CJRC: A Reliable Human-Annotated Benchmark DataSet for Chinese Judicial Reading Comprehension

Xingyi Duan[1(✉)], Baoxin Wang[1], Ziyue Wang[1], Wentao Ma[1], Yiming Cui[1,2], Dayong Wu[1], Shijin Wang[1], Ting Liu[2], Tianxiang Huo[3], Zhen Hu[3], Heng Wang[3], and Zhiyuan Liu[4]

[1] Joint Laboratory of HIT and iFLYTEK (HFL), iFLYTEK Research, Beijing, China
{xyduan,bxwang2,zywang27,wtma,ymcui,dywu2,sjwang3}@iflytek.com
[2] Research Center for Social Computing and Information Retrieval (SCIR), Harbin Institute of Technology, Harbin, China
{ymcui,tliu}@ir.hit.edu.cn
[3] China Justice Big Data Institute, Beijing, China
{huotianxiang,huzhen,wangheng}@cjbdi.com
[4] Department of Computer Science and Technology, Tsinghua University, Beijing, China
lzy@tsinghua.edu.cn

Abstract. We present a Chinese judicial reading comprehension (CJRC) dataset which contains approximately 10K documents and almost 50K questions with answers. The documents come from judgment documents and the questions are annotated by law experts. The CJRC dataset can help researchers extract elements by reading comprehension technology. Element extraction is an important task in the legal field. However, it is difficult to predefine the element types completely due to the diversity of document types and causes of action. By contrast, machine reading comprehension technology can quickly extract elements by answering various questions from the long document. We build two strong baseline models based on BERT and BiDAF. The experimental results show that there is enough space for improvement compared to human annotators.

1 Introduction

Law is closely related to people's daily life. Almost every country in the world has laws, and everyone must abide by the law, thereby enjoying rights and fulfilling obligations. Tens of thousands of cases such as traffic accidents, private lending and divorce disputes occurs every day. At the same time, many judgment documents will be formed in the process of handling these cases. The judgment document is usually a summary of the entire case, involving the fact description, the court's opinion, the verdict, etc. The relatively small number of legal staff and the uneven level of judges may lead to wrong judgments. Even the judgments

M. Sun et al. (Eds.): CCL 2019, LNAI 11856, pp. 439–451, 2019.
https://doi.org/10.1007/978-3-030-32381-3_36

Cause of Action	变更抚养关系纠纷
Case Description	经审理查明,原告王x0与被告张1原系夫妻关系,2011年3月16日生育一女王某雯**2016年1月20日**,原告王x0与被告**张1**协议离婚,约定婚生女王某雯由被告张1抚养,原告王x0每月支付**1000.00元**抚养费直到婚生女王某雯满十八周岁另查明,婚生女王某雯随被告张1现居住在**保定市莲池区永华园小区**,被告张1现在保定吉轩商贸有限公司工作
QA Pairs	Q1: 原告与被告何时离婚? A1: 2016年1月20日 Q2: 王某雯是否是原被告双方亲生女儿? A2: YES Q3: 约定王某雯由谁抚养? A3: 张1 Q4: 原告每个月需要支付多少抚养费? A4: 1000.00元 Q5: 王某雯现在住在哪里? A5: 保定市莲池区永华园小区

Fig. 1. An example from the CJRC dataset. Each case contains cause of action (or called charge for criminal cases), context, and some QA pairs where yes/no and unanswerable question types are included.

in similar cases can be very different sometimes. Moreover, a large number of documents make it challenging to extract information from them. Thus, it will be helpful to introduce artificial intelligence to the legal field for helping judges make better decisions and work more effectively.

Currently, researchers have done amounts of work on the field of Chinese legal instruments, involving a wide variety of research aspects. Law prediction [1,20] and charge prediction [8,13,25] have been widely studied, especially, CAIL2018 (Chinese AI and Law challenge, 2018) [22,26] was held to predict the judgment results of legal cases including relevant law articles, charges and prison terms. Some other researches include text summarization for legal documents [11], legal consultation [15,24] and legal entity identification [23]. There also exists some systems for similar cases search, legal documents correction and so on.

Information retrieval usually only returns a batch of documents in a coarse-grained manner. It still takes a lot of effort for the judges to read and extract information from document. Elements extraction often requires pre-defining element types. Different element types need to be defined for different cases or

Table 1. Comparison of CJRC with existing reading comprehension datasets

	Lang	#Que	Domain	Answer type
CNN/Daily Mail	ENG	1.4M	News	Fill in entity
RACE	ENG	870K	English Exam	Multi. choices
NewsQA	ENG	100K	CNN	Span of words
SQuAD	ENG	100K	Wiki	Span of words, Unanswerable
CoQA	ENG	127K	Children's Sto. etc.	Span of words, yes/no, unanswerable
TriviaQA	ENG	40K	Wiki/Web doc	Span/substring of words
HFL-RC	CHN	100K	Fairy/News	Fill in word
DuReader	CHN	200K	Baidu Search/Baidu Zhidao	Manual summary
CJRC	**CHN**	**50K**	**Law**	**Span of words, yes/no, unanswerable**

crimes. Manual definition and labeling processes are time consuming and labor intensive. These two technologies cannot cater for the fine-grained, unconstrained information extraction requirements. By contrast, reading comprehension technology can naturally extract fine-grained and unconstrained information.

In this paper, we present the first Chinese judicial reading comprehension dataset (CJRC). CJRC consists of about 10K documents which are collected from http://wenshu.court.gov.cn/ published by the Supreme People's Court of China. We mainly extract the fact description from the judgment document and ask law experts to annotate four to five question-answer pairs based on the fact. Eventually, our dataset contain around 50K questions with answers. Since some of the questions cannot be directly answered from the fact description, we have asked law experts to annotate some unanswerable and yes/no questions similar to SQuAD2.0 and CoQA datasets (Fig. 1 shows an example). In view of the fact that the civil and criminal judgment documents greatly differ in the fact description, the corresponding types of questions are not the same. This dataset covers the two types of documents and thereby covers most of the judgment documents, involving various types of charge and cause of action (in the following parts, we will use *casename* to refer to civil cases and criminal charges.).

The main contribution of our work can be concluded as follows:

- CJRC is the first Chinese judicial reading comprehension dataset to fill gaps in the field of legal research.
- Our proposed dataset includes a wide range of areas, specifically 188 causes of action and 138 criminal charges. Moreover, the research results obtained through this dataset can be widely applied, such as information retrieval and factor extraction.
- The performance of some powerful baselines indicates there is enough space for improvement compared to human annotators.

Fig. 2. Annotate platform interface

2 Related Work

2.1 Reading Comprehension Datasets

Machine reading comprehension (MRC) has emerged a few datasets for researches. Among these data sets, English reading comprehension datasets occupy a large proportion. Almost each of the mainstream datasets is designed to cater for demands of requiring specific scenes or domains corpus, or to solve one or more certain problems. CNN/Daily mail [7] and NewsQA [21] refer to news field, SQuAD 2.0 [16] focuses on Wikipedia, and RACE [12] concentrates on Chinese middle school students' English reading comprehension examination questions. SQuAD 2.0 [16] mainly introduces the unanswerable questions due to the real situations that we sometimes cannot find a favourable answer according to a given context. CoQA [17] is a large-scale reading comprehension dataset which contains questions that depend on a conversation history. TriviaQA [21] and SQuAD 2.0 [9] pay attention to complex reasoning questions, which means that we need to jointly infer the answers via multiple sentences.

Compared with English datasets, Chinese reading comprehension datasets are quite rare. HFL-RC [3] is the first Chinese Cloze-style reading comprehension dataset, and it is collected from People Daily and Children's Fairy Tale. DuReader [6] is an open-domain Chinese reading comprehension dataset, and it is based on Baidu Search and Baidu Zhidao. Our dataset is the first Chinese judicial reading comprehension dataset, and contains multiple types of

questions. Table 1 compares the above datasets with ours, mainly considering the four dimensions: language, scale of questions, domain, and answer type.

2.2 Reading Comprehension Models

Cloze-style and span-extraction are two of the most widely studied tasks of MRC. Cloze-style models are usually designed as classification models to predict which word has the maximum probability. Generally, models need to encode query and document respectively into a sequence of vectors, where each vector denotes a token's representation. The next operations lead to different methods. Stanford Attentive Reader [2] firstly obtains the query vector, and then exploits it to calculate the attention weights on all the contextual embeddings. The final document representation is computed by the weighted contextual embeddings and is used for the final classification. Some other models [5,10,19] are similar with Stanford Attentive Reader.

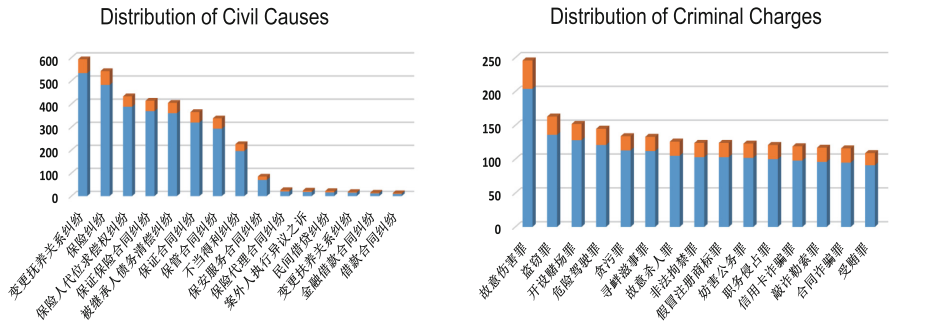
Span-extraction based reading comprehension models are basically consistent in terms of the goal of calculating the start position and the end position. Some classic models are R-Net [14], BiDAF [18], BERT [4], etc. BERT is a powerful pre-trained model and performs well on many NLP tasks. It is worth noting that almost all the top models on the SQuAD 2.0 leaderboard are integrated with BERT. In this paper, we use BERT and BiDAF as two strong baselines. The gap between human and BERT is 15.2%, indicating that models still have enough room for improvement.

3 CJRC: A New Benchmark Dataset

Our legal documents are all collected from China Judgments Online[1]. We select from a batch of judgment documents, obeying the standard that the length of fact description or plaintiff's claim is not less than 150 words, where both of the two parts are extracted with regular rules. We obtain 5858 criminal documents and 5737 civil documents. We build a data annotation platform (Fig. 2) and ask law experts to annotate QA pairs. In the following subsections, we detail how to confirm the training, development, and test sets by several steps.

In-Domain and Out-of-Domain. Referring to CoQA dataset, we divide the dataset into in-domain and out-of-domain. In-domain means that the data type of test data exists in train sets, and conversely, out-of-domain means the absence. Taking into account that *casename* can be regarded as the natural segmentation attribute, we firstly determine which *casenames* should be included in the training set. Then development set and test set should contain *casenames* in the training set and *casenames* not in the training set. Finally, we obtain totally 8000 cases for training set and 1000 cases respectively for development set and

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



(a) civil (b) criminal

Fig. 3. (a) Distribution of the top 15 civil causes. (b) Distribution of the top 15 criminal charges. Blue area denotes the training set and red area denotes the development set. (Color figure online)

test set. For development and test set, the number of cases is the same whether it is divided by civil and criminal, or by in-domain and out-of-domain. The distribution of *casenames* on the training set is shown in Fig. 3.

Annotate Development and Test Sets. After splitting the dataset, we ask annotators to annotate two extra answers for each question of each example in development and test sets. We obtain three standard answers for each question.

Redefine the Task. Through preliminary experiments, we discovered that the distinction between in-domain and out-of-domain is not obvious. It means that performance of the model trained on training set is almost the same regarding in-domain and out-of-domain, and it is even likely that the latter works better. The possible reasons are as follows:

- *Casenames* inside and outside the domain are similar. In other words, the corresponding cases show some similar case issues. For example, two cases related to the contract, housing sales contract disputes and house lease contract disputes, may involve same issues such as housing agency or housing quality.
- Questions about time, place, etc. are more common. Moreover, due to the existence of the "similar *casenames*" phenomenon, the corresponding questions would also be similar.

However, as we all known, there are remarkable differences between civil and criminal cases. As mentioned in the module "**In-domain and out-of-domain**", the corpus would be divided by domain or type of cases (civil and criminal). Although we no longer consider the division of in-domain and out-of-domain, it would also make sense to train a model to perform well on both civil and criminal data.

Table 2. Dataset statistics of CJRC

	Civil	Criminal	Total
Train			
Total Cases	4000	4000	8000
Total *Casenames*	126	53	179
Total Questions	19333	20000	40000
Total Unanswerable Questions	617	617	1901
Total Yes/No Questions	3015	2093	5108
Development			
Total Cases	500	500	1000
Total *Casenames*	188	138	326
Total Questions	3000	3000	6000
Total Unanswerable Questions	685	561	1246
Total Yes/No Questions	404	251	655
Test			
Total Cases	500	500	1000
Total *Casenames*	188	138	326
Total Questions	3000	3000	6000
Total Unanswerable Questions	685	577	1262
Total Yes/No Questions	392	245	637

Adjust Data Distribution. Through preliminary experiments, we also discovered that the unanswerable questions are more challenging than the other two types of questions. To increase the difficulty of the dataset, we have increased the number of unanswerable questions in development set and test set. Related experiments will be presented in the experimental section.

Via the processing of the above steps, we get the final data. Statistics of the data are shown in Table 2. The subsequent experiments will be performed on the final data.

4 Experiments

4.1 Evaluation Metric

We use macro-average F1 as our evaluation metric which is consistent with the CoQA competition. For each question, n F1 scores need to be calculated with n standard human answers, and the maximum value is taken as its F1 score. However, in assessing human performance, each standard answer needs to be compared to $n-1$ other standard answers to calculate the F1 score. In order to compare human indicators more fairly, n standard answers need to be divided into n groups, where each group contains $n-1$ answers. Finally, the F1 score

Table 3. Experimental results

	Civil	Criminal	Overall
Human	94.9	92.7	93.8
BiDAF	61.1	62.7	61.9
BERT	80.1	77.2	78.6

Table 4. Experimental results of in-domain and out-of-domain on development set and test set

Method	Development			Test		
	Civil	Criminal	Overall	Civil	Criminal	Overall
In-Domain	82.1	78.6	80.3	84.7	80.2	82.5
Out-of-Domain	**82.3**	**83.9**	**83.1**	80.9	**82.9**	81.9

of each question is the average of the n groups' F1. The F1 score of the entire dataset is the average of all questions' F1. The formula is as follow:

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

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

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

$$precision = \frac{Lc}{Lp} \tag{4}$$

$$recall = \frac{Lc}{Lg} \tag{5}$$

$$f1(gold, pred) = \frac{2 * precision * recall}{precision + recall} \tag{6}$$

$$Avef1 = \frac{\sum_{i=0}^{Count_{ref}} (max(f1(gold_{\rightarrow i}, pred))}{Count_{ref}} \tag{7}$$

$$F1_{macro} = \frac{\sum_{i=1}^{N} (Avef1_i)}{N} \tag{8}$$

Where *gold* denotes standard answers, *pred* denotes answers predicted by models, *len* means to calculate length, $InterSec$ means to calculate the number of overlap chars. $Count_{ref}$ represents the total references, $\rightarrow i$ represents that the predicted answer is compared to all standard answers except the current one in a single group described as above.

4.2 Baselines

We implement and evaluate two powerful and typical model architectures: BiDAF proposed by [18] and BERT proposed by [4]. Both of the two models

are designed to deal with these three types of questions. These two models learn to predict the probability which is used to judge whether the question is unanswerable. In addition to the way of dealing with unanswerable questions, we concatenate [YES] and [NO] as two tokens with the context for BERT, and concatenate "KYN" as three chars with the context for BiDAF where 'K' denoting "Unknown" means cannot answer the question according to the context. Taking BiDAF for example, during the prediction stage, if start index is equal to 1, then model outputs "YES", and if it is equal to 2, then model outputs "NO".

Some other implementation details: for BERT, we choose the Bert-Base Chinese pre-trained model[2], and then fine-tuning on it with our train data. It is trained on Tesla P30G24, and batch size is set to 8, max sequence length is set to 512, number of epoch is set to 2. For BiDAF, we remove the char embedding, and split string into a sequence of chars, which roles as word in English, like "2 0 1 9 年 5 月 3 0 日". We set embedding size to 300, and other parameters follow the setting in [4].

4.3 Result and Analysis

Experimental results on test set are shown in Table 3. From this table, it is obvious that BERT is 14.5–19% points higher than BiDAF, and Human performance is 14.8–15.5% points higher that BERT. This implies that models could be improved markedly in future research.

Experimental Effect of In-Domain and Out-of-Domain. In this section, we mainly explain why we no loner consider the division of in-domain and out-of-domain described in Sect. 2. We adopts the dataset before adjusting data distribution and select BERT model to verify. Notice that we only train data belong to civil for "Civil", train data belong to criminal for "Criminal", and train all data for "Overall". And type of cases on development set and test set is corresponding to the training corpus. It can be seen from Table 4 that the F1 score of out-of-domain is even higher than that of in-domain, which obviously does not meet the expected result of setting in-domain and out-of-domain.

Comparisons of Different Types of Questions. Table 5 presents fine-grained results of models and humans on the development set and test set, where both of the two sets are not adjusted. We observe that humans maintain high consistency on all types of questions, especially on the "YES" questions. The human agreement on criminal data is lower than that on civil data. This is partly because that we firstly annotate the criminal data, and then have more experience when marking the civil data. It could result in a more consistent granularity of the selected segments on the "Span" questions.

Among the different question types, unanswerable questions are the hardest, and "No" questions are second. We analyze why the performance of unanswer-

[2] https://github.com/google-research/bert.

Table 5. Comparisons of different types of questions.

	Bert			BiDAF			Human		
	Civil	Criminal	Overall	Civil	Criminal	Overall	Civil	Criminal	Overall
	Development								
Unanswerable	69.5	63.3	68.0	7.6	11.4	8.5	92.0	87.1	90.8
YES	91.7	93.2	92.4	83.5	91.2	86.9	96.9	96.2	96.6
NO	78.0	59.0	73.2	57.9	44.9	54.6	94.2	87.8	92.6
Span	84.8	81.8	83.2	80.1	76.0	77.9	91.6	88.4	89.9
	Test								
Unanswerable	67.7	65.6	67.1	10.6	16.0	12.2	91.5	87.7	90.4
YES	91.8	95.6	93.4	77.3	92.8	83.7	97.3	96.5	96.9
NO	72.9	69.7	71.8	47.8	43.3	46.3	96.3	92.5	95.0
Span	84.3	82.4	83.3	79.1	76.2	77.6	93.5	90.9	92.2

Table 6. Comparison data of unanswerable questions and "NO" questions, where unanswerable+ denotes adding extra unanswerable questions on the training set of the civil data.

	Number of Questions (Training set)		Number of Questions (Test set)		Performance (Test set)	
	Civil	Criminal	Civil	Criminal	Civil	Criminal
Unanswerable	617	617	186	77	67.7	65.6
NO	1058	485	134	67	72.9	69.7
Unanswerable+	1284	617	186	77	77.3	67.1
NO	1058	485	134	67	81.6	71.1

able questions is the lowest, and conclude two possible causes: (1) the total number of unanswerable questions on the training set is few; (2) the unanswerable questions are more troublesome than the others.

It is easy to verify the first cause via observing the corpus. To verify the second point, we compare the unanswerable questions and the "NO" questions. Table 6 shows some comparison data of the two types of questions. The first two rows show that unanswerable questions presents a lower performance than the other on the criminal data, even though the former owns more questions. This has basically illustrated that the unanswerable questions are more hard. We have further experimented with increasing the number of unanswerable questions of civil data on the training set. The last two rows in Table 6 demonstrates that increasing unanswerable questions' quantity has an significant impact on performance. However, despite having a larger amount of questions for unanswerable questions, it presents a lower score than "NO" questions.

The above experiments could explain that the unanswerable questions are more challenging than other types of questions. To increase the difficulty of the corpus, we adjusts data distribution through controlling the number of unanswer-

Table 7. Influence of unanswerable questions. Implement BERT and BiDAF on development set and test set. +Train stands for increasing the number of unanswerable questions on the training set. −Dev-Test means no adjusting the number of unanswerable questions on the development set and the test set.

	Bert			BiDAF		
	Civil	Criminal	Overall	Civil	Criminal	Overall
	Development					
Human (before adjust)	92.3	89.0	90.7	–	–	–
Human (after adjust)	93.6	90.8	92.2	–	–	–
CJRC+Train	83.7	77.3	80.5	63.3	62.5	62.9
CJRC−Dev-Test	84.0	81.8	82.9	73.7	75.0	74.3
CJRC+Train−Dev-Test	84.8	81.7	83.3	73.8	74.9	74.4
CJRC	82.0	76.4	79.2	62.8	63.1	63.0
	Test					
Human (before adjust)	93.9	91.3	92.6	–	–	–
Human (after adjust)	94.9	92.7	93.8	–	–	–
CJRC+Train	82.3	77.9	80.1	61.3	61.9	61.6
CJRC−Dev-Test	83.2	82.5	82.8	72.2	74.6	73.4
CJRC+Train−Dev-Test	84.5	82.1	83.3	72.6	74.0	73.3
CJRC	80.1	77.2	78.6	61.1	62.7	61.9

able questions. The following section would show details about the influence of unanswerable questions.

Influence of Unanswerable Questions. In this section, we mainly discuss the impact of the number of unanswerable questions on the difficulty of the entire dataset. **CJRC** represents that we only increase the number of unanswerable answers on the development and the test set without changes on the training set. **CJRC+Train** stands for adjusting all the datasets. **CJRC−Dev-Test** means no adjusting any of the datasets. **CJRC+Train−Dev-Test** means only increasing the number of unanswerable questions of the training set. From Table 7, we can observe the following phenomenon:

- Increasing the number of unanswerable questions in development and test sets can effectively increase the difficulty of the dataset. In terms of BERT, before adjustment, the gap with human indicator is 9.8%, but after adjustment, the gap increases to 15.2%.
- By comparing CJRC+Train and CJRC (or comparing CJRC+Train−Dev-Test and CJRC−Dev-Test), we can conclude that BiDAF cannot handle unanswerable questions effectively.
- Increasing the proportion of unanswerable questions in development and test sets is more effective in increasing the difficulty of the dataset, compared with

reducing the number of unanswerable questions of the training set (get the conclusion by observing CJRC, CJRC+Train and CJRC−Dev-Test).

5 Conclusion

In this paper, we construct a benchmark dataset named CJRC (Chinese Judicial Reading Comprehension). CJRC is the first Chinese judicial reading comprehension, and could fill gaps in the field of legal research. In terms of the types of questions, it involves three types of questions, namely span-extraction, YES/NO and unanswerable questions. In terms of the types of cases, it contains civil data and criminal data, where various of criminal charges and civil causes are included. We hope that researches on the dataset could improve the efficiency of judges' work. Integrating Machine reading comprehension with Information extraction or information retrieval would produce great practical value. We describe in detail the construction process of the dataset, which aims to prove that the dataset is reliable and valuable. Experimental results illustrate that there is still enough space for improvement on this dataset.

Acknowledgements. This work is supported by the National Key R&D Program of China under Grant No. 2018YFC0832103.

References

1. Fawei, B., Pan, J.Z., Kollingbaum, M., Wyner, A.Z.: A methodology for a criminal law and procedure ontology for legal question answering. In: Ichise, R., Lecue, F., Kawamura, T., Zhao, D., Muggleton, S., Kozaki, K. (eds.) JIST 2018. LNCS, vol. 11341, pp. 198–214. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-04284-4_14
2. Chen, D., Bolton, J., Manning, C.D.: A thorough examination of the CNN/daily mail reading comprehension task. CoRR abs/1606.02858 (2016). http://arxiv.org/abs/1606.02858
3. Cui, Y., Liu, T., Chen, Z., Wang, S., Hu, G.: Consensus attention-based neural networks for Chinese reading comprehension, July 2016
4. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. CoRR abs/1810.04805 (2018). http://arxiv.org/abs/1810.04805
5. Dhingra, B., Liu, H., Cohen, W.W., Salakhutdinov, R.: Gated-attention readers for text comprehension. CoRR abs/1606.01549 (2016). http://arxiv.org/abs/1606.01549
6. He, W., et al.: DuReader: a Chinese machine reading comprehension dataset from real-world applications. CoRR abs/1711.05073 (2017). http://arxiv.org/abs/1711.05073
7. Hill, F., Bordes, A., Chopra, S., Weston, J.: The goldilocks principle: reading children's books with explicit memory representations, November 2015
8. Hu, Z., Li, X., Tu, C., Liu, Z., Sun, M.: Few-shot charge prediction with discriminative legal attributes. In: Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, 20–26 August 2018, pp. 487–498 (2018). https://aclanthology.info/papers/C18-1041/c18-1041

9. Joshi, M., Choi, E., Weld, D.S., Zettlemoyer, L.: TriviaQA: a large scale distantly supervised challenge dataset for reading comprehension. CoRR abs/1705.03551 (2017). http://arxiv.org/abs/1705.03551
10. Kadlec, R., Schmid, M., Bajgar, O., Kleindienst, J.: Text understanding with the attention sum reader network. CoRR abs/1603.01547 (2016). http://arxiv.org/abs/1603.01547
11. Kanapala, A., Pal, S., Pamula, R.: Text summarization from legal documents: a survey. Artif. Intell. Rev. 1–32 (2017)
12. Lai, G., Xie, Q., Liu, H., Yang, Y., Hovy, E.H.: Race: large-scale reading comprehension dataset from examinations. CoRR abs/1704.04683 (2017). http://arxiv.org/abs/1704.04683
13. Luo, B., Feng, Y., Xu, J., Zhang, X., Zhao, D.: Learning to predict charges for criminal cases with legal basis. In: Proceedings of EMNLP (2017)
14. Natural Language Computing Group, Microsoft Research Asia: R-net: machine reading comprehension with self-matching networks. In: Proceedings of ACL (2017)
15. Quaresma, P., Rodrigues, I.P.: A question answer system for legal information retrieval (2005)
16. Rajpurkar, P., Jia, R., Liang, P.: Know what you don't know: unanswerable questions for squad. CoRR abs/1806.03822 (2018). http://arxiv.org/abs/1806.03822
17. Reddy, S., Chen, D., Manning, C.D.: CoQA: a conversational question answering challenge. CoRR abs/1808.07042 (2018). http://arxiv.org/abs/1808.07042
18. Seo, M.J., Kembhavi, A., Farhadi, A., Hajishirzi, H.: Bidirectional attention flow for machine comprehension. CoRR abs/1611.01603 (2016). http://arxiv.org/abs/1611.01603
19. Sukhbaatar, S., Szlam, A., Weston, J., Fergus, R.: Weakly supervised memory networks. CoRR abs/1503.08895 (2015). http://arxiv.org/abs/1503.08895
20. Tran, A.H.N.: Applying deep neural network to retrieve relevant civil law articles. In: Proceedings of the Student Research Workshop Associated with RANLP 2017, pp. 46–48. INCOMA Ltd., Varna, September 2017. https://doi.org/10.26615/issn.1314-9156.2017_007
21. Trischler, A., et al.: NewsQA: a machine comprehension dataset. CoRR abs/1611.09830 (2016). http://arxiv.org/abs/1611.09830
22. Xiao, C., et al.: CAIL 2018: a large-scale legal dataset for judgment prediction. CoRR abs/1807.02478 (2018). http://arxiv.org/abs/1807.02478
23. Yin, X., Zheng, D., Lu, Z., Liu, R.: Neural entity reasoner for global consistency in NER (2018)
24. Zhang, N., Pu, Y.F., Yang, S.Q., Zhou, J.L., Gao, J.K.: An ontological Chinese legal consultation system. IEEE Access **5**, 18250–18261 (2017)
25. Zhong, H., Guo, Z., Tu, C., Xiao, C., Liu, Z., Sun, M.: Legal judgment prediction via topological learning. In: Proceedings of EMNLP (2018)
26. Zhong, H., et al.: Overview of CAIL 2018: legal judgment prediction competition. CoRR abs/1810.05851 (2018). http://arxiv.org/abs/1810.05851