Artificial Intelligence and Law
https://doi.org/10.1007/s10506-023-09354-x

ORIGINAL RESEARCH

Semantic matching based legal information retrieval system for COVID-19 pandemic

Junlin Zhu[1] · Jiaye Wu[1] · Xudong Luo[1] · Jie Liu[1]

Accepted: 15 February 2023

Abstract

Recently, the pandemic caused by COVID-19 is severe in the entire world. The prevention and control of crimes associated with COVID-19 are critical for controlling the pandemic. Therefore, to provide efficient and convenient intelligent legal knowledge services during the pandemic, we develop an intelligent system for legal information retrieval on the WeChat platform in this paper. The data source we used for training our system is "The typical cases of national procuratorial authorities handling crimes against the prevention and control of the new coronary pneumonia pandemic following the law", which is published online by the Supreme People's Procuratorate of the People's Republic of China. We base our system on convolutional neural network and use the semantic matching mechanism to capture inter-sentence relationship information and make a prediction. Moreover, we introduce an auxiliary learning process to help the network better distinguish the relation between two sentences. Finally, the system uses the trained model to identify the information entered by a user and responds to the user with a reference case similar to the query case and gives the reference legal gist applicable to the query case.

Keywords Information retrieval · Semantic matching · Convolutional neural network · Legal knowledge · COVID-19

This manuscript is an extended version of our prior work: Wu et al. "Few-Shot Legal Knowledge Question Answering System for COVID-19 Epidemic", 2020 3rd International Conference on Algorithms, Computing and Artificial Intelligence.

✉ Xudong Luo
luoxd@mailbox.gxnu.edu.cn

Junlin Zhu
zhujlfree@163.com

Jiaye Wu
kennywu96@163.com

Jie Liu
625650997@qq.com

[1] Guangxi Key Lab of Multi-Source Information Mining & Security, College of Computer Science and Engineering, Guangxi Normal University, Guilin 541004, China

Published online: 14 March 2023

1 Introduction

A recent series of pneumonia outbreaks caused by a novel coronavirus has spread across China and the rest of the world. During the pandemic, many criminals took the opportunity to break the law and commit crimes, endangering people's health and safety while challenging to prevent and control the pandemic by the inspecting authorities. Because of the unprecedented outbreak of the pandemic and its wide-ranging impact, the inspecting authorities need to be case-specific when confronted with cases against the prevention and control of the pandemic at the beginning. The increasing number of criminal cases poses a significant challenge to the work of the authorities. On the other hand, people's knowledge of the laws related to the pandemic is also lacking. In artificial intelligence, there has been a lot of research work that has explored the possibility of AI tools to aid legal judgments (Huang and Luo 2018). Moreover, such tools cannot only make legal work more efficient. However, they can also help ordinary people lacking legal knowledge.

For example, Weber et al. (1998), Zhong et al. (2019), and Liu et al. (2019) develop three case-based reasoning systems on legal documents to solve real-world problems. And Liu and Hsieh (2006) use the K-nearest neighbour as a classifier to process Chinese judicial documents. With the success of deep learning in natural language processing, deep neural networks have also been used to predict crime type in legal cases (Hu et al. 2018). Many researchers studied the application of these methods in the legal field with significant results. However, few attempts have been made to apply these methods to the legal cases involved in the COVID-19 pandemic and develop an operational system.

To date, the lack of open, high-quality data of legal cases regarding the pandemic, the diversity of charges, and the complexity of cases make achieving a reliable information retrieval model a challenging task. To the end, in this paper, we try to use the small amount of available data to implement a WeChat-based information retrieval system of legal knowledge. The system can identify similar cases regarding outbreaks. Based on the description of the case entered by the user, the system can give the corresponding similar case or the reference legal gist that applies to the description. In this paper, we collect and construct the relevant dataset by ourselves. We also propose a semantic matching network based on the siamese structure as the basis for the system's functional implementation. Moreover, we propose a relation learning module to help the network better model the relationship between the two sentences.

Figure 1 shows how we implement our system. Firstly, we use crawlers to collect data from the target websites and perform pre-processing operations on the data. Then, we use the data collected in the first step to train a neural network model to support our system's core functions. Finally, we deploy the model trained in the second step to the developer server, receive user input, and provide the user with the corresponding result. To realise the function of providing a reference to the legal gist, because of the small amount of available dataset and the diversity of crimes, we use text matching instead of case classification to provide the legal gist applied by similar cases as the target output.

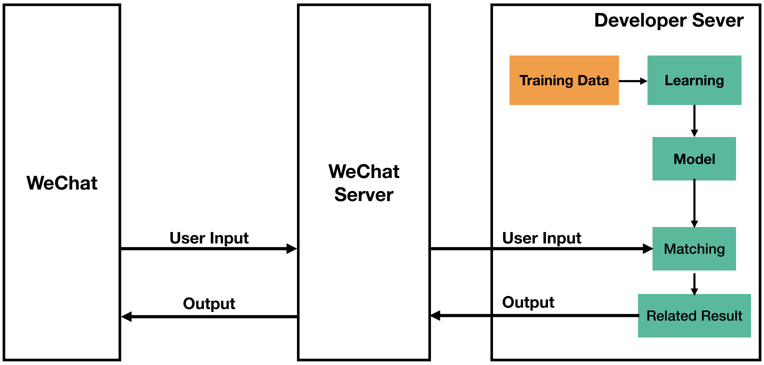


Fig. 1 The flow chart of our system

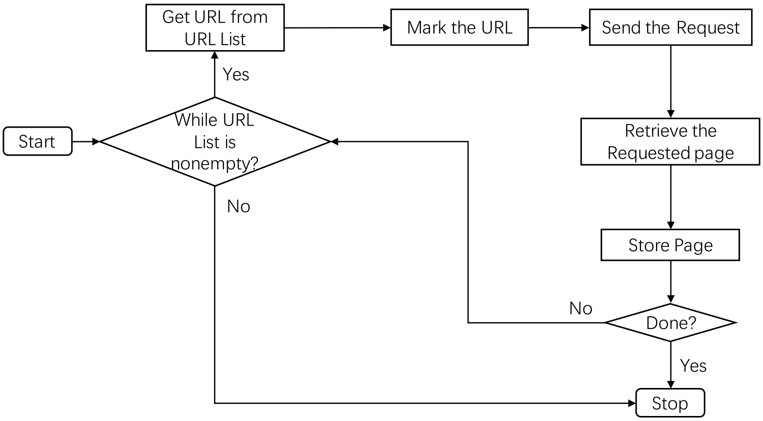


Fig. 2 Typical flow of crawling process

The main contributions of this paper are as follows. (1) We propose an effective semantic matching network for implementing our system. (2) We design and implement a system to assist legal practitioners and ordinary people during the pandemic. And (3) the system implemented is based on the WeChat public platform for deployment, with a low access and user-friendly use threshold.

The rest of this paper is organised as follows. Section 2 presents the definition of the problem we are going to solve in this paper. Section 3 describes the procedure of data acquisition and reprocessing. Section 4 gives the details of the text encoder of our network. Section 5 discusses the relation learning module in our network. Section 6 presents the experimental analysis on our model. Section 7 discusses the deployment of our system. Section 8 compare our wok with the related work. Finally, Sect. 9 concludes this paper with future work.

2 Problem definition

For each case entered by a user, we treat the factual description as a sequence of words, *i.e.*,

$$X = \{x_1, \dots, x_N\},$$

where N is the number of words. Our system's input is a sequence of words, and its output is a reference case that is most similar to the input case or the reference legal gist applicable to that case. We implement the two functions based on similar case matching (we will detail their implementations in later sections).

3 Data preparation

This section will describes the data acquisition and reprocessing.

3.1 Data acquisition

Algorithm 1: Acquire and prepare data

```
  Data: homepage $L$.
  Result: structured dataset $S$.
1 StartURL = $L$.
2 ListOfURLs = fetch(URLOf(StartURL)).
3 $S$ = CreateNewDatasetFile().
4 while notEmpty(ListOfURLs) do
5     for each url $\in$ ListOfURLs do
6         Content = fetch(ContentOf(url));
7         Records = Parse(Content);
8         Info = Extract(Records);
9         Store(S, Info).
```

We use a crawler to acquire the relevant data online. Figure 2 shows a typical flow of the crawling process. Algorithm 1 shows data acquisition and preparation steps based on this crawling process. On its lines 1–3, we derive the URL List used in this work from the online release office page of the Supreme People's Procuratorate of the People's Republic of China.[1] Specifically, on line 1, we treat the online release office page as homepage L. Then, on line 2, we filter the URLs with the keyword "typical cases of epidemic prevention and control crimes" in the title to form the URL List. Finally, on line 3, we initialise the dataset file of S for subsequent dataset information storage. On lines 4–9, we iterate over each URL in the URL list to construct our structured dataset file. Specifically, on line 6, we use the request library in Python to retrieve each URL's page source in the URL list. On line 7, we use the BeautifulSoup library of Python to parse the content from these page-sources and save the parsed contents as HTML records for URL List. On line 8, we

[1] https://www.spp.gov.cn/spp/xwfbh/wsfbt/.

Table 1 The statistics of the raw dataset

Crime type	Number
Hindering the prevention and treatment of infectious diseases	9
Disrupting public service	10
Defiance and affray	10
Fraud	10
Selling fake and inferior products	2
Illegal business	2
Robbery	14
Rumour-mongering	29
Illegal acquisition and sale of rare and endangered wildlife	6
Illegal hunting and killing of rare and endangered wildlife protected by the state	5

The table shows number of cases for each crime type

use a simple rule-based filter to extract the case information we want from each case in the HTML records, including the fact and type of crime.

Table 1 shows the statistics of collected data. We manually annotate each case with a legal gist according to the law applicable to the crime type based on the stored information. The constructed structured dataset consists of 97 items, which is stored in the format of "Crime", "Law Gist", and "Fact" (see Table 2).

Table 2 Examples of formatted data after our processing

Crime	Law gist	Fact
Hindering the prevention and treatment of infectious diseases	...does not comply with the provisions of articles 114 and 115, paragraph 1, of the Penal Code.	January 20, 2020, a hospital in Wuhan City, Hubei Province,.
Disrupting public service	...with the law to prevent and control the pandemic shall be convicted and punished for obstructing official duties in accordance with the provisions of article 277, paragraphs 1 and 3, of the Penal Law	...the suspect Liu Mou disregarded the government ban, from Zhushan County, Zhushan Town.
…	…	…

Algorithm 2: Acquisition of Final Training Data

```
    Input: all facts $F$, all crime type $C$.
    Output: final training data $\hat{X}$.
1   $\hat{X}$ = [].
2   Initialise a list $inds$ range from 0 to len($C$).
3   for $i$=0:len($C$)-1 do
4       $cases = F[C[i]]$;
5       remove $i$ from inds;
6       $cInds$ = copy(inds);
7       $j = \text{len}(C) - i - 1$;
8       for item $k, cs \in enumerate(cases)$ do
9           shuffle($cInds$);
10          $selectedInds$ = topJ($cInds$);
            /* get negative example */
11          for item $ind \in selectedInds$ do
12              $key = C[ind]$;
13              $SC = F[key]$;
14              $si$ = randomInt(0, len($SC$)-1);
15              $d$ = createDict(0, $cs$, $SC[si]$);
16              $\hat{X}$.add(d).
            /* get positive example */
17          for $l$=$k$+1:len($cases$)-1 do
18              $d$ = createDict(1, $cs$, $cases[l]$);
19              $\hat{X}$.add(d).
20  return $\hat{X}$.
```

Since we collected a relatively small number of available case samples, we follow a slightly different approach to constructing positive and negative case samples, i.e., we use the relative similarity between cases for classification. More specifically, we are based on the assumption that, for a case, the similarity of the same type is higher than the different type. We will verify the validity of this hypothesis in Sect. 6.5 later.

Algorithm 2 shows the method we use to construct the final training data. On line 1, we initialise the list to store final training data. On line 2, we initialise a list of indexes to filter crime types for data enhancement. On lines 3–19, we use data augmentation techniques to obtain the final training data. Specifically, on line 4 of the algorithm, we select all the cases from F that match the crime type $C[i]$. On line 5, we remove index i from the list of indexes to exclude crime type $C[i]$ from the data enhancement process later. On line 6, we create a temporary copy of the index list to facilitate subsequent data enhancement operations. On line 7, we obtain the number of crime types remaining after excluding i crime types. On line 8, we iterate through the crime cases and begin the data enhancement operation. On lines 9–16, for each case cs matching with $C[i]$, we construct a negative sample by randomly selecting one case from each of the crime types ranked after $C[i]$ combined with cs. On lines 17–19, we construct positive samples by combining the current cs with each case in the *cases* ranked after cs.

The following is an piece of the final training data:

```
{
     label: 0,
     A: January 20,2020, a hospital in Wuhan City, Hubei...,
     B: January 23,2020, Desheng Town, Zhushan County, Hubei Province...
}
...
```

As shown above, a piece of the final training data is a 3-tuple (y, X_1, X_2), where X_1 is a description of one case, X_2 is a description of another case, and y is the similarity label. If X_1 is similar to X_2, label y is 1 (see line 18 of Algorithm 2); otherwise, $y = 0$ (see line 15 of Algorithm 2).

3.2 Data augmentation by back translation

To increase the sample number and intra-class sample diversity, we use a complementary data augmentation technique to extend our dataset. The data augmentation method uses two translation tools: one of which translates Chinese into English, and the other translates English back into Chinese. To avoid the method affecting the sample balance of the dataset, we only apply it to the categories with less than nine sample sizes. As shown in Table 1, these categories are: “Selling Fake and Inferior Products”, “Illegal Business”, “Illegal Hunting and Killing of Rare and Endangered Wildlife Protected By The State”, and “Illegal Acquisition and Sale of Rare and Endangered Wildlife”.

In this work, we employ Google Translate and Baidu Translate as translation tools for our data augmentation operations. The former is a famous translation engine worldwide, while the latter is a top-ranked translation engine in China. We use Google Translate as our Chinese-English system and Baidu Translate as our English-Chinese system. The process of data augmentation is illustrated in Fig. 3, where Chinese is the source language. The number distribution of the two categories is shown in Fig. 4. We can see that our data construction method yields a more balanced training dataset, enabling us to learn a better performance model.

As shown in Fig. 5, our network consists of two modules: text encoder and relation learning module. We will discuss them in turn in the following two sections.

4 Text encoder

This section will present the text encoder of our network model.

Support there is a pair of case descriptions:

$$X_1 = \{w_{1,1}, \dots, w_{1,m}\}, \tag{1}$$

$$X_2 = \{w_{2,1}, \dots, w_{2,n}\}, \tag{2}$$

where $w_{i,j}$ indicates the jth word in the ith sequence, and m and n are the length of the two cases, respectively. Then we convert the words in the text to their corresponding word embeddings:

$$E_1 = \{e_{1,1}, \dots, e_{1,m}\} = W_e[V(X_1)], \tag{3}$$

$$E_2 = \{e_{2,1}, \dots, e_{2,n}\} = W_e[V(X_2)], \tag{4}$$

where $V(\cdot)$ denotes the function that converts the sequence to its indices in the corresponding vocabulary, W_e is the corresponding embedding matrix, and $e_{i,j}$ is the vector corresponding to the jth word in the ith sentence.

The trainable word embedding is initialised from 300-dimension pre-trained Chinese Word Vectors (Li et al. 2018). For Chinese, the accuracy of word segmentation affects the subsequent semantic parsing process. Therefore, in practice, we use character embedding, which also can convey strong semantics. All out-of-vocabulary words are mapped to a <UNK> token, randomly initialised to a 300-dimension vector.

We then use Convolutional Neural Network (CNN) (Krizhevsky et al. 2017) as a shared text encoder to produce the high-level semantic representation vector of all words for two cases. We have tried CNN, Bi-LSTM (Graves and Schmidhuber 2005), Bi-GRU (Cao et al. 2019) and BERT (Devlin et al. 2019) as the encoder in experiments (see Sect. 6.4), finding CNN is the best, so finally we choose CNN as t he encoder in our system.

Figure 6 shows the CNN structure we used.

1. *Convolution* Suppose the input sentence is X_1 (see formula (1)), and the embedding is E_1 (see formula (3)) as the sentence matrix. Thus, we have a sentence matrix of $m \times d$. Then the convolution operation is to apply a filter $W_k \in \mathbb{R}^{k \times d}$ to a window of k words to produce a new feature. For example, we consider subsequence $S_j = [e_{1,j}; \cdots ; e_{1,j+k}]$ where [;] denotes the concatenation operation, then the convolution operation takes the dot product of W_k with S_j followed by a non-linear function to generate a new feature $c_{k,j}$, *i.e.*,

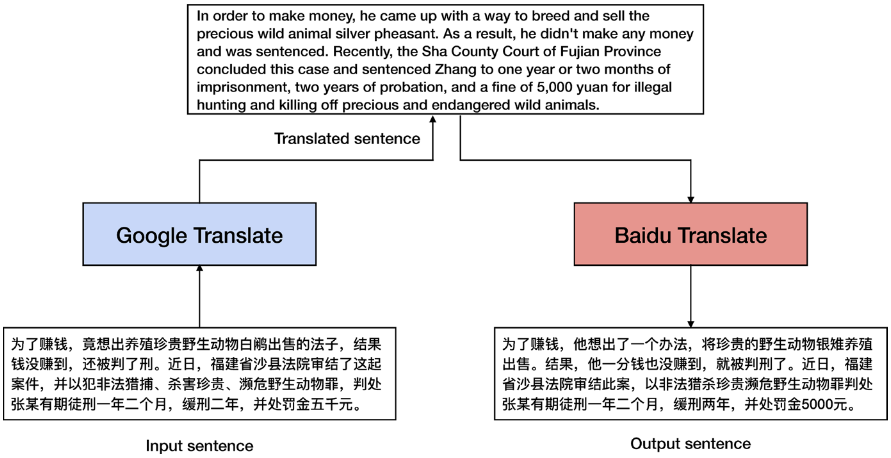
$$c_{k,j} = \mathrm{ReLU}(W_k \cdot S_j + b_k), \tag{5}$$

where b_k is the bias. Thus, the filter is applied to each possible window of k words in E_1 to produce a feature map:

$$\vec{c}_k = (c_{k,1}, \dots, c_{k,m-k+1}). \tag{6}$$

To capture the different features, we use l filters (*i.e.*, $W = \{w_{1,k}, \cdots, w_{l,k}\}$) in the convolution operation. Then, the convolution result is a matrix:

$$C_k = [c_{k,1}; \cdots ; c_{k,l}] \in \mathbb{R}^{l \times (m-k+1)}.$$

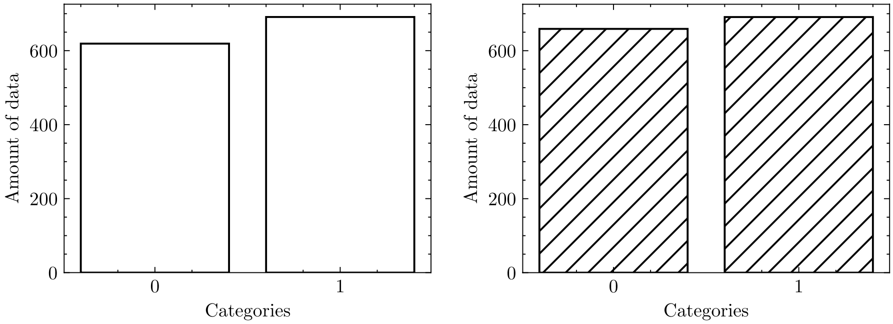


In order to make money, he came up with a way to breed and sell the precious wild animal silver pheasant. As a result, he didn't make any money and was sentenced. Recently, the Sha County Court of Fujian Province concluded this case and sentenced Zhang to one year or two months of imprisonment, two years of probation, and a fine of 5,000 yuan for illegal hunting and killing off precious and endangered wild animals.

为了赚钱，竟想出养殖珍贵野生动物白鹇出售的法子，结果钱没赚到，还被判了刑。近日，福建省沙县法院审结了这起案件，并以犯非法猎捕、杀害珍贵、濒危野生动物罪，判处张某有期徒刑一年二个月，缓刑二年，并处罚金五千元。

为了赚钱，他想出了一个办法，将珍贵的野生动物银雉养殖出售。结果，他一分钱也没赚到，就被判刑了。近日，福建省沙县法院审结此案，以非法猎杀珍贵濒危野生动物罪判处张某有期徒刑一年二个月，缓刑两年，并处罚金5000元。

Fig. 3 The process of data augmentation operation



(a) The number distribution of two categories before data augmentation

(b) The number distribution of two categories after data augmentation

Fig. 4 The number distributions of two categories before and after data augmentation

According to our experimental results, in practice, we use three kinds of convolution kernels with widths of 2, 3 and 4 (*i.e.*, $k = 2, 3, 4$). So we end up with three convolution results of C_2, C_3, and C_4.

2. *Max pooling* After obtaining futures, we perform the max pooling operation on each feature map in the convolution result, *i.e.*,

$$\overline{c}_k = \max\{c_{k,i} \mid i \in \{1, \dots, l\}\}, \tag{7}$$

meaning to use the maximum value as the feature for that region. Putting them togather, we have:

$$\overline{C} = \{\overline{c}_1, \dots, \overline{c}_l\}. \tag{8}$$

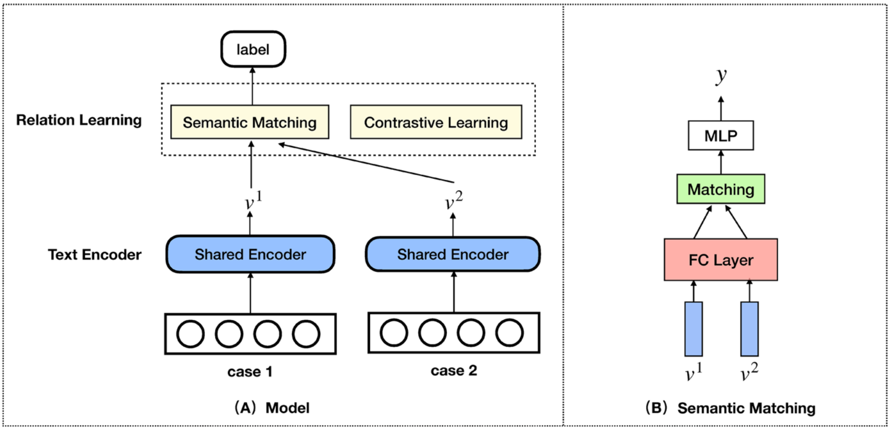


Fig. 5 The architecture of proposed network

The idea behind max pooling is to capture the essential features in each feature mapping.

Finally, we concatenate all the results obtained after max-pooling as the output of the CNN:

$$\vec{v}_1 = [\overline{C}_2; \overline{C}_3; \overline{C}_4]. \tag{9}$$

Similarly, we can also perform the above operation to obtain the output v_2 of input X_2.

5 Relation learning

This section will discuss relation learning (the second part of our network model). After passing the text encoder, the original input sequence pairs $\langle X_1, X_2 \rangle$ are transformed into a new representation $\langle v_1, v_2 \rangle$. To enhance the model's ability to recognise the relationship of input sentence pairs, we propose two modules: semantic matching (Sect. 5.1) and contrast learning (Sect. 5.2). We use joint learning (Sect. 5.3) to train these two modules together.

5.1 Semantic matching

A semantic matching module is to learn the semantic relations between sentence pairs. Specifically, we first perform a nonlinear transformation on $\vec{v}_1$ and $\vec{v}_2$ as follows:

$$\vec{v}'_1 = \mathrm{ReLU}(W_s\vec{v}_1 + b_s),$$
$$\vec{v}'_2 = \mathrm{ReLU}(W_s\vec{v}_2 + b_s),$$

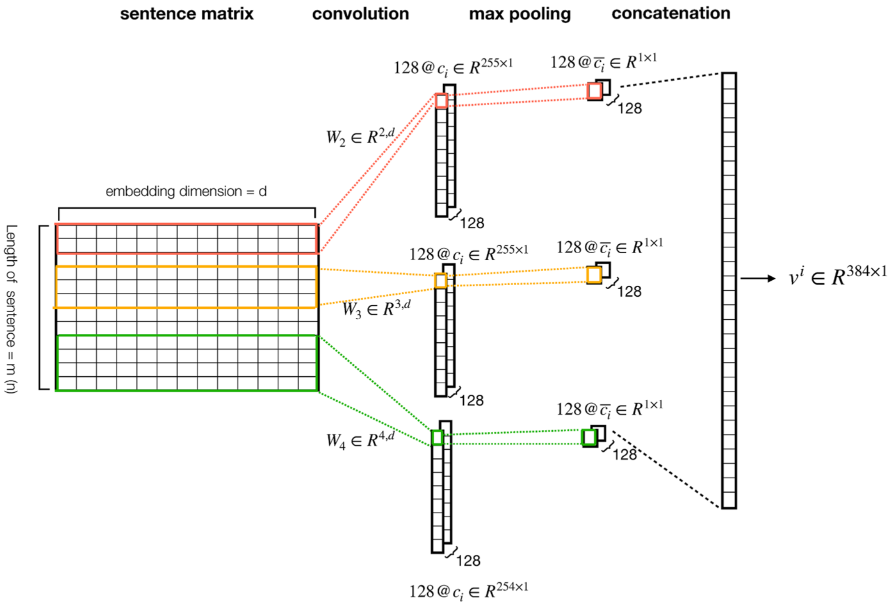

sentence matrix
convolution
max pooling
concatenation
$128 @ c_i \in R^{255\times1}$
$128 @ \overline{c_i} \in R^{1\times1}$
$W_2 \in R^{2,d}$
128
embedding dimension = d
Length of sentence = m (n)
$128 @ c_i \in R^{255\times1}$
$128 @ \overline{c_i} \in R^{1\times1}$
$W_3 \in R^{3,d}$
$v^i \in R^{384\times1}$
$128 @ \overline{c_i} \in R^{1\times1}$
$W_4 \in R^{4,d}$
$128 @ c_i \in R^{254\times1}$

Fig. 6 The architecture of a CNN for s word input sentence

where ReLU (Rectified Linear Unit) (Nair and Hinton 2010) is a non-linear activation function as follows:

$$f(x) = \max\{0, x\}.$$

Then, we use a heuristic matching technique (Wang et al. 2018) to model the relationship between sentence pairs:

$$\vec{z} = [\vec{v}'_1; \vec{v}'_2; (\vec{v}'_1 \circ \vec{v}'_2); (\vec{v}'_1 - \vec{v}'_2)],$$

where [;] denotes the concatenation operation, and ∘ denotes the element-wise product. Heuristic matching is considered more effective when combining multiple representations than simple concatenation and addition operations (Mou et al. 2016). The element-wise product and difference can be viewed as capturing the different aspects of the relationship between the two sentences. Next, we employ a MLP (Multi-Layer Perceptron) (Gardner and Dorling 1998) with one hidden layer for label prediction:

$$P(\hat{y} \mid X_1, X_2) = \mathrm{MLP}(\vec{z}).$$

5.2 Contrastive learning

To enhance the model's ability to distinguish between sentences, we propose an auxiliary task *contrastive learning*. In this task, the model first measures the Euclidean distance between vectors $\vec{v}_1$ and $\vec{v}_2$ by

$$D(\vec{v}_1, \vec{v}_2) = \sqrt{(v_{1,1} - v_{2,1})^2 + \cdots + (v_{1,n} - v_{2,n})^2}. \tag{10}$$

Then, to achieve the goal of this task, we use the loss function of contrastive learning (Hadsell et al. 2006), which can ensure that semantically similar samples are closer in the embedding space. As a distance-based loss function, contrastive loss aims to maximise the distance of features from different class and minimise the distance of features from the same class. The contrastive loss is calculated as follows:

$$L_c = \frac{1}{K}\sum_{i=1}^{K} L(y_i, \vec{v}_{i,1}, \vec{v}_{i,2}), \tag{11}$$

where K is the number of training pairs, and

$$L(y_i, \vec{v}_{i,1}, \vec{v}_{i,2}) = y_i(D(\vec{v}_{i,1}, \vec{v}_{i,2}))^2 + (1 - y_i)\big(\max\{\alpha - D(\vec{v}_{i,1}, \vec{v}_{i,2}), 0\}\big)^2,$$

where $(y_i, \vec{v}_{i,1}, \vec{v}_{i,2})$ is the ith labeled sample pair, $y_i \in \{0, 1\}$ is the similarity label of the ith sample pair, $D(\vec{v}_{i,1}, \vec{v}_{i,2})$ is the Euclidean distance of the ith sample pair, and the α is a margin value that represents our tolerance for the distance between dissimilar features.

The Euclidean distance is affected by the vector dimensionality and fluctuates widely in its range of values. Therefore, to bypass the above problems, we first perform L2-normalisation of the vectors $\vec{v}^1$ and $\vec{v}^2$ before calculating the distance as follows:

$$\vec{u}_1 = \frac{\vec{v}_1}{\sqrt{\sum_{i=1}^{n}(v_{1,i})^2}},$$

$$\vec{u}_2 = \frac{\vec{v}_2}{\sqrt{\sum_{i=1}^{n}(v_{2,i})^2}}.$$

Then, we replace vectors $\vec{v}_1$ and $\vec{v}_2$ in formula (10) by $\vec{u}_1$ and $\vec{u}_2$.

5.3 Joint learning

For the semantic matching module, we use Cross-Entropy as the loss function:

$$L_s = -yP(\hat{y} \mid X_1, X_2).$$

Table 3 The statistics of our dataset

Split	Similarity-label = 0	Similarity-label = 1
Train	395	415
Dev	132	138
Test	132	138

Then we integrate the above loss function with the contrastive loss function (*i.e.*, formula (11)) as follows:

$$\mathcal{L} = (1 - \lambda)L_s + \lambda L_c, \tag{12}$$

where λ is the hyper-parameter that controls the weight of auxiliary loss.

Finally, the whole network is trained by minimising the joint objective function (12).

6 Experiment

This section will present the experimental evaluations on our system.

6.1 Datasets and experimental settings

The whole datasets contain 1350 items. Table 3 shows the detailed statistics of the dataset. In our experiments, the max length is set to 512. The model is implemented with the PyTorch framework. For the BERT model, we use the BertAdam optimiser with a learning rate of $5e^{-4}$ to minimise the loss during training and set the batch size to 16. For the rest of the model, we use Adam optimiser with a learning rate of $1e^{-3}$ and set the batch size to 16. We use 128-dimensional filters for CNN with a width of 2, 3, and 4, respectively. We set the hidden size as 300 for all the GRU and LSTM layers and set the layer's number as 1. The margin α is set to 1. We use the F1-score on the validation set to achieve early stopping for all experiments.

6.2 Baselines

We compare our network with the following four semantic matching models on the dataset we have constructed.

- **ConvNet** (Severyn and Moschitti 2015): ConvNet uses a convolutional network to derive features from text pairs. It then uses various pooling techniques to map the text pair features. A trainable matrix is used to compute similarity features.

Finally, sentence pair features, similarity features and additional manual features are concatenated and passed into the fully connected layer for label prediction.

- **MatchPyramid** (Pang et al. 2016): MatchPyramid models text matching task as image recognition. It first uses input text pairs to obtain a Matching Matrix and then uses Hierarchical Convolution to capture rich information about matching patterns. Finally, an MLP is used for label prediction.
- **ESIM** (Chen et al. 2017): ESIM is a hybrid neural inference model containing four components: Input Encoding, Local Inference Modelling, Inference Composition and Prediction. The Input Encoding module is used to encode the input text. Local inference modelling computes the differences between the representations obtained from Input Encoding to achieve local information inference for the sequence. Inference Composition performs combinatorial learning of the local inference information and its context. Finally, the final vector v is used to do label prediction.
- **Sentence-BERT (SBERT)** (Reimers and Gurevych 2019): SBERT is a variant of the pre-trained BERT network, which uses siamese structures to obtain sentence embeddings with semantic information. It performs a pooling operation on each of the sentence representations obtained from BERT to obtain two fixed-size sentence vectors u and v. Then the vectors u, v, and $|u - v|$ are concatenated for label prediction.

6.3 Evaluation criteria

According to criteria *Accuracy*, *Precision*, *Recall*, and *F1-score*, we evaluate the performance of the above models. Their specific calculations are as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}, \tag{13}$$

$$Precision = \frac{TP}{TP + FP}, \tag{14}$$

$$Recall = \frac{TP}{TP + FN}, \tag{15}$$

$$F1 = \frac{2 \times Recall \times Precision}{Recall + Precision}, \tag{16}$$

where

- *TP* denotes the correctly predicted values with label 1;
- *TN* denotes the correctly predicted values with label 0;
- *FP* denotes the scenario where the predicted value is 1, and the true value is 0; and
- *FN* denotes the scenario where the predicted value is 0, and the true value is 1.

Furthermore, we compute the inference time for each model to compare the efficiency of the models. In this paper, we use the time taken by the models to process each pair of input texts as the inference time in milliseconds.

6.4 Results and analysis

Table 4 shows the overall experimental results of the comparison. Figure 7 shows that our proposed network significantly outperforms baseline models with respect to almost all the evaluation criteria. This means that our relation learning module in the network is effective. Although SBERT outperforms our network with respect to criterion Recall, our network has better overall performance and takes much less time to infer than SBERT. We also perform an ablation study to analyse the impact of the contrastive learning on network performance, and the experimental result is shown in Fig. 8. The results show that the introduction

Table 4 Experimental result of different models on our test set, where CL stands for Contrastive Learning

Model	Accuracy	Precision	Recall	F1-score	Inference time (ms)
ConvNet	91.85	92.65	91.30	91.97	4.56
MatchPyramid	68.15	64.29	84.78	73.12	3.68
ESIM	87.04	82.39	94.93	88.22	29.41
SBERT	95.56	93.15	**98.55**	95.77	44.12
Ours	**97.78**	**97.83**	97.83	**97.83**	4.41
Ours w/o CL	92.59	92.75	92.75	92.75	**4.27**
Ours w/Bi-LSTM	91.85	89.73	94.93	92.25	–
Ours w/Bi-GRU	94.81	93.66	96.38	95.00	–
Ours w/BERT	97.41	96.45	98.55	97.49	–

Bold highlights the best score in the column

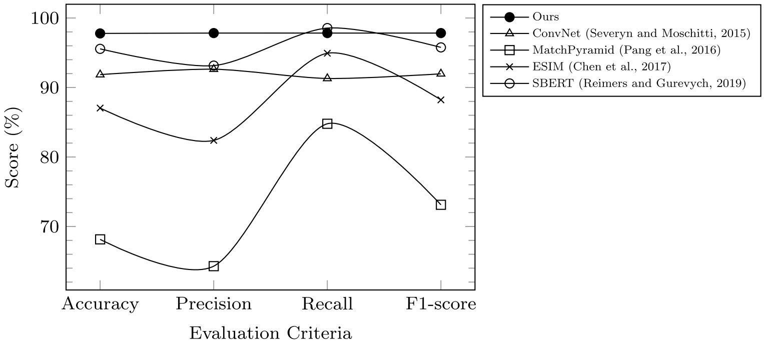


Fig. 7 Experimental result of different models on our test set

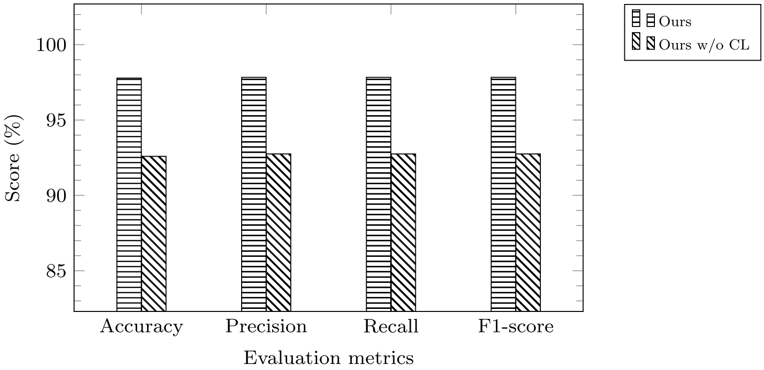


Fig. 8 Ablation result of our model on test set

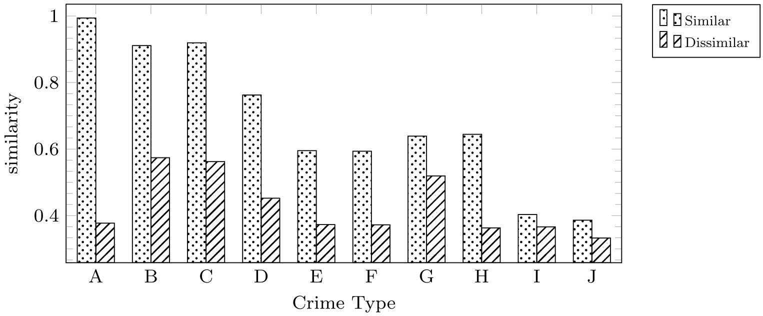


Fig. 9 Cosine similarity of positive and negative samples in the test set (the higher similarity the more similar the two samples)

of contrastive learning leads to performance improvements. We also analyse the model performance when using Bi-LSTM, Bi-GRU and BERT as encoders and find that the best results are obtained using CNN.

The above experimental observations may imply:

1. When applying deep neural networks to real-world scenarios, some SOTA neural networks may not perform as well as we expect. We should deal with problems case by case.
2. In the case of few-shot dataset, the contrastive learning might be a good way to improve the representation learning ability of the model.

6.5 Distance distribution of samples

Figure 9 visualises the cosine similarity of intra-class and inter-class samples for all the crime types of the test data. The bars with stars indicate the similarity of intra-class (similar) samples, and the bars with slashes indicates the similarity of inter-class (dissimilar) samples. Specifically, for X_1 (one case) in each 3-tuple data

Table 5 Mapping relationships between the crime types involved and the letters in Fig. 9

Crime type	Letter
Hindering the prevention and treatment of infectious diseases	A
Disrupting public service	B
Defiance and affray	C
Fraud	D
Selling fake and inferior products	E
Illegal business	F
Robbery	G
Rumour-mongering	H
Illegal acquisition and sale of rare and endangered wildlife	I
Illegal hunting and killing of rare and endangered wildlife protected by the state	J

(y, X_1, X_2) of each type of criminal cases, we calculate its average similarity with all the cases of the same crime type and its average similarity with all the cases of different crime types. For the convenience of visualisation, we map each crime type to the corresponding letter. Table 5 shows their corresponding relations.

The similarity between $\vec{u}_1$ and $\vec{u}_2$ is calculated by:

$$\text{similarity}\ (\vec{u}_1, \vec{u}_2) = \frac{\sum_{i=1}^{n} u_{1,i} \cdot u_{2,i}}{\sqrt{\sum_{i=1}^{n} u_{1,i}^2} \times \sqrt{\sum_{i=1}^{n} u_{2,i}^2}}, \tag{17}$$

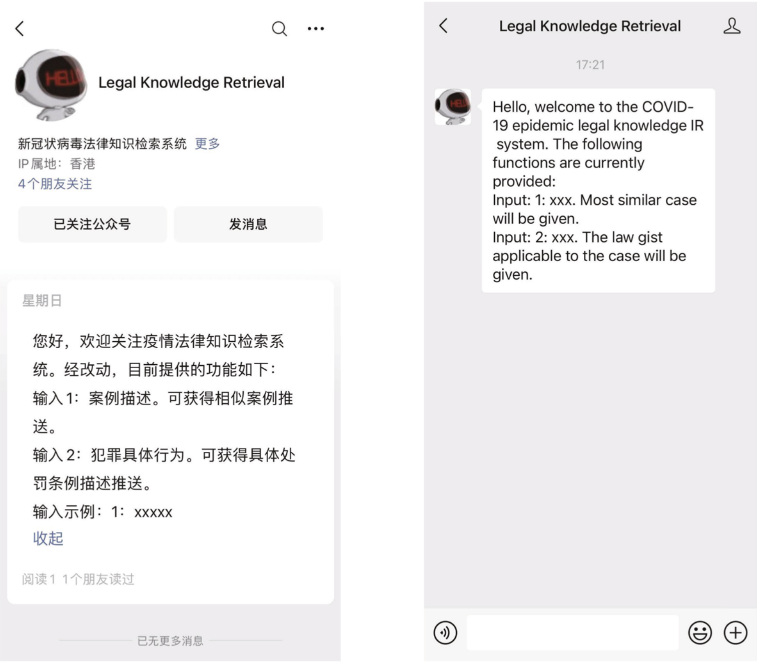
which is cosine similarity. We take the average of the cosine similarities calculated for each class's positive and negative samples separately to represent the intra-class and inter-class sample similarity. The results show that the hypothesis that the intra-class samples are more similar than the inter-class samples is reasonable, and our method can distinguish the two texts well in most cases.

Algorithm 3: Get Response

```
Input: user's input $X = \{w_1, \cdots, w_n\}$.
Output: corresponding case gist $G$ or similar case $C$.
if $int(\ X[0]\ ) == 1$ then
    $Inp$ = getTensor($X$[2 :]);
    $R$ = getResponse($Inp$, mode='case');
    return $C$.
else
    $Inp$ = getTensor($X$[2 :]);
    $G$ = getResponse($Inp$, mode='gist');
    return $G$.
```

7 Deployment

This section will present the details of how we deploy our system.



您好，欢迎关注疫情法律知识检索系统。经改动，目前提供的功能如下：
输入1：案例描述。可获得相似案例推送。
输入2：犯罪具体行为。可获得具体处罚条例描述推送。
输入示例：1：xxxxx

Hello, welcome to the COVID-19 epidemic legal knowledge IR system. The following functions are currently provided:
Input: 1: xxx. Most similar case will be given.
Input: 2: xxx. The law gist applicable to the case will be given.

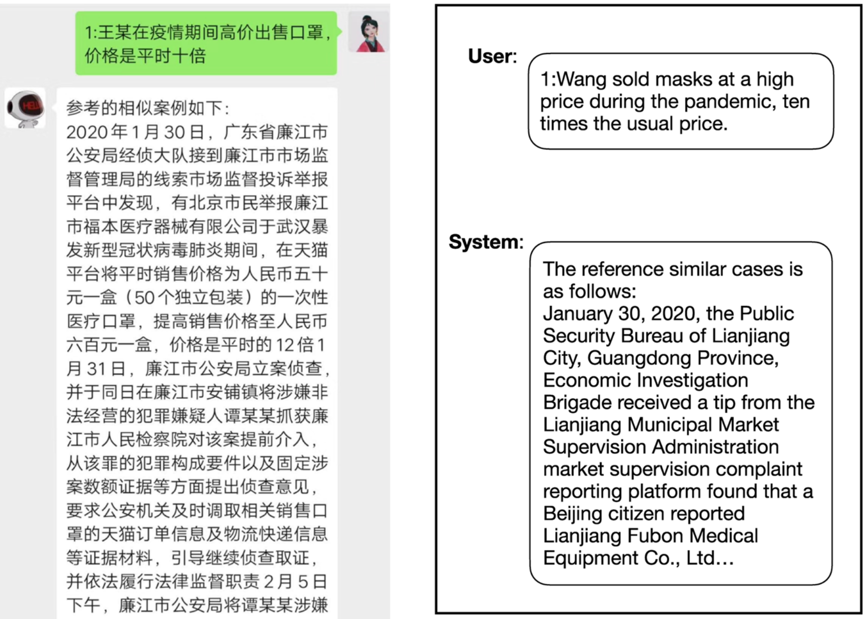
(a) The tips for new followers in Chinese

(b) The translation of the tips for new followers in Chinese

Fig. 10 The tips for new followers in Chinese and English. When new users follow our WeChat Official Account, the system will give relevant instructions on using the system

```
Algorithm 4: Retrieve Answer
Input: query vector $Inp$, Answer type $mode$.
Output: corresponding case gist $G$ or similar case $C$.
$coarseVec$ = MatchingDatabase[‘coarse’].
$idxCoarse$ = findNearest(Inp, coarseVec).
$key$ = ‘crime_’ + str($idxCoarse$).
$c$ = getCrime(MatchingDatabase, $key$).
if $mode$ == ‘$gist$’ then
    $G$ = getGist(MatchingDatabase, $key$);
    return $G$.
else
    $sInd$, $eInd$ = getRange(MatchingDatabase, $key$);
    $fineVecs$ = LocalDocVecs[sInd:eInd];
    $idxFine$ = findNearest(Inp, fineVecs);
    $cases$ = getCases(MatchingDatabase, $key$);
    $C$ = cases[idxFine];
    return $C$.
```

Once the model is trained, we deploy it on the developer server to provide the users’ service. As shown in Fig. 10, we customised the “Followed Reply” function



1:王某在疫情期间高价出售口罩，价格是平时十倍

参考的相似案例如下：
2020年1月30日，广东省廉江市公安局经侦大队接到廉江市市场监督管理局的线索市场监督投诉举报平台中发现，有北京市民举报廉江市福本医疗器械有限公司于武汉暴发新型冠状病毒肺炎期间，在天猫平台将平时销售价格为人民币五十元一盒（50个独立包装）的一次性医疗口罩，提高销售价格至人民币六百元一盒，价格是平时的12倍1月31日，廉江市公安局立案侦查，并于同日在廉江市安铺镇将涉嫌非法经营的犯罪嫌疑人谭某某抓获廉江市人民检察院对该案提前介入，从该罪的犯罪构成要件以及固定涉案数额证据等方面提出侦查意见，要求公安机关及时调取相关销售口罩的天猫订单信息及物流快递信息等证据材料，引导继续侦查取证，并依法履行法律监督职责2月5日下午，廉江市公安局将谭某某涉嫌

User:

1:Wang sold masks at a high price during the pandemic, ten times the usual price.

System:

The reference similar cases is as follows:
January 30, 2020, the Public Security Bureau of Lianjiang City, Guangdong Province, Economic Investigation Brigade received a tip from the Lianjiang Municipal Market Supervision Administration market supervision complaint reporting platform found that a Beijing citizen reported Lianjiang Fubon Medical Equipment Co., Ltd…

(a) An example that a real user uses our system in Chinese

(b) The translation of the example in (a)

Fig. 11 An example that a real user uses our system. A real user entered a case about selling masks at an excessive price, and the system responded with a similar case for that user

of the WeChat Official Account. When new users enter the dialogue window, the system automatically gives a brief introduction of system functions and usage examples. Then users can enter content according to the system's prompts and get the appropriate responses.

Algorithm 3 shows how the system process user's input and provides the corresponding result to the user. Specifically, we first use string operations to extract the case description part and a character representing the type of response the user wants of the user's input. If the first character of the user's original input is 1, we feed the matching case with the highest similarity to the user as a result (see lines 1–4). Otherwise, the matching legal gist is fed back to the user as the reference result (see lines 5–8). We then pre-process the user's case descriptions and encode them using the trained model to obtain a vector representation. Next, we obtain the desired return result by matching the vector representation of the input case with the vectors in the matching database. The whole matching database can be found in "Appendix". Figure 11 shows an example of using our system.

As shown in Algorithm 4, to speed up the response time of the system, our matching process is divided into two parts: coarse-sorting and fine-sorting. We firstly perform a coarse sorting process. On lines 1 and 2, we take all the vectors corresponding to the *coarse* fields in the matching database as the set of vectors to be retrieved and retrieve the index number *idxCoarse* corresponding to the nearest neighbour vector of the query vector *Inp*. The index number *idxCoarse* reflects the crime type to the input case. On line 3, we use the index *idxCoarse* to construct the primary key *key* so that we can query the matching database. On line 4, we extract the corresponding crime type from the matching database based on the *key*. If parameter *mode* is equal to 'gist', on lines 5–7, we simply extract the corresponding case gist from the matching database based on the *key* as the return result, without entering the refined sorting process. Otherwise, on lines 8–13, we proceed with the fine-sorting process. On line 9, based on the *key*, we extract the retrieval range of the local doc vectors corresponding to the crime type *c* from the matching database. On line 10, we extract the new set of vectors to be retrieved from the local doc vectors. On line 11, we obtain the new index number *idxFine* of the nearest neighbour vector of the query vector. On line 12, we extract all cases corresponding to crime type *c* from the matching database. On line13, using the index numbers *idxFine*, we can get the most similar cases from the *cases* as the return result.

8 Related work

This section will discuss the related work to show how our work advances the state of the art in relevant research areas.

8.1 Legal information retrieval

Legal IR has become one of the essential topics in the research area of Artificial Intelligence and Law (AI & Law) (Westermann et al. 2019; Šavelka and Ashley 2022).

- Some researchers focus on developing easy-to-use legal IR systems. For example, Nejadgholi et al. (2017) developed a semi-supervised learning-based legal IR system for immigration cases. The system can find cases in which the fact description is the most similar to the query. Sugathadasa et al. (2018) also developed a legal IR system. It uses a page rank graph network with TF-IDF (Mao and Chu 2002), which builds document embeddings by constructing a vector space to perform document retrieval tasks.
- Some emphasise proposing effective IR methods for specific legal cases. For examople, Savelka et al. (2019) and Novotná (2020) employ static word embeddings and cosine similarity to calculate the similarity between a judgement and the query. Šavelka and Ashley (2022) further leverages static word embeddings and various topic models to retrieve proper sentences from a case law database based on the phrase in a given clause.

However, our work in this paper is different from the above work in the following aspects.

(1) These methods usually retrieve specific cases (*e.g.*, immigration, veterans' claims, or others) because they are trained on specific case datasets. As a result,

few can retrieve legal cases of the COVID-19 pandemic. Instead, our system focuses on retrieving legal cases related to the COVID-19 pandemic.

(2) Furthermore, although these methods are effective and interpretable, they cannot utilize the semantic information of sentences to identify similarities between sentences because they use Boolean rules or word embedding techniques to calculate the similarity between sentences. Unlike them, our system uses a convolutional neural network model trained on the COVID-19 pandemic dataset to capture the semantic information between the data.

In addition, unlike our work in this paper, few existing studies developed datasets on legal cases of the COVID-19 pandemic. For example, Ma et al. (2021c) developed a Chinese legal case retrieval dataset on general criminal cases. Rabelo et al. (2022) produced a common law case retrieval dataset based on the Federal Court of Canada case law database. They are both well-known large-scale legal case retrieval datasets. However, neither of them involves legal cases related to COVID-19.

8.2 CNN-based methods for legal information retrieval

Some researchers use CNN to improve the retrieval performance of legal IR, but our work in this paper is different from them. For example, Tran et al. (2019) used CNNs to implement document and sentence-level pooling, achieving the state-of-art result on the COLIEE dataset (Kano et al. 2018). Wan and Ye (2021) proposed a TinyBERT-based Siamese-CNN model to calculate the similarity between judgment documents. Specifically, they first used TinyBERT to obtain the embedding of documents and then extracted document features by a CNN to calculate the similarity of judgment documents. Sampath and Durairaj (2022) used CNNs to obtain similarity features of substructures (*i.e.* parts of the whole legal document) and used the similarity features to improve the retrieval performance of the model further. However, although we also used CNNs to complete the pooling step, unlike them, we propose a semantic matching network for pairwise relationship learning based on CNN and use contrastive learning to help the network distinguish between sentences.

8.3 Pre-trained language model based methods for legal information retrieval

Some researchers apply pretrained language models to legal IR. Rossi and Kanoulas (2019) used the pre-trained language model BERT to reformulate the ranking problem in the retrieval problem as a pairwise relevance score problem to improve legal IR. Shao et al. (2020) used BERT to build a paragraph-level interaction model to calculate the relevance between two cases and to complete document matching. Xiao et al. (2021) developed a pretrained language model, called Lawformer, for long legal document understanding. The model is based on pretrained language model Longformer, integrating local and global attention to capturing long-range dependencies between words in a lengthy legal document. The model can be used for similar legal case retrieval. Askari et al. (2021) integrates lexical and neural ranking models for similar legal case retrieval. They used multiple methods (including term extraction and automatic summarisation based on longformer encoder-decoder) for longer query documents to create a shorter query document. Their experiments show the excellent performance of their model on the COLIEE retrieval benchmarks. Wehnert et al. (2021) fine-tuned a BERT classifier for the statute law retrieval task and combined it with TF-IDF to vectorise the documents. Their approach outperforms most baseline methods. Zhu et al. (2022) proposed a two-stage BERT-based ranking method and integrated it with multi-task learning to improve the retrieval performance of the model further. Abolghasemi et al. (2022) used multi-task learning to fine-tune the BERT model to learn document-level representation. Their approach improves the efficiency of the BERT re-ranker in similar legal case retrieval.

However, although these methods based on pretrained language models have improved legal IR performance, they require substantial computational resources, making them difficult to implement for online deployment. Rather, the core of our system is a lightweight CNN-based retrieval model that is easily deployed online and responds in real-time in retrieval.

8.4 COVID-19 information retrieval

With the impact of the COVID-19 pandemic in recent years, there has been an explosion of information about With the impact of the COVID-19 pandemic in recent years, there has been an explosion of information about COVID-19. To this end, some researchers study information retrieval of COVID-19. For example, Wise et al. (2020) first proposed a heterogeneous knowledge graph on COVID-19 for extracting and visualising the relationships among COVID-19 scientific articles. Then they develop a document similarity engine by combining graph topology information with semantic information, which utilises low-dimensional graph embeddings and semantic embeddings from the knowledge graph for similar article retrieval. Esteva et al. (2021) designed a semantic and multi-stage search engine for the COVID-19 lature, which helps healthcare workers to find scientific answers and avoid misinformation in times of crisis. Alzubi et al. (2021) developed a retriever-reader dual algorithm called COBERT, which can retrieve literature on COVID-19 pneumonia. Tran et al. (2021) developed a scientific paper retrieval system for supporting users to efficiently access knowledge in the large number of COVID-19 scientific papers published rapidly. Aonillah et al. (2022) developed a COVID-19 question answering system in Bahasa Indonesia using recognising question entailment.

However, they do not involve relevant legal information but only some simple general knowledge retrieval. Unlike them, the system we developed mainly provides users with legal knowledge related to COVID-19 pneumonia and retrieves some specific real legal cases for users' reference.

8.5 Deep neural network based information retrieval

Due to the successful application of deep neural networks in natural language processing, many deep neural network based information retrieval methods have been proposed for document retrieval. Liu et al. (2019) used a hybrid approach to calculate the similarity between texts. They first extract keywords from the text to obtain the corresponding word vectors. Then, they combine the obtained word vectors with the statistical vectors of predefined text feature words to synthesise the similarity between two legal cases. Khattab and Zaharia (2020) proposed a ranking model called ColBERT, which uses BERT to encode queries and documents independently to accomplish end-to-end retrieval of millions of documents. Liu et al. (2021) developed a four-tower BERT model that uses the distance between hard negative instances and simple negative instances to enhance the performance of the retriever during training. The studies mentioned above are based on the siamese structure, consisting of two or more consistent sub-networks (*e.g.*, CNN, LSTM or BERT). These sub-networks extract features from two or more inputs separately during training. This structure obtains the maximum retrieval efficiency and is suitable for deployment in actual production environments.

Similar to them, we also employ a siamese structure for online retrieval tasks. However, they have not yet use self-supervised learning based on the siamese structure to enhance the task effect. In contrast, we used self-supervised learning in this paper to enhance the semantic correlation between legal cases. Also, we analyse the corresponding effects of the self-supervised learning network based on the siamese structure and demonstrate that the network performs better than the baseline model.

8.6 BERT-based information retrieval

Since the BERT model was born in 2018, many researchers have presented many BERT-based methods for various IR tasks and achieved impressive results. These methods are in three categories according to the size of the datasets. The first category of methods is large-scale IR, which uses enormous data resources to train models for optimal performance. For example, a BERT-based CogLTX framework proposed by Ding et al. (2020) shows good performance on four large datasets (including NewsQA (Trischler et al. 2017), HotpotQA (Wharton et al. 1994), 20NewsGroups (Lang 1995), and Alibaba), each of which contains tens of thousands of documents. Li and Gaussier (2021) utilises 500,000 news articles and 25 million web documents to train a standard BERT model for long document retrieval. Liu et al. (2021) proposed a four-tower model focusing on the retrieval phase and demonstrated its effectiveness in large-scale retrieval. Guo et al. (2022) adopts the same transformer architecture as BERT and utilises nearly three million web pages with HTML sources and their tree structure to pre-train an information retrieval model called Webformer. Their model outperforms existing pre-trained models in web page retrieval tasks. However, these methods require large training data for feature extraction and representations, which acquisition is complex, and on which the cost of performing manual annotation is prohibitive. Instead, our work investigates the similarity computation performance of deep learning networks for complex datasets with small samples and explores a feasible solution in practice, especially in the legal domain.

The second category of BERT-based IR methods are few-shot IR. They use the small number of available data to fine-tune BERT to accomplish the retrieval task. Maillard et al. (2021) developed a single general-purpose “universal" BERT-based retrieval model that uses a multi-task approach for training and performs compatible with or even better than specialised retrievers in few-sample retrieval. Yu et al. (2021) proposed a session-intensive retrieval system, ConvDR, which learns the embedding dot product to retrieve documents. They also use a teacher-student BERT-based framework to grant ConvDR the ability to learn from very few samples to address the problem of data-hungriness. Mokrii et al. (2021) investigated the transferability of BERT-based ranking models and comprehensively compared transfer learning with models trained on pseudo-labels generated using the BM25 scoring function. Their study shows that fine-tuning models trained on pseudo labels by using a small number of annotated queries can match or exceed the performance of transferred models. All the above methods primarily use BERT to improve the performance of the models for retrieval in scenarios with a small number of training data. These models show that BERT is good at dealing with few-sample scenarios. However, in our ablation experiments, we replaced the CNN encoder by BERT and found that BERT did not lead to a performance improvement but a degradation. Therefore, BERT is not always effective in few-sample real-world scenarios of IR. Furthermore, large language models tend to be overfitting because of insufficient data for fine-tuning. Therefore, we should handle it according to the specific task.

The third category of BERT-based IR methods is zero-shot IR, such as cross-language zero-shot IR (MacAvaney et al. 2020; Wang et al. 2021; Nair et al. 2022), zero-shot self-supervised learning (Assareh 2022) and zero-shot neural paragraph retrieval (Ma et al. 2021a). However, these methods are not as effective as the first two above in practical applications. Therefore, we do not discuss them here.

8.7 Intelligent legal system

Benefiting from a large number of available high-quality textual datasets (Locke and Zuccon 2018; Duan et al. 2019; Zhong et al. 2020b) and web-based public data in the legal field, researchers recently studied extensively how natural language processing can empower the legal field on many topic, such as:

(1) *Legal opinion generation* Ye et al. (2018) propose a model for generating legal opinions based on factual descriptions and uses the generated legal opinions to improve the interpretability of their crime prediction system.

(2) *Legal judgment prediction* Zhong et al. (2020a) proposed a reinforcement learning-based model for predicting interpretable judgments. Ma et al. (2021b) used

a real legal dataset to reasonably predict the court judgment of a case according to the plaintiff's claims and their court debate. Feng et al. (2022) proposed an event-based prediction model with constraints for legal judgments, which addresses the failure of existing methods to locate the critical event information that determines the judgment.

(3) *Legal document summarisation*Nguyen et al. (2021) used reinforcement learning to train a model for legal document summarisation. Specifically, they used a proximal policy optimisation approach and proposed a new reward function that encourages the generation of summarisation candidates that satisfy lexical and semantic criteria. Klaus et al. (2022) fine-tuned a transformer-based extractive summarisation model to simplify the statutes, so non-jurists can easily understand them.

However, although our system is an intelligent legal system, ours is different from the above systems. They explore task-specific improvements on large datasets and demonstrate the feasibility of the methods under specific designs. Rather, in this paper, we aim to explore the usability of neural network models, including the currently popular pre-trained language models, in implementing a matching-style legal IR system for the COVID-19 pandemic with samples of small size.

9 Conclusion

We propose a semantic matching network for pairwise relation learning. Moreover, we introduce auxiliary contrastive learning to help network better distinguish the sentences. Our experiments show that our method leads to performance improvements under a variety of encoder designs.

Based on the network we proposed, we design and implement a legal IR system for the COVID-19 pandemic. The system can identify: (1) the crime cases entered by the user and find the most similar cases to be pushed to the user as answers, and (2) the crime cases documented by the user and give the reference legal gist applicable to the case. Meanwhile, the study could benefit developing a more comprehensive legal IR system and similar systems. In the future, it is worth doing more experiments on more data set to analyse the effect of various neural models for such tasks and accordingly improving the system.

Appendix: matching database

The matching database we used is shown in Table 6. For the sake of simplicity, the contents of the 'facts' field are not listed. In fact, the 'facts' field is composed of all the factual descriptions contained in each crime type. The 'coarse' field holds a text vector representation for each crime type only, which is a vector of the first case out of all the case descriptions corresponding to each crime type.

Table 6 Matching database used to get system's response

Key	Crime	Gist	Coarse	Facts	Start	End
crime_0	Hindering the Prevention and Treatment of Infectious Diseases	... In accordance with Article 330 of the Criminal Law, the crime of.	[0.0, 1.656654953956604, ⋯]	[fact1, fact2, ⋯]	0	9
crime_1	Disrupting Public Service	... In accordance with Article 277 (1) and (3) of the Criminal Law, the crime of.	[0.0, 0.5752912759780884, ⋯]	[fact1, fact2, ⋯]	9	19
crime_2	Defiance and Affray	... disrupt the normal work of epidemic prevention and control is to be considered as an aggravating circumstance.	[0.0, 1.035813331604004, ⋯]	[fact1, fact2, ⋯]	19	29
crime_3	Fraud	... In accordance with the provisions of Article 266 of the Criminal Law.	[0.0, 1.4794719219207764, ⋯]	[fact1, fact2, ⋯]	29	39
crime_4	Selling Fake and Inferior Products	... If the provisions of Article 140 of the Criminal Law.	[0.0, 2.0093281269073486, ⋯]	[fact1, fact2, ⋯]	39	43
crime_5	Illegal Business	... In accordance with the provisions of Article 225(4) of the Criminal Law.	[0.0, 1.699977159500122, ⋯]	[fact1, fact2, ⋯]	43	47
crime_6	Robbery	... disrupt the normal work of epidemic prevention and control is to be considered as an aggravating circumstance.	[0.0, 1.0496822595596313, ⋯]	[fact1, fact2, ⋯]	47	61
crime_7	Rumour-mongering	... In accordance with Article 291.1(2) of the Criminal Law.	[0.0, 0.09105069190263748, ⋯]	[fact1, fact2, ⋯]	61	90
crime_8	Illegal Hunting and Killing of Rare and Endangered Wildlife Protected By The State	... In accordance with Article 341(1) of the Criminal Code.	[0.0, 0.826084554195404, ⋯]	[fact1, fact2, ⋯]	90	102
crime_9	Illegal Acquisition and Sale of Rare and Endangered Wildlife	... In accordance with Article 341(1) of the Criminal.	[0.0, 2.1975739002227783, ⋯]	[fact1, fact2, ⋯]	102	112

Acknowledgements Deep thanks are due to the reviewer of this paper, whose insightful comments greatly helped us significantly improve the quality of the paper. This work was supported by the National Natural Science Foundation of China (No. 61762016) and Guangxi Key Lab of Multi-Source Information Mining & Security (No. 19-A-01-01).

References

Abolghasemi A, Verberne S, Azzopardi L (2022) Improving BERT-based query-by-document retrieval with multi-task optimization. European Conference on Information Retrieval, Lecture Notes in Computer Science 13186:3–12

Alzubi JA, Jain R, Singh A, Parwekar P, Gupta M (2021) COBERT: COVID-19 question answering system using BERT. Arab J Sci Eng 46:1–11

Aonillah MZ, Hasmawati H, Romadhony A (2022) Question entailment on developing Indonesian COVID-19 question answering system. J Comput Syst Inform 3(4):269–276

Askari A, Verberne S, Alonso O, Marchesin S, Najork M, Silvello G (2021) Combining lexical and neural retrieval with longformer-based summarization for effective case law retrieva. In: Proceedings of the 2nd International Conference on Design of Experimental Search & Information Retrieval Systems, pp 162–170

Assareh A (2022) Information retrieval from alternative data using zero-shot self-supervised learning. In: 2022 IEEE Symposium on Computational Intelligence for Financial Engineering and Economics, pp 1–5

Cao Y, Li T, Jia Z, Yin C (2019) BiGRU: new method of Chinese text sentiment analysis. Comput Sci Explor 13(6):973–981

Chen Q, Zhu X, Ling ZH, Wei S, Jiang H, Inkpen D (2017) Enhanced LSTM for natural language inference. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, vol 1, pp 1657–1668

Devlin J, Chang MW, Lee K, Toutanova K (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol 1, pp 4171–4186

Ding M, Zhou C, Yang H, Tang J (2020) CogLTX: applying BERT to long texts. Adv Neural Inf Process Syst 33:12792–12804

Duan X, Wang B, Wang Z, Ma W, Cui Y, Wu D, Wang S, Liu T, Huo T, Hu Z, Wang H, Liu Z (2019) CJRC: A reliable human-annotated benchmark dataset for Chinese judicial reading comprehension. In: Chinese Computational Linguistics. Lecture Notes in Computer Science, vol 11856, pp 439–451

Esteva A, Kale A, Paulus R, Hashimoto K, Yin W, Radev D, Socher R (2021) COVID-19 information retrieval with deep-learning based semantic search, question answering, and abstractive summarization. NPJ Digit Med 4(1):1–9

Feng Y, Li C, Ng V (2022) Legal judgment prediction via event extraction with constraints. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, vol 1, pp 648–664

Gardner MW, Dorling S (1998) Artificial neural networks (the multilayer perceptron): a review of applications in the atmospheric sciences. Atmos Environ 32(14–15):2627–2636

Graves A, Schmidhuber J (2005) Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Netw 18(5–6):602–610

Guo Y, Ma Z, Mao J, Qian H, Zhang X, Jiang H, Cao Z, Dou Z (2022) Webformer: Pre-training with web pages for information retrieval. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 1502–1512

Hadsell R, Chopra S, LeCun Y (2006) Dimensionality reduction by learning an invariant mapping. In: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol 2, pp 1735–1742

Hu Z, Li X, Tu C, Liu Z, Sun M (2018) Few-shot charge prediction with discriminative legal attributes. In: Proceedings of the 27th International Conference on Computational Linguistics, pp 487–498

Huang Q, Luo X (2018) State-of-the-art and development trend of artificial intelligence combined with law. Comput Sci 45(12):1–11

Kano Y, Kim MY, Yoshioka M, Lu Y, Rabelo J, Kiyota N, Goebel R, Satoh K (2018) COLIEE-2018: Evaluation of the competition on legal information extraction and entailment. JSAI International Symposium on Artificial Intelligence, Lecture Notes in Computer Science 11717:177–192

Khattab O, Zaharia M (2020) ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 39–48

Klaus S, Van Hecke R, Djafari Naini K, Altingovde IS, Bernabé-Moreno J, Herrera-Viedma E (2022) Summarizing legal regulatory documents using transformers. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 2426–2430

Krizhevsky A, Sutskever I, Hinton GE (2017) Imagenet classification with deep convolutional neural networks. Commun ACM 60:84–90

Lang K (1995) Newsweeder: Learning to filter netnews. In: Proceedings of the 12th International Conference on International Conference on Machine Learning, pp 331–339

Li M, Gaussier E (2021) KeyBLD: Selecting key blocks with local pre-ranking for long document information retrieval. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 2207–2211

Li S, Zhao Z, Hu R, Li W, Liu T, Du X (2018) Analogical reasoning on Chinese morphological and semantic relations. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, vol 2, pp 138–143

Liu CL, Hsieh CD (2006) Exploring phrase-based classification of judicial documents for criminal charges in Chinese. In: Foundations of Intelligent Systems. Lecture Notes in Computer Science, vol 4203, pp 681–690

Liu P, Wang S, Wang X, Ye W, Zhang S (2021) QuadrupletBERT: An efficient model for embedding-based large-scale retrieval. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp 3734–3739

Liu Y, Luo X, Yang X (2019) Semantics and structure based recommendation of similar legal cases. In: Proceedings of the 2019 IEEE 14th International Conference on Intelligent Systems and Knowledge Engineering, pp 388–395

Locke D, Zuccon G (2018) A test collection for evaluating legal case law search. In: Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, pp 1261–1264

Ma J, Korotkov I, Yang Y, Hall K, McDonald R (2021a) Zero-shot neural passage retrieval via domain-targeted synthetic question generation. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pp 1075–1088

Ma L, Zhang Y, Wang T, Liu X, Ye W, Sun C, Zhang S (2021b) Legal judgment prediction with multi-stage case representation learning in the real court setting. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 993–1002

Ma Y, Shao Y, Wu Y, Liu Y, Zhang R, Zhang M, Ma S (2021c) LeCaRD: A legal case retrieval dataset for Chinese law system. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 2342–2348

MacAvaney S, Soldaini L, Goharian N (2020) Teaching a new dog old tricks: Resurrecting multilingual retrieval using zero-shot learning. European Conference on Information Retrieval, Lecture Notes in Computer Science 12036:246–254

Maillard J, Karpukhin V, Petroni F, Yih Wt, Oguz B, Stoyanov V, Ghosh G (2021) Multi-task retrieval for knowledge-intensive tasks. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, vol 1, pp 1098–1111

Mao W, Chu WW (2002) Free-text medical document retrieval via phrase-based vector space model. In: Proceedings of the AMIA Symposium, pp 489–493

Mokrii I, Boytsov L, Braslavski P (2021) A systematic evaluation of transfer learning and pseudo-labeling with BERT-based ranking models. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 2081–2085

Mou L, Men R, Li G, Xu Y, Zhang L, Yan R, Jin Z (2016) Natural language inference by tree-based convolution and heuristic matching. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, vol 2, pp 130–136

Nair S, Yang E, Lawrie D, Duh K, McNamee P, Murray K, Mayfield J, Oard DW (2022) Transfer learning approaches for building cross-language dense retrieval models. European Conference on Information Retrieval, Lecture Notes in Computer Science 13185:382–396

Nair V, Hinton GE (2010) Rectified linear units improve restricted Boltzmann machines. In: Proceedings of the 27th International Conference on International Conference on Machine Learning, pp 807–814

Nejadgholi I, Bougueng R, Witherspoon S (2017) A semi-supervised training method for semantic search of legal facts in canadian immigration cases. In: Proceedings of the 30th International Conference on Legal Knowledge and Information Systems, pp 125–134

Nguyen DH, Nguyen BS, Nghiem NVD, Le DT, Khatun MA, Nguyen MT, Le H (2021) Robust deep reinforcement learning for extractive legal summarization. In: Proceedings of the 28th International Conference on Neural Information Processing, Communications in Computer and Information Science, vol 1517, pp 597–604

Novotná T (2020) Document similarity of Czech supreme court decisions. Masaryk Univ J Law Technol 14(1):105–122

Pang L, Lan Y, Guo J, Xu J, Wan S, Cheng X (2016) Text matching as image recognition. In: Proceedings of the 30th AAAI Conference on Artificial Intelligence, pp 2793–2799

Rabelo J, Goebel R, Kim MY, Kano Y, Yoshioka M, Satoh K (2022) Overview and discussion of the competition on legal information extraction/entailment (COLIEE) 2021. Rev Socionetw Strateg 16(1):111–133

Reimers N, Gurevych I (2019) Sentence-BERT: Sentence embeddings using siamese BERT-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp 3973–3983

Rossi J, Kanoulas E (2019) Legal search in case law and statute law. In: Proceedings of the 32th International Conference on Legal Knowledge and Information Systems, vol 322, pp 83–92

Sampath K, Durairaj T (2022) PReLCaP: precedence retrieval from legal documents using catch phrases. Neural Process Lett 54:1–19

Šavelka J, Ashley KD (2022) Legal information retrieval for understanding statutory terms. Artif Intell Law 30(2):245–289

Savelka J, Xu H, Ashley KD (2019) Improving sentence retrieval from case law for statutory interpretation. In: Proceedings of the 17th International Conference on Artificial Intelligence and Law, pp 113–122

Severyn A, Moschitti A (2015) Learning to rank short text pairs with convolutional deep neural networks. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 373–382

Shao Y, Mao J, Liu Y, Ma W, Satoh K, Zhang M, Ma S (2020) BERT-PLI: Modeling paragraph-level interactions for legal case retrieval. In: Proceedings of the 29th International Joint Conferences on Artificial Intelligence, pp 3501–3507

Sugathadasa K, Ayesha B, Silva Nd, Perera AS, Jayawardana V, Lakmal D, Perera M (2018) Legal document retrieval using document vector embeddings and deep learning. Science and Information Conference, Advances in Intelligent Systems and Computing 857:160–175

Tran V, Nguyen ML, Satoh K (2019) Building legal case retrieval systems with lexical matching and summarization using a pre-trained phrase scoring model. In: Proceedings of the 17th International Conference on Artificial Intelligence and Law, pp 275–282

Tran V, Tran VH, Nguyen P, Nguyen C, Satoh K, Matsumoto Y, Nguyen M (2021) CovRelex: A COVID-19 retrieval system with relation extraction. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, pp 24–31

Trischler A, Wang T, Yuan X, Harris J, Sordoni A, Bachman P, Suleman K (2017) NewsQA: A machine comprehension dataset. In: Proceedings of the 2nd Workshop on Representation Learning for NLP, pp 191–200

Wan Z, Ye N (2021) Similarity calculation method of siamese-CNN judgment document based on Tiny-BERT. In: 2021 International Conference on Intelligent Computing, Automation and Applications, pp 27–32

Wang R, Zhang Z, Zhuang F, Gao D, Wei Y, He Q (2021) Adversarial domain adaptation for cross-lingual information retrieval with multilingual BERT. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, pp 3498–3502

Wang W, Yan M, Wu C (2018) Multi-granularity hierarchical attention fusion networks for reading comprehension and question answering. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, pp 1705–1714

Weber R, Martins A, Barcia R (1998) On legal texts and cases. Federal University of Santa Catarina, Tech. rep

Wehnert S, Sudhi V, Dureja S, Kutty L, Shahania S, De Luca EW (2021) Legal norm retrieval with variations of the BERT model combined with TF-IDF vectorization. In: Proceedings of the 18th International Conference on Artificial Intelligence and Law, pp 285–294

Westermann H, Šavelka J, Walker VR, Ashley KD, Benyekhlef K (2019) Computer-assisted creation of boolean search rules for text classification in the legal domain. In: Proceedings of the 32th International Conference on Legal Knowledge and Information Systems, pp 123–132

Wharton CM, Holyoak KJ, Downing PE, Lange TE, Wickens TD, Melz ER (1994) Below the surface: analogical similarity and retrieval competition in reminding. Cogn Psychol 26(1):64–101

Wise C, Calvo MR, Bhatia P, Ioannidis V, Karypus G, Price G, Song X, Brand R, Kulkani N (2020) COVID-19 knowledge graph: Accelerating information retrieval and discovery for scientific literature. In: Proceedings of Knowledgeable NLP: the First Workshop on Integrating Structured Knowledge and Neural Networks for NLP, pp 1–10

Xiao C, Hu X, Liu Z, Tu C, Sun M (2021) Lawformer: a pre-trained language model for Chinese legal long documents. AI Open 2:79–84

Ye H, Jiang X, Luo Z, Chao W (2018) Interpretable charge predictions for criminal cases: Learning to generate court views from fact descriptions. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018, vol 1, pp 1854–1864

Yu S, Liu Z, Xiong C, Feng T, Liu Z (2021) Few-shot conversational dense retrieval. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 829–838

Zhong H, Wang Y, Tu C, Zhang T, Liu Z, Sun M (2020) Iteratively questioning and answering for interpretable legal judgment prediction. In: Proceedings of the AAAI Conference on Artificial Intelligence 34:1250–1257

Zhong H, Xiao C, Tu C, Zhang T, Liu Z, Sun M (2020b) JEC-QA: A legal-domain question answering dataset. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence, pp 9701–9708

Zhong Q, Fan X, Luo X, Toni F (2019) An explainable multi-attribute decision model based on argumentation. Expert Syst Appl 117:42–61

Zhu J, Luo X, Wu J (2022) A BERT-based two-stage ranking method for legal case retrieval. Knowledge Science, Engineering and Management, Lecture Notes in Computer Science 13369:534–546

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.