EQUALS: A Real-world Dataset for Legal Question Answering via Reading Chinese Laws

Andong Chen*
Alibaba Group
Hangzhou, China
ands691119@gmail.com

Feng Yao*
Tsinghua University
Beijing, China
yaof20@mails.tsinghua.edu.cn

Xinyan Zhao
Alibaba Group
Hangzhou, China
zhaoxinyan.zx@alibaba-inc.com

Yating Zhang
Alibaba Group
Hangzhou, China
yatingz89@gmail.com

Changlong Sun
Alibaba Group
Hangzhou, China
changlong.scl@taobao.com

Yun Liu
Tsinghua University
Beijing, China
liuyun89@tsinghua.edu.cn

Weixing Shen
Tsinghua University
Beijing, China
wxshen@tsinghua.edu.cn

ABSTRACT

Legal Question Answering (LQA) is a promising artificial intelligence application with high practical value. A professional and effective legal question answering (QA) agent can assist in reducing the workload of lawyers and judges, and help to achieve judicial accessibility. However, the NLP community lacks a large-scale LQA dataset with high quality, making it difficult to develop practical data-driven LQA agents. To tackle this bottleneck, this work presents EQUALS, a well-annotated real-world dataset for lEgal QUestion Answering via reading Chinese LawS. EQUALS contains 6,914 {question, article, answer} triplets as well as a pool of articles of laws that covers 10 different collections of Chinese Laws. Questions and the corresponding answers in EQUALS are collected from a professional law consultation forum. More importantly, the exact spans of law articles are annotated by senior law students as the answers. In this way, we could assure the quality and professionalism of EQUALS. Furthermore, this work proposes a QA framework that encompasses a law article retrieval module and a machine reading comprehension module for extracting accurate answers from the law article. We conduct thorough experiments with representative baselines on EQUALS, and the results indicate that EQUALS is a challenging question answering task. To the best of our knowledge, EQUALS is the largest real-world LQA dataset which shall significantly promote the research of LQA tasks. The work has been open-sourced and is available at: https://github.com/andongBlue/EQUALS.

*Indicates equal contribution

ICAIL 2023, June 19–23, 2023, Braga, Portugal

ACM ISBN 979-8-4007-0197-9/23/06…$15.00
https://doi.org/10.1145/3594536.3595159

CCS CONCEPTS

• **Information systems** → *Question answering*; *Recommender systems*; • **Computing methodologies** → Language resources.

KEYWORDS

Legal Dataset, Legal Question Answering, Question Answering Framework

ACM Reference Format:
Andong Chen, Feng Yao*, Xinyan Zhao, Yating Zhang, Changlong Sun, Yun Liu, and Weixing Shen. 2023. EQUALS: A Real-world Dataset for Legal Question Answering via Reading Chinese Laws. In *Nineteenth International Conference on Artificial Intelligence and Law (ICAIL 2023), June 19–23, 2023, Braga, Portugal.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3594536.3595159

1 INTRODUCTION

With the progress of artificial intelligence (AI) and data accumulation in the legal domain, tremendous advances have been made in Legal AI [19], such as legal information extraction [6, 23], relevant case retrieval [5, 17], legal judgment prediction [4, 11, 22], and so on. However, in the field of legal question answering (LQA), since legal questions usually require professional answers with domain knowledge, and the scope of legal questions is very broad, less attention has been paid to this field. On the other hand, in fact, there is a huge number of people in need of professional legal knowledge as well as their own legal assistant who is required to effectively answer the legal questions they encountered in daily life. For example, some legal websites offer legal expert resources where users can ask questions and lawyers can answer them[1].

Therefore, developing an LQA AI agent which enables to the provision of professional consulting services to the general public can not only help legal professionals to reduce the heavy and redundant workload but also is beneficial for realizing the vision of equal access to justice for all. Despite the lack of attention in the LQA field, there are still a number of potential benefits that

[1]https://www.lawtime.cn/ and https://answers.justia.com/

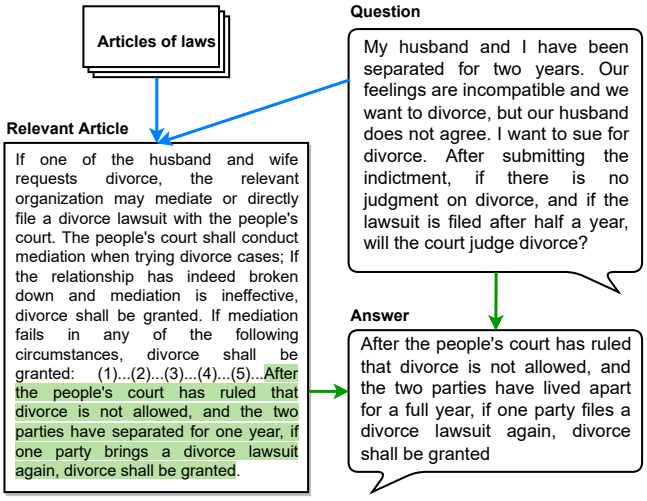

Figure 1: An example in EQUALS. All examples shown in the paper are translated from Chinese for illustration. Limited by space, the article of law does not show all the contents. The blue arrow indicates the retrieval process, and the green arrow indicates the answer extraction process.

can be realized through the development of advanced LQA AI systems. For instance, by automating the process of answering legal questions, LQA AI systems can greatly improve the efficiency and accuracy of legal services. This can not only help to reduce the workload of legal professionals but also make legal services more accessible to the general public. Furthermore, the ability of LQA AI systems to process vast amounts of legal data and knowledge can also make legal information more widely available and easier to understand for non-experts. In conclusion, despite the challenges faced in the field of legal question answering, the potential benefits of LQA AI systems are numerous and far-reaching. By focusing on the development of these systems, we can help to promote equal access to justice and make legal services more efficient, accurate, and accessible to the general public.

High-quality LQA datasets are a crucial prerequisite and cornerstone for advancing the development of the LQA field. In recent years, there have been efforts to construct public LQA datasets, and Table1 presents the statistics of the state-of-the-art LQA-related datasets. COLIEE-2018[8] released a QA dataset in Japanese, constructed from legal examination, which enables the determination of the entailment relationship between a given problem sentence and an article sentence through a legal entailment task. Similarly, Zhong et al.[25] proposed an MRC dataset based on the National Judicial Examination of China. However, these two datasets are derived from judicial examinations, which are far from real-world legal question answering scenarios. BSARD[10] is a valuable resource for conducting Belgian statutory article retrieval tasks. However, the answer granularity of BSARD is relatively coarse since the answer is limited to a single article. In contrast, EQUALS provides more precise answers, with a median answer extraction of 39. Moreover, the median article length retrieved by BSARD was 77, while our median length was 81, indicating that EQUALS articles have a higher length. As such, presenting the entire article as an answer may not be conducive to users obtaining accurate answers, as the answer needs to be more precise. Therefore, while BSARD is a useful resource, it may benefit from expanding the granularity of the answer. Furthermore, not all legal questions can be answered using law articles alone[10]. For example, the question "My friend has never paid me back the money he borrowed, I exposed him on social media. Is it illegal?" is not answerable using prior LQA datasets since these datasets do not filter such questions. These limitations emphasize the necessity of a well-annotated dataset that closely resembles real-world consultation dialogues between legal professionals and the general public in the LQA community.

To overcome these limitations, there is a need for more comprehensive and high-quality LQA datasets that are representative of real-world legal question-answering scenarios[12]. These datasets should not only cover a wide range of legal topics, but also provide fine-grained and detailed answers that can help LQA systems to better understand and respond to legal questions. Moreover, the data in these datasets should be well-annotated, taking into account the context and nuances of each legal question, so as to accurately capture the nuances of the legal domain. One approach to creating a high-quality LQA dataset is to collaborate with legal professionals, who can provide expert annotations and help to ensure that the data in the dataset accurately represents the legal domain. Another approach is to use data from real-world legal consultation dialogues, such as those that occur on legal advice websites or forums. This type of data can help to capture the conversational nature of legal questions and provide a more representative sample of the types of legal questions that people ask in everyday life. In summary, high-quality LQA datasets are crucial for the development of advanced LQA systems, and the creation of these datasets should be a priority for researchers in the field. By focusing on creating datasets that accurately represent real-world legal question-answering scenarios, and by leveraging the expertise of legal professionals, we can help to ensure that LQA systems are able to provide accurate, high-quality answers to legal questions.

To drive the progress of research and application in the field of LQA, in this work, we introduce a large-scale and well-annotated dataset for real-world lEgal QUestion Answering via reading Chinese LawS (EQUALS). EQUALS consists of 6, 914 {question, article, answer} triplets involving 10 Chinese laws and a repository containing 3081 law articles. As illustrated in Figure 1, EQUALS not only provides the question-answer pair but also provides the reference article of the law with the annotated span for the answer. EQUALS has three properties: (1) questions in EQUALS are collected from real-world posts on an online legal question answering community, all raw data were desensitized; (2) we invited senior law students to write answers, and an answer is a fragment of a law article relevant to the question, which will support to build a more precise LQA system; (3) by manual screening, all questions in EQUALS are fit to answer with an article of law. We believe this type of question could be learned by a machine learning model, and the learned model tends to be adopted in real-world applications because the answer is supported by a law article. Besides, we provide two labels (knowledge-driven type and case-analysis type) for questions in EQUALS. In summary, to perform well on EQUALS, the model needs first to find the relevant articles of law from an article pool, then to extract a pinpoint answer from the retrieved top-ranking

Dataset	Data source	Question filtering*	Annotated answer type	Size	Language
COLIEE-2018[8]	Legal exam	✗	Law article	720	Japanese
JEC-QA[25]	Legal exam	✗	Option	26,365	Chinese
BSARD[10]	Legal support service team	✗	Law article	1,108	French
QAS4CQAR[24]	The Chinese building regulations	✗	Span in law article	3,500	Chinese
MBE dataset[3]	Legal exam	✗	Law article	400	English
EQUALS	Legal Q&A community	✓	Span in law article	6,914	Chinese

Table 1: Comparison of recent LQA dataset. *: Filtering questions that can not be answered with an article of law.

articles of law. This pipeline is consistent with the process of a lawyer answering a question of her/his client.

This work provides a detailed illustration of the construction process of EQUALS and the analysis of data statistics. To show the potential and usefulness of the proposed EQUALS corpus, following the classical open-domain question answering systems [1], we utilize BM25 and Sentence-BERT [14] for retrieving relevant law articles, respectively. The answer extraction task is formalized as a span prediction problem and a BERT-based machine reading comprehension (MRC) model is introduced for this sub-task. Experimental results show that EQUALS is a more challenging dataset than similar open-domain QA datasets. The complexity of legal events and the diversity of expressions in questions increase the difficulty for both retrieval tasks and MRC tasks. We believe that EQUALS will be beneficial for the research of intelligent justice and question answering. The work has been open-sourced[2].

2 RELATED WORK

2.1 Legal Question Answering

Legal Question Answering (LQA) is a promising artificial intelligence application. Kien et al. [9] propose a retrieval-based model for retrieving all appropriate law articles for a legal question, their experiments were conducted on a non-public Vietnamese dataset. QAS4CQAR [24] is a retrieval-based QA system for building regulations-related questions. BSARD [10] also propose a retrieval model for answering users' question with the retrieved article of law. Therefore, most current LQA models focus on how to retrieve an appropriate article of law for users. However, the granularity of the answer is too coarse, the information may be not clear. Differing from previous work, we further extract the accurate answer from the retrieved article of law.

To drive the progress of the data-driven LQA approach, some LQA datasets have been proposed. Fawei et al. [3] present a corpus in the form of textual entailment from the question to an answer, which is derived from a USA national bar exam. COLIEE-2018 [8] has a legal entailment sub-task to determine the entailment relationships between a given problem sentence and an article sentence. Zhong et al. [25] propose an MRC dataset based on the National Judicial Examination of China. These datasets are derived from judicial examinations, which are far from the real-world scenario in terms of the form of questions and answers. Recently, Louis and Spanakis presented a Belgian statutory article retrieval dataset BSARD that consists of native legal questions with relevant articles[10]. However, taking an article as the reply is relatively coarse in terms of the granularity of the answer, and the size of BSARD is relatively small. These limitations suggest that the community needs a large-scale dataset that is close to the real-world consultation dialogues between the general public and legal professionals.

2.2 Open-domain Question Answering

This work is also inspired by the open-domain question answering (OpenQA) task [7, 21], which aims to answer a question in the form of natural language based on a large-scale unstructured document repository. The typical pipeline of OpenQA is to first search for the relevant documents as the context, and then to extract an answer from the retrieved document. The document retrieval component is usually implemented with a probabilistic model such as BM25 [16] or deep learning model [14, 18]. The answer extraction component is usually implemented with information extraction techniques such as the machine reading comprehension (MRC) model [2, 20]. Since OpenQA and LQA are very similar in the process of question answering, one could reference OpenQA models when tackling EQUALS.

The EQUALS dataset presents unique challenges for traditional OpenQA models and algorithms due to its structured format and specialized legal language. Legal questions often require a deep understanding of legal concepts and answers may need to be extracted from multiple sources. To address these challenges, specialized models and algorithms must be developed for the legal domain to understand legal concepts and language, extract relevant information, and adapt to changes. This will contribute to the advancement of intelligent justice and legal question answering systems.

3 DATASET CONSTRUCTION

The annotations and extractions in the EQUALS dataset were carried out by a team of 20 senior law students, who have received extensive training in the field. This guarantees a high standard of accuracy and expertise in the annotated data. Furthermore, the quality of the responses has been rigorously evaluated by two quality inspectors to ensure relevance, coherence, and conformity with academic standards. This meticulous process, combined with the user-friendly format of the dataset, makes the EQUALS dataset a valuable resource for legal question answering researchers and practitioners alike. The information contained in the dataset can be leveraged to develop and enhance legal QA systems, as well as to increase the accessibility of legal knowledge to the general public.

[2]https://github.com/andongBlue/EQUALS

Error Type	Original Question	Modified Question
Ambiguous/non-factual entity	A dog lost its owner and was taken in by Xiao Zhang and took good care of. A few days later, the dog owner came to pick it up. Can Xiao Ming ask the dog owner for the feeding fee?	A dog lost its owner and was taken in by Xiao Zhang and took good care of. A few days later, the dog owner came to pick it up. Can **Xiao Zhang** ask the dog owner for the feeding fee?
Incomplete question	Separated for more than three years. Divorce for two years. The child's father has not paid child support.	Separated for more than three years. Divorce for two years. The child's father has not paid child support. **What should I do?**
Multiple questions	How can I cancel an exclusive entrustment agreement? What are the ways to terminate the exclusive entrustment agreement?	How can I cancel an exclusive entrustment agreement?
Missing punctuation	We were together in 1997, no marriage certificate but one child. How can we get a divorce	We were together in 1997, no marriage certificate but one child. How can we get a divorce **?**
Redundant information	Hello lawyer, I would like to ask if a man always likes to beat people, can a woman file for a divorce?	Can a woman file for divorce if a man always likes to beat people?

Table 2: Grammatical error types in raw questions and revision method.

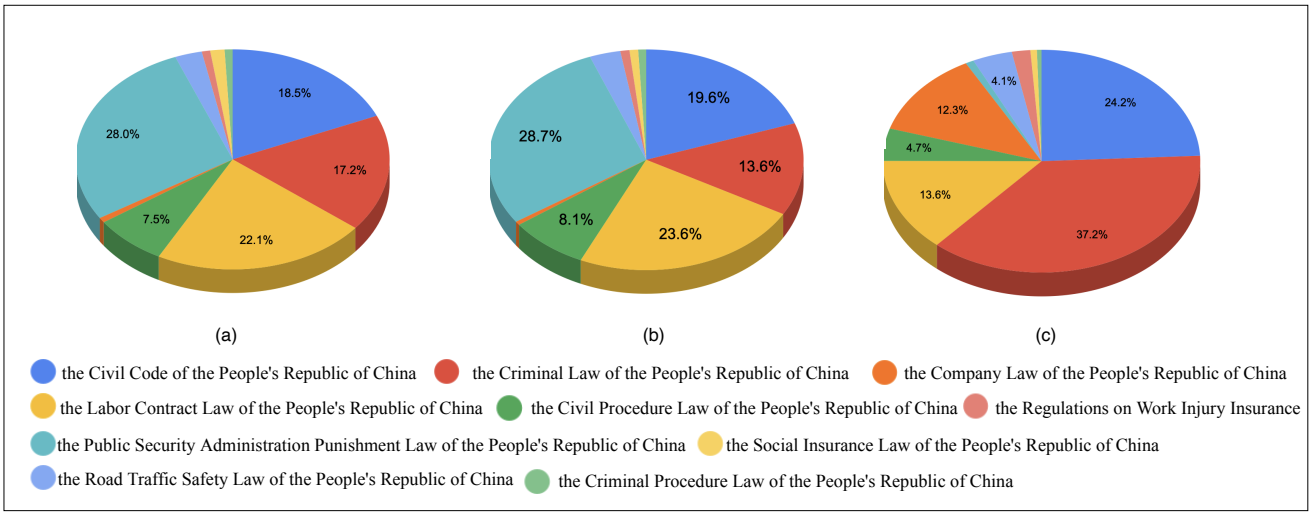

Figure 2: Distribution of the domain of questions in EQUALS: (a) represents the domain distribution of all questions in EQUALS, (b) represents the domain distribution of the case-analysis questions, and (c) represents the domain distribution of the knowledge-driven questions.

3.1 Data Source

3.1.1 Articles of Laws. For the articles of laws, we obtained 10 Chinese laws from PKULAW[3]. The names of each law and the number of law articles within each law can be found in Table 3. In total, we collected 3,081 law articles across all 10 laws, with an average of 308 articles per law. Furthermore, the law with the most articles contains 1,260 articles, which is 1,193 articles more than the law with the least number of articles.

3.1.2 Question-answer Pairs. Lawtime[4], which is a free Chinese legal advice forum that features real user-generated questions. The forum contains over 20 million pieces of legal advice across more than 60 legal fields. Using the 10 Chinese laws previously mentioned, we crawled 43,105 authentic legal consultation questions and ensured that all data is de-identified.

To improve the accuracy of the answers, we collaborated with senior law students who annotated the most relevant law article and the specific portion of the article that serves as the answer to

[3]https://pkulaw.com/

[4]https://www.lawtime.cn/default.php

Law	Number of articles
the Regulations on Work Injury Insurance	67
the Criminal Law of the People's Republic of China	505
the Company Law of the People's Republic of China	218
the Civil Code of the People's Republic of China	1260
the Criminal Procedure Law of the People's Republic of China	308
the Labor Contract Law of the People's Republic of China	98
the Civil Procedure Law of the People's Republic of China	284
the Social Insurance Law of the People's Republic of China	98
the Public Security Administration Punishment Law of the People's Republic of China	119
the Road Traffic Safety Law of the People's Republic of China	124
ALL Laws	3081

Table 3: Statistics of law articles in the retrieval repository.

each question. This annotation process is further explained in the subsequent section, providing a comprehensive understanding of the steps taken to ensure the accuracy of the answers.

3.2 Data Annotation

The objective of this data annotation task is to identify the most relevant law article for each real legal question and extract the precise answer from that article. The quality of the data annotation is ensured by using senior law students to annotate the law articles, thereby ensuring their professionalism and expertise in the legal domain. The dataset and baseline models will be made publicly available for research purposes. This section outlines the specific annotation guidelines, which involve four steps: *Question Discard*, *Question Revision*, *Question Classification*, and *Answer Annotation and Evaluation*.

3.2.1 Questions Discard. In order to guarantee the quality of our dataset, it is necessary to filter out any unqualified legal questions present in the raw data. This is due to the fact that the legal questions in our dataset were generated by real users, leading to the possibility of noise in the data. We initially refined a total of 20,000 questions, and after careful screening, only 6,914 questions were kept for further annotation, while the rest were ultimately discarded. The process of filtering out unqualified questions involves a thorough evaluation by our annotators. They determine if a question is unqualified based on factors such as its unclear intention or its broadness, rendering it unanswerable. For example, questions like "I want to consult about divorce?" and "Why is my contract not legally binding?" are considered unqualified and are therefore removed from the dataset. This screening process allowed us to ensure that only high-quality, well-defined questions remain in the final dataset, making it easier to obtain accurate answers.

3.2.2 Questions Revision. Among the 6,914 questions that were kept for annotation, 4,102 were revised to different extents by our annotators. Additionally, the annotators are required to maintain the original intent and meaning of the question while revising the grammar. The revised question must also be easy to understand and accurately reflect the legal issue at hand. The revised questions are then verified by a senior annotator to ensure their accuracy and clarity. This process ensures that the questions in our dataset are of high quality and suitable for use in building a legal question answering system. Table 2 lists common (but not exhaustive) situations that require revisions by annotators.

3.2.3 Question Classification. Based on the questions that are crawled, we classify them into two types: knowledge-driven questions and case-analysis questions. Knowledge-driven questions are mainly aimed at directly asking about a specific legal concept or law, such as "What is the content of the sales contract?". On the other hand, case-analysis questions are based on legal issues that the users have encountered, as shown in the example in Figure 1.

3.2.4 Answer Annotation and Evaluation. We recruited 20 senior law students to participate in the annotation task. Prior to the task, all annotators received basic training on annotation techniques. The task involved providing annotators with a question and asking them to locate a relevant law article and extract a continuous span from the article as the answer. Once the annotation task was completed, the quality of the answers was evaluated by two quality inspectors, who assessed the answers from two different perspectives.

First, they evaluated the relevance of the law article and the specific portion of the article to the given question. Second, they assessed the completeness and preciseness of the information contained in the extracted span. Any discrepancies between the annotations were discussed and resolved through consensus among the annotators and quality inspectors. This multi-step quality control process helped ensure the high accuracy and reliability of the annotated legal questions and answers in our dataset.

- **Relevance:** The created triple of question, answer, article is logically relevant in the law article.
- **Preciseness:** Extracted answer is *correct* to the question and is the *most compact* continuous span.

Finally, due to the large scale of the data, we conducted a random sampling of 60% of the questions to be annotated twice. The kappa score of 0.87 achieved by the team was satisfactory, considering the complexity of the task. Therefore, we have successfully constructed a well-annotated and high-quality real-world dataset for legal question answering through the examination of Chinese laws. A sample from the dataset is presented in Table 4.

4 DATASET ANALYSIS

Table 5 presents a detailed statistical analysis of the EQUALS dataset. To gain a deeper understanding of the characteristics of the dataset, we also introduce the widely-used Chinese answer span extraction dataset, CMRC[5]. It can be observed that the average length of

[5]https://paperswithcode.com/dataset/cmrc

Question	I want to divorce my husband officially. But now he doesn't agree. What can I do to get a divorce?
Article	If one party in a marriage requests a divorce, the relevant organizations can mediate or directly file a divorce lawsuit with the people's court. The People's Court should mediate in divorce cases, if the relationship has indeed broken down and mediation is ineffective, the divorce should be approved. If one of the following situations occurs, the mediation is ineffective and the divorce should be approved: (1) remarriage or cohabitation with another person; (2) committing domestic violence or abuse, abandoning family members; (3) having persistent vices such as gambling or drug use; (4) living apart for more than two years due to a lack of affection; (5) other situations leading to the breakdown of the marital relationship. If one party is declared missing and the other party files a divorce lawsuit, the divorce should be approved. If the people's court decides not to approve the divorce, and the two parties have been separated for more than one year after the judgment, if one party files a divorce lawsuit again, the divorce should be approved.
Answer	If one party in a marriage requests a divorce, the relevant organizations can mediate or directly file a divorce lawsuit with the people's court. The people's court should mediate in divorce cases and, if the relationship has indeed broken down and mediation is ineffective.
Answer Span	[0, 273]
Question Type	Case-Analysis Question Type

Table 4: An example in EQUALS.

questions and answers in the EQUALS dataset is significantly higher than that of CMRC, which suggests that the EQUALS dataset poses a more challenging task, as we will verify in Section 7.

Following the methodology of SQUAD[13], we have defined the task of the EQUALS dataset as question and answer extraction. We have thus annotated the beginning of the answers in the dataset. Additionally, to assist researchers in understanding the relationship between questions, answers, and law articles, we have also annotated the relevant laws and regulations for each data point. After annotation and data cleaning, the EQUALS dataset consists of a total of 6914 question, answer, article triples. An example of this is presented in Table 4. The percentage of laws in EQUALS is shown in Figure 2. We observed that the three laws that accounted for the largest proportion of total data and case-analysis question type data were: 'The Civil Code of the People's Republic of China, "Labor Contract Law of the People's Republic of China" and 'The Regulations on Work Injury Insurance', respectively Accounted for 27.99%, 22.09%, and 18.51%. However, the top three laws in knowledge-driven question Type data: 'Criminal Law of the People's Republic of China', 'Civil Code of the People's Republic of China' and 'Labor Contract Law of the People's Republic of China', respectively 37.23%, 24.22 and 13.57%.

These findings highlight the diversity of legal areas covered in the EQUALS dataset, as well as the different types of questions asked by users. Additionally, the fact that the top three laws in the case-analysis question type are different from the top three laws in the knowledge-driven question type demonstrates the different nature of these two types of questions and the importance of classifying questions in this way. This information can be used by researchers and practitioners to understand the coverage and distribution of legal knowledge in the dataset, which can inform their selection of data for specific research projects and use cases.

Furthermore, the results of this analysis can also provide insights into the needs and interests of users who are seeking legal information, and inform the development of legal question answering systems. By understanding the most common laws and regulations being referred to in both types of questions, it can help improve the accuracy and efficiency of legal question answering systems in meeting the needs of users. Additionally, the dataset can be used as a benchmark for evaluating the performance of legal question answering systems and for developing new algorithms and models. Overall, the EQUALS dataset and its analysis can contribute to the advancement of legal technology and the development of more effective tools for accessing and understanding the law.

		Question	Answer	Law article
Case	Avg length	35	70	251
	Min length	5	5	195
	Max length	98	261	622
	Count	5853	5853	5853
Concept	Avg length	15	63	262
	Min length	6	3	195
	Max length	60	258	952
	Count	1061	1061	1061
All	Avg length	32	69	252
	Min length	5	3	195
	Max length	98	261	952
	Count	6914	6914	6914
CMRC	Avg length	15	17	452
	Min length	5	1	195
	Max length	89	100	962
	Count	10321	10321	10321

Table 5: The statistics of questions, answers, and law articles in EQUALS.

5 QUESTION ANSWERING FRAMEWORK

Inspired by the OpenQA task, we employed a similar approach to simulate real-world legal question answering tasks in the EQUALS dataset. As depicted in Figure 3, the process involves understanding the questions of real users, identifying the relevant articles of laws, and answering the questions through machine reading comprehension (MRC) model. In this section, we divide the process into two parts: "Retrieval of Relevant Law Article" and "Machine Reading Comprehension Task".

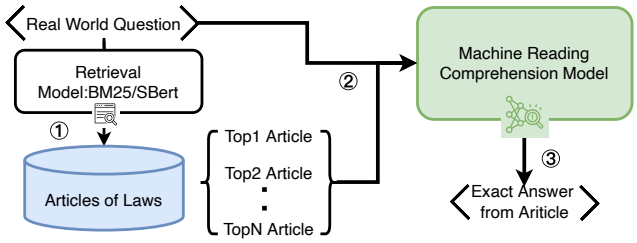

Figure 3: An overview of our question answering system.

5.1 Retrieval of Relevant Law Articles

In this subsection, we simulate the process of identifying legal issues and retrieving relevant law articles. We investigate the current mainstream methods in information retrieval, which can be broadly categorized into traditional methods, such as BM25, and deep learning-based methods, such as Sentence-BERT [15]. Our goal is to evaluate the effectiveness of these methods in retrieving relevant law articles.

5.1.1 BM25. We first employed the traditional information retrieval algorithm, BM25, to calculate the similarity between a legal question and an article of law. The algorithm computes a similarity score using the following formula:

$$\begin{aligned} S &= \sum_{i}^{n} IDF\left(q_i\right) * Relevance \\ &= \sum_{i}^{n} IDF\left(q_i\right) * \frac{f\left(q_i, D\right) * (k1+1)}{f\left(q_i, D\right) + k1 * \left(1 - b + b * \frac{\text{fieldLen}}{\text{avgFieldLen}}\right)} \end{aligned} \tag{1}$$

Among them, $IDF(q_i)$ represents the weight of the i_{th} word in the question, IDF is the inverse document frequency of the i_{th} query term. The second part of the formula represents the relevance of each word in the question to the legal data. In the formula, $k1$ and b are adjustable parameters (the default value is $k1 = 1.2$, $b = 0.75$). Additionally, $f(q_i, D)$ represents the number of occurrences of the i_{th} word in the question within the legal data, while $fieldLen$ and $avgFieldLen$ represent the length and average length of articles in the legal data set, respectively.

5.1.2 Sentence-BERT. In addition, we map the legal question data and articles of laws to hidden vectors and compute their similarity using vector calculations. The Sentence-BERT model demonstrates promising results on common semantic textual similarity datasets. Sentence-BERT incorporates a pooling operation on the output of BERT to derive a fixed-size sentence embedding. The objective function employs a triplet loss approach, which fine-tunes the Sentence-BERT model such that the distance between an anchor legal question (a), a positive article (p), and a negative article (n) is minimized. Mathematically, the following loss function is minimized:

$$Loss = \max\left(\left\|e_a - e_p\right\| - \left\|e_a - e_n\right\| + \epsilon, 0\right) \tag{2}$$

The equation is defining the embeddings for a, p, and n, represented by e_a, e_p, and e_n, respectively. The $|\cdot|$ symbol represents a distance metric, and the margin ϵ ensures that the similarity between e_p and e_a is at least greater than the similarity between e_p and e_n. The distance metric used is Cosine similarity, and the value of ϵ is set to 1.

5.2 Machine Reading Comprehension Task

The subfield of Natural Language Processing (NLP), Machine Reading Comprehension (MRC), aims to advance algorithms that can comprehend and provide answers to questions based on a given text. MRC can be further categorized into two main types: open-domain MRC, where the answer to a question can be derived from any text, and closed-domain MRC, in which the text to be used for answering questions is pre-determined.

The present research endeavors to tackle the challenge of open-domain MRC by answering legal questions using open law articles as inputs. In choosing an appropriate model for this task, the characteristics of the Bidirectional Encoder Representations from Transformers (BERT) model were taken into consideration. BERT comprises multiple bidirectional Transformer layers and has been pre-trained on a massive corpus, enabling it to exhibit state-of-the-art performance on several reading comprehension datasets, including the Stanford Question Answering Dataset (SQUAD). The pre-training process grants BERT a general understanding of language, which can then be fine-tuned to the specific task of legal MRC. In this study, we fine-tune Bert-base-chinese[6] on the EQUALS dataset to develop a machine reading comprehension model specialized in the legal domain.

6 EXPERIMENTS SETTINGS

6.1 Training Details

6.1.1 Retrieval Task of Relevant Law Article: In order to ensure the accuracy and efficiency of the retrieval method, we use both discrete and deep learning-based methods. For the discrete retrieval method, we use the BM25 algorithm from the gensim[7] library and the Chinese word segmentation tool, jieba[8]. To improve the performance of the deep learning-based retrieval method, proper design of training data and parameters is essential. The data splitting for the retrieval task follows the same approach as the Machine Reading Comprehension Task, with specific splitting ratios provided in Section 6.1.2. To achieve this, we use the questions and laws marked by the EQUALS dataset as positive examples and randomly replace the laws with other laws to form negative examples with a positive to negative sample ratio of 1:5.

We utilize the Sentence-BERT model[9], which is encapsulated by the SentenceTransformer tool, for the deep learning-based retrieval method. In the fine-tuning stage, we set the batch size to 16, use the Adam optimizer with a learning rate of 2e-5, and employ a linear learning rate warm-up for 10% of the training data. Additionally, we use the pooling strategy in Sentence-BERT.

To further improve the performance of the model, we also explore the impact of article enrichment on its performance by using a Top_N strategy. Specifically, we concatenate the Top_N articles that are most relevant to the legal question and use them as input for

[6]https://huggingface.co/bert-base-chinese
[7]https://github.com/RaRe-Technologies/gensim
[8]https://github.com/fxsjy/jieba
[9]https://www.sbert.net

	Case-Analysis Type		Knowledge-Driven Type		All EQUALS	
	EM	F1	EM	F1	EM	F1
Gold+MRC	57.21	57.70	62.55	63.07	61.23	62.10
BM25+MRC	31.13	31.96	41.12	41.32	39.21	39.98
Sentence-BERT+MRC	37.26	37.59	43.48	43.19	41.21	42.11

Table 6: Evaluation results on the EQUALS test set. Gold+MRC: using the ground truth law article as the context for MRC model. BM25+MRC: when employing the BM25 retrieval model and MRC model for answer extraction, a Top1 strategy was utilized. Sentence-BERT+MRC: when employing the Sentence-BERT retrieval model and MRC model for answer extraction, a Top1 strategy was utilized

the machine reading comprehension model. In this experiment, we test various values of N, including $Top1$, $Top3$, $Top5$, and $Top10$.

6.1.2 Machine Reading Comprehension Task: For the machine reading comprehension model EQUALS, we divide the data set into a training set (80%) and a test set (20%). During the training of Bert-base-Chinese, we set the batch size to 4, use the Adam optimizer with a learning rate of 2e-5, and limit the maximum answer length to 261. In addition, we use a linear learning rate warm-up for 10% of the training data and employ a dropout rate of 0.1 to prevent overfitting.

6.2 Evaluation Metrics

We used EM and F1 scores as evaluation metrics to measure the EQUALS quality and the performance of the machine reading comprehension framework.

EM (Exact Match): Calculate whether the predicted outcome is an exact match to the standard answer. The formula for calculating EM is as follows:

$$EM = \frac{N_{\text{real}}}{N_{\text{all}}} \tag{3}$$

where N_{real} represents the number of answers predicted that exactly match the true answer. The N_{all} represents the total number of true answers.

F1: Calculate the degree of word-level match between the predicted outcome and the standard answer. The formula for calculating F1 is as follows:

$$P = \frac{N_{\text{overlap}}}{N_{\text{all answer}}} \tag{4}$$

$$R = \frac{N_{\text{overlap}}}{N_{\text{truth answer}}} \tag{5}$$

$$F1 = \frac{2 \times P \times R}{P + R} \tag{6}$$

where $N_{overlap}$ denotes the number of words/characters predicted correctly, i.e. the lexical overlap between the predicted answer and the true answer; $N_{allanswer}$ denotes the number of words/characters of the predicted answer; and $N_{truthanswer}$ denotes the number of words/characters of the true answer.

7 RESULT DISCUSSIONS

7.1 Quantitative analysis

7.1.1 The importance of the effectiveness of law article retrieval. The experimental results are shown in Table 6. The Gold MRC method achieved the best results when using the same data. When utilizing the complete range of test sets for EQUALS, the values of Gold MRC in EM and F1 were 61.32 and 62.10, respectively. Under the evaluation metric of EM, Gold MRC exceeded BM25+MRC and Sentence-BERT+MRC by 22.02 and 20.02 points, respectively, and it also achieved the same result under the F1 metric. Additionally, the results of Sentence-BERT+MRC were better than BM25+MRC, indicating that the retrieval model has a significant impact on the final experimental results and that the traditional BM25 algorithm is less effective than the deep learning method (Sentence-BERT). The reasons why different retrieval methods affect the results will be discussed in further detail in the Case Study section.

It is worth mentioning that while Gold MRC was the best performer in the experiments, it is not the only method that achieved good results. Both BM25+MRC and Sentence-BERT+MRC demonstrated their effectiveness and potential in solving machine reading comprehension problems. The comparison of these three methods shows that the choice of retrieval method is crucial in achieving good results and that the combination of retrieval and machine reading comprehension methods can lead to further improvement in performance. Furthermore, the choice of evaluation metrics also has an impact on the results and a comprehensive evaluation should consider multiple metrics such as EM and F1. In conclusion, the experimental results provide valuable insights into the performance of different machine reading comprehension models and the importance of proper retrieval method design.

7.1.2 Performance comparison of knowledge-driven and other question types in legal domain QA. In table 6, the experimental results for the different question types showed that the knowledge-driven question type performed much better than the other types of data. Using the evaluation metric of EM, the Sentence-BERT+MRC method obtained a score of 43.48 for knowledge-driven question type data, which is 6.22 and 2.27 higher than the case-analysis question type data and the full data of EQUALS, respectively. This suggests that knowledge-driven question type data is relatively simpler to comprehend.

EQUALS		CMRC	
EM	F1	EM	F1
45.47	45.80	64.52	87.10

Table 7: MRC performance of answer extraction results on EQUALS and a widely used general domain Chinese MRC dataset (CMRC). We use the ground truth law article as context for EQUALS.

The results in Figure 4 show that as Top_N increases, the length and richness of the articles also increases, resulting in an improvement in the final EM and F1 experimental results. The impact of the Top_N strategy on the results will be further discussed in the Case Study section of the paper.

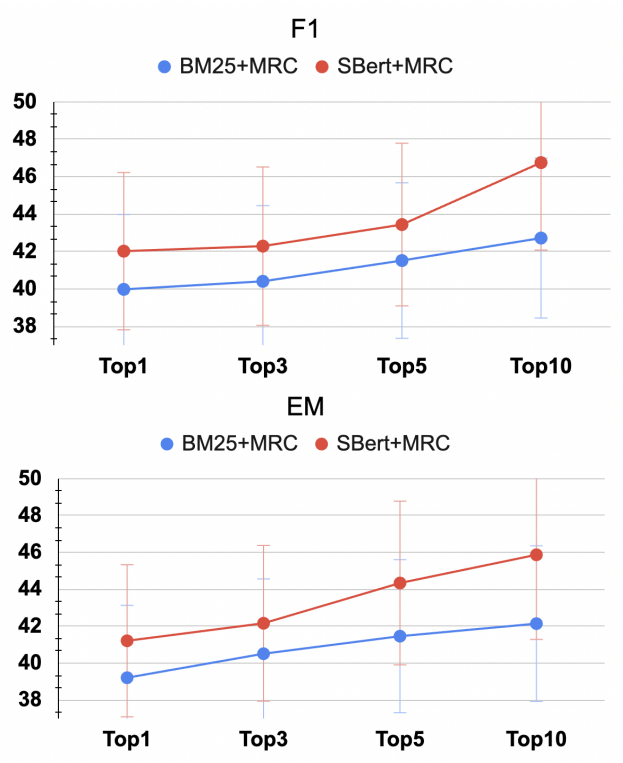

Figure 4: Results of using different numbers of retrieved articles as the MRC context on EQUALS.

In this paragraph, we compared the experimental results of the general domain QA dataset (CMRC) with those of the legal domain dataset, EQUALS. We conducted separate model training on both EQUALS and CMRC datasets and the statistical results of these datasets can be found in Table 7. Our analysis showed that the QA model performed better on the CMRC dataset when compared to the legal domain dataset, EQUALS. These experimental findings confirm that the QA task for the legal domain is indeed more challenging compared to the open-domain QA task, as we observed during our model training on both EQUALS and CMRC datasets.

This difference in performance can be attributed to several factors, such as the complexity of legal language, the lack of sufficient high-quality training data, and the specialized knowledge required to answer legal questions. Despite these challenges, our results indicate that Sentence-BERT+MRC method performs well on the legal domain dataset, and the knowledge-driven question type data is relatively easier to comprehend compared to the other question types. Moreover, the impact of article enrichment through the Top_N strategy on the final results is noteworthy and suggests that this strategy can effectively improve the performance of the model.

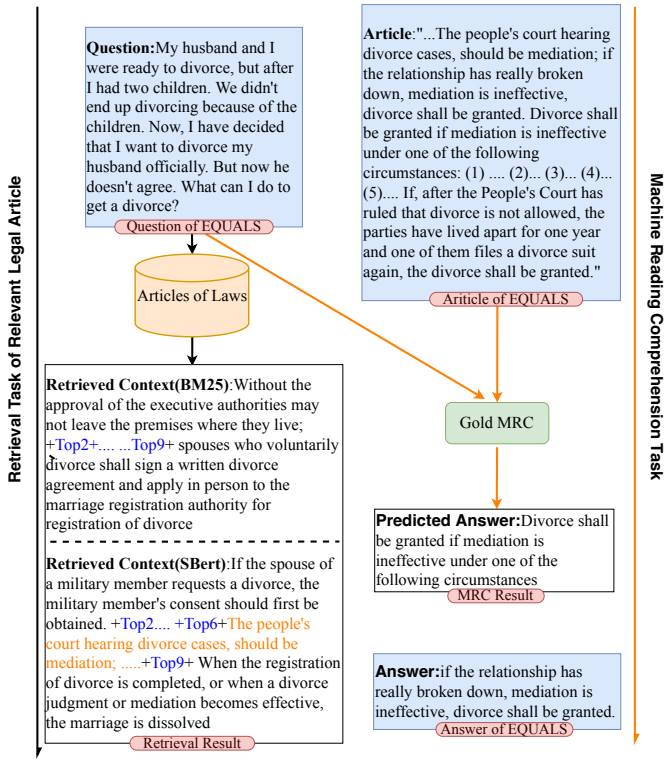

Figure 5: Case Study Example: blue words indicate top10 retrieved results stitched together, and orange words represent the same part as the ground truth law article.

7.2 Case Study

In this section, we aim to intuitively demonstrate the unique characteristics of the EQUALS dataset and the specifics of our proposed Question Answering (QA) framework using an example illustrated in Figure 5.

The `Retrieval Task of Relevant Law Article` was conducted by first selecting a legal question from the EQUALS dataset, then utilizing the BM25 and Sentence-BERT methods to retrieve articles from the `Articles of Laws`. The Top_1 to Top_{10} most relevant articles to the legal question were retrieved using both methods. Our results showed that all articles retrieved (Top1 to Top10) by the Sentence-BERT method were relevant to the topic of marriage, whereas only the top 10 results from the BM25 method were relevant to the same topic. This suggests that the Sentence-BERT method outperforms the BM25 method in domain-specific QA tasks as its retrieval results are more aligned with the topic of the question. Additionally, the top result retrieved by the Sentence-BERT method was not incorrect, and the correct article was found in the top 7 results retrieved using the same method. This implies that the larger the value of N in the Top_N strategy, the higher the probability of finding the correct article.

In the `Machine Reading Comprehension Task`, we selected a question and an annotated article as input for the Gold MRC model on the validation set. As depicted in the right part of Figure 5, the MRC model produced results that were not entirely correct, but

still had some relevance to the annotated answer. This highlights the challenging nature of the EQUALS dataset, as the Gold MRC model trained on annotated data struggles to comprehend the actual question and article, even with a high level of accuracy. This can be attributed to the complexity of the legal domain and the specificity of the language used in the EQUALS dataset.

8 CONCLUSION

The EQUALS dataset is a legal question answering dataset that includes annotated questions, answers, and corresponding legal articles. It was annotated by legal professionals and used to establish a retrieval and machine reading comprehension framework. The dataset's complexity and variability make it an ideal resource for developing end-to-end question answering systems and advancing intelligent justice systems. Future plans for the EQUALS dataset include improving models and algorithms, integrating large language models and knowledge graph methods, and exploring real-world applications with the goal of establishing an accurate and dependable question answering system for legal professionals and laypeople.

ACKNOWLEDGMENTS

This work is supported by National Key RD Program of China (2020YFC0832505, 2022YFC3340904). This work was also supported by Alibaba Group through Alibaba Research Intern Program

REFERENCES

[1] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to Answer Open-Domain Questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*. Association for Computational Linguistics, 1870–1879.

[2] Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2019. A Span-Extraction Dataset for Chinese Machine Reading Comprehension. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, 5882–5888. https://doi.org/10.18653/v1/D19-1600

[3] Biralatei Fawei, Adam Z. Wyner, and Jeff Z. Pan. 2016. Passing a USA National Bar Exam: a First Corpus for Experimentation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*. European Language Resources Association (ELRA).

[4] Yi Feng, Chuanyi Li, and Vincent Ng. 2022. Legal Judgment Prediction via Event Extraction with Constraints. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*. Association for Computational Linguistics, 648–664.

[5] Changzhen Ji, Xin Zhou, Yating Zhang, Xiaozhong Liu, Changlong Sun, Conghui Zhu, and Tiejun Zhao. 2020. Cross Copy Network for Dialogue Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 1900–1910. https://doi.org/10.18653/v1/2020.emnlp-main.149

[6] Donghong Ji, Peng Tao, Hao Fei, and Yafeng Ren. 2020. An end-to-end joint model for evidence information extraction from court record document. *Inf. Process. Manag.* 57, 6 (2020), 102305.

[7] Bert F. Green Jr., Alice K. Wolf, Carol Chomsky, and Kenneth Laughery. 1961. Baseball: an automatic question-answerer. In *Papers presented at the 1961 western joint IRE-AIEE-ACM computer conference, IRE-AIEE-ACM 1961 (Western), Los Angeles, California, USA, May 9-11, 1961*, Walter F. Bauer (Ed.). ACM, 219–224. https://doi.org/10.1145/1460690.1460714

[8] 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 *New Frontiers in Artificial Intelligence - JSAI-isAI 2018 Workshops, JURISIN, AI-Biz, SKL, LENLS, IDAA, Yokohama, Japan, November 12-14, 2018, Revised Selected Papers*, Vol. 11717. Springer, 177–192.

[9] Phi Manh Kien, Ha-Thanh Nguyen, Ngo Xuan Bach, Vu Tran, Minh Le Nguyen, and Tu Minh Phuong. 2020. Answering Legal Questions by Learning Neural Attentive Text Representation. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*. International Committee on Computational Linguistics, 988–998.

[10] Antoine Louis and Gerasimos Spanakis. 2022. A Statutory Article Retrieval Dataset in French. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*. Association for Computational Linguistics, 6789–6803.

[11] Luyao Ma, Yating Zhang, Tianyi Wang, Xiaozhong Liu, Wei Ye, Changlong Sun, and Shikun Zhang. 2021. Legal Judgment Prediction with Multi-Stage Case Representation Learning in the Real Court Setting. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*. ACM, 993–1002.

[12] Jorge Martinez-Gil. 2021. A survey on legal question answering systems. *arXiv preprint arXiv:2110.07333* (2021).

[13] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, 2383–2392. https://doi.org/10.18653/v1/D16-1264

[14] Nils Reimers and Iryna Gurevych. 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, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*. Association for Computational Linguistics, 3980–3990.

[15] Nils Reimers and Iryna Gurevych. 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, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, 3980–3990. https://doi.org/10.18653/v1/D19-1410

[16] Stephen E. Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.* 3, 4 (2009), 333–389.

[17] Yunqiu Shao, Jiaxin Mao, Yiqun Liu, Weizhi Ma, Ken Satoh, Min Zhang, and Shaoping Ma. 2020. BERT-PLI: Modeling Paragraph-Level Interactions for Legal Case Retrieval. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*. ijcai.org, 3501–3507.

[18] Börkur Sigurbjörnsson. 2004. Language Modeling for Information Retrieval. *J. Log. Lang. Inf.* 13, 4 (2004), 531–534.

[19] Changlong Sun, Yating Zhang, Qiong Zhang, and Xiaozhong Liu. 2020. Legal Artificial Intelligence - Have You Lost a Piece from Jigsaw Puzzle?. In *Proceedings of the AAAI 2020 Spring Symposium on Combining Machine Learning and Knowledge Engineering in Practice, AAAI-MAKE 2020, Palo Alto, CA, USA, March 23-25, 2020, Volume I (CEUR Workshop Proceedings, Vol. 2600)*. CEUR-WS.org.

[20] Yuan Sun, Chaofan Chen, Andong Chen, and Xiaobing Zhao. 2021. Tibetan question generation based on sequence to sequence model. *Comput. Mater. Continua* 68, 3 (2021), 3203–3213.

[21] Shuohang Wang, Mo Yu, Xiaoxiao Guo, Zhiguo Wang, Tim Klinger, Wei Zhang, Shiyu Chang, Gerry Tesauro, Bowen Zhou, and Jing Jiang. 2018. R^3: Reinforced Ranker-Reader for Open-Domain Question Answering. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*. AAAI Press, 5981–5988.

[22] Yiquan Wu, Kun Kuang, Yating Zhang, Xiaozhong Liu, Changlong Sun, Jun Xiao, Yueting Zhuang, Luo Si, and Fei Wu. 2020. De-Biased Court's View Generation with Causality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*. Association for Computational Linguistics, 763–780.

[23] Feng Yao, Chaojun Xiao, Xiaozhi Wang, Zhiyuan Liu, Lei Hou, Cunchao Tu, Juanzi Li, Yun Liu, Weixing Shen, and Maosong Sun. 2022. LEVEN: A Large-Scale Chinese Legal Event Detection Dataset. In *Findings of the Association for Computational Linguistics: ACL 2022*. Association for Computational Linguistics, Dublin, Ireland, 183–201. https://doi.org/10.18653/v1/2022.findings-acl.17

[24] Botao Zhong, Wanlei He, Ziwei Huang, Peter E. D. Love, Junqing Tang, and Hanbin Luo. 2020. A building regulation question answering system: A deep learning methodology. *Adv. Eng. Informatics* 46 (2020), 101195. https://doi.org/10.1016/j.aei.2020.101195

[25] Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. JEC-QA: A Legal-Domain Question Answering Dataset. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, 9701–9708.