AI Open 3 (2022) 193–199

Contents lists available at ScienceDirect

AI Open

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

Augmented and challenging datasets with multi-step reasoning and multi-span questions for Chinese judicial reading comprehension

Qingye Meng [a,*], Ziyue Wang [a], Hang Chen [a], Xianzhen Luo [a], Baoxin Wang [a,c], Zhipeng Chen [a], Yiming Cui [a,c], Dayong Wu [a], Zhigang Chen [a], Shijin Wang [a,b]

[a] State Key Laboratory of Cognitive Intelligence, iFLYTEK Research, China
[b] iFLYTEK AI Research (Hebei), LangFang, China
[c] Research Center for SCIR, Harbin Institute of Technology, Harbin, China

ARTICLE INFO

Keywords:
Dataset
Reading comprehension
Legal AI

ABSTRACT

The existing judicial reading comprehension datasets are relatively simple, and the answers to the questions can be obtained through single-step reasoning. However, the content of legal documents in actual scenarios is complex, making it problematic to infer correct results merely by single-step reasoning. To solve this type of issue, we promote the difficulties of questions included in Chinese Judicial Reading Comprehension (CJRC) dataset and propose two augmented versions, CJRC2.0 and CJRC3.0. These datasets are derived from Chinese judicial judgment documents in different fields and annotated by judicial professionals. Compared to CJRC, there are more types of judgment documents in the two datasets, and the questions become are more challenging to answer. For CJRC2.0, we only preserve complex questions that require to be solved by multi-step reasoning. Besides, we provide additional supporting facts to the answers. For CJRC3.0, we introduce a new question type, the multi-span question, which should be answered by extracting and combining multiple spans in the documents. We implement two powerful baselines to evaluate the difficulty of our proposed datasets. Our proposed datasets fill gaps in the field of explainable legal machine reading comprehension.

1. Introduction

With the popularization of legal knowledge, legal tasks have drawn growing attention from academic research and industrial applications than ever. The judgment document is the basis of many legal tasks. It summarizes the background information, the fact description, the court's opinion, the verdicts and the legal basis. In recent years, researchers managed to assist legal tasks with Artificial Intelligence (AI) techniques and proposed a set of AI research tasks in the legal domain, such as the judgment prediction task (Hu et al., 2018; Kang et al., 2019; Xiao et al., 2018; Chen et al., 2019), the similar case retrieval (Kano et al., 2018; Locke and Zuccon, 2018; Xiao et al., 2019; Tran et al., 2019), the legal text summarization task (Merchant and Pande, 2018; Kanapala et al., 2019; Bhattacharya et al., 2019), and the legal information extraction task (Cardellino et al., 2017; Yin et al., 2018; Vacek and Schilder, 2017; Wang et al., 2021).

One of the principle objects of these legal tasks is to understand the context of the input judgment document and automatically make decisions towards predefined targets, such as retrieving similar cases and locating a piece of key information from the input. Some of the tasks can be fulfilled by the information retrieval technology that returns a batch of candidates through semantic matching and statistical analysis (Locke and Zuccon, 2018; Tran et al., 2019). Others can be solved by information extraction (Cardellino et al., 2017; Vacek and Schilder, 2017). This requires manual definition on the types of target information with respect to different cases and crimes. However, these methods depend too much on the handcrafted data and cannot generalize to unseen cases or crimes.

To this end, researchers propose to treat these tasks in the machine reading comprehension (MRC) manner. The MRC task requires the machine to answer questions according to the context of given passages. It can extract fine-grained and unconstrained information, and answers various questions related to the given passages. Following this idea, Duan et al. (2019) proposed a human-annotated benchmark for Chinese judicial reading comprehension task, call CJRC. This benchmark involves two types of judgment documents in CJRC, criminal and civil, and requires model to answer three types of questions corresponding to

* Corresponding author.
E-mail addresses: qymeng5@iflytek.com (Q. Meng), zywang27@iflytek.com (Z. Wang), hangchen3@iflytek.com (H. Chen), xzluo@iflytek.com (X. Luo), bxwang2@iflytek.com (B. Wang), zpchen@iflytek.com (Z. Chen), ymcui@iflytek.com (Y. Cui), dywu2@iflytek.com (D. Wu), zgchen@iflytek.com (Z. Chen), sjwang3@iflytek.com (S. Wang).

https://doi.org/10.1016/j.aiopen.2022.12.001
Received 22 June 2022; Received in revised form 13 November 2022; Accepted 1 December 2022
Available online 5 December 2022

the given documents. However, CJRC has the following shortcomings: (1) the answers to the questions in CJRC are straightforward and can be found in single sentences in the given documents; (2) although being able to infer the correct answers, the benchmark fails to examine if the learnt models are explainable. In practice, many of the questions require complex reasoning, and legal staff expect for a trustworthy system that provides explanations to the predictions.

To address the above issue, we propose two enhanced Chinese judicial reading comprehension datasets CJRC2.0 and CJRC3.0, which are more similar to practical scenarios and are more challenging than CJRC. Following CJRC, we collect judgment documents from the official website, China Judgment Online[1] and hire legal experts to annotate question-answer (QA) pairs. We include a new type of case, the administration, in CJRC2.0 and CJRC3.0. The questions annotated for CJRC2.0 are more complex and require to answer through multi-step reasoning. For each of the answers, experts also provide one or more sentence-level supporting facts that lead to the answer. Supervised by the supporting facts, the learnt model will be able to explain its predictions. CJRC3.0 involves both single-step reasoning and multi-step reasoning. For some challenging questions in CJRC3.0, the answers are composites of multiple spans extracted from the documents. Compared with the CJRC dataset, CJRC2.0 and CJRC3.0 have the following updates:

- We broaden the types of judgment documents for CJRC2.0 and CJRC3.0. Apart from the criminal and the civil cases, the administration cases are included.
- The difficulty of questions is increased by incorporating the multi-step reasoning. All the questions in CJRC2.0 require to be solved through multi-step reasoning instead of single-step reasoning. This enables the trained model to perform better in practical scenarios. For CJRC3.0, we keep both single-step reasoning and multi-step reasoning.
- For CJRC2.0, we provide extra supporting facts to the answers. For each of the answers, there exists one or more sentence-level support facts in the given documents. This provides extra supervision and helps the model to make predictions with reasonable grounds.
- For CJRC3.0, we introduce a new type of QA pairs, the multi-span type. CJRC and CJRC2.0 only contain three types of QA pairs: single-span, YES/NO and unanswerable. For the multi-span questions, the answers to the questions are extracted from multiple inconsecutive segments in the original text.

2. Related work

2.1. Legal reading comprehension dataset

One type of machine reading comprehension task is to ask the machine to answer the corresponding questions according to the context of given passages. It can be divided into four categories: cloze test, multiple choice, span extraction, and free-form answering. Recently, a number of scholars proposed reading comprehension datasets for legal AI tasks. As for the span extraction task, Duan et al. (2019) proposed the CJRC dataset, which is the first Chinese legal reading comprehension dataset, referring to the data format of SQUAD 2.0 (Rajpurkar et al., 2018). CJRC consists of about 50k QA pairs, including three question-and-answer types: single-span extraction, YES/NO, and unanswerable. For the free-form answering task, Zhong et al. (2020) proposed the JEC-QA dataset. The dataset is obtained from China Judicial Exam and the questions are divided into knowledge-driven questions (KD-Questions) and case-analysis questions. In addition, legal reading comprehension datasets for private international law and tax law were proposed by Sovrano et al. (2021) and Holzenberger et al. (2020) respectively. In this paper, we following the idea of CJRC and propose two augmented and more challenging datasets for Chinese legal reading comprehension.

2.2. Methods of legal reading comprehension

Conventionally, rule-based methods (Kim et al., 2013; Kim and Goebel, 2017) are most widely used for legal reading comprehension. The key idea of these methods is to select different features through feature extraction technology, construct and learn a ternary scoring function based on these features. To enhance the model performance, Oanh (Tran et al., 2013) proposed a method based on graph matching, which converts the entire article and query into a graph structure. It also considers the matching degree between the graph structure of the article and the query. Based on this, Fawei et al. (2015) introduced conceptual interpretation to instantiate an ontology relative to concepts and relations. Subsequently, methods based on machine learning and deep learning techniques have gradually become the mainstream, such as SVM (Do et al., 2017), CRF (Bach et al., 2017), CNN (Kim et al., 2015), and BiDAF (Seo et al., 2016). With the proposal of deep pre-trained language models such as BERT (Devlin et al., 2018), researchers use them for the legal reading comprehension task and achieve better performance than the traditional machine learning and deep learning methods (Xiao et al., 2021). In this paper, we also employ pre-trained language models as baselines for our proposed datasets.

3. CJRC2.0 and CJRC3.0

CJRC (Duan et al., 2019) is the first Chinese legal reading comprehension dataset. It contains 5858 criminal judgment documents and 5737 civil judgment documents. These documents are annotated by legal professionals according to unified annotation rules. The final dataset contains a total of 51,333 QA pairs of three question-and-answer types: single-span extraction, YES/NO, and unanswerable. The data format is shown in Fig. 1. CJRC serves as the benchmark for CAIL2019 legal reading comprehension competition . The results of the competition show that the F1 score of the best submitted model is 4.6% higher than BERT baseline, which is significantly improved. However, the score is not as competitive as human performance, which is 9.3% lower.

In order to continuously promote the development of legal AI technology and further improve the performance of the reading comprehension model in Chinese judicial field, we propose two augmented and challenging datasets for Chinese judicial reading comprehension, the CJRC2.0 and CJRC3.0. In this section, we will introduce the setting and construction procedure of these datasets.

3.1. CJRC2.0 dataset

In the CJRC dataset, the answers to single-span questions are obtained by single-step reasoning from the judgment documents. However, in practice, most of the cases and questions are more complex, and the answers produced by single-step reasoning could not solve the corresponding problem properly. Therefore, we design more complicated questions for the CJRC2.0 dataset. In concrete, the answers to single-span and YES/NO questions require inferring through multi-step reasoning. In addition to predicting the answers, the supporting facts that are used to infer the answers are also provided. Fig. 2 shows a sample QA pair and the corresponding support facts in CJRC2.0.

Similar to CJRC, we design the following rules and hire professional judicial personnel to annotate:

1. First of all, we decide whether the text is suitable for labeling. If the content of the text is too simple to label, mark “No”, and skip to the next document. If not, mark as follows.

[1] http://wenshu.court.gov.cn/.

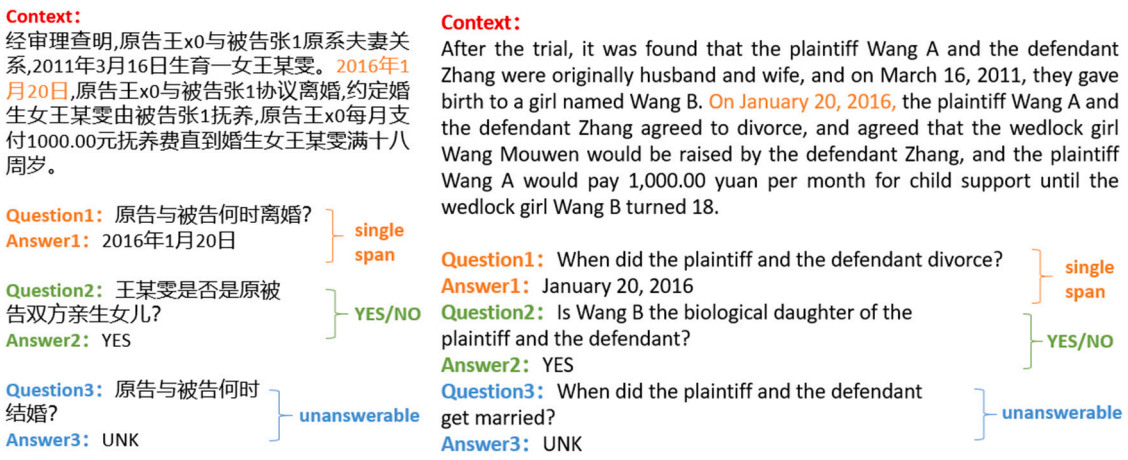

Fig. 1. The demonstrations of three QA types in CJRC.

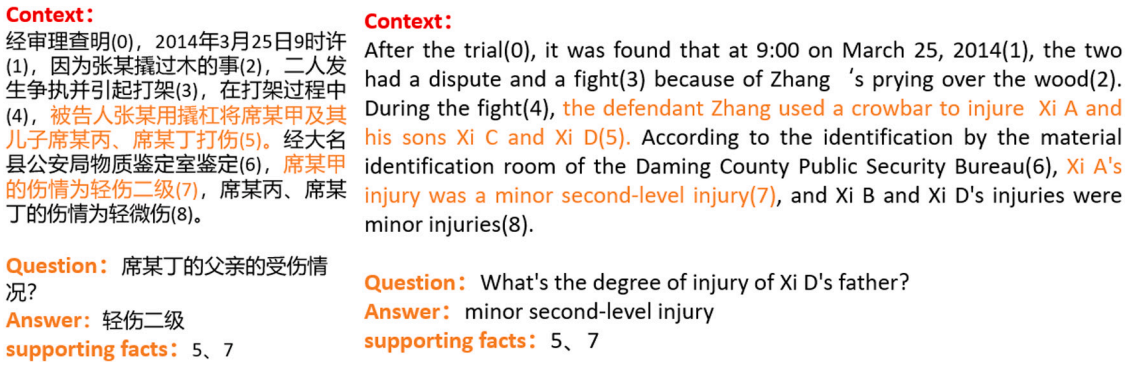

Fig. 2. Examples of the QA pair and supporting facts in CJRC2.0. The order of index of the supporting facts is slightly different due to the translation.

2. For each suitable text, ask a question and give an answer. The answer to span-based questions must meet the following requirements: being a continuous segment from the text; and been given by reasoning over multiple sentences (at least two sentences).
3. In addition to Rule 2, the type of YES/NO questions can be asked. The answer should be annotated as "YES" or "NO".
4. The questions without correct answers are allowed. If we cannot infer a correct answer from the given text, mark the answer to such questions as "UNK". This means the answer does not appear in the text and the question is unanswerable.
5. The indices of sentences used for inferring the answer, i.e., the indices of supporting facts, needs to be provided. The sentences are separated by commas, and the indexing starts from 0. For the unanswerable questions, fill in "-1" as the index of the supporting fact.

We make strict regulations for labeling to ensure the legal professionals answer questions in the same manner which are stated below.

- Ask questions in the field of time, place, person, amount, weight, tools and motivation of the crime.
- Ask more concrete questions such as:
  - Who was more seriously injured?
  - Did the defendant steal more money in the second time?
- The answers should not simply repeat the questions and better to be formed in different words, for example:
  paragraph A: X and Y are married.
  paragraph B: X and Z are mother–child relationships.
  Question: What did **Z's father** ?
- The answer cannot be found directly from the context, it can only be extracted and inducted from the information of multiple sentences, such as:
  paragraph A: The plaintiff applied for government **information disclosure** to the **defendant** by registered mail.
  paragraph B: Defendant C received the plaintiff's information disclosure application.
  Question: In what way did the plaintiff send the **information disclosure** application to **C**?
- If the content of the paragraph are few, containing only one statement, do not question from here.

Moreover, we expand the types of documents in the dataset. The types of judgment documents in CJRC2.0 are expanded to three categories: civil cases, criminal cases, and administrative cases. In total, the CJRC2.0 dataset contains 9532 QA pairs consisting of question, answer, supporting facts, etc. The number of questions of each type and the division of the dataset are shown in Table 1.

3.2. CJRC3.0 dataset

As for CJRC3.0, comparing with CJRC2.0, the difficulty of the problem is increased. We define a new question type, the multi-span type. It requires the answer to be extracted from multiple fragments in the original text. Meanwhile, for the questions of this type, we further divide it into three sub-types:

- **Apparent**: The questions can be split into at least two sub-questions in the literal sense. Keyword "**and**" is always in this type of questions, each sub-question can be answered individually. As shown in Fig. 3, the question, "When the plaintiff and

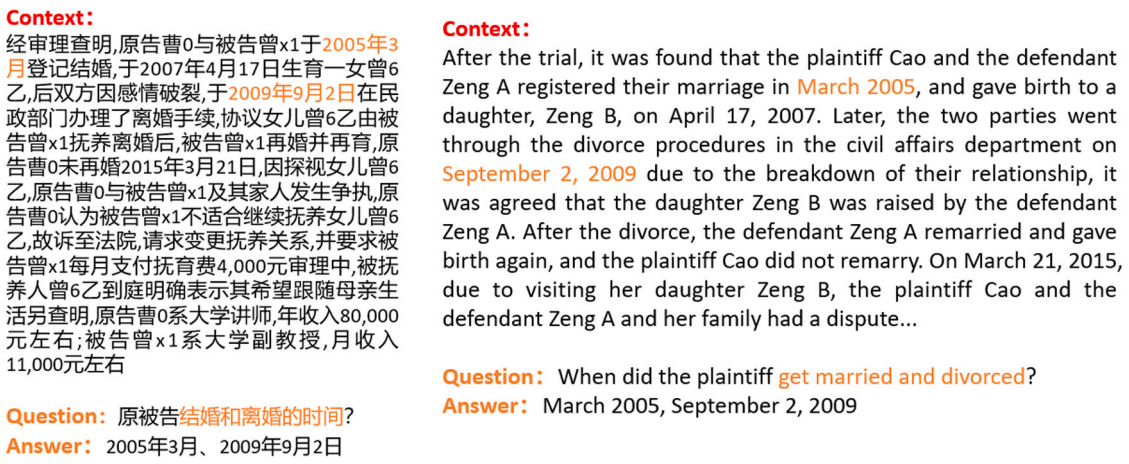

Fig. 3. An example of sub-type *apparent* in CJRC3.0. (The multi-span questions that can be directly split into multiple sub-problems).

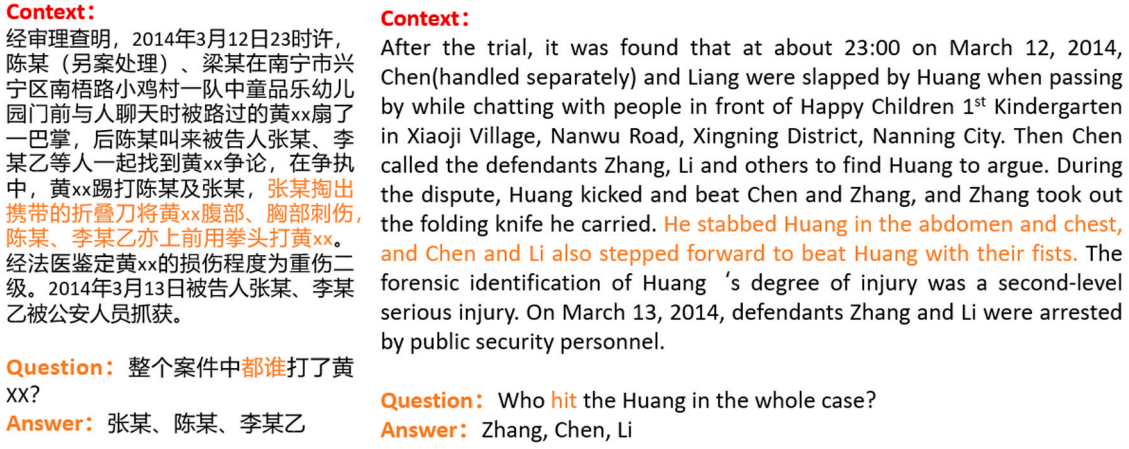

Fig. 4. An example of sub-type *explicit* in CJRC3.0. (The multi-span questions that contain certain keywords but cannot be split into sub-questions literally.)

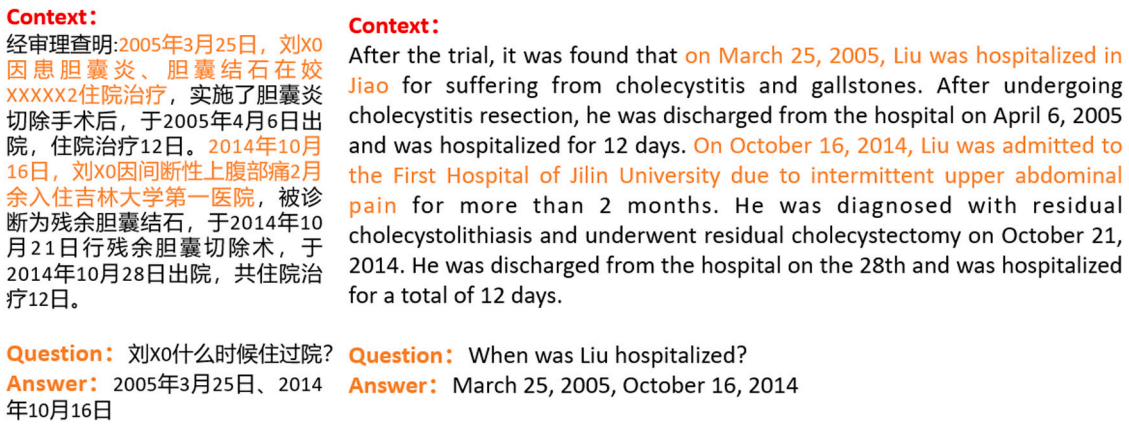

Fig. 5. An example of sub-type *implicit* in CJRC3.0. (The multi-span questions without indicative keywords and cannot be split into sub-questions.) The answer to the question appear in two inconsecutive spans in the content.

defendant were married and divorced?", was rewritten into two sub-questions, "When the plaintiff and defendant were married?" and "When the plaintiff and defendant were divorced?".

- **Explicit**: The question cannot be directly split into multiple sub-questions. Nevertheless, the descriptions of this type of questions contain keywords or key phrases. It indicates that the answers are supposed to appear in multiple spans, such as "respectively" and "which". An example is demonstrated in Fig. 4.
- **Implicit**: The question cannot be split into sub-questions, and the indicative keywords are not presented. The answer can be speculated from several separated fragments in the given document and an example is shown in Fig. 5. It is easy to distinguish the implicit question from the explicit question. First, there are no keyword hints, for example "respectively", in the implicit questions. Second, although forming in a similar way with single-span questions, it can be further validated from the multi-span question if combining the full content from a paragraph.

Table 1
The distribution of different types and the dataset division of CJRC2.0.

	Train	Dev	Test
Single-span	2784	1620	2288
Yes/No	1512	191	189
Unanswerable	758	95	95

Table 2
The statistic of CJRC3.0. To answer single-span, Yes/No and unanswerable questions, the training set of CJRC should be included in addition to the training set of CJRC3.0.

	Train	Dev	Test
Single-span	–	637	637
Multi-span	4200	613	613
Yes/No	–	85	64
Unanswerable	–	151	150

Table 3
The actual ratios of different types/sub-types of questions in CJRC3.0. The ratios are calculated according to 100 randomly sampled data. "Single-step": single-step reasoning; "Multi-step": multi-step reasoning.

Type	Detail type	Actual ratio
Single-span	Single-step	16%
	Multi-step	14%
Multi-span	Apparent	20%
	Explicit	8%
	Implicit	26%
Yes/No	Single-step	6%
	Multi-step	4%
Unanswerable	Unanswerable	6%

Table 4
The consistency ratio of legal professionals and senior legal advisor on CJRC2.0 and CJRC3.0 Datasets. The ratios are calculated according to 100 randomly sampled data.

	Consistency ratio
CJRC2.0	0.92
CJRC3.0	0.95

We design the following rules for the labeling process:

1. First of all, we decide whether the text is suitable for labeling. If the content of the text is too simple to label, mark "No", and skip to the next document. If not, mark as follows.
2. For each suitable text, ask a question and give an answer. The answer can be a continuous segment of the text, or multiple segments.
3. In addition to Rule 2, the type of YES/NO questions can be asked. The answer should be annotated as "YES" or "NO".
4. The questions without correct answers are allowed, which means the answer does not appear in the text and the question is unanswerable. Mark the answer to such questions as "UNK".
5. Single-span questions and YES/NO questions need to include two types: the answer obtained by single-step reasoning and the answer obtained by multi-step reasoning.
6. Multi-span questions need to include three sub-types, *apparent*, *explicit*, and *implicit*. The final ratios of different types of questions are shown in Table 3.

Eventually, we obtain the CJRC3.0 dataset containing the data of judgment documents in the three fields of civil cases, criminal cases, and administrative cases, with a total of 7149 question-and-answer pairs. The number of questions of each type and the division of the dataset are shown in Table 2.

To verify the consistency of the labeling, 100 pieces of data are randomly selected from the CJRC2.0 and CJRC3.0. Subsequently, these data are relabeled by senior legal advisors. The consistency ratio are shown in Table 4.

4. Experiments

4.1. Evaluation metric

We apply different F1 measurements for CJRC2.0 and CJRC3.0 since the prediction targets are different. For CJRC2.0, we jointly calculate F1 scores of the answer and the supporting fact as the final score. While for CJRC3.0, we use the F1 of the answer as the final evaluation metric.

For CJRC2.0, we first calculate the precision $P^{(ans)}$ and the recall $R^{(ans)}$ of the answers. len(·) represents the character length, $gold$ represents the standard answer, $pred$ represents the model prediction result, and InterSec(·) represents the number of overlapping characters. Specific,

$$Lg = \text{len}(gold) \tag{1}$$

$$Lp = \text{len}(pred) \tag{2}$$

$$Lc = \text{InterSec}(gold, pred) \tag{3}$$

$$P^{(ans)} = \frac{Lc}{Lp} \tag{4}$$

$$R^{(ans)} = \frac{Lc}{Lg} \tag{5}$$

Then, we calculate the precision $P^{(sup)}$ and the recall $R^{(sup)}$ of supporting facts as follows:

$$P^{(sup)} = \frac{TP}{TP + FP} \tag{6}$$

$$R^{(sup)} = \frac{TP}{TP + FN} \tag{7}$$

where TP represents the number of correct predictions of supporting facts; FP represents the number of incorrect predictions; and FN is the amount of gold supporting facts that the model fails to predict.

Finally, the Joint F1 is a combination of the precision and the recall of the answer and the supporting fact. Specific,

$$P^{(joint)} = P^{(ans)} P^{(sup)} \tag{8}$$

$$R^{(joint)} = R^{(ans)} R^{(sup)} \tag{9}$$

$$\text{Joint F}_1 = \frac{2P^{(\text{joint})} R^{(\text{joint})}}{P^{(\text{joint})} + R^{(\text{joint})}} \tag{10}$$

For CJRC3.0, we only adopt the F1 score of the answer as the final evaluation metric. For the calculation the multi-span question, we divide the answer into multiple single-span answers. The calculation of $P^{(ans)}$ and $R^{(ans)}$ is shown in formula (1)–(5). The Answer F1 is as follow:

$$\text{Answer F}_1 = \frac{2P^{(\text{ans})} R^{(\text{ans})}}{P^{(\text{ans})} + R^{(\text{ans})}} \tag{11}$$

4.2. Baseline models

We implement two powerful pre-trained language models based on BERT structure: RoBERTa-wwm-ext (Cui et al., 2021) and Chinese ELECTRA. RoBERTa-wwm-ext is a Chinese RoBERTa (Liu et al., 2019) pre-training model using the whole word masking (wwm) technology. The pre-training process involves about 5.4B tokens of data from diverse resources, including Chinese Wikipedia, other encyclopedias, news, etc. We use LTP (Che et al., 2020), a word segmentation tool, to tokenize the data, and mask all Chinese characters that form the same word. In addition, following the ELECTRA model structure (Clark et al., 2020), we pre-train a Chinese-legal-ELECTRA model by using a large amount of Chinese judicial corpora. The discriminator of this pre-trained Chinese-legal-ELECTRA model is employed as the Chinese ELECTRA baseline.

Table 5
Answer F1, supporting facts F1 and Joint F1 of the CJRC2.0 test set.

Model	Dataset	Answer F1	Supporting facts F1	Joint F1
Chinese ELECTRA	CJRC	0.561	0.417	0.296
	CJRC2.0	0.690	0.736	0.549
	CJRC + CJRC2.0	0.718	0.741	0.582
RoBERTa-wwm-ext	CJRC	0.546	0.423	0.292
	CJRC2.0	0.699	0.742	0.561
	CJRC + CJRC2.0	0.725	0.739	0.591

Table 6
Answer F1 of the CJRC3.0 test set.

Model	Dataset	Answer F1
Chinese ELECTRA	CJRC	0.628
	CJRC3.0	0.717
	CJRC + CJRC3.0	0.774
RoBERTa-wwm-ext	CJRC	0.696
	CJRC3.0	0.742
	CJRC + CJRC3.0	0.786

During fine-tuning and inferring, the instance data, containing questions, answers and paragraphs, are converted into a unified input format. This includes input ids, token type ids, the start and the end positions of the answer, and other features. Since the maximum input length of these baseline models is restricted by 512, the input tokens exceed the length will be ignored. We apply a sliding window to divide the article into multiple paragraphs to avoid the over-length problem. Finally, the model learns the probability of the starting and the ending position of the answer or the probability of unanswerable questions. The proposed CJRC2.0 and CJRC3.0 datasets can be merged with the original CJRC dataset to acquire better performance. This results in three different training sets for each baseline: CJRC, CJRC2.0/CJRC3.0, and CJRR+CJRC2.0/CJRC3.0. We adopt the multi-task joint training method for CJRC2.0, which includes three modules: span extraction, answer type classification, and supporting facts discrimination.

For the settings of hyperparameters, we use the base version with 12 layer, 768 hidden and 12 heads. Both models are trained using Tesla V100 32G GPU. The batch size of the RoBERTa-wwm-ext baseline is 2; the learning rate is 1e–5; and the number of training epochs is 10. The batch size of Chinese ELECTRA baseline is 8; the learning rate is 7e–5; and the number of training epochs is 5. For more implementation details and codes of the baseline models, please refer to the websites of judicial reading comprehension task of CAIL2020[2] and CAIL2021.[3]

5. Results and analysis

The experimental results on the test set of CJRC2.0 are shown in Table 5. We report the performances of two baseline models, Chinese ELECTRA and RoBERTa-wwm-ext. In detail, each of the baseline model is fine-tuned on three different training sets, i.e., CJRC, CJRC2.0 and CJRC+CJRC2.0. The CJRC dataset is not annotated with supporting facts. Thus, we depend on the location of predicted spans to evaluate the F1 of supporting facts of models trained on CJRC. We index sentences in the document starting from 0 and take the indices of predicted spans as the predictions for supporting facts. For Chinese ELECTRA, we find that the participation of CJRC2.0 in the training set markedly boosts the performance of the model. To be specific, by comparing the results of training on CJRC and training on CJRC+CJRC2.0, we can conclude that CJRC2.0 improves the F1 scores to a great extent, especially for the F1 of supporting facts. The improvement is +15.7% for F1 of answer predictions, while the improvement is expanded to +32.4% for F1 of supporting fact predictions. Similar trends can be observed for the other baseline, the RoBERTa-wwm-ext, when comparing between the results of CJRC and CJRC+CJRC2.0. However, the trends on the results between baselines are inconsistent when merging CJRC into CJRC2.0. For Chinese ELECTRA, the results are improved just slightly. CJRC+CJRC2.0 improves the results of CJRC2.0 by +2.8% and +0.5% for Answer F1 and Supporting facts F1 respectively. While for RoBERTa-wwm-ext, adding CJRC to CJRC2.0 even impairs the F1 of supporting facts by 0.3%.

The experimental results on the test set of CJRC3.0 are shown in Table 6. We list the results of training on CJRC, CJRC3.0 and CJRC+CJRC3.0. The CJRC3.0 test set includes 4 types of QA pairs, while the training set of CJRC3.0 contains merely the multi-span questions. However, models trained on CJRC3.0 still outperform those trained on CJRC. This indicates that CJRC3.0 provides better supervision than CJRC does under the practical multi-span setting.

6. Conclusion

This paper introduces two Chinese judicial reading comprehension datasets CJRC2.0 and CJRC3.0. These datasets enrich the resource of Chinese judicial datasets and keep challenging the existing models. Our proposed datasets expand the document types, increase the difficulty of corresponding questions, and provide explanations to the model's predictions. It requires multi-step reasoning to answer the questions CJRC2.0 and CJRC3.0 rather than single-step reasoning, and asks for additional supporting facts to the answers. To further increase the difficulty, CJRC3.0 adds a multi-span question type where a question should be answered through at least two different spans from the text. We build two powerful baseline models for these two datasets respectively and show that the model could be markedly improved given more research on CJRC2.0 and CJRC3.0 datasets. We believe that the proposed datasets can boost the model's interpretability in the legal field, and consequently make machine reading comprehension technology more applicable to actual judicial scenarios.

Declaration of competing interest

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

References

Bach, N.X., Thien, T.H.N., Phuong, T.M., et al., 2017. Question analysis for vietnamese legal question answering. In: 2017 9th International Conference on Knowledge and Systems Engineering. KSE, IEEE, pp. 154–159.

Bhattacharya, P., Hiware, K., Rajgaria, S., Pochhi, N., Ghosh, K., Ghosh, S., 2019. A comparative study of summarization algorithms applied to legal case judgments. In: European Conference on Information Retrieval. Springer, pp. 413–428.

Cardellino, C., Teruel, M., Alemany, L., Villata, S., 2017. Legal NERC with ontologies, Wikipedia and curriculum learning. In: 15th European Chapter of the Association for Computational Linguistics (EACL 2017). pp. 254–259.

Che, W., Feng, Y., Qin, L., Liu, T., 2020. N-LTP: A open-source neural Chinese language technology platform with pretrained models. arXiv preprint arXiv:2009.11616.

Chen, H., Cai, D., Dai, W., Dai, Z., Ding, Y., 2019. Charge-based prison term prediction with deep gating network. arXiv preprint arXiv:1908.11521.

Clark, K., Luong, M.-T., Le, Q.V., Manning, C.D., 2020. Electra: Pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:2003.10555.

Cui, Y., Che, W., Liu, T., Qin, B., Yang, Z., 2021. Pre-training with whole word masking for chinese bert. IEEE/ACM Trans. Audio Speech Lang. Process. 29, 3504–3514.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Do, P.-K., Nguyen, H.-T., Tran, C.-X., Nguyen, M.-T., Nguyen, M.-L., 2017. Legal question answering using ranking SVM and deep convolutional neural network. arXiv preprint arXiv:1703.05320.

[2] https://github.com/china-ai-law-challenge/CAIL2020/tree/master/ydlj/baseline.

[3] https://github.com/china-ai-law-challenge/CAIL2021/tree/main/ydlj/baseline.

Duan, X., Wang, B., Wang, Z., Ma, W., Cui, Y., Wu, D., Wang, S., Liu, T., Huo, T., Hu, Z., et al., 2019. Cjrc: A reliable human-annotated benchmark dataset for chinese judicial reading comprehension. In: China National Conference on Chinese Computational Linguistics. Springer, pp. 439–451.

Fawei, B., Wyner, A., Pan, J.Z., Kollingbaum, M., 2015. Using legal ontologies with rules for legal textual entailment. In: AI Approaches to the Complexity of Legal Systems. Springer, pp. 317–324.

Holzenberger, N., Blair-Stanek, A., Van Durme, B., 2020. A dataset for statutory reasoning in tax law entailment and question answering. arXiv preprint arXiv:2005.05257.

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

Kanapala, A., Pal, S., Pamula, R., 2019. Text summarization from legal documents: a survey. Artif. Intell. Rev. 51 (3), 371–402.

Kang, L., Liu, J., Liu, L., Shi, Q., Ye, D., 2019. Creating auxiliary representations from charge definitions for criminal charge prediction. arXiv preprint arXiv:1911.05202.

Kano, Y., Kim, M.-Y., Yoshioka, M., Lu, Y., Rabelo, J., Kiyota, N., Goebel, R., Satoh, K., 2018. Coliee-2018: Evaluation of the competition on legal information extraction and entailment. In: JSAI International Symposium on Artificial Intelligence. Springer, pp. 177–192.

Kim, M.-Y., Goebel, R., 2017. Two-step cascaded textual entailment for legal bar exam question answering. In: Proceedings of the 16th Edition of the International Conference on Articial Intelligence and Law. pp. 283–290.

Kim, M.-Y., Xu, Y., Goebel, R., 2015. A convolutional neural network in legal question answering. In: JURISIN Workshop.

Kim, M.-Y., Xu, Y., Goebel, R., Satoh, K., 2013. Answering yes/no questions in legal bar exams. In: JSAI International Symposium on Artificial Intelligence. Springer, pp. 199–213.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

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

Merchant, K., Pande, Y., 2018. Nlp based latent semantic analysis for legal text summarization. In: 2018 International Conference on Advances in Computing, Communications and Informatics. ICACCI, IEEE, pp. 1803–1807.

Rajpurkar, P., Jia, R., Liang, P., 2018. Know what you don't know: Unanswerable questions for squad. arXiv preprint arXiv:1806.03822.

Seo, M., Kembhavi, A., Farhadi, A., Hajishirzi, H., 2016. Bidirectional attention flow for machine comprehension. arXiv preprint arXiv:1611.01603.

Sovrano, F., Palmirani, M., Distefano, B., Sapienza, S., Vitali, F., 2021. A dataset for evaluating legal question answering on private international law. In: Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law. pp. 230–234.

Tran, O.T., Ngo, B.X., Nguyen, M.L., Shimazu, A., 2013. Answering legal questions by mining reference information. In: JSAI International Symposium on Artificial Intelligence. Springer, pp. 214–229.

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

Vacek, T., Schilder, F., 2017. A sequence approach to case outcome detection. In: Proceedings of the 16th Edition of the International Conference on Articial Intelligence and Law. pp. 209–215.

Wang, B., Wang, Z., Wang, B., Wu, D., Chen, Z., Wang, S., Hu, G., 2021. Various legal factors extraction based on machine reading comprehension. In: Lin, H., Zhang, M., Pang, L. (Eds.), Information Retrieval. Springer International Publishing, Cham, pp. 16–31.

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

Xiao, C., Zhong, H., Guo, Z., Tu, C., Liu, Z., Sun, M., Feng, Y., Han, X., Hu, Z., Wang, H., et al., 2018. Cail2018: A large-scale legal dataset for judgment prediction. arXiv preprint arXiv:1807.02478.

Xiao, C., Zhong, H., Guo, Z., Tu, C., Liu, Z., Sun, M., Zhang, T., Han, X., Hu, Z., Wang, H., et al., 2019. Cail2019-scm: A dataset of similar case matching in legal domain. arXiv preprint arXiv:1911.08962.

Yin, X., Zheng, D., Lu, Z., Liu, R., 2018. Neural entity reasoner for global consistency in ner. arXiv preprint arXiv:1810.00347.

Zhong, H., Xiao, C., Tu, C., Zhang, T., Liu, Z., Sun, M., 2020. Jec-qa: A legal-domain question answering dataset. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 9701–9708.