*.

4. Experiments

4.1. Dataset

The dataset used in this task comes from JEC-QA (Zhong et al., 2020). Each sample in this dataset contains the following fields (the test set does not contain the "answer" field), as shown in Table 3. In addition, individual data fields were flawed, which we corrected.

The total number of samples in the dataset is 21,072, of which 51% are single-choice questions, 21% are questions with two correct

Table 3
Dataset fields and meanings. Only part of the data contains the type field.

Field	Meaning
Answer	A list of real answers to the question.
Id	Unique identifier for the issue.
Option list	List of candidate options.
Statement	The question stem.
Topic	The law involved in the question.
Type	0 for KD-questions, 1 for CA-questions.

Table 4
Experimental environment.

Parameter	Configuration
OS	Ubuntu 20.04.3
CPU	Intel(R) Core(TM) i7-11700 @ 2.50 GHz
GPU	NVIDIA GeForce RTX 3090
Python	3.7.13
PyTorch	1.11.0
Memory	32G

Table 5
Experimental hyperparameter settings.

Name	KDM	CAM	PM
Batch size	4	4	16
Learn rate	3e−4	3e−4	2e−5
Dropout	0.3	0.3	0.5
Epoch	30	30	50

options, 20% are questions with three correct options, and 8% are questions with four correct options. Regarding the types of questions, KD-questions accounted for 36.9%, and CA-questions accounted for 73.1%. In order to better verify the effect of the model, we set the train-validation-test ratio to 7:1:2.

4.2. Experiment setting

All experiments in this paper were carried out in the environment shown in Table 4. The training time of BM is about 15 h, the training time of PCM is about 1 h, and the cumulative total training time is about 16 h.

The hyperparameter settings of the experimental part are shown in Table 5.

4.3. Pre-trained model

BERT (Devlin et al., 2018). The bidirectional encoding representation of the Transformer is adopted to improve the architecture fine-tuning-based approach. Masked language model (MLM), which overcomes the one-way limitation and "next sentence prediction" (NSP), can be used to pre-train text pair representations. It is the state-of-the-art framework for many NLP tasks.

RoBERTa (Liu et al., 2019). It is a fine-tuning of the BERT model, using 1,000% more datasets and computing power than BERT. Removed the NSP, introduced a dynamic mask, and found that a larger batch size is more useful during training. The pre-trained language model used by KDM and CAM is it.

ALBERT (Lan et al., 2019). The parameters of the model can be reduced based on maintaining performance. It uses the concept of parameter sharing across layers. Another task "sentence order prediction" is proposed. The pre-trained language model used by PCM is it.

Lawformer (Xiao et al., 2021). A pre-trained model trained on Longformer using many legal long texts. It can achieve significant performance improvement on tasks with long sequence input.

4.4. Result and discussion

Table 6 shows the top 7 scores of the third stage of the CAIL2021 judicial examination task, whose test set in this stage is the questions of the 2020 judicial examination. Although our model ranked fifth among all 15 teams in the third stage, we ranked third in the second stage and fourth in the final overall ranking. Consequently, this proves that by using BM to analyse and predict according to the question type and DAM, the model's recognition accuracy for the correct answer can be improved. The accuracy evaluation indicators used in this task are as follows.

Table 6
The score of the third stage of the CAIL2021 judicial examination task. Ours: (BM+DAM+Random).

Stage	Model	KD_{acc}	CA_{acc}	Score
Third	1st place team	43.21	24.62	30.00
	2nd place team	35.80	24.62	27.86
	3rd place team	29.63	19.60	22.50
	4th place team	23.46	20.60	21.43
	Ours	**22.22**	**19.10**	**20.00**
	6th place team	16.05	19.60	18.57
	7th place team	20.99	15.58	17.14
			Ranking	
Second	**Ours**		**3rd**	
Final			**4th**	

Table 7
Online ablation experiments.

Model	KD_{acc}	CA_{acc}	Score
BM+DAM+Random	37.93	20.47	23.00
BM+DAM	34.48	16.37	19.00
BM	31.03	16.37	18.50
KDM	34.48	12.28	15.50
CAM	13.79	16.37	16.00
RoBERTa	17.24	8.77	10.00
Baseline random	6.90	5.26	5.50

$$acc(f;D) = \frac{1}{n}\sum_{i=1}^{n}\left(f(x_i) = label_i\right) * 100 \tag{24}$$

where f is a model and D is a test set of size n. Furthermore, for KD-questions and CA-questions, the prediction of the question is correct if and only if the predicted answer list is entirely consistent with the standard answer list. For online scoring, " Score" represents the accuracy of the test set.

As shown in Table 7, we conduct an online ablation experiment whose results are from the second stage of the CAIL2021 judicial examination task. Removing the random post-processing module reduces the model's accuracy on the test set by 4%, proving that the module can correct some phenomena that violate common sense in the prediction results of the test set. Then, DAM is removed to reduce the model's accuracy on the test set by 0.5%, proving that using data augmentation to expand the training data helps improve the model's generalization ability on the test set. Next, replacing BM with KDM, Score is reduced by 3%, and the accuracy rate of CA-questions is reduced by more than 4% compared with BM. So it is proved that the CAM can better compare and grasp the semantic differences between different options in CA-questions and is more suitable for the characteristics of CA-questions. Accordingly, replacing BM with CAM has a negative impact, with a 2.5% drop in accuracy. In addition, we also compared RoBERTa's binary classification of sentence pairs and the official baseline, and the results show that our online model has advantages.

Besides, we also conduct offline experiments to demonstrate the effectiveness of BCA. We randomly selected 20% data of JEC-QA (Zhong et al., 2020) and the objective questions of other judicial examinations to form a set D questions for testing. The calculation formula of the evaluation index is as follows:

$$EAR(f;D) = \frac{1}{n}\sum_{i=1}^{n}\left(f(x_i) = None\right) * 100 \tag{25}$$

where "None" represents an empty list, and EAR represents the proportion of the number of questions with empty prediction results.

As shown in Table 8, our method can achieve a Score of 32.9, which gives an upper bound for our model performance. Additionally, the table shows the results of several baseline experiments on JEC-QA. Our

Table 8
Evaluation results of different models on JEC-QA. Bold marks the highest score. † marks results quoted directly from the original papers.

Model	KD_{acc}	CA_{acc}	Score
Ours	**27.59**	**35.98**	**32.90**
Multi-Matching (Tang et al., 2019)†	23.63	29.06	28.63
Co-matching (Wang et al., 2018)†	25.37	28.61	26.06
ABGN (Wu and Luo, 2021)†	20.47	20.69	20.5
BiRNN	19.54	15.44	16.94
RoBERTa	12.70	14.58	13.89
Random	10.56	11.54	11.18

Table 9
Contrast experiments of ablation offline. Bold marks the highest score.

Model	KD_{acc}	CA_{acc}	Score	EAR
BCA(BM+DAM+PCM)	**27.59**	**35.98**	**32.90**	**0**
BM+DAM+CM	25.06	32.62	29.85	0
BM+DAM+Random	22.81	30.09	27.42	0
BM+DAM	21.34	28.97	26.17	15.80
BM(Lawformer)+DAM	16.34	20.56	19.01	24.54
KDM	21.50	14.73	17.21	19.44
CAM	16.52	27.43	23.43	13.11
CAM-Attention	–	22.75	–	21.83

Table 10
Comparison of different data augmentation methods. Bold marks the highest score.

Model	KD_{acc}	CA_{acc}	Score	EAR
Ours(BM+DAM)	**21.34**	**28.97**	**26.17**	**15.80**
BM+A	20.53	21.76	21.31	18.52
BM+B	17.15	18.54	18.03	15.46
BM+C	19.01	21.72	20.55	20.09

model outperforms multi-matching by more than 4%, especially on CA-questions our model compares to the best model by more than 6% and exceeds the best model by 2% on KD-questions.

To study the main components in our model, we conducted an ablation study. The results are reported in Table 9. The offline experiment was compared using Lawformer with RoBERTa. While Lawformer works well on long legal texts, this pre-trained model does not work well in this task. In addition, we redesigned the experiments, including KDM, CAM, and CAM without the attention module. The results were consistent with the online experiment, confirming the role of each part of the model. Furthermore, the ablation experiments once again proved the positive effect of PCM on the experimental results so that the offline results could reach state-of-the-art results.

As shown in Table 10, to verify the effectiveness of DAM, we use Method C without data augmentation to compare it with the other three data augmentation methods. Method B represents that negative sentence pairs are enhanced. We can see that the data after Method B reduces the accuracy rate by more than 2%, which shows that this method does not fully enable the model to learn the semantic relationship between question–answer pairs but reduces the generalization ability. Method A represents that both positive and negative sentence pairs are enhanced, the average accuracy is increased by 0.76%, and the enhanced part of the data improves the model's generalization. DAM is the data enhancement method proposed in this paper, and the average accuracy can be increased by 5.52%, which has the best performance of the three data enhancement methods. In conclusion, BM can grasp the implicit semantic relationship between the question stem and the correct options through data enhancement and improve the model's generalization ability.

In the online experiment, the role of the random module is to randomly set an answer for a sample once it is detected that the answer list is empty. Although the final result is improved, this is an imprecise operation. After statistical analysis of the ratios of single-choice and multiple-choice questions for KD-questions and CA-questions above, it can be seen that the questions whose predicted answers are empty are not necessarily all single-choice questions. Then, we improved the post-processing module and designed the PCM, which includes PM that can pre-judgment the answer list and CM that can comprehensively judge and select.

Table 11
Contrastive experiment of PM. Bold marks the highest score.

Method	Dev_{acc}	$Test_{acc}$
Ours (PM)	**79.87**	**79.78**
P-tuning (Liu et al., 2021b)	78.20	76.34
Other template[a]	77.93	78.55
AGN	78.46	74.15
ALBERT	71.50	77.00
BERT+LSTM	64.55	65.00
Transformer	63.01	64.50
Hand-crafted rules	–	43.35

[a]Represents the best result of other templates.

Table 12
Comparison experiment of CM. Bold marks the highest score.

Method	**Our (CM)**	Random	TF-IDF	Bow	K-means
Test	**35.06**	17.93	13.79	12.83	11.56

In Table 11, to verify the performance of PM, we extract the "sentence" field of the dataset samples and supplement it with the "label" field, combining them into a new dataset S, which constitutes a binary classification task. The "label" fields are set as follows:

$$label = \begin{cases} 0 & \text{length(answer)} = 1 \\ 1 & \text{length(answer)} \geq 2 \end{cases} \tag{26}$$

where length(.) means calculating the list length.

Then, we compared AGN (Li et al., 2021), BERT (Devlin et al., 2018), Transformer (Vaswani et al., 2017) and other prompt templates. The results show that the recognition ability of PM is competitive. Other Prompt templates are as follows:

Template 1: $[X1]$, 这是一道{单,多}选题

Template 2: 它是一个{单,多}选题, $[X1]$

As shown in Table 12, we verify the effectiveness of CM by comparing the accuracy of random and clustering methods. We form a small dataset L by choosing question samples whose predicted answers are empty in the test set. To ensure fairness and better simulate the random post-processing of online experiments, we conducted five consecutive random experiments, and the accuracy rates of the five experiments were (17.24%, 27.58%, 10.34%, 13.79%,20.69%), with an average accuracy of 17.93%. Methods such as bow, TF-IDF, and K-Means are less accurate and equally uncompetitive. The results show that CM has an advantage.

4.5. Case analysis

To better compare and test our method, as shown in Appendix, we selected several typical questions to compare the differences between our model and other models.

The PLM here refers to a basic pre-training model that predicts most of the options for the correct first question but lacks an understanding of "environmental factors". That is because pre-training cannot capture the implicit semantic relationship of "environment" in a legal context. The PLM-finetuning model represents a fine-tuned pre-training language model that can capture the implicit semantics of "environment", but it lacks the specific meaning of the system "in society". The prediction results of BCA show that it can obtain special semantics in specific situations and make up for the deficiencies of the former two. Similarly, in the second question, PLM chose utterly wrong. Although the prediction result of PLM-finetuning contained the correct answer, it chose other options incorrectly, and BCA could get all the correct answers. The above prediction results show that the fine-tuned

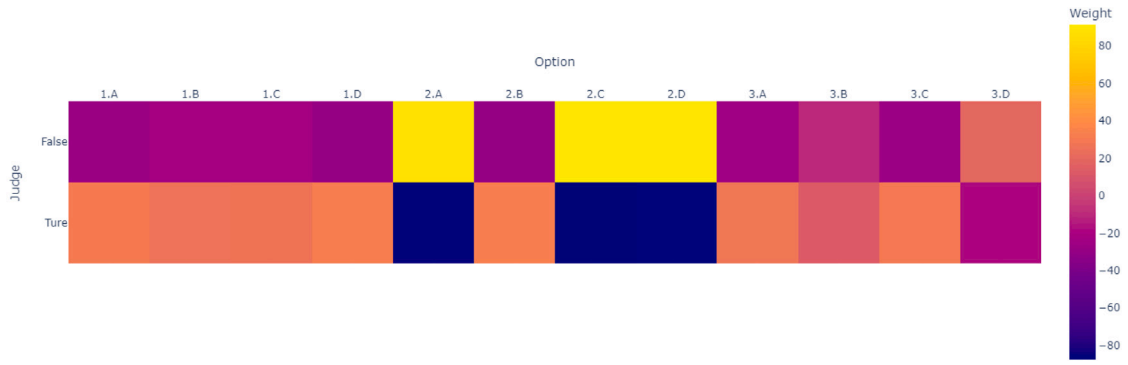

Fig. 4. Distribution of attention weights. The numbers and letters represent the options for the questions in Appendix. "True" represents the selected weight of this option. "False" represents the weight of the option not selected. "Weight" represents the weight value of the model. The higher the value, the higher the attention will be, and vice versa.

Table A.1
Case analysis.

Statement	PLM	PLM-FT	BCA	Candidate options
法律秩序是人们在社会生活中依法行事而形成的行为有规则和有序的状态。影响法律秩序的因素是多方面的,主要包括下列哪些选项? *The legal order is a state where people act under the law in social life with regular and orderly behaviour. Many factors affect the legal order. Which options are belonged to this factor?*	✓ ✓ ✓ ✓	 ✓ ✓ ✓	✓ ✓ ✓ ✓	✓A: 体制方面的因素 *Institutional factors* ✓B: 个人方面的因素 *Personal factors* ✓C: 环境方面的因素 *Environmental factors* ✓D: 法律本身的因素 *Legal factors*
兹有四个事例:①张某驾车违章发生交通事故致搭车的李某残疾;②唐某参加王某组织的自助登山活动因雪崩死亡;③吴某与人打赌举重物因用力过猛致残;④赵某心情不好邀好友郑某喝酒,郑某畅饮后驾车撞树致死。根据公平正义的法治理念和民法有关规定,下列哪一观点可以成立? *Here are four cases: 1. Zhang violated the rules and caused a traffic accident, which resulted in the disability of Li, who was taking a ride; 2. Tang died of an avalanche when he participated in a self-help mountaineering activity organized by Wang; 3. Wu was disabled due to excessive force in a bet with someone. 4. Zhao was in a bad mood and invited his friend Zheng to drink. After drinking, Zheng drove into a tree and died. According to the rule of law concept of fairness and justice and the relevant provisions of civil law, which of the following viewpoints can be established?*	 ✓	✓ ✓ ✓	 	✗ A:①张某与李某未形成民事法律关系合意,如让张某承担赔偿责任,是惩善扬恶,显属不当 *Zhang and Li have not formed a civil legal relationship agreement. If Zhang is liable for compensation, it is to punish the good and promote evil, which is inappropriate.* ✓B:② 唐某应自担风险,如让王某承担赔偿责任,有违公平 *Tang should bear the risk at his own risk. It would be against fairness to make Wang bear the responsibility for compensation.* ✗ C:③ 吴某有完整意思能力,其自担损失,是非清楚 *Wu has a complete ability to make sense, he is responsible for the loss, and it is clear whether it is right or wrong.* ✗ D:④ 赵某虽有召集但未劝酒,无需承担责任,方能兼顾法理与情理 *Although Zhao had summoned but did not persuade him to drink, he does not need to take responsibility so that he can consider legal and rational.*
甲、乙经共谋后到丙的住所对其实施了强奸,事后,甲趁丙不注意之机,将丙的钱包拿走。第二天,甲发现丙的钱包里有一张已经中了5万元的彩票,即兑了奖。就甲拿走被害人钱包和私自兑奖的行为而言,下列哪些选项是正确的? *After Party A and Party B conspired, they went to Party C's residence to rape him. Afterwards, Party A took advantage of Party C's inattentiveness and took Party C's wallet away. The next day, Party A found that Party C's wallet contained a lottery ticket that had already won 50,000 yuan, and Party A redeemed the prize. Which of the following options is correct regarding Party A taking the victim's wallet and redeeming the prize privately?*	 ✓	✓ ✓	 ✓ ✓	✗ A: 甲和乙成立盗窃罪的共同犯罪 *Party A and B establish a joint crime of theft.* ✓B: 甲单独对自己的行为承担刑事责任 *Party A is solely criminally responsible for his actions.* ✓C: 甲的行为构成盗窃罪 *Party A's conduct constitutes theft.* ✗ D: 甲的行为构成盗窃罪和诈骗罪,应实行数罪并罚 *Party A's behaviour constitutes the crime of theft and fraud and should be punished for several crimes.*

pre-trained model has a more vital learning ability than the original model. In addition, BCA can obtain the implicit semantic relationship between options and correct the illogical answer list, making it more suitable for judicial examination tasks. However, in the third question, the three models are not satisfactory. In contrast, BCA is slightly better because its answer list contains correct answers, while both PLM and PLM-finetuning are missing, and options A and B have obvious contradictory statements, which BCA can distinguish. Although, to a certain extent, it shows that BCA can capture deep legal semantic relations. However, there are still some deficiencies implicit in obtaining the semantic relationship.

In order to better demonstrate the effect of the model's weight on sentence level, as shown in 4, the model's prediction of the data in Appendix after BM is intercepted. We can see that the correct options for the first two questions have higher attention weights, while the wrong options have relatively low or negative ones. In addition, for the third question, It is worth noting that after PCM processing, option A was excluded, and finally, although the correct options B and C were not excluded, another wrong option was added. It also illustrates the effectiveness and limitations of BCA from another aspect. Our subsequent work will explore this issue in depth.

5. Conclusion

The BCA proposed in this paper is a type-based solution supplemented by a data augmentation module and a post-processing module for the judicial examination task. Online and offline experimental results show that it has good performance. However, we find that the lack of external knowledge in the legal domain, the incorporation of external knowledge into the model, and the multi-hop reasoning

applying knowledge are significant challenges for this task. Therefore, we will consider introducing external data to construct knowledge graphs in the future and introduce knowledge graphs and multi-hop reasoning technology into existing methods.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work is supported by the National Key Research and Development Program of China (No. 2022YFC3301801), the Fundamental Research Funds for the Central Universities (No. DUT22ZD205).

Appendix. Case analysis

(See Table A.1.)

References

Adiwardana, Daniel, Luong, Minh-Thang, So, David R, Hall, Jamie, Fiedel, Noah, Thoppilan, Romal, Yang, Zi, Kulshreshtha, Apoorv, Nemade, Gaurav, Lu, Yifeng, et al., 2020. Towards a human-like open-domain chatbot. arXiv preprint arXiv:2001.09977.

Brown, Tom, Mann, Benjamin, Ryder, Nick, Subbiah, Melanie, Kaplan, Jared D, Dhariwal, Prafulla, Neelakantan, Arvind, Shyam, Pranav, Sastry, Girish, Askell, Amanda, et al., 2020. Language models are few-shot learners. Adv. Neural Inf. Process. Syst. 33, 1877–1901.

Clark, Christopher, Gardner, Matt, 2017. Simple and effective multi-paragraph reading comprehension. arXiv preprint arXiv:1710.10723.

Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, Toutanova, Kristina, 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Dhingra, Bhuwan, Liu, Hanxiao, Yang, Zhilin, Cohen, William W, Salakhutdinov, Ruslan, 2016. Gated-attention readers for text comprehension. arXiv preprint arXiv:1606.01549.

Ding, Ning, Wang, Xiaobin, Fu, Yao, Xu, Guangwei, Wang, Rui, Xie, Pengjun, Shen, Ying, Huang, Fei, Zheng, Hai-Tao, Zhang, Rui, 2021. Prototypical representation learning for relation extraction. arXiv preprint arXiv:2103.11647.

Goldberg, Yoav, 2019. Assessing BERT's syntactic abilities. arXiv preprint arXiv:1901.05287.

Han, Xu, Zhao, Weilin, Ding, Ning, Liu, Zhiyuan, Sun, Maosong, 2021. Ptr: Prompt tuning with rules for text classification. arXiv preprint arXiv:2105.11259.

He, Wei, Liu, Kai, Liu, Jing, Lyu, Yajuan, Zhao, Shiqi, Xiao, Xinyan, Liu, Yuan, Wang, Yizhong, Wu, Hua, She, Qiaoqiao, et al., 2017. Dureader: a chinese machine reading comprehension dataset from real-world applications. arXiv preprint arXiv:1711.05073.

Hermann, Karl Moritz, Kocisky, Tomas, Grefenstette, Edward, Espeholt, Lasse, Kay, Will, Suleyman, Mustafa, Blunsom, Phil, 2015. Teaching machines to read and comprehend. Adv. Neural Inf. Process. Syst. 28.

Jawahar, Ganesh, Sagot, Benoît, Seddah, Djamé, 2019. What does BERT learn about the structure of language? In: ACL 2019-57th Annual Meeting of the Association for Computational Linguistics.

Jin, Di, Gao, Shuyang, Kao, Jiun-Yu, Chung, Tagyoung, Hakkani-tur, Dilek, 2020. Mmm: Multi-stage multi-task learning for multi-choice reading comprehension. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34. pp. 8010–8017.

Kano, Yoshinobu, Hoshino, Reina, Taniguchi, Ryosuke, 2017. Analyzable legal yes/no question answering system using linguistic structures. In: COLIEE@ ICAIL. pp. 57–67.

Kim, Mi-Young, Xu, Ying, Goebel, Randy, 2014. Legal question answering using ranking svm and syntactic/semantic similarity. In: JSAI International Symposium on Artificial Intelligence. Springer, pp. 244–258.

Kim, Mi-Young, Xu, Ying, Goebel, Randy, Satoh, Ken, 2013. Answering yes/no questions in legal bar exams. In: JSAI International Symposium on Artificial Intelligence. Springer, pp. 199–213.

Kim, Mi-Young, Xu, Ying, Lu, Yao, Goebel, Randy, 2016. Question answering of bar exams by paraphrasing and legal text analysis. In: JSAI International Symposium on Artificial Intelligence. Springer, pp. 299–313.

Krizhevsky, Alex, Sutskever, Ilya, Hinton, Geoffrey E., 2012. Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. 25.

Lai, Guokun, Xie, Qizhe, Liu, Hanxiao, Yang, Yiming, Hovy, Eduard, 2017. Race: Large-scale reading comprehension dataset from examinations. arXiv preprint arXiv:1704.04683.

Lan, Zhenzhong, Chen, Mingda, Goodman, Sebastian, Gimpel, Kevin, Sharma, Piyush, Soricut, Radu, 2019. Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942.

Lester, Brian, Al-Rfou, Rami, Constant, Noah, 2021. The power of scale for parameter-efficient prompt tuning. arXiv preprint arXiv:2104.08691.

Li, Xianming, Li, Zongxi, Xie, Haoran, Li, Qing, 2021. Merging statistical feature via adaptive gate for improved text classification. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 35. pp. 13288–13296.

Li, Xiang Lisa, Liang, Percy, 2021. Prefix-tuning: Optimizing continuous prompts for generation. arXiv preprint arXiv:2101.00190.

Lin, Yankai, Ji, Haozhe, Liu, Zhiyuan, Sun, Maosong, 2018. Denoising distantly supervised open-domain question answering. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Volume 1: Long Papers, pp. 1736–1745.

Liu, Yinhan, Ott, Myle, Goyal, Naman, Du, Jingfei, Joshi, Mandar, Chen, Danqi, Levy, Omer, Lewis, Mike, Zettlemoyer, Luke, Stoyanov, Veselin, 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Liu, Pengfei, Yuan, Weizhe, Fu, Jinlan, Jiang, Zhengbao, Hayashi, Hiroaki, Neubig, Graham, 2021a. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. arXiv preprint arXiv:2107.13586.

Liu, Xiao, Zheng, Yanan, Du, Zhengxiao, Ding, Ming, Qian, Yujie, Yang, Zhilin, Tang, Jie, 2021b. GPT understands, too. arXiv preprint arXiv:2103.10385.

Ma, Xiaofei, Wang, Zhiguo, Ng, Patrick, Nallapati, Ramesh, Xiang, Bing, 2019. Universal text representation from BERT: an empirical study. arXiv preprint arXiv:1910.07973.

Martinez-Gil, Jorge, 2021. A survey on legal question answering systems. arXiv preprint arXiv:2110.07333.

Minaee, Shervin, Kalchbrenner, Nal, Cambria, Erik, Nikzad, Narjes, Chenaghlu, Meysam, Gao, Jianfeng, 2021. Deep learning–based text classification: a comprehensive review. ACM Comput. Surv. 54 (3), 1–40.

Rajpurkar, Pranav, Zhang, Jian, Lopyrev, Konstantin, Liang, Percy, 2016. Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250.

Rakhlin, A., 2016. Convolutional neural networks for sentence classification. GitHub.

Richardson, Matthew, Burges, Christopher J.C., Renshaw, Erin, 2013. Mctest: A challenge dataset for the open-domain machine comprehension of text. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. pp. 193–203.

Seo, Minjoon, Kembhavi, Aniruddha, Farhadi, Ali, Hajishirzi, Hannaneh, 2016. Bidirectional attention flow for machine comprehension. arXiv preprint arXiv:1611.01603.

Shin, Taylor, Razeghi, Yasaman, Logan IV, Robert L, Wallace, Eric, Singh, Sameer, 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. arXiv preprint arXiv:2010.15980.

Tang, Min, Cai, Jiaran, Zhuo, Hankz Hankui, 2019. Multi-matching network for multiple choice reading comprehension. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 7088–7095.

Taniguchi, Ryosuke, Hoshino, Reina, Kano, Yoshinobu, 2018. Legal question answering system using framenet. In: JSAI International Symposium on Artificial Intelligence. Springer, pp. 193–206.

Taniguchi, Ryosuke, Kano, Yoshinobu, 2016. Legal yes/no question answering system using case-role analysis. In: JSAI International Symposium on Artificial Intelligence. Springer, pp. 284–298.

Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N, Kaiser, Łukasz, Polosukhin, Illia, 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.

Verma, Aayushi, Morato, Jorge, Jain, Arti, Arora, Anuja, 2020. Relevant subsection retrieval for law domain question answer system. In: Data Visualization and Knowledge Engineering. p. 299.

Wang, Shuohang, Jiang, Jing, 2016. Machine comprehension using match-lstm and answer pointer. arXiv preprint arXiv:1608.07905.

Wang, Wenhui, Yang, Nan, Wei, Furu, Chang, Baobao, Zhou, Ming, 2017a. Gated self-matching networks for reading comprehension and question answering. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 189–198.

Wang, Shuohang, Yu, Mo, Chang, Shiyu, Jiang, Jing, 2018. A co-matching model for multi-choice reading comprehension. arXiv preprint arXiv:1806.04068.

Wang, Shuohang, Yu, Mo, Jiang, Jing, Zhang, Wei, Guo, Xiaoxiao, Chang, Shiyu, Wang, Zhiguo, Klinger, Tim, Tesauro, Gerald, Campbell, Murray, 2017b. Evidence aggregation for answer re-ranking in open-domain question answering. arXiv preprint arXiv:1711.05116.

Wu, Jiaye, Luo, Xudong, 2021. Alignment-based graph network for judicial examination task. In: International Conference on Knowledge Science, Engineering and Management. Springer, pp. 386–400.

Xiao, Chaojun, Hu, Xueyu, Liu, Zhiyuan, Tu, Cunchao, Sun, Maosong, 2021. Lawformer: A pre-trained language model for Chinese legal long documents. AI Open 2, 79–84.

Yang, Wei, Xie, Yuqing, Tan, Luchen, Xiong, Kun, Li, Ming, Lin, Jimmy, 2019. Data augmentation for bert fine-tuning in open-domain question answering. arXiv preprint arXiv:1904.06652.

Yang, Zichao, Yang, Diyi, Dyer, Chris, He, Xiaodong, Smola, Alex, Hovy, Eduard, 2016. Hierarchical attention networks for document classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 1480–1489.

Yin, Wenpeng, Ebert, Sebastian, Schütze, Hinrich, 2016. Attention-based convolutional neural network for machine comprehension. arXiv preprint arXiv:1602.04341.

Zhang, Shuailiang, Zhao, Hai, Wu, Yuwei, Zhang, Zhuosheng, Zhou, Xi, Zhou, Xiang, 2019. Dual co-matching network for multi-choice reading comprehension. arXiv preprint arXiv:1901.09381.

Zheng, Wenjiao, 2021. Exploration on the training path of compound talents in artificial intelligence and law in China. Education and Teaching Forum.

Zhong, Haoxi, Xiao, Chaojun, Tu, Cunchao, Zhang, Tianyang, Liu, Zhiyuan, Sun, Maosong, 2020. Jec-qa: A legal-domain question answering dataset. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34. pp. 9701–9708.

Zhu, Yunchang, Pang, Liang, Lan, Yanyan, Shen, Huawei, Cheng, Xueqi, 2021a. Adaptive information seeking for open-domain question answering. arXiv preprint arXiv:2109.06747.

Zhu, Haichao, Wei, Furu, Qin, Bing, Liu, Ting, 2018. Hierarchical attention flow for multiple-choice reading comprehension. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 32.

Zhu, Pengfei, Zhang, Zhuosheng, Zhao, Hai, Li, Xiaoguang, 2021b. Duma: Reading comprehension with transposition thinking. IEEE/ACM Trans. Audio Speech Lang. Process. 30, 269–279.