Future Generation Computer Systems 113 (2020) 218–227

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Multi-task reading for intelligent legal services

Yujie Li [b], Gang Hu [c], Jinyang Du [c], Haider Abbas [d], Yin Zhang [a,*]

[a] School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu, China
[b] Yangzhou University, Yangzhou, China
[c] School of Information and Safety Engineering, Zhongnan University of Economics and Law, Wuhan, China
[d] National University of Sciences and Technology, Islamabad, Pakistan

ARTICLE INFO

Article history:
Received 1 March 2020
Received in revised form 26 June 2020
Accepted 1 July 2020
Available online 8 July 2020

Keywords:
Legal intelligence
Legal empirical research
Machine reading comprehension
Bert
Multi-task learning

ABSTRACT

Since legal data contains both structured data and unstructured data, it is a great challenge to implement machine reading comprehension technology in empirical analysis of law. This paper proposes a multi-tasking reading for intelligent legal services, which applies statistical analysis and machine reading comprehension techniques, and can process both structured and unstructured data. At the same time, this paper proposes a machine reading comprehension model that can perform multi-task learning, LegalSelfReader, which can solve the problem of diversity of questions. In the experiment of the legal reading comprehension dataset CJRC, the model proposed in this paper is far superior to the two classic models of BIDAF and Bert in three evaluation indicators. And our model is also better than some models published by HFL(Harbin Institute of Technology and iFly Joint Lab), and has also achieved lower consumption in training costs. At the same time, in the experiment of visualizing the attention value, it also demonstrates that the model proposed in this paper has a stronger ability to extract evidence.

1. Introduction

Applying artificial intelligence technology to the legal field can speed up and improve the legal research process and reduce the time cost and capital of legal research [1], which makes legal intelligence research [2] a very promising field. Katz pointed out in the 2012 study that with the rapid development of artificial intelligence, traditional legal tasks such as the generation of legal documents and the prediction of case results will usher in change [3]. This change can also be seen in three other aspects. First, speech recognition technology is used for trial records [4]. Second, using machine learning methods to assist lawyers in the review of legal documents [5]. Furthermore, some machine learning methods have also been applied to build intelligent referee systems [6,7].

It can be seen that when artificial intelligence is applied to legal research, natural language understanding has become the most promising technology for successful application, because legal research contains a large amount of text data. For example, in data-driven legal empirical analysis [8], researchers need to manually read a large number of referee documents and organize the data, which is a rather time-consuming and laborious process. If you use machine reading comprehension technology and build an auxiliary reading system, you can alleviate the burden of researchers in the process of summarizing data.

Machine reading comprehension tasks are usually defined as given a passage and a passage-related question. After reading the passage, the model gives the answer to the question [9]. After the emergence of deep learning, machine reading comprehension technology has made great progress. Some researchers in previous years have been working on solving the problem of cloze or one-way selection, and models such as Attentive Reader [10], Stanford AR [11], GA Reader [12], and AOA Reader [13] have appeared. Later, the development of machine reading comprehension tends to solve problems close to real-life scenarios, such as span extraction problems or multi-task type problems, such as BiDAF [14], Match-LSTM [15], S-net [9] and other models. In 2018, Google released a powerful language model—Bert [16]. The model has been successful in 11 natural language processing tasks, and to some extent has improved the most advanced performance of machine reading comprehension. With the continuous development of machine reading comprehension technology, it demonstrates the ability to deal with long text and multitasking problems, which makes machine reading comprehension technology possible to apply to legal empirical analysis. However, there are still many challenges in applying machine reading comprehension to the process of legal empirical analysis.

* Corresponding author.
E-mail addresses: yzyjli@gmail.com (Y. Li), hugang@stu.zuel.edu.cn (G. Hu), 201912200075@stu.zuel.edu.cn (J. Du), dr.h.abbas@ieee.org (H. Abbas), yin.zhang.cn@ieee.org (Y. Zhang).

https://doi.org/10.1016/j.future.2020.07.001
0167-739X/

1. Structured and unstructured legal data:
   The legal empirical analysis process is a relatively complex process. It contains many forms of data, including structured data such as statistical yearbooks, as well as unstructured data such as interview records and referee documents. Therefore, it is a research method that requires both structured data analysis and unstructured analysis;
2. Diversity of questions:
   For a referee documents, the researcher may ask questions that have answers, such as the sentence of the offender, the cause of the crime, and whether there is a gang crime. At the same time, the researcher may also ask questions that cannot be answered according to the referee documents. The machine reading comprehension model of the traditional fragment extraction class does not deal with this complex type of problem;

In this context, this paper investigates the process of legal researchers in the empirical research of law, proposes a solution of our own, and designs a multi-task reading for intelligent legal services.

1. The system first uses statistical analysis methods to satisfy the structured data measurement part of empirical research. Such measurements can measure structured data in empirical research, such as statistical yearbooks. For unstructured data such as interview records and referee documents, the system uses machine reading comprehension technology to analyze and meet the measurement part of unstructured data in empirical research;
2. In order to solve the problem diversity, this paper designs a machine reading comprehension model that can perform multi-task learning-LegalSelfReader. The model can deal with three types of problems: span extraction, yes or no judgment, and unanswerable question. This basically satisfies the problem type requirements in the empirical analysis of law.

The remaining chapters of this article will be arranged as follows. Section 2 will introduce the related research on machine reading comprehension. Section 3 will show the legal empirical research system designed in this paper. Section 4 will demonstrate the performance of the proposed model on a legal reading comprehension dataset. The fifth chapter introduces the conclusions of this paper and looks forward to it at the same time.

2. Related research

One of the difficulties faced by empirical research is that for unstructured data, such as text data, researchers can only rely on humans to make measurements. This is a slow and time-consuming process. A feasible solution is to let the machine reading comprehension model in natural language processing replace the researcher, read the legal judgment documents, legal dossier and other data to complete the qualitative measurement task.

Machine reading comprehension is an important task in natural language processing. In recent years, it has received more and more scholars' attention and made some good progress.

2.1. Evaluation method of reading ability

One difficulty of machine reading comprehension is how to test the model's ability to understand text. This is a difficulty in evaluating indicators, and it is also a difficulty in datasets. Using datasets to simulate realistic reading comprehension scenarios requires taking into account both the authenticity of the simulation and the reading ability of the current model. In some previous studies, due to lack of model capabilities, researchers usually construct relatively simple datasets. For example, the CNN/Daily mail [10] dataset proposed by Hermann in 2014 belongs to the cloze type dataset. Then there is the TOEFL listening dataset constructed by Tseng et al. [17], which belongs to the single selection category. The prediction results of this type of dataset usually exist in a relatively small result set.

At the same time, some other researchers believe that the reading comprehension tasks of cloze and single-choice categories do not meet the real reading comprehension tasks in reality. The real reading comprehension task should be to give one or more documents, then give a question, and then find the corresponding content from the original text according to the question. This produces a reading comprehension task of segment extraction. In 2017, Stanford released the SQuAD dataset [18] and organized an online competition. In the subsequent SQuAD2.0, it also contains two types of questions: span extraction and unanswerable, which is more in line with the real scene. In 2018, Baidu also released a larger Chinese reading comprehension dataset-DuReader [19],. Their task is similar to SQuAD, and it further considers the diversity of reports in the Internet scenario, constructed multiple documents corresponding to the tasks of multiple types of problems.

Moreover, some researchers believe that pure fragment extraction may not be a real task. They believe that reading comprehension in reality should be a process of extracting evidence, that is, first find several evidences that can answer questions from the chapter—this is still a task of extracting fragments. However, it is necessary to synthesize these evidences into one answer later rather than directly as an answer. Microsoft's MS-MARCO dataset [20], is a representative of this category.

Structured and unstructured legal data: Using a dataset that is closer to the actual reading scenario, a model with stronger reading ability can be trained. Therefore, the simulation of real reading scenarios is the top priority in machine reading comprehension.

In the previous dataset research, the CJRC dataset released by HFL(Harbin Institute of Technology and iFly Joint Lab) is a typical multi-tasking dataset for legal data. The types of questions include three types of questions: span extraction, yes or no judgment, and unanswerable question. It is a dataset that is more in line with the actual reading scenarios of legal documents. However, there is still a considerable gap between the dataset and the reading scenarios in actual legal empirical research. In legal empirical research, legal researchers may get structured data, such as statistical yearbooks, or they may get unstructured data, such as judgment documents and files. Their dataset ignores the fact that structured data reading exists in empirical research in law. Therefore, the model trained using the CJRC dataset cannot complete the reading task of legal empirical research end-to-end.

2.2. Models

At the same time, the machine reading comprehension model can also be divided into several types:

1. Attention-based model:
   Attention-based models can also be divided into three subtypes. The first type is the traditional attention-based model, which includes the Stanford AR [11] model proposed by Hermann et al. Attentive Reader [10] and Chen et al. Among them, Attentive Reader uses the feedforward neural network as the attention function, and Stanford AR based on the Attentive Reader study, using the bilinear function as the attention function.

The second type is the multi-hop attention model. The multi-hop attention mechanism simulates human multi-step reading, hoping to obtain a deeper understanding of the passage through the multiple attention accumulation of the model. This includes the AMRNN model proposed by Tseng et al. [17] and the GA Reader proposed by Dhingra et al. [12].

The third type is a Pointer-wise based model. For the cloze-type question, the answer will only come from the corresponding chapter. The traditional machine comprehension model constructs the full dictionary greatly increases the search space of the answer. If you use the Pointer-wise mechanism, you can limit the answer to the chapter. This type of model includes the AS Reader model proposed by Kadlec et al. [21] and AOA Reader [13] proposed by Cui and other improved AS Reader.

2. Span extraction model:
   In the span extraction model, there are two typical models. One is the BiDAF model proposed by Seo et al. [14], which uses a bi-attention mechanism from question to passage and passage to the question, so that the model can capture more important information. The other is the match-LSTM model proposed by Wang et al. [15]. The attention layer of the model uses the match-LSTM attention mechanism that they proposed in the text implied task, and the output layer uses the Pointer network. The Pointer network can limit the answer only from the passage.
3. Some heuristic multi-stage models:
   Tanh et al. believe that human beings are an extraction-and-synthesis process when answering questions with multiple evidences. Thus, they proposed an extract-and-synthesis model, S-net, which uses an improved match-LSTM for evidence presentation and then uses a seq2seq model to synthesize answers [9]. At the same time, for the unanswerable type problem, Hu et al. proposed a read-then-verify model. The reader part of the model is still a model of the span extraction class. The reader model outputs the start and end position probability of the span and the probability of being unanswerable. The verify part verifies the questionability of the question by verifying the legitimacy of the answer [22].
4. Bert's related model:
   In 2018, scholars from Google proposed the Bert pre-training model. In the published papers, they showed that Bert achieved good performance improvement on 11 NLP tasks such as machine reading comprehension [16]. Later, some scholars discussed the problem of Bert's handling of Chinese NLP tasks, and made some good progress, including the ERNIE [23] proposed by Baidu and the Bert-WWM model proposed by HFL(Harbin Institute of Technology and iFly Joint Lab) [24].

Diversity of questions in legal reading: In conventional machine reading comprehension tasks, there are multiple reports of text content, and multiple types of questions need to be answered for multiple documents. However, in the reading comprehension of legal documents, there is uniqueness in the reporting of the content of the documents. Usually answer multiple types of questions for a single document. This is the simpler part of legal reading comprehension. But the difficulty is that while the problems are diverse, the legal reading comprehension tends to be evidence extraction type problems, such as yes or no problems. To deal with this type of problem requires the model to have good evidence extraction capabilities. In the known research, no researchers have discussed the possibility of constructing this model in legal reading comprehension. In this case, this paper proposes a legal document reading model—LegalSelfReader, which can handle multiple types of problems at the same time, and the subsequent experimental results prove that it has good evidence extraction ability. At the same time, in the previous study, this article chose the more classic BiDAF and Bert models as the benchmark of this article.

3. Model design and system implementation

3.1. The overall structure of the legal empirical research system

Due to the structured and unstructured legal data, the existing machine reading comprehension models trained on the existing datasets cannot directly construct a legal reading comprehension system that conforms to real-life scenarios. Therefore, this article considers a more reasonable way to construct a legal reading comprehension system, that is, using a machine reading comprehension model to process unstructured data such as referee documents and other statistical analysis models to process structured data such as statistical yearbooks. According to the actual needs of legal researchers, this paper designs a legal empirical research system as shown in Fig. 1.

The system is based on the machine reading comprehension technology in NLP, which can quantitatively measure and qualitatively measure the legal data at the same time. The system is mainly divided into five parts.

1. Data input module:
   After a legal researcher determines a legal study, he or she needs to collect certain legal data according to the research needs. These data include: existing statistical data, legal documents, case materials, court records, etc. Statistical data such as statistical yearbooks are available and can be processed using statistical methods. Data such as legal documents, case materials, and court records are unstructured data, which requires more advanced processing. In this system, machine reading comprehension models will be used to process qualitative data. The data can be input by a legal researcher to sort out the input by himself or by inputting a keyword to allow the system-assisted legal researcher to craw relevant data from the Internet.
2. Data preprocessing module:
   After the original data is obtained, data cleaning and data conversion are required to be input to the subsequent processing module. For the original statistical data, it is necessary to fill in the missing items in the data, delete or replace the abnormal items, and perform statistics on the outliers. This requires a complete statistical data preprocessing module. For textual data such as the original legal documents, case materials, and court records, the original data is first filtered to process the null data. Then, it is necessary to segment the text type data, and then map the word sequence data after the word segmentation into digital sequence data to facilitate the processing of the subsequent model.
3. Data analysis module:
   For the statistical data that has been processed, this paper designs a simple statistical processing module to analyze and process them. These analysis functions include: mean, variance, median, regression analysis, cluster analysis and so on. For text data that has already been processed, this paper uses a machine reading comprehension model to analyze it. In the empirical study of law, a concept generation process overlaps with the measurement process, so the measurement requirement may be analyzed after the researcher obtains the text. At this time, the researcher is required to give a specific measurement problem for the specific text data, and the model will find the answer in

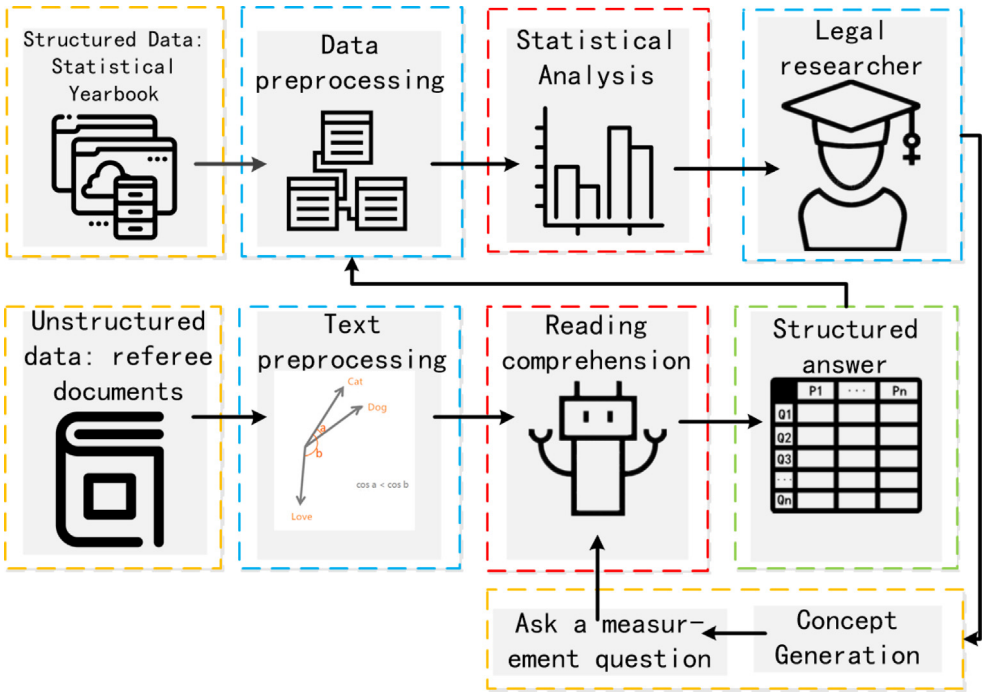


Fig. 1. The overall structure of the legal empirical research system.

the text data according to the problem, and then feedback to the researcher. The internal details of the model will be shown in the next summary.

4. Post-processing module for reading results:
   In this module, the data will be further integrated, and through the steps of question classification and answer extraction, a structured reading result data will be formed. The analysis process of structured reading results is similar to that of structured data, and both will be sent to the data preprocessing module and statistical analysis module for further analysis.
5. Result push module:
   The integrated data will be fed back to the legal researchers, which will help the researchers complete the background research part of the legal research. At the same time, the data will be simply analyzed by regression analysis and cluster analysis to help the legal researchers find the truth from the complicated data. The existing legal problems, the researchers then based on their theoretical knowledge, propose solutions to the problem.

3.2. LegalSelfReader

To address the diversity of questions, this paper defines the task of reading and understanding legal documents as three subtasks, which are span extraction task, yes or no judgment, and unanswerable type, as show in Fig. 2. Dealing with these three types of problems at the same time requires the model to satisfy a multi-tasking training architecture, so that the coding layers can be shared between different outputs of the model, but relatively independent training. Therefore, this paper establishes a three-output legal document self-reading model, which is unanswerable probability prediction, non-probability prediction and span probability prediction.

3.2.1. Bert layer

The span extraction prediction part of the model is divided into 5 layers: The first layer is the Bert layer, which uses the Bert-Chinese implementation proposed by Google to encode the input chapters and questions. The preprocessed chapters and problem sequences are organized into three sequences. Dictionary mapping sequence for words:

$$\{E_{[CLS]}, E_{token_1}, \ldots, E_{token_n}, E_{[SEP]}, \tag{1}$$

$$E_{token_1}, \ldots, E_{token_m}, E_{[SEP]}\} \tag{2}$$

Where $[CLS]$ and $[SEP]$ is a separator, the sequence of questions is $token_1, \ldots, token_n$, the length is n, the chapter sequence is $token_1, \ldots, token_m$, and the length is m.

The sequence of front and back marks sentence, the problem sequence is marked as A, and the chapter sequence is marked as B:

$$\{E_A, \ldots, E_A, E_B, \ldots, E_B\} \tag{3}$$

The sequence of word positions, the sequence of positions of the question is $E_0, \ldots, E_n$, and the sequence of positions of the passage is $E_1, \ldots, E_m$:

$$\{E_0, \ldots, E_n, E_0, \ldots, E_m\} \tag{4}$$

The three sequences inside Bert are summed and encoded to get the final result.

$$\{E'_{[CLS]}, E'_{token_1}, \ldots, E'_{token_n}, E'_{[SEP]}, \tag{5}$$

$$E'_{token_1}, \ldots, E'_{token_m}, E'_{[SEP]}\} \tag{6}$$

3.2.2. Feature fusion layer

The second layer is a feature fusion layer. Some previous studies have shown that adding some a priori features related to words can improve the performance of the model [9,25]. The same applies to the scenario of legal judgment documents. Adding a named entity identification vector helps the model to identify the name of the offender, the location of the crime, the name of the criminal gang (such as the company name, etc.). Adding a partofspeech vector helps the model to identify some entity words, quantifiers, and so on. Therefore, after obtaining the semantic coding vector, the model further fuses the part-of-speech tagging and the named entity tagging vector C_t to obtain

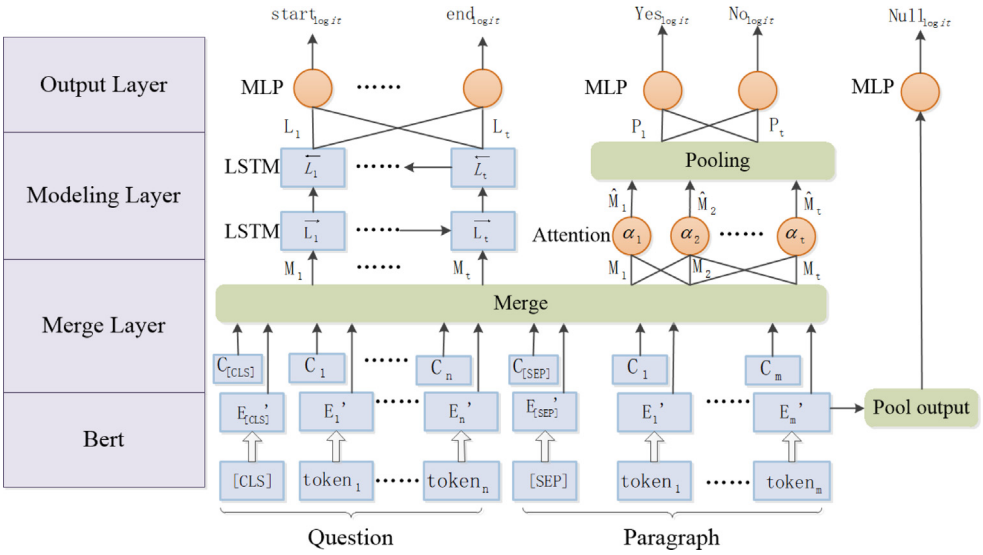


Fig. 2. LegalSelfReader.

the semantic coding vector M_t containing the rich features.

$$M_t = [E'_t; C_t] \quad t \in [1, n+m+3] \tag{7}$$

3.2.3. Modeling layer

In the reading task of the real scene, if the question type is yes or no, then we usually choose to skim the full text and care more about the macro information of the article. At this time, we can get more important in the article through the attention layer and the pooling layer. section. If the problem type is a elaboration class topic (span extraction is similar to this type of topic), we usually read the full text in order to get more detailed context information. At this time we choose to keep the complete word vector representation and then use the bidirectional LSTM to go further. The context information is extracted and then mapped using a layer of MLP.

1. Modeling for span extraction prediction:
   This layer uses a bidirectional LSTM network to process the forward semantic coding vector $\overrightarrow{M_t}$ and the backward semantic coding vector $\overleftarrow{M_t}$ on a time-step t basis, obtaining forward and backward context vector $\overrightarrow{L_t}$ and $\overleftarrow{L_t}$, and connecting the two to obtain the final context vector L_t.

$$\overrightarrow{L_t} = LSTM(\overrightarrow{M_t}) \tag{8}$$

$$\overleftarrow{L_t} = LSTM(\overleftarrow{M_t}) \tag{9}$$

$$L_t = [\overrightarrow{L_t}; \overleftarrow{L_t}] \tag{10}$$

2. Modeling for yes or no judgments:
   The first layer is the self-attention layer. The self-attention mechanism can discover the autocorrelation in the sequence, get the front and back dependence of the sequence elements, and output the implicit representation of the dependency. For the fusion representation $M = M_1, M_2, \ldots, M_t$ passed by the feature fusion layer, the attention layer uses the feedforward network whose activation function is relu, and obtains the attention value for the front and rear elements M_i and M_j in the M sequence. And use this attention value to get a new fusion representation $\hat{M}$.

$$S_M^t = relu(V^t \cdot [W_M^i \cdot M_i + W_M^j \cdot M_j]) \tag{11}$$

$$\alpha_j^i = relu(S_M^t) \tag{12}$$

$$\hat{M}_j = \sum_{k=1}^{n} \alpha_j^i \cdot M_j \tag{13}$$

The second layer is the pooling layer. This article uses an average pooling layer to get a final pooled output P_i:

$$P_i = average\_pooling(\hat{M}_i) \tag{14}$$

3.2.4. Output layer

The fifth layer is the output layer, which is implemented using MLP (MultiLayer Perceptron):

For the output of the span prediction:

$$\begin{bmatrix} start_{logit} \\ end_{logit} \end{bmatrix} = W_{span} \cdot M_t + b_{span} \tag{15}$$

Where $start_{logit}$ is the chapter token as the starting position probability of the answer, end_{logit} is the probability that the chapter token is the ending position of the answer, and W_{span} and b_{span} are the weights and offsets of the output layer. For non-predicted output:

$$\begin{bmatrix} Yes_{logit} \\ No_{logit} \end{bmatrix} = W_{yes_no} \cdot P_t + b_{yes_no} \tag{16}$$

Where Yes_{logit} is the probability that the answer is "Yes", No_{logit} is the probability that the answer is "No", and W_{yes_no} and b_{yes_no} are weights and offsets. For the output of the unanswerable probability:

$$Null_{logit} = W_{null} \cdot P_t + b_{null} \tag{17}$$

Where $Null_{logit}$ is the probability that the question has unanswerable, W_{null} and b_{null} are weights and offsets.

3.2.5. Loss calculation

To facilitate the calculation of the loss, the final output is shaped into two new probability outputs, namely:

$$\acute{start}_{logit} = [start_{logit}, Null_{logit}, Yes_{logit}, No_{logit}] \tag{18}$$

$$\acute{end}_{logit} = [end_{logit}, Null_{logit}, Yes_{logit}, No_{logit}] \tag{19}$$

The actual output with the same format and containing non-probability, refusal probability is y_{start}, y_{end}, using the cross entropy as the loss function to calculate the loss $loss_{start}$, $loss_{end}$, and

then the mean of the two losses can be obtained to obtain the overall loss $loss_{all}$. Expressed as:

$$loss_{start} = -\frac{1}{N}\sum_{i=1}^{N}[y^i_{start} \cdot \log \acute{start}^i_{logit} + (1 - y^i_{start}) \cdot \log(1 - \acute{start}^i_{logit})] \tag{20}$$

$$loss_{end} = -\frac{1}{N}\sum_{i=1}^{N}[y^i_{end} \cdot \log \acute{end}^i_{logit} + (1 - y^i_{end}) \cdot \log(1 - \acute{end}^i_{logit})] \tag{21}$$

$$loss_{all} = \frac{loss_{logit} + loss_{end}}{2} \tag{22}$$

Where N is the number of samples.

4. Experiment

4.1. Experimental environment

In this paper, experiments are performed on a machine with a 64-bit Windows system. The machine's external memory size is 930 GB, memory space is 48 GB, CPU type is single-core Intel i7-8700K, GPU type is NVIDA GeForce GTX 1080Ti, and GPU size is 11 GB. All the experimental programs in this article are written in python. The deep learning framework used is Pytorch and the version number is 1.13.0.

The original data used in this article comes from the CAIL 2019 legal reading comprehension competition.[1] This dataset is published by HFL (Harbin Institute of Technology and iFly Joint Lab). It is a multi-task machine reading comprehension dataset for the judicial field. The passage of the dataset come from China Referee Documents Network. Questions and answers are hand-written by legal experts. The types of questions include span extraction, yes or no judgment, and unanswerable question. The answers are span of corresponding passage. After a simple pre-processing of the original dataset, each sample is determined as a five-tuple, including the chapter, question, answer text, start and end positions of the answer in the span, and yes/no question marks. The training set contains 40,000 samples and the test set contains 7000 samples.

4.2. Experimental design

4.2.1. Evaluation index

This paper uses Rouge-L, macro average F1 score, and EM (Exact Match) score to evaluate the proposed system. The F1 score is a commonly used classification evaluation index, which takes into account both the accuracy and recall of classification problems. The macro average F1 score is a variant of the F1 score. When the evaluation dataset contains multiple reference answers, the predicted answer and the multiple answers are separately obtained to obtain the F1 score, and the average is used to obtain the macro average F1 score.

$$Avf1 = \frac{\sum_{i=1}^{Count_{ref}} max(f1(gold_i, pred))}{Count_{ref}} \tag{23}$$

$$F1_{macro} = \frac{\sum_{i=1}^{N} Avf1_i}{N} \tag{24}$$

Among them, $gold_i$ represents the ith reference answer, $pred$ represents the predicted answer, $f1(\cdot)$ is the original F1 score solving function, and $Count_{ref}$ is the number of reference answers. Eqs. (4)–(1) solves to get the average F1 score $Avf1$ of the predicted answer and all the reference answers in a single sample. N is the number of samples, and $F1_{macro}$ is the macro average F1 score obtained for all samples.

Rouge and Bleu scores are commonly used indicators for machine translation, but recently researchers have pointed out that when Bleu scores are evaluated in machine reading comprehension, there is a large deviation from Rouge scores. One possible reason is that Bleu has set a penalty term for long answers This makes Bleu more inclined to choose shorter answers, which has an impact on the evaluation of machine reading comprehension [9]. Therefore, this article did not choose Bleu score as the evaluation index, only Rouge-L was used. Rouge-L mainly compares the longest common subsequence between the predicted answer and the reference answer, and finds the Rouge-L score, in order to obtain the "similarity" between the two through the Rouge-L score. The formula of Rouge-L is:

$$R_{lcs} = \frac{LCS(X, Y)}{m} \tag{25}$$

$$P_{lcs} = \frac{LCS(X, Y)}{n} \tag{26}$$

$$F_{lcs} = \frac{(1 + \beta^2)R_{lcs}P_{lcs}}{R_{lcs} + \beta^2 P_{lcs}} \tag{27}$$

Among them, $LCS(X, Y)$ is the length of the longest common subsequence of the reference digest X and the candidate digest Y. m and n are the lengths of the reference answer X and the candidate answer Y, R_{lcs} and P_{lcs} respectively represent the recall rate and accuracy, and F_{lcs} is the Rouge-L score.

EM is the proportion of predicted answers that are exactly the same as the gold standard answers among all predicted answers.

4.2.2. Experimental design

Two benchmarks were selected for this article. The first benchmark was BIDAF. This model was proposed by Seo et al. [14]. It was mainly used to handle reading and comprehension tasks of span extraction. It had achieved good results on the SQuAD1.1 ranking. In order to adapt to the multi-tasking dataset of this paper, in the preprocessing part, the three characters "KYN" are added to all the span. Yes or no type questions and unanswerable questions are also treated as span extraction problems. Among them, "Y" and "N" span are extracted for yes or no type questions, "K" span are extracted for unanswerable questions, and other span extraction questions are predicted from the fourth character.

Another benchmark is the Bert model proposed by Google [16]. This model is almost a milestone in the study of machine reading comprehension in recent years, and has caused huge repercussions in the field of natural language understanding. The Bert model is used for multi-tasking reading comprehension of two types of questions: segment extraction and rejection. The Bert used in this paper has made a small modification to the original Bert model, so that the position of the probability of unanswerable probability was previously predicted, and now the three probability of unanswerable, yes and no are predicted simultaneously.

At the same time, in the experimental part, three experiments are designed in this paper. The first experiment compares the model proposed in this article with BiDAF and Bert on the mentioned dataset, and uses F1 and Rouge-L, EM to evaluate at the same time to determine whether the model proposed in this article is advanced. At the same time, this paper also uses a five-fold cross-validation to construct an integrated model and shows the results.

The second experiment is the ablation study of the model proposed in this paper. The experimental method is to ablate the

[1] http://cail.cipsc.org.cn/

Table 1
Control experiment result.

Model	Macro average F1	Rouge-L	EM
BiDAF	0.6243	0.3139	0.2322
Bert	0.7430	0.7150	0.5840
Bert-wwm [24]	0.7520	–	0.5470
Bert-wwm-ext*	0.7600	–	0.5560
RoBERTa-wwm-ext*	0.7910	–	0.5870
RoBERTa-wwm-large-ext*	0.8240	–	0.6210
LegalSelfReader	**0.8480**	**0.8300**	**0.6820**

Table 2
Training cost comparison.

Model	Devices	Training steps
BERT-wwm	TPU v3	200K
BERT-wwm-ext	TPU v3	1.4M
RoBERTa-wwm-ext	TPU v3	1M
RoBERTa-wwm-ext-large	TPU Pod v3–32	2M
LegalSelfReader	GTX 1080Ti	60.3K

Table 3
Problem type experiment result.

Model	Macro average F1	Rouge-L	EM
When	**0.9350**	**0.9330**	**0.8600**
Who	0.7580	0.7470	0.6000
Where	0.8600	0.8400	0.6800
Entity	0.8090	0.8050	0.6100
Why	0.7320	0.6920	0.4200
What	0.8500	0.8350	0.5600
How	0.8140	0.7800	0.5600
How many/How much	0.8520	0.8220	0.6800
Yes/No	**0.9000**	**0.8900**	**0.9000**

various components of the model to determine the effectiveness of each model component.

The third experiment is to evaluate according to different types of problems to determine that the model proposed in this paper is better at that type of problem.

4.3. Comparative experiments

In order to verify the effectiveness of the system designed in this paper, two baselines are set up in this paper: BIDAF and Bert. Together with the model proposed in this paper, experiments are performed on the data described in Section 4.1. The experimental results are shown in Table 1. Among them, the ones with "*" are the experimental results on the development set published by HFL on github.[2]

It can be seen from the experimental results that compared with the traditional BIDAF and Bert models, our model has improved greatly in three indicators. Traditional BIDAF uses word2vec pre-trained word vectors to get fixed semantic word vectors. Our model uses the Bert model to obtain word vectors. The resulting word vectors are context-dependent, so they have been greatly improved. Although we made some adjustments to the original Bert model to enable it to answer yes or no questions, the original Bert model did not have the ability to answer yes and no questions, so it performed poorly on our multi-task machine reading comprehension dataset. At the same time, when compared with some of the more novel Transformer-like models proposed by HFL, our model also has a significant performance improvement. In addition, our model is only a single hidden layer model based on the original Bert-Chinese. There is no large-scale retraining of the Bert model. Only a BiLSTM layer is added to the span extraction output part, and only one is added to the right or wrong judgment output layer Attention layer and a pooling layer. Compared to rebuilding a new Transformer model and retraining with new expectations, our model still obtains better results at such a low cost. The Table 2 shows the training cost of our model compared with HFL. The data comes from their public results on GitHub.

4.4. Analysis of problem diversity

In order to analyze the types of problems that the model proposed in this paper is good at, we use some specific problem keywords in Chinese to design a heuristic problem category classification, which could be divided the following categories: 'When', 'Who', 'Where', 'Entity', 'Why', 'What', 'How', 'How many/much' and 'Yes/No'.

Based on the keywords shown above, we randomly selected the test set. Each question category was screened to obtain 100 chapter-question-answer pairs, and then evaluated using the trained model. The results in Table 3 were obtained:

From the experimental results, it can be seen that the model proposed in this paper has excellent performance on all problem types. Among them, the model proposed in this paper is better at dealing with time type problems and yes or no type problems, and has achieved a score of 0.9 or more on the average F1 score of the macro. In most cases, the format of the time type answer is relatively fixed. The model only needs to learn this fixed time format and then perform simple matching to basically get the final answer. The yes or no type problem is more complicated. It is not a problem of matching types. It requires the model to deeply understand the overall semantics of the article, and then make a judgment of yes or no. The model proposed in this paper can achieve better performance in yes or no types, indicating that our multi-task training for yes or no type problems has been successful, and it enables the model to perform deep passage semantic understanding.

At the same time, the model obtained poor results on Who and Why types of problems. We looked at the data on Who types of issues, and finally found that in order to protect personal privacy, the names of people in the data were anonymized, which may cause the named entity recognition vectors we added to deviate, which makes the model's effect worse. For Why-type questions, it can usually organize multiple answers (the correct answer may be more than one gold standard answer). This is because Why-type questions often show multiple relevance in the article. In addition to the passage span of the correct answer, the model may also explore other span related to the question, and these span may become the answer. This makes the effective range of the attention value enlarged, so that the model cannot give an exact answer, and reduces the performance of the model.

In order to verify our conjecture, this paper conducted a visual analysis of the attention value.

4.5. Visualization of attention values

We show attention data for three types of problems, including time-type problems, Why-type problems, and yes or no type problems. Due to the length of the passage, we will show some of the attention values and intercept the part that can reflect important information. The darker the color in the graph, the higher the attention value.

In the sample of the time type problem, as in Fig. 3, the problem is "When was ... ?", And the attention value is darker in

[2] https://github.com/ymcui/Chinese-BERT-wwm#%E4%B8%AD%E6%96%87%E6%A8%A1%E5%9E%8B%E4%B8%8B%E8%BD%BD

Fig. 3. Time type problem visualization result.

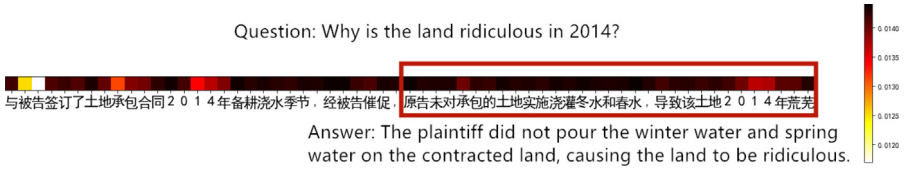


Fig. 4. Cause type problem visualization result.

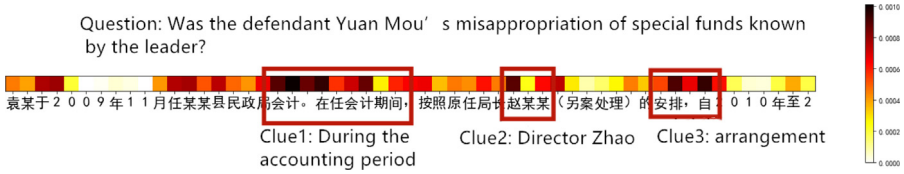


Fig. 5. Yes/No type problems visualization result.

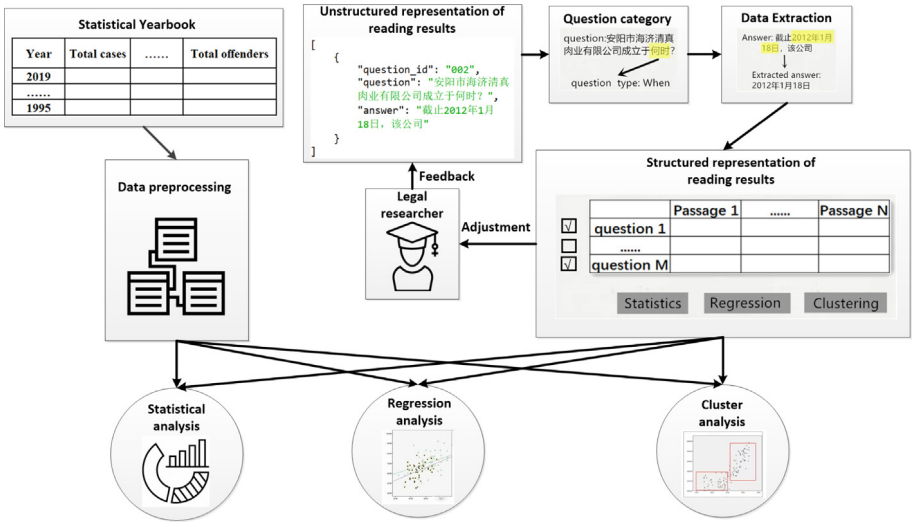


Fig. 6. System post-processing flow chart.

"January 18, 2012". It can be clearly seen that the model assigns attention weights above the context to span of the time type. It does not focus on time-independent information, which greatly improves the prediction ability of the model.

In the sample of Why type questions, as in Fig. 4, the question is "Why is the land ridiculous in 2014?". The attention value is darker in the exact answer "the plaintiff did not irrigate the contracted land with winter water and spring water, which caused the land to be barren in 2014", but it also gave higher attention weight to the nearby location. The attention value does not show a big difference in context. It is difficult for the model to obtain a more effective answer when using this distribution of attention values. This is basically in line with our explanation of the poor performance of Why type problems in the previous section.

At the same time, in Fig. 5, for yes or no type problems, the model must have a comparative reasoning ability. One of the most important points of the reasoning ability is that the model needs to be able to find clues by itself. What is more interesting is that we find that the model has a good ability to "find clues". As shown in the figure, it is a sample of non-judgment type. The question is "Was the defendant Yuan Mou's misappropriation of spacial funds known by the leader?". The part with higher attention value in the figure is a more important clue to answer this question. For example, it says Yuan "as an accountant", and his leader "Zhao Moumou", and his director's behavior " arrangement". This shows that the model has a good ability to extract evidence.

4.6. Discussion on structured and unstructured legal data

In the course of legal empirical research, the ultimate purpose of reading legal documents is to transform unstructured text data into structured tabular data. Afterward, statistical analysis was performed on these data to observe the realistic picture of

legal issues. Therefore, the post-processing of unstructured legal text reading is actually similar to the processing of a structured statistical yearbook. In Fig. 6, we give a detailed post-processing stage of the system. As with structured data such as yearbooks, some structured data obtained after reading comprehension can also be processed using statistical analysis.

1. The model is in the post-processing stage of giving the predicted answer:
   The model cannot be used directly in the legal analysis for the given prediction passage span, because such passage span contain a lot of other information, which requires a post-processing stage. For each given prediction span, the original problem is classified again. For example, if a question is a "When" type question, the post-processing module will extract time-formatted data from the predicted span. If the question is a "How much" or "How many" type question, the post-processing module will extract Numeric type data.
   After getting cleaner data, the system helps law researchers make certain statistics. First, the system can perform statistics. For example, if we read the sentence data in all the judgment documents, we can analyze the average sentence and statistics of each sentence period. Second, for some numerical data, the system can perform regression analysis on them to determine the correlation between different data. For example, in cybercrime, we can analyze the linear relationship between the number of stolen user information and the sentence. Furthermore, for a large amount of data of the same type of adjudication documents, the system can perform cluster analysis on these data to determine different sub-categories in the same type of crime, so that legal researchers can find places where it is possible to refine legal provisions and propose the opinion of.
2. The system requires participation of law researchers:
   From recent studies and experiments in this article, we can see that the current machine reading comprehension cannot achieve better performance on all types of problems, especially "Why" type problems. Therefore, this still requires the participation of law researchers. When the system gives plausible answers, legal researchers need to check the answers again, correct some of the wrong predictions given by the model, and then use the model to train these data heavily. In the foreseeable future, the system may also use better-performing models. One benefit of this is that the participation of legal researchers in the system's self-reading process will gradually decrease until the system fully realizes self-reading.

5. Conclusion

As legal research moves toward the legal intelligence research stage, artificial intelligence technology will be used on a large scale in legal research.

At the same time, the rise of legal empirical research with text data as the core has brought opportunities for the application of natural language processing technologies such as machine reading comprehension in legal research. However, structured data such as statistical yearbooks and unstructured data such as legal documents may be used in empirical research in law. There is structured and unstructured legal data. At the same time, in the process of reading legal documents, there may also be questions of time, substance, right and wrong, etc., and there is a diversity of questions. Based on this problem, this paper designs a multi-task reading system for intelligent legal services. The analysis module of the system uses both statistical analysis and machine reading comprehension technology, and can simultaneously process structured data such as statistical yearbooks, as well as unstructured data such as judgment documents, file materials, and interview text records, it solves the problem of structured and unstructured legal data. At the same time, this article has designed a legal document self-reading model - LegalSelfReader, which can be applied to empirical research in law, and can answer three types of questions: segment extraction, judgment of right and wrong, and refusal to answer.

Empirical research data can be precise, fuzzy, quantitative, non-quantitative, and can be physical, political, biological, sociological, economic, etc. [26]. This illustrates the diversity and multi-modality of empirical data. In addition to textual information such as referee documents and file materials, may also be video and audio information such as court trial videos and relevant person comments. The direction that this article can continue to study is to re-integrate the video and audio question answering system in the original system, so that the system can process more modal data.

CRediT authorship contribution statement

Yujie Li: Methodology, Resources. **Gang Hu:** Conceptualization, Software, Writing - original draft. **Jinyang Du:** Data curation, Visualization. **Haider Abbas:** Investigation, Formal analysis. **Yin Zhang:** Writing - review & editing, Supervision, Project administration, Funding acquisition.

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.

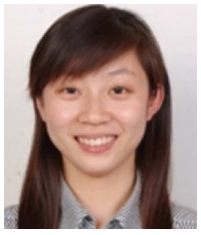
Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grant 61702553.

References

[1] N. Netten, S.W. van den Braak, M.S. Bargh, S. Choenni, F.L. Leeuw, Legal logistics: A framework to unify data centric services for smart and open justice, Int. J. E-Plann. Res. 7 (2) (2018) 51–69.
[2] M. Hildebrandt, Law as computation in the era of artificial legal intelligence: Speaking law to the power of statistics, Univ. Toronto Law J. 68 (Suppl. 1) (2018) 12–35.
[3] D.M. Katz, Quantitative legal prediction-or-how i learned to stop worrying and start preparing for the data-driven future of the legal services industry, Emory Law J. 62 (2012) 909.
[4] P. Kenne, M. O'Kane, H.G. Pearcy, Language modeling of spontaneous speech in a court context, in: Fourth European Conference on Speech Communication and Technology, 1995.
[5] M. Mills, Using AI in law practice: It's practical now, Law Pract. 42 (2016) 48.
[6] D.M. Katz, I. Bommarito, J. Michael, J. Blackman, Predicting the behavior of the supreme court of the united states: A general approach, 2014, arXiv preprint arXiv:1407.6333.
[7] M. Riesen, G. Serpen, Validation of a bayesian belief network representation for posterior probability calculations on national crime victimization survey, Artif. Intell. Law 16 (3) (2008) 245–276.
[8] D.C. Baldus, G. Woodworth, D. Zuckerman, N.A. Weiner, The use of peremptory challenges in capital murder trials: A legal and empirical analysis, University of Pennsylvania Journal of Constitutional Law 3 (2001) 3.
[9] C. Tan, F. Wei, N. Yang, B. Du, W. Lv, M. Zhou, S-net: From answer extraction to answer synthesis for machine reading comprehension, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018.
[10] K.M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, P. Blunsom, Teaching machines to read and comprehend, in: Advances in Neural Information Processing Systems, 2015, pp. 1693–1701.

[11] D. Chen, J. Bolton, C.D. Manning, A thorough examination of the cnn/daily mail reading comprehension task, 2016, arXiv preprint arXiv:1606.02858.
[12] B. Dhingra, H. Liu, Z. Yang, W. Cohen, R. Salakhutdinov, Gated-attention readers for text comprehension, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vol. 1: Long Papers, 2017, pp. 1832–1846.
[13] Y. Cui, Z. Chen, S. Wei, S. Wang, T. Liu, G. Hu, Attention-over-attention neural networks for reading comprehension, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vol. 1: Long Papers, 2017, pp. 593–602.
[14] M. Seo, A. Kembhavi, A. Farhadi, H. Hajishirzi, Bidirectional attention flow for machine comprehension, 2016, arXiv preprint arXiv:1611.01603.
[15] S. Wang, J. Jiang, Machine comprehension using match-LSTM and answer pointer, in: ICLR 2017: International Conference on Learning Representations, Toulon, France, April 24–26: Proceedings, 2017, pp. 1–15.
[16] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018, arXiv preprint arXiv:1810.04805.
[17] B.-H. Tseng, S.-s. Shen, H.-Y. Lee, L.-S. Lee, Towards machine comprehension of spoken content: Initial TOEFL listening comprehension test by machine, in: Interspeech 2016, 2016, pp. 2731–2735.
[18] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, SQuAD: 100, 000+ questions for machine comprehension of text, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016, pp. 2383–2392.
[19] W. He, K. Liu, J. Liu, Y. Lyu, S. Zhao, X. Xiao, Y. Liu, Y. Wang, H. Wu, Q. She, et al., DuReader: A Chinese machine reading comprehension dataset from real-world applications, in: Proceedings of the Workshop on Machine Reading for Question Answering, 2018, pp. 37–46.
[20] P. Bajaj, D. Campos, N. Craswell, L. Deng, J. Gao, X. Liu, R. Majumder, A. McNamara, B. Mitra, T. Nguyen, et al., MS MARCO: A human generated MAchine Reading COmprehension dataset, 2016, arXiv preprint arXiv:1611.09268.
[21] R. Kadlec, M. Schmid, O. Bajgar, J. Kleindienst, Text understanding with the attention sum reader network, 2016, arXiv preprint arXiv:1603.01547.
[22] M. Hu, F. Wei, Y. Peng, Z. Huang, N. Yang, D. Li, Read+ verify: Machine reading comprehension with unanswerable questions, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 6529–6537.
[23] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, Q. Liu, ERNIE: Enhanced language representation with informative entities, 2019, arXiv preprint arXiv:1905.07129.
[24] Y. Cui, W. Che, T. Liu, B. Qin, Z. Yang, S. Wang, G. Hu, Pre-training with whole word masking for Chinese BERT, 2019, arXiv preprint arXiv:1906.08101.
[25] M. Hu, Y. Peng, Z. Huang, X. Qiu, F. Wei, M. Zhou, Reinforced mnemonic reader for machine reading comprehension, in: Proceedings of the 27th International Joint Conference on Artificial Intelligence, AAAI Press, 2018, pp. 4099–4106.
[26] L. Epstein, G. King, Exchange: Empirical research and the goals of legal scholarship, Univ. Chicago Law Rev. 69 (1) (2002) 191–209.



Yujie Li received the B.S. degree in Computer Science and Technology from Yangzhou University in 2009. She received M.S. degrees in Electrical Engineering from Kyushu Institute of Technology and Yangzhou University in 2012, respectively. She received a Ph.D. degree from Kyushu Institute of Technology in 2015. From 2016 to 2017, she was a Lecturer in Yangzhou University. Currently, she is a JSPS research fellow (FPD) at Kyushu Institute of Technology and an Assistant Professor in Fukuoka University, Japan. Her research interests include computer vision, sensors, Internet of Things, and image segmentation.

Gang Hu received the B.S. degree from Hubei University of Chinese Medicine, China, in 2018. He is now a master degree candidate of the School of Information and Safety Engineering, Zhongnan University of Economics and Law (ZUEL), China. He is an IEEE student member. His research interests include machine learning and deep learning.

Jinyang Du received the B.S. degree from Taiyuan University of Technology, China, in 2019. He is now a master degree candidate of the School of Information and Safety Engineering, Zhongnan University of Economics and Law (ZUEL), China. He is an IEEE student member. His research interests include machine learning and deep learning.

Haider Abbas is currently heading the R &D Department at Military College of Signals, NUST, Pakistan. He is the director of National Cyber Security Auditing and Evaluation Lab (NCSAEL) at MCS NUST. Dr. Abbas is a Cyber Security professional, academician, researcher and industry consultant who took professional trainings and certifications from Massachusetts Institute of Technology (MIT), United States; Stockholm University, Sweden; Stockholm School of Entrepreneurship, Sweden; IBM, USA and Certified Ethical Hacker from EC-Council. He received his MS in Engineering and Management of Information Systems (2006) and Ph.D. in Information Security (2010) from KTH—Royal Institute of Technology, Stockholm, Sweden. His professional career consists of activities ranging from R&D and Industry Consultations (Government & Private), through multi-national research projects, research fellowships, doctoral studies advisory services, International Journal Editorships, Conferences/Workshops Chair, Invited/Keynote Speaker, Technical Program Committee Member and reviewer for several international journals and conferences. He has also been appointed by Springer as Full-time Regional Editor for all submissions from Pakistan and Iran for Neural Computing and Applications (ISI-Indexed, IF 4.6, JCR-Ranking Q1). He is an adjunct faculty and doctoral studies advisor at Al-Farabi Kazakh National University, Almaty, Kazakhstan, Manchester Metropolitan University, United Kingdom and Florida Institute of Technology, USA.

Dr. Abbas has won several awards from National and International organizations in recognition of his professional excellence and services to the international research community. He has also been awarded one of the youngest Fellows of the Institution of Engineering and Technology, U.K.; a fellow of the British Computer Society, U.K.; and a fellow of the Institute of Science and Technology, U.K. He has been elected to the grade of Senior Member of the Institute of Electrical and Electronics Engineers (IEEE), USA. He has been appointed as a Member of the Board of Governors for National Information Technology Board (NITB), Government of Pakistan.

Yin Zhang is a Professor of the School of Information and Communication Engineering, University of Electronic Science and Technology of China. He is Co-chair of IEEE Computer Society Big Data STC. He serves as editor or associate editor for IEEE Network, IEEE Access, Journal of Information Processing Systems, etc. He is a Guest Editor for Future Generation Computer Systems, IEEE IoT Journal, ACM/Springer Mobile Networks & Applications, Sensors, Neural Computing and Applications, Multimedia Tools and Applications, Wireless Communications and Mobile Computing, Electronic Markets, Journal of Medical Systems, New Review of Hypermedia and Multimedia, etc. He also served as Track Chair of IEEE CSCN 2017, TPC Co-Chair of CloudComp 2015 and TRIDENTCOM 2017, etc. He has published more than 100 prestigious conference and journal papers, including 14 ESI Highly Cited Papers. He is an IEEE Senior Member since 2016. He got the Systems Journal Best Paper Award of the IEEE Systems Council in 2018. He was named in Clarivate Analytics Highly Cited Researchers List in 2019. His research interests include mobile computing, edge intelligence, cognitive wireless communications, etc.