# LeCaRD: A Legal Case Retrieval Dataset for Chinese Law System

Yixiao Ma, Yunqiu Shao, Yueyue Wu, Yiqun Liu*, Ruizhe Zhang, Min Zhang, Shaoping Ma
yiqunliu@tsinghua.edu.cn
Department of Computer Science and Technology, Institute for Artificial Intelligence,
Beijing National Research Center for Information Science and Technology,
Tsinghua University, Beijing 100084, China

## ABSTRACT

Legal case retrieval is of vital importance for ensuring justice in different kinds of law systems and has recently received increasing attention in information retrieval (IR) research. However, the relevance judgment criteria of previous retrieval datasets are either not applicable to non-cited relationship cases or not instructive enough for future datasets to follow. Besides, most existing benchmark datasets do not focus on the selection of queries. In this paper, we construct the Chinese **Le**gal **Ca**se **R**etrieval **D**ataset (LeCaRD), which contains 107 query cases and over 43,000 candidate cases. Queries and results are adopted from criminal cases published by the Supreme People's Court of China. In particular, to address the difficulty in relevance definition, we propose a series of relevance judgment criteria designed by our legal team and corresponding candidate case annotations are conducted by legal experts. Also, we develop a novel query sampling strategy that takes both query difficulty and diversity into consideration. For dataset evaluation, we implemented several existing retrieval models on LeCaRD as baselines. The dataset is now available to the public together with the complete data processing details.

## CCS CONCEPTS

• **Information systems** → **Test collections**.

## KEYWORDS

legal case retrieval, relevance judgment criteria, query sampling

**ACM Reference Format:**
Yixiao Ma, Yunqiu Shao, Yueyue Wu, Yiqun Liu*, Ruizhe Zhang, Min Zhang, Shaoping Ma. 2021. 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 (SIGIR '21), July 11–15, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3404835.3463250

## 1 INTRODUCTION

Legal case retrieval is significant to ensure legal justice in different law systems. Following the doctrine of *stare decisis*, precedents (prior cases decided in courts of law) are cited in common law jurisdictions to support arguments [20]. Meanwhile, although the prior cases are not directly involved in the final judgment in some other law systems (e.g., China, Japan, Germany), they are still crucial references during the decision-making process [9]. With the growing number of digitized legal documents, automatic legal case retrieval has received increasing attention in the broad research fields of Information Retrieval (IR) [2, 22].

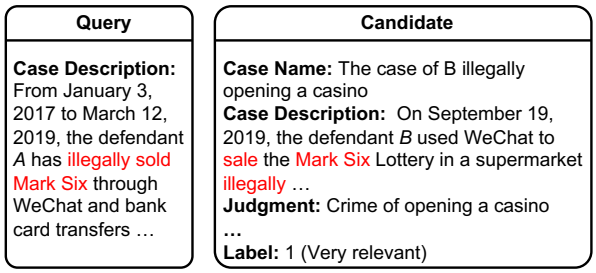


**Figure 1: An example of the query and candidate case.**

In recent years, researchers have been working on legal information retrieval and constructed several datasets. For the application of legal case retrieval datasets, relevance judgment is one of the crucial issues. There are two kinds of relevance judgments: cited-based and expert-based. Some datasets objectively define relevance judgment criteria as supportive precedents cited in the query document. For instance, Kano et al. [10] provided a benchmark dataset composed of Federal Court of Canada case law. However, such criteria can not be applied to other law systems without stare decisis (e.g. Chinese law system) due to the lack of the citation structure in case documents. Meanwhile, for search tasks with a particular topic, it is hard to take a citation as relevant without legal experts' help [12]. In other datasets such as AILA [3], relevant cases are assessed by legal experts with their professional knowledge. However, it is hard for assessors to have an absolutely unified understanding of relevance without a specific definition. Consequently, expert-based relevance judgment relies on assessors' professional ability. Such criteria are not instructive enough for future datasets to follow.

In addition to relevance judgment, query sampling is also significant in legal case retrieval. Users of real search systems are more concerned about the retrieval results of complicated or controversial queries than simple queries. Nevertheless, the difficulty of query set is often ignored by previous datasets which randomly sample queries in the case pool [10, 12]. Moreover, because online search users include parties litigants, judges, and the public with insufficient legal knowledge, search results need to provide relevant


*SIGIR '21, July 11–15, 2021, Virtual Event, Canada*

ACM ISBN 978-1-4503-8037-9/21/07…$15.00
https://doi.org/10.1145/3404835.3463250

*Corresponding author

cases with different aspects to meet all kinds of user needs. For instance, both procurators and lawyers expect to search for cases to support their different litigation strategies, while the public with unclear search intent may expect searched cases to provide as much basic information as possible. Therefore, the query diversity of the dataset is important as well.

In this paper, we present **LeCaRD**, the first **Le**gal **Ca**se **R**etrieval **D**ataset based on the Chinese law system. LeCaRD composes of 107 query cases and 10,700 candidate cases selected from a corpus of over 43,000 Chinese criminal judgements. A legenda of a query and candidate case is shown in Figure 1. Unlike previous works' relevance judgements that recognize relevant cases through either supportive cases in citations or expert knowledge, we propose a series of relevance judgement criteria based on critical factors combining subjective and objective evaluation. Our criteria are designed generally under the guidance of the official document published by the Supreme People's Court of China. All the assessments are made by multiple legal experts who are Masters in criminal law.

Furthermore, to cover queries of different difficulties and categories, we propose a novel query sampling strategy composed of two parts. The first part of queries contain queries of common charges sampled by a difficulty selection algorithm, while the second part queries include controversial cases to further increase the diversity of the dataset. To this end, we collect the complete set cases from Chinese second trial or retrial documents that have once revised their first-trial charges.

Several typical retrieval models are implemented on LeCaRD as baselines in this paper. The dataset and its preprocessing files are available to the public at https://github.com/myx666/LeCaRD. We believe this dataset can further facilitate the research on legal case retrieval.

## 2 EXISTING DATASETS

### 2.1 Case Law Collection

Case Law Collection [12] was presented in 2018 as a standard test dataset for evaluating case law search. It comprises 2,572 judgments over 12 queries (also known as topics in this dataset) in total. All documents are obtained from American judicial decisions. Different from the legal case retrieval task in this paper, Locke and Zuccon [12] extract the topic from the single question presented to the United States Supreme Court [13]. Specifically, the query in this collection is a question without detailed information. Although Case Law Collection fills the gap in evaluating case law retrieval, it is difficult to popularize on a large scale. Extracting the key issues in a case is also a challenging task since it requires professional knowledge.

In the relevance assessment process, assessors made up of two lawyers and one paralegal annotated the 2,572 documents with four-level relevance, from the least relevant level (*not relevant*) to the most (*on point*). However, the guidance of relevance assessments is too vague to follow, which is hard to apply to constructing a larger-scale dataset.

### 2.2 COLIEE

The Competition on Legal Information Extraction/Entailment (COLIEE) [10] is a well-known competition held annually since 2014 to improve the development of state-of-the-art information retrieval and entailment methods in the legal field. In particular, it involves a legal case retrieval task based on Canadian case law. The corresponding dataset in COLIEE 2020[1] consists of 650 query cases (520 for training, 130 for testing) in total, and each query case has 200 candidate cases. Participants are required to identify relevant cases that can support the query case decision.

The COLIEE dataset promotes the process of relevant case retrieval in the common law system. However, the common law system is different from the Chinese law system in many ways. As illustrated in Section 1, the definition and application of relevant cases vary with law systems. For instance, in COLIEE, relevant cases are identified according to case citations. However, in the Chinese law system, no such citations are included in judgments. Therefore, the construction of the dataset in COLIEE cannot be well applied to the case retrieval task in the Chinese law system.

### 2.3 CAIL

The Chinese AI and Law challenge dataset (CAIL2018) [23] is a large-scale Chinese legal dataset for judgment prediction. It has over 2.6 million criminal cases annotated with 183 criminal law articles and 202 criminal charges. All the criminal documents are collected from China Judgments Online website. In 2019, the Chinese AI and Law 2019 Similar Case Matching dataset (CAIL2019-SCM), which contains 8,964 triplets, was released with its corresponding relevant case matching task. However, there still exists a non-trivial gap between the task definition of CAIL2019-SCM and real needs in legal practice. Specifically, the triplet $(A, B, C)$ composed of three case descriptions is the basic unit of CAIL2019-SCM. In other words, every query case $A$ has only two candidate cases $B$ and $C$. Participants only need to determine which candidate case is more similar to the query case than the other one. Besides, the dataset only consists of documents in three legal fields, i.e., private lending, intellectual property disputes, and maritime affairs. Therefore, the coverage of Chinese criminal law is limited.

## 3 TASK DEFINITION

Given a query case, our task is to retrieve relevant cases from a pool of candidate cases. To be specific, given a query case $q$ and a set of candidate cases $C = \{c_1, c_2, ..., c_M\}, M \in \mathbb{N}^+$, the task is to identify all relevant cases $S = \{r_1, r_2, ..., r_k | r_i \in C \wedge support(r_i, q)\}$, where $support(s_i, q)$ denotes case $r_i$ is a relevant case supporting query case $q$ in at least one aspect.

Statistics of the dataset are shown in Table 1. Among all 107 queries, 77 queries are selected from common cases and 30 are selected from controversial cases which will be introduced in the following section. Each candidate case pool has at least one relevant case with respect to the query.

## 4 DATASET CONSTRUCTION

In this section, we elaborate on the complete process of constructing our dataset. The rest of this section is organized as follows: Section 4.1 illustrates how the corpus is collected and preprocessed, Section 4.2 introduces our query sampling strategy, Section 4.3

[1] https://sites.ualberta.ca/~rabelo/COLIEE2021/COLIEE_2020_summary.pdf

**Table 1: Dataset statistics of LeCaRD.**

| Statistic | Number |
|---|---|
| Total documents | 43823 |
| Total queries | 107 |
| Candidate cases per query | 100 |
| Avg. relevant cases per query | 10.33 |
| Charges of query cases | 20 |

**Table 2: Definitions of keys in the corpus document.**

| Key | Description |
|---|---|
| ajID | Unique case ID |
| ajjbqk | Basic case information |
| cpfxgc | Court analysis |
| pjjg | Judgment |
| qw | Full text |
| writID | Unique document ID |
| writName | Document title |

presents our candidate pooling method, Subsection 4.4 demonstrates how the relevance judgment criteria are developed, and Section 4.5 introduces details of the annotation process and then analyzes the annotation results.

## 4.1 Corpus and Preprocessing

Following Xiao et al. [23], we collect over 46,000 documents as the raw corpus from China judgments Online [2] published by the Supreme People's Court of China. All documents are criminal decisions within 20 years randomly selected from the overall six million documents. The sampled raw corpus are then constructed as various $(key, value)$ pairs, of which the meanings are shown in Table 2. Note that $ajID$ and $writID$ are different because one case can have several documents if †he case has a second trial or retrial. A document has a unique document ID but may share its case ID with other documents.

During preprocessing, we remove documents without $writName$, $ajjbqk$, or other important keys from the raw corpus. Documents with a too long (more than ten pages) or too short (one paragraph) $qw$ are not taken into consideration. We also replace names and other identity information with placeholders. Finally, we collect a corpus containing 43823 documents.

## 4.2 Query

As mentioned in Section 1, the queries in LeCaRD consist of two parts: common queries and controversial queries. In the following two subsections, we introduce our query sampling strategies regarding these two types of queries.

*4.2.1 Common Query.* Previous works [12, 24] ignore the uneven distribution of both charges and difficulties of sampled queries. Models training on these skewed datasets can not retrieve a relevant case comprehensively. To tackle this problem, we count the charge

[2]https://wenshu.court.gov.cn/

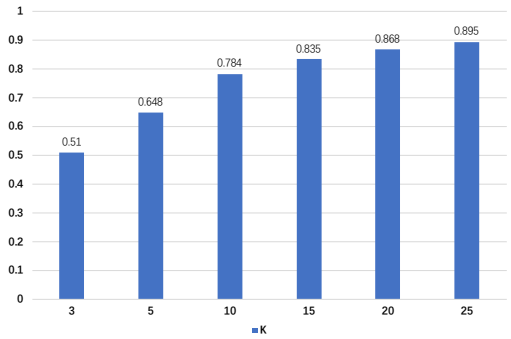


**Figure 2: Ratio of top-K frequent charge to the total charges.**

**Table 3: Four categories of query cases.**

| | Categories | | | |
|---|---|---|---|---|
| **prediction correctness** | correct | | wrong | |
| **prediction entropy** | high | low | high | low |

distribution of all Chinese criminal cases in 20 years. As shown in Fig 2, top-20 frequent charges account for 86.8% of all cases. Considering both case coverage and time consumption, we choose the first part of queries from the top-20 frequent charges.

Query difficulty is also significant to the quality of our first part of queries. We aim to averagely cover the top-20 frequent charges while keeping each charge contains queries with diversified difficulty. To be specific, we first adopted a legal judgment prediction model [15] to predict each case's criminal charges. Based on the predicted probability of each charge, we calculated the prediction entropy of a case by

$$H(c_i) = -\sum_{j=1}^{N} p_{ij} \log p_{ij} \tag{1}$$

, where $H(c_i)$ denotes the entropy of the $i$-th case, $p_{i,j}$ is the predicted probability of $j$-th charge in $i$-th case, and $N$ is the total number of criminal charges[3]. We assume that the higher entropy indicates the lower confidence of the prediction model. In other words, the case is more difficult for judgment if it has a higher entropy. According to the prediction correctness and prediction entropy, we group cases into four categories, as shown in Table 3. The 'correct-high' category means the model predicts the criminal charge correctly but uncertainly, while the 'wrong-low' denotes the model predicts the charge incorrectly but confidently. For each frequent charge, we sample one case in each category as the query case. Theoretically, there are $20 \times 4 = 80$ queries as the first part. Three of these queries are further removed because of their inadequate length or low quality. Therefore the final amount of common case queries is 77.

*4.2.2 Controversial Query.* For the second part of queries, we focus on controversial cases. An apparent and straightforward method is to select cases from second trial or retrial documents where crime judgments were changed from their first trial judgments. In

[3]$N = 272$ in our dataset

hearing this type of case, the judge either makes judgments based on inappropriate articles or has difficulty making a convincing decision. In this paper, we collect the complete set of Chinese second trial and retrial judgments denoted as $R$. Based on this judgment set, we calculate the probability of revising judgments. Suppose $|A|$ is the number of cases convicted of crime $A$ in the first trial, $W_c(A, B)$ the weight of changing charge $A \in S$ to charge $B \in T$ in a specific case $c$, $S$ the set of charges in the first trial of case $c$, $T$ the set of charges in the second trial or retrial of case $c$, and $P_{A \to B}$ the probability of changing charge $A$ into charge $B$. We have:

$$P_{A \to B} = \frac{\sum_{c \in R} W_c(A, B)}{|A|} \tag{2}$$

$$W_c(A, B) = \begin{cases} \frac{1}{|S|} & S \subset T \\ \frac{1}{|S \setminus T|} & S \not\subset T \text{ and } A, B \notin S \cap T \\ 0 & Others \end{cases} \tag{3}$$

Then, we can select controversial case queries of charge $C_0$ in the following way: suppose $C_0$ is changed to other $p$ charges in total, and the sequence $P_{C_0 \to C_1}$, $P_{C_0 \to C_2}$, ... $P_{C_0 \to C_p}$ is in a descending order. Controversial cases are selected from top-q charges, where $q$ is the smallest number that satisfies:

$$\sum_{i=1}^{q} P_{C_0 \to C_i} \geq 0.5 \tag{4}$$

Of all top-20 frequent charges, we sample 30 queries from the second trial and retrial judgment set. Each charge has up to three queries.

## 4.3 Pooling

Arora et al. [1] made a 50-document pool for each query through merging the top 100 retrieved documents using four standard IR models. Similarly, we adopt three retrieval models for pooling: TF-IDF[18], BM25[17], and Language Modeling[14]. Before pooling, we first use THULAC [21] to split Chinese sentences into words. Then, we remove stop words from our corpus according to the Chinese Stop Words List [4]. Finally, each retrieval model retrieves top 100 cases separately from the collected corpus. Unlike Arora et al. [1], we do not simply merge three top 100 cases because this will cause bias to the final pool. Instead, we divided the candidate case pool into three strata [4, 11]:

- **Strata 1**: Cases occurring in the top 100 cases in at least two retrieval models.
- **Strata 2**: Cases occurring in the top 100 cases in only one retrieval models.
- **Strata 3**: Cases not occurring in any of the top 100 cases.

For each query, the candidate pool consists of 30 cases from Strata 1, 30 cases from Strata 2, and 40 cases from Strata 3. If Strata 1 does not have 30 cases, the leftover cases will be sampled from Strata 2. Also, cases from Strata 3 will be added to the candidate pool if Strata 2 does not have enough cases. Cases from Strata 1 or Strata 2 are not limited to the charge of its query, so there are no top-20 charge restrictions to the candidate pool.

[4]https://github.com/yinzm/ChineseStopWords/blob/master/ChineseStopWords.txt

## 4.4 Relevance judgment Criteria

On July 27th, 2020, the Supreme People's Court of China published a guidance document [5] about relevant case retrieval under the Chinese law system. Regarding the definition of relevant case, the document addresses three aspects: application of law, focus of disputes, and basic fact.

Among three aspects, application of law is the theoretical basis of judgment, but can not meet the dataset users' inquiry needs because different circumstances may apply to the same article. Consequently, application of law is not an ideal criterion for relevance assessment. Another aspect, focus of dispute, is the core issue of arguments between the parties to the case. It makes up the main content of judgments and is critical for the judge to summarize the evidence and the application of articles. However, focus of dispute is not adopted in our criteria for two reasons. First, common case law systems are concerned about focus of dispute because they comply with the principle 'no trial without complaint'. In China, however, the comprehensive review of the authority is not restricted by the focus of dispute. Besides, focus of dispute as a relevance criterion is more suitable for scenarios where there is a clear query of the issue, rather than a conviction judgment based on the facts of cases.

As discussed above, we mainly focus on the basic fact. Unlike the guidance document which clearly states the charge of a query case before retrieval, the query in our legal case search scenario only contains fact description without its charge. Assessors need to determine whether the query case constitutes a crime and what crime it constitutes before annotation. Therefore, our relevance judgment criteria do not directly follow the concept of key circumstances illustrated in the guidance document. Instead, we propose new criteria based on the critical factor. Critical factor is directly related to the application of the law and the results of the judgment. It has a substantial impact on the trial of the case. In criminal proceedings, critical factors are sufficient to influence conviction and sentencing by relevant laws and regulations. They determine whether the defendant's action constitutes a charge, what it constitutes, and how severe it is.

The relevance judgment criteria are defined as:

> *Two cases are defined as relevant if the similarity between their critical factors is high.*

where the 'critical factors' consist of key circumstances and key constitutive elements of crime (key elements). Key elements are the legal concept abstraction of key circumstances. Cases without key elements or with different key elements will have a different judgment. The criteria do not require critical factors of two relevant cases to be completely relevant. Two cases with partial relevant critical factors are also considered relevant. An example of a query case is:

> **Query 1:** *... Defendant Zhang XX and Liu XX ran into conflicts with the victim Shi XX. They provoked the victim with excuses and beat the victim Shi XX for no reason, resulting in minor injuries of victim Shi XX. Their behavior disrupted social order ... therefore it constitutes the crime of quarreling and provoking trouble and is a joint crime. ...*

[5]https://www.chinacourt.org/article/detail/2020/07/id/5375599.shtml

In this example, the key circumstances include *'ran into conflicts with the victim Shi XX'*, *'provoked the victim with excuses and beat the victim Shi XX for no reason'*, *'resulting in minor injuries of victim Shi XX'*, and *disrupted social order*. The key elements include 'Beating others at will', 'Causing minor injuries of others', and 'The disruption of social order'. In this example, non-critical factors are the identities of defendants *Zhang*, *Liu* and the victim *Shi*.

Notably, relevant cases may involve different criminal charges. Given Query 1, an example of the candidate case is:

> **Candidate 1:** *... The defendant Chen XX had a grudge against the victim Wang XX. On DD/MM/YYYY, when Wang XX was on his way home from work, Chen XX together with another defendant Li XX beat Wang XX, causing Wang XX to be second-level slightly injured. ...*

where the key circumstances include *'bear Wang XX'* and *'second-level slightly injured'*. The key elements include 'Beating others or intentional harm to others' and 'Causing minor injuries of others'. Therefore, Candidate 1 is assessed relevant to Query 1. However, the defendant Chen XX and Li XX are both convicted of intentional injury, which is different from the charge of Query 1.

On the other hand, cases of the same charge can be assessed as irrelevant because of the difference between critical factors. An example of an irrelevant candidate case is:

> **Candidate 2:** *... In order to vent his emotions, defendant Chen XX, acted aggressively, and forcibly took other people's property. He also arbitrarily damaged other people's belongings worth over 1,000 RMB for more than three times. His behavior violated the 293rd article of the Criminal Law of People's Republic of China, ... therefore it constitutes the crime of quarreling and provoking trouble. ...*

Although Candidate 2 has the same charge as Query 1, the key circumstances of this candidate case are *'forcibly took other people's property'*, *'arbitrarily damaged other people's belongings'*, and *'worth over 1,000 RMB for more than three times'*. The key elements include 'Forcibly taking or arbitrarily destroying or occupying public and private property' and 'Serious crime'. In this example, neither key circumstances nor key elements are similar to those of Query 1. Therefore Candidate 2 is assessed as irrelevant to Query 1.

## 4.5 Annotation and Analysis

Our relevance assessment team contains one expert and nine assessors. The relevance judgment criteria are mainly designed by the expert (Ph.D. in Law). All assessors are masters in Chinese criminal law who are familiar with cases in our dataset. Before annotation, our expert introduced the judgment criteria to assessors. The expert also illustrated how to label example candidate cases to ensure all assessors understand the concepts in the criteria well. All annotation tasks are repeatedly annotated by three different assessors.

According to the relevance judgment criteria in Section 4.4, all $(query, candidate)$ pairs have a four-level relevance label shown in Table 4. Assessors may skip a case if they are not sure about the relevance of the case. After annotation, all cases remained unlabelled will be annotated by the expert or other assessors. The final annotation result is the average value of three annotation results.

**Table 4: Relevance label and corresponding descriptions.**

| Label | Description |
|---|---|
| 1 | Both key facts and key circumstances are relevant. |
| 2 | Key facts are relevant but key circumstances are irrelevant. |
| 3 | Key facts are irrelevant but key circumstances are relevant. |
| 4 | Both key facts and key circumstances are irrelevant. |

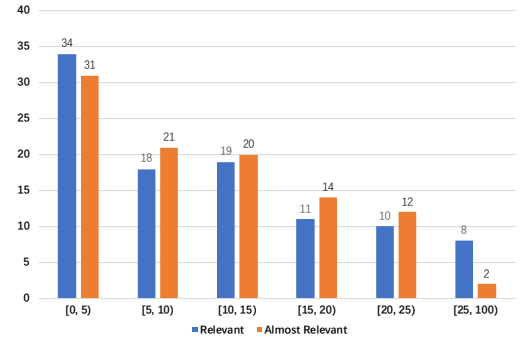


**Figure 3: Distribution of relevant and almost relevant case numbers per query.**

Therefore, we divide nine assessors into three groups. Each group containing three assessors annotates one-third of the candidate cases. As a result, the distribution of relevant cases and almost relevant case numbers per query is shown in Figure 3. The Fleiss's kappa [5, 8] value between three assessors is 0.500, which indicates a moderate agreement ($(0.41, 0.60)$) between three assessors. Notably, among all 107 queries, seven queries do not have any relevant case after annotation. This is because the relevant cases of such queries have almost nothing similar to their queries in the semantic level. Therefore standard retrieval models can not retrieve relevant cases from the corpus. We remove these seven queries from the query set before the experiments in Section 5. In the latest update of LeCaRD, our Ph.D. expert manually retrieve 18 relevant cases from the China Judgments Online website for these queries and add them to our dataset.

# 5 EXPERIMENT

We implement several typical retrieval models on LeCaRD, which can further function as baselines for comparing models in the legal case retrieval task. Two types of retrieval models are involved, i.e., traditional bag-of-words IR models and deep neural models. In particular, we include BM25 [17], LMIR [14], and TF-IDF [18] as traditional bag-of-words IR models, following previous work [19]. As for the neural ranking model, we consider BERT [7] since it has made significant improvements in various NLP tasks and has also been applied to current IR tasks, e.g., ad-hoc retrieval [6].

Traditional retrieval models including BM25, TF-IDF, and LMIR are implemented by the existing package [16], and all parameters are set to default values. As for BERT, we adopt a criminal law-specific BERT, which was pre-trained using 663 million Chinese criminal judgments [25]. We fine-tune our BERT with a sentence pair classification task in an end-to-end fashion. This task input

Table 5: Evaluation of baseline models on different query sets of LeCaRD.

| | Model | P@5 | P@10 | MAP | NDCG@10 | NDCG@20 | NDCG@30 |
|---|---|---|---|---|---|---|---|
| Common query set | BM25 | 0.423 | 0.410 | 0.490 | 0.726 | 0.790 | 0.883 |
| | TF-IDF | 0.348 | 0.305 | 0.480 | **0.789** | **0.830** | 0.847 |
| | LMIR | **0.460** | **0.430** | **0.511** | 0.766 | 0.813 | **0.896** |
| Controversial query set | BM25 | 0.348 | 0.283 | **0.463** | 0.745 | 0.821 | 0.903 |
| | TF-IDF | 0.157 | 0.113 | 0.379 | **0.812** | **0.841** | 0.852 |
| | LMIR | **0.357** | **0.326** | 0.443 | 0.779 | 0.837 | **0.911** |
| Overall query set | BM25 | 0.406 | 0.381 | 0.484 | 0.731 | 0.797 | 0.888 |
| | TF-IDF | 0.304 | 0.261 | 0.457 | **0.795** | **0.832** | 0.848 |
| | LMIR | **0.436** | **0.406** | **0.495** | 0.769 | 0.818 | **0.900** |
| Test set | BM25 | 0.380 | 0.350 | 0.498 | 0.739 | 0.804 | 0.894 |
| | TF-IDF | 0.270 | 0.215 | 0.459 | **0.817** | **0.836** | 0.853 |
| | LMIR | 0.450 | **0.435** | 0.512 | 0.769 | 0.807 | 0.896 |
| | BERT | **0.470** | 0.430 | **0.568** | 0.774 | 0.821 | **0.899** |

contains the query case and its candidate cases. Due to the input length limit of BERT, the task input only includes case descriptions instead of the full text. A [*SEP*] token separates the input text pair, and a [*CLS*] token is added to the end of the input. After BERT outputs hidden state vectors, the first vector is fed into a fully-connected layer for final relevance classification.

Experiments are conducted on different types of query sets. Traditional retrieval baselines mentioned above are mainly evaluated on three types of queries: common queries, controversial queries, and the overall queries (i.e., common + controversial). The fine-tuned BERT is trained on the 80% of the overall queries and evaluated on the rest 20% as a test set. Test set selection details can be found on our project website. To further compare the fine-tuned BERT with retrieval models, we also test the performances of three retrieval baselines on the test set. We utilize precision metrics, including P@5, P@10, Mean Average Precision (MAP), and ranking metrics, including NDCG@10, NDCG@20, and NDCG@30 for evaluation. Particularly, the BERT is fine-tuned on a classification task, and we rank the candidates according to the predicted scores when calculating these evaluation metrics. In detail, the candidates are ranked by sorting $\Delta = pred_1 - pred_0$ in a descending order, where $pred_1$ is the probability of the given candidate to be predicted as 1 and $pred_0$ is the probability to be predicted as 2, 3, or 4. In this way, candidates with higher confidence to be relevant have higher rankings in the retrieved list.

All results are shown in Table 5. Among the bag-of-words IR models, LMIR achieves better performances on the precision metrics, including P@5, P@10, and Mean Average Precision (MAP), while TF-IDF performs poorly on these metrics. This comparison result is also consistent with that on the COLIEE dataset [19]. Furthermore, we find that TF-IDF gives a better ranking, especially on the top of the retrieved result list (e.g., NDCG@10 and NDCG@20).

Comparing among different types of query sets (i.e., Common and Controversial), these three models show various performances. Specifically, according to MAP, while LMIR performs best on the common query set, BM25 performs better on the controversial one. Meanwhile, although these models show better performance on the common query set than on the controversial one when measured with the precision metrics, opposite results are observed when measured with NDCG. These results indicate the differences between common query cases and controversial ones.

The last part of Table 5 compares the performance of BERT with other traditional IR models. As a result, the fine-tuned BERT achieves better performances on a proportion of metrics, especially the precision ones, e.g., P@5 and MAP. However, its improvement is not stable considering various metrics in Table 5. It also suggests that legal case retrieval is a quite challenging IR task. The development of better retrieval models for legal case retrieval is worth future investigating.

## 6 CONCLUSION

In this paper, we present LeCaRD, a legal case retrieval dataset for Chinese law systems. We develop new relevance judgment criteria considering both subjective and objective evaluation. Also, compared with other legal case retrieval datasets that randomly sample queries, we propose a novel query sampling strategy to generate a query set comprising both common queries and controversial queries. Further experiments prove the challenge of our dataset.

In the future, we will continue working on this dataset, including expanding the number of queries and the depth of candidate pools by different methods. The latest information and resources will be updated to our project website https://github.com/myx666/LeCaRD.

## ACKNOWLEDGMENTS

This work is supported by the National Key Research and Development Program of China (2018YFC0831700), Natural Science Foundation of China (Grant No. 61732008, 61532011, 62002194), Beijing Academy of Artificial Intelligence (BAAI), and Tsinghua University Guoqiang Research Institute.

## REFERENCES

[1] Piyush Arora, Murhaf Hossari, Alfredo Maldonado, Clare Conran, and Gareth JF Jones. 2018. Challenges in the development of effective systems for professional legal search. In *ProfS/KG4IR/Data: Search@ SIGIR*.

[2] Trevor Bench-Capon, Michał Araszkiewicz, Kevin Ashley, Katie Atkinson, Floris Bex, Filipe Borges, Daniele Bourcier, Paul Bourgine, Jack G Conrad, Enrico Francesconi, et al. 2012. A history of AI and Law in 50 papers: 25 years of

the international conference on AI and Law. *Artificial Intelligence and Law* 20, 3 (2012), 215–319.

[3] Paheli Bhattacharya, Kripabandhu Ghosh, Saptarshi Ghosh, Arindam Pal, Parth Mehta, Arnab Bhattacharya, and Prasenjit Majumder. 2019. Overview of the FIRE 2019 AILA Track: Artificial Intelligence for Legal Assistance.. In *FIRE (Working Notes)*. 1–12.

[4] WG Cochran. 1977. Double sampling. *Cochran WG. Sampling techniques. 3rd ed. New York: John Wiley & Sons, Inc* (1977), 327–58.

[5] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.

[6] Zhuyun Dai and Jamie Callan. 2019. Deeper text understanding for IR with contextual neural language modeling. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 985–988.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[8] Joseph L Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement* 33, 3 (1973), 613–619.

[9] Hanjo Hamann. 2019. The German Federal Courts Dataset 1950–2019: From Paper Archives to Linked Open Data. *Journal of Empirical Legal Studies* 16, 3 (2019), 671–688.

[10] Yoshinobu Kano, Mi-Young Kim, Masaharu Yoshioka, Yao Lu, Juliano Rabelo, Naoki Kiyota, Randy Goebel, and Ken Satoh. 2018. Coliee-2018: Evaluation of the competition on legal information extraction and entailment. In *JSAI International Symposium on Artificial Intelligence*. Springer, 177–192.

[11] D Lewis. 1996. The TREC-5 filtering track, TREC-5.

[12] Daniel Locke and Guido Zuccon. 2018. A test collection for evaluating legal case law search. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 1261–1264.

[13] Daniel Locke, Guido Zuccon, and Harrisen Scells. 2017. Automatic query generation from legal texts for case law retrieval. In *Asia Information Retrieval Symposium*. Springer, 181–193.

[14] Jay M Ponte and W Bruce Croft. 1998. A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. 275–281.

[15] A Rakhlin. 2016. Convolutional Neural Networks for Sentence Classification. *GitHub* (2016).

[16] Radim Rehurek, Petr Sojka, et al. 2011. Gensim—statistical semantics in python. *Retrieved from genism. org* (2011).

[17] Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at TREC-3. *Nist Special Publication Sp* 109 (1995), 109.

[18] Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management* 24, 5 (1988), 513–523.

[19] Yunqiu Shao, Jiaxin Mao, Yiqun Liu, Weizhi Ma, Ken Satoh, Min Zhang, and Shaoping Ma. [n.d.]. BERT-PLI: Modeling Paragraph-Level Interactions for Legal Case Retrieval.

[20] Olga Shulayeva, Advaith Siddharthan, and Adam Wyner. 2017. Recognizing cited facts and principles in legal judgements. *Artificial Intelligence and Law* 25, 1 (2017), 107–126.

[21] Maosong Sun, Xinxiong Chen, Kaixu Zhang, Zhipeng Guo, and Zhiyuan Liu. 2016. Thulac: An efficient lexical analyzer for chinese.

[22] Marc Van Opijnen and Cristiana Santos. 2017. On the concept of relevance in legal information retrieval. *Artificial Intelligence and Law* 25, 1 (2017), 65–87.

[23] Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, et al. 2018. Cail2018: A large-scale legal dataset for judgment prediction. *arXiv preprint arXiv:1807.02478* (2018).

[24] Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Tianyang Zhang, Xianpei Han, Zhen Hu, Heng Wang, et al. 2019. Cail2019-scm: A dataset of similar case matching in legal domain. *arXiv preprint arXiv:1911.08962* (2019).

[25] Haoxi Zhong, Zhengyan Zhang, Zhiyuan Liu, and Maosong Sun. 2019. *Open Chinese Language Pre-trained Model Zoo*. Technical Report. https://github.com/thunlp/openclap