CDD: A Large Scale Dataset for Legal Intelligence Research

Changzhen Ji[1]**, Yating Zhang, Adam Jatowt**[2] **and Haipang Wu**[1]
[1]Hithink RoyalFlush Information Network, Hangzhou, China
[2]University of Innsbruck, Innsbruck, Austria
czji_hit@outlook.com, yatingz89@gmail.com
adam.jatowt@uibk.ac.at, wuhaipang@myhexin.com

Abstract

As an important application of Artificial Intelligence, legal intelligence has recently attracted the attention of many researchers. Previous works investigated diverse issues like predicting crimes, predicting outcomes of judicial debates, or extracting information/knowledge from various kinds of legal documents. Although many advances have been made, the research on supporting prediction of court judgments remains relatively scarce, while the lack of large-scale data resources limits the development of this research. In this paper, we present a novel, large-size Court Debate Dataset (CDD), which includes $30,481$ court cases, totaling $1,144,425$ utterances. CDD contains real-world conversations involving judges, plaintiffs and defendants in court trials. To construct this dataset we have invited experienced judges to design appropriate labels for data records. We then asked law school students to provide annotations based on the defined labels. The dataset can be applied to several downstream tasks, such as text summarization, dialogue generation, text classification, etc. We introduce the details of the different tasks in the rapidly developing field of legal intelligence, the research of which can be fostered thanks to our dataset, and we provide the corresponding benchmark performance.

1 Introduction

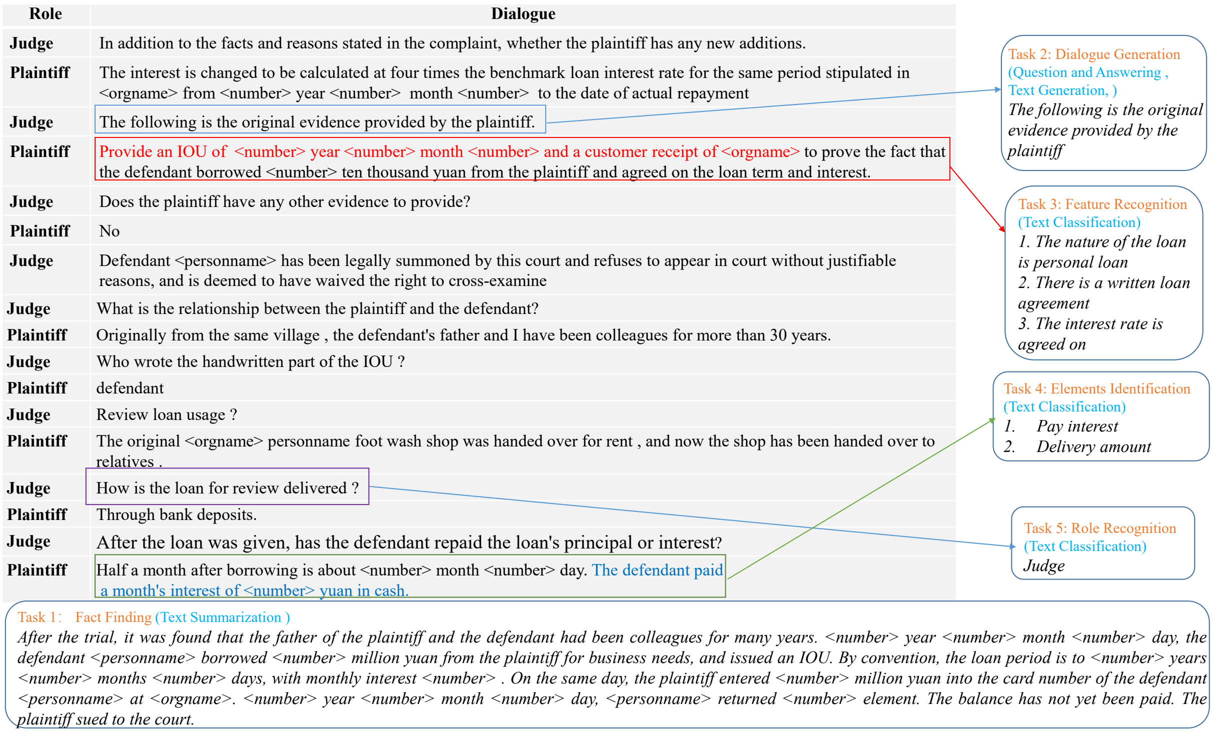
The increasing needs for efficient, high quality judicial service and the shortage of judicial personnel have become important concerns in the current society. The use of Artificial Intelligence (AI) technology to assist judges in effectively adjudicating cases is a research area that has potential to help improve judicial efficiency. In the real world, Legal Intelligence (LI) (Gray, 1997) could be applied in many scenarios, such as supporting management of court trials, legal judgment prediction, case information extraction, etc. The use of Artificial Intelligence technology to provide judicial services could not only alleviate the pressure on judges, but might also improve the efficiency of delivering judicial decisions.

In the recent years, judicial intelligence has gradually entered into the field of interest of many researchers, resulting in some explorations in this field ranging from legal judgment prediction (Xu et al., 2020; Zhong et al., 2020), analyzing trial cases, predicting particular laws that apply to a given case (Luo et al., 2017; Li et al., 2022), through court trialing to predicting the type of committed crimes. The advancement of Legal Intelligence research is however closely related to the availability of public datasets. Two well-known public available datasets that are currently in use are especially worth mentioning here: CAIL and ECHR. Chinese AI and Law challenge (CAIL) (Xiao et al., 2018) contains more than 2.6 million verdicts of criminal cases published by the Supreme People's Court of China[1], where each verdict consists of the identified facts given by the judge and the applicable law articles, charges, and prison terms, supporting the task of judgment prediction. ECHR (Chalkidis et al., 2019), on the other hand, is the first English legal judgment prediction dataset, containing cases from the European Court of Human Rights. Although previous research has made significant progress on the track of judgment prediction, the lack of effective and diverse datasets has become a considerable obstacle to the development of the Legal Intelligence field.

Legal intelligence involves a wide range of scenarios and is not just limited to legal judgment prediction or crime prediction. It can provide judges with more efficient and transparent trials in more ways. In this context, we provide a large-scale judicial dataset, which contains the real-world dia-

[1] China Judgement Online: https://wenshu.court.gov.cn/

Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 66–73
December 6-10, 2023 ©2023 Association for Computational Linguistics



Role	Dialogue
Judge	In addition to the facts and reasons stated in the complaint, whether the plaintiff has any new additions.
Plaintiff	The interest is changed to be calculated at four times the benchmark loan interest rate for the same period stipulated in <orgname> from <number> year <number> month <number> to the date of actual repayment
Judge	The following is the original evidence provided by the plaintiff.
Plaintiff	Provide an IOU of <number> year <number> month <number> and a customer receipt of <orgname> to prove the fact that the defendant borrowed <number> ten thousand yuan from the plaintiff and agreed on the loan term and interest.
Judge	Does the plaintiff have any other evidence to provide?
Plaintiff	No
Judge	Defendant <personname> has been legally summoned by this court and refuses to appear in court without justifiable reasons, and is deemed to have waived the right to cross-examine
Judge	What is the relationship between the plaintiff and the defendant?
Plaintiff	Originally from the same village , the defendant's father and I have been colleagues for more than 30 years.
Judge	Who wrote the handwritten part of the IOU ?
Plaintiff	defendant
Judge	Review loan usage ?
Plaintiff	The original <orgname> personname foot wash shop was handed over for rent , and now the shop has been handed over to relatives .
Judge	How is the loan for review delivered ?
Plaintiff	Through bank deposits.
Judge	After the loan was given, has the defendant repaid the loan's principal or interest?
Plaintiff	Half a month after borrowing is about <number> month <number> day. The defendant paid a month's interest of <number> yuan in cash.

Task 2: Dialogue Generation (Question and Answering , Text Generation,)
The following is the original evidence provided by the plaintiff

Task 3: Feature Recognition (Text Classification)
1. The nature of the loan is personal loan
2. There is a written loan agreement
3. The interest rate is agreed on

Task 4: Elements Identification (Text Classification)
1. Pay interest
2. Delivery amount

Task 5: Role Recognition (Text Classification)
Judge

Task 1: Fact Finding (Text Summarization)
After the trial, it was found that the father of the plaintiff and the defendant had been colleagues for many years. <number> year <number> month <number> day, the defendant <personname> borrowed <number> million yuan from the plaintiff for business needs, and issued an IOU. By convention, the loan period is to <number> years <number> months <number> days, with monthly interest <number> . On the same day, the plaintiff entered <number> million yuan into the card number of the defendant <personname> at <orgname>. <number> year <number> month <number> day, <personname> returned <number> element. The balance has not yet been paid. The plaintiff sued to the court.

Figure 1: Example Dialog in Court Debate Dataset.

logues between judges, plaintiffs and defendants in court trials of private lending cases. Figure 1 gives an example, where the judge is inquiring about the details of the case and the party being questioned answers them[2]. We invited experienced judges to define judicial features and elements which mark key characteristics of debates, and we asked a large pool of judicial practitioners from law schools to provide the corresponding annotations. In the end, our annotated dataset has multiple dimensions including: facts, features, elements and roles. It can be then applied to multiple downstream tasks. As shown in Figure 1, it can be used to foster research in Fact Finding, Dialogue Generation, Feature Recognition, Elements Identification and Role Recognition. In total, we introduce five downstream tasks and discuss their associated application scenarios as well as provide baseline models to establish reference performance. Based on the proposed dataset, one can thus conduct research focusing on multiple application scenarios. We will describe the details of those tasks in Section 3.

2 Related work

Legal Intelligence research has been initiated in 1960s (Nagel, 1960). Nagel (1960) proposed the use of algebraic calculations to determine the judgement of the court case. Especially, in the recent years, legal intelligence has emerged as a popular topic attracting attention of many researchers (McElvain et al., 2019; Biega et al., 2020; Bhattacharya et al., 2020; Shao, 2020; Dong and Niu, 2021; Ma et al., 2021).

Dong and Niu (2021) proposed to predict the outcome of trials based on the facts of the judicial cases. Zhong et al. (2018) introduced a topological multi-task learning framework (TOP-JUDGE) that incorporates multi-task learning and DAG dependencies into judgment prediction. Zhou et al. (2019) leveraged multi-view dispute representation for e-commerce judgement result prediction while Wang et al. (2019) utilized fact, law and article information to build a hierarchical matching network for crime classification. Li et al. (2022) extracted objective elements from factual descriptions for crime prediction.

The release of relevant datasets often provides important stimuli for a field. Xiao et al. (2018) published CAIL to foster research in judgment prediction. Duan et al. (2019) proposed a Chinese Judicial Reading Comprehension (CJRC) datasets[3]. Xiao et al. (2019) published CAIL2019-SCM, a similar case matching dataset. Chalkidis et al. (2019) released the first English legal judgment prediction dataset (ECHR), containing cases from the European Court of Human Rights. Malik et al.

[2] Sensitive information (e.g., person's name) has been removed for privacy issue.

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

Table 1: The publicly available large-scale judicial datasets.

Dataset	**Scale**	**Language**	**Supported downstream task**
CAIL2018 (Xiao et al., 2018)	2.6 million criminal cases	Chinese	☑ Text Classification ▨ Question and Answering ▨ Text Generation ▨ Text Summarization
CAIL2019 (Xiao et al., 2019)	8,964 triplets of cases	Chinese	☑ Text Classification ▨ Question and Answering ▨ Text Generation ▨ Text Summarization
CJRC (Duan et al., 2019)	10K documents and 50K questions with answers	Chinese	▨ Text Classification ☑ Question and Answering ▨ Text Generation ▨ Text Summarization
ECHR (Chalkidis et al., 2019)	11.5k cases from European Court of Human Rights public database	English	☑ Text Classification ▨ Question and Answering ☑ Text Generation ▨ Text Summarization
ILDC (Malik et al., 2021)	35k Indian Supreme Court Cases	English	☑ Text Classification ▨ Question and Answering ☑ Text Generation ▨ Text Summarization
CDD	30,481 court dialogue cases, twelve feature and fourteen Judicial Elements	Chinese/English	☑ Text Classification ☑ Question and Answering ☑ Text Generation ☑ Text Summarization

Table 2: Basic Statistics of Court Debate Dataset

Total cases	30,481
Total utterances	1,144,425
Total words	18,590,439
Average turns	37.62
Max turns of case	461
Average length of utterance	162.44
Max length of utterance	2382

(2021) provided ILDC for Court Judgment Prediction and Explanation (CJPE) tasks. The current large-scale judicial disclosed datasets are compared in Table 1.

Note that the current judicial research focuses more on classification tasks such as case outcome prediction, crime prediction, etc. It is difficult to carry out richer and multi-scenario tasks due to insufficient resources. Our work fills this gap aiming to provide a comprehensive dataset for researchers to promote the progress of legal intelligence.

3 Court Debate Dataset

3.1 Data Collection

The data comes from the actual records of court trial procedures of private lending cases[4]. It contains $30,481$ trial cases, $1,144,425$ utterances and $18,590,439$ words. Each case is a multi-turn dialogue between judges, plaintiffs, and defendants. The average number of turns of the dialogue in a case is 37 and the maximum is 461. For processing the raw conversation data, we use jieba[5] toolkit for word segmentation. The overall dataset statistics are shown in Table 2. In particular, we remove sensitive information (e.g., replacing all numbers, person names, and organization names with specific tokens `<number>`, `<personname>`, `<orgname>`, respectively). In addition, we also align the trial of a case to its final verdict so that the fact description summarized by the judge can be regarded as the summary of the court debate transcript. In order to enable any researchers to freely use our dataset, we have translated the original content into English using professional translators[6].

3.2 Data Definition

To make the data available for academic research, we asked experienced judges to define features and elements to indicate the important aspects of trials.

The features are defined as the qualitative evidential features of the case that can help to determine the judgment result. As for the case of private lending, which is the type of trials that CDD contains, during the initial review of a case, a judge usually needs to consider some qualitative features of the case, such as: "whether there is a written loan agreement", "whether the interest rate has been agreed on", etc. Following such logic, the judge is usually able to issue the verdict. We asked 6 experienced judges for this and they have defined 12 qualitative essential features. The 12 features are listed in Appendix A.1.

In order to determine the facts of the case, the judge needs to also investigate and inquire about the factual elements, such as: "loan amount", "loan period", etc. Therefore, in order to clarify the facts of the case, the experienced judges helped us to define 14 elements for the case of private lending. Note that these 14 elements do not necessarily appear in all the loan cases. In some simple cases, only a few of these elements appear in the conversation. The 14 element tags are listed in Appendix A.2.

[4] The dataset is provided by the High People's Court of a province in China. All the court transcripts have been manually recorded by the court clerk.

[5] `https://github.com/fxsjy/jieba`

[6] `https://github.com/jichangzhen/CDD`

3.3 Data Annotation

Following the judges' provision of the definition of key evidential features and factual elements, we hired students from law schools to annotate the court debate data based on the provided label settings.

The annotation process was conducted as follows:

- For features: the annotators need to give qualitative judgment. Take the label "whether there is a written loan agreement" as an example. An annotator is asked to first find out if there exists any mention about the loan agreement and he/she has to determine whether it is a written loan agreement rather than a verbal one. If so, this label will be marked as 'yes', otherwise as 'no'. The annotator needs to read the dialogue between the judge, the plaintiff and the defendant, and then provide the annotation based on the factual information found in the dialogue.
- For elements: an annotator labels whether or not each element appeared in the conversation. Therefore, for labeling elements, the annotators only need to focus on the mentions of the elements. If the element is mentioned in particular context then it is marked as 'yes' for this element label, otherwise is annotated as 'no'.
- For speaking roles: each utterance is marked with the role of its speaker (plaintiff, defendant or judge).
- For summary: as mentioned in Sec. 3.1, the fact description in the verdict is regarded as the summary of the court debate transcript.

3.4 Task Definitions

According to the data described in Sec 3.2 and the annotation outlined in Sec 3.3, we define five tasks for our dataset: (1) Fact Finding (FF), (2) Dialogue Generation (DG), (3) Feature Recognition (FR), (4) Elements Identification (EI) and (5) Role Recognition (RR).

• **Fact Finding (FF)** is a text summarization task. After trial, a judge summarizes the facts based on the answers of the plaintiff and defendant. These facts include the key notes extracted from the case, which record who, when, and where, as well as the cause, the course and the result of the incident. In this task, the entire dialogue is regarded as an input, and the fact description in the corresponding verdict is treated as the output.

• **Dialogue Generation (DG)** is a fundamental task of natural language processing. Considering judicial scenarios, the generation of judge's utterance has potential to support intelligent solutions towards more effective court trials. To fully use the entire court debate data for the task of dialogue generation, we divide each trial debate into smaller units. Specifically, due to the different lengths of judicial cases, some cases have more than 400 dialogue rounds, and some cases less than 10 dialogue rounds. We divide each case into multiple dialogue samples, so that each dialogue sample has only 5-10 dialogue rounds[7]. The last sentence of each dialogue sample is always the judge's utterance. With this setting, we assume the prior utterances before the last utterance of each dialogue sample as an input, while the last sentence is considered as an output that needs to be generated. Note that one objective for such setting is to investigate the application of an intelligent assistance for judges for the next question formulation. The basic statistics of the dataset for the task of dialogue generation are given in the Table 3.

Table 3: Statistics of the dataset for dialogue generation.

Dialogue sample	133,268
Average length	37.62
Average turns	8.5
Max turns of case	10
Min turns of case	5

• **Feature Recognition (FR)** is a multi-label classification task where 12 factual features are in advance defined by an experienced judge and each case is annotated with the above 12 factual features. Since the annotation is conducted over the entire dialogue, therefore for each sample, the input is the entire dialogue and the output are the binary choices over the 12 feature labels.

• **Elements Identification (EI)** is also a multi-label classification task. As mentioned in Sec 3.2, 14 elements tags are predefined by the judges. Different from Feature Recognition, the task of Elements Identification relies on gathering the detailed information of the case. For each sample, the input is the entire conversation of a case and the task is to predict whether the information related to each element appeared in the court record or not.

• **Role Recognition (RR)** is a conventional multi-classification task. In the conventional trial process, there are usually three roles: judge, plaintiff and defendant. We use the utterances in the trial to predict the speakers' roles. Therefore, Role Recognition is a three-class classification task.

[7] For example, if a case has 20 rounds of dialogue, the annotator should divide it into 2-4 dialogue samples.

Table 4: Dataset distribution: the number of dialogues, sentences, words, divided into the training set, development set and test set.

dataset	dialogue	sentence	word
train	27,432	1,029,528	16,725,537
dev	1,524	56,941	924,952
test	1,525	57,956	939,950
total	30,481	1,144,425	18,590,439

Studying this task could help us in better understanding of trial debate (see Section 5.4 for specific practical implications).

4 Experiments

In this section, we describe the experiments conducted on CCD, and we introduce classical baseline models tested for the above-discussed tasks.

4.1 Baselines

The entire dataset is divided into the training set, development set and test set. The division of the dataset is summarized in Table 4.

We group the the five judicial tasks discussed before into three categories of NLP tasks. These are Fact Finding as text summarization task; Dialogue Generation, as text generation task; Feature Recognition, Elements Identification and Role Recognition as text classification tasks.

For text summarization and text generation tasks we test the following models:

- **S2S+attention** (Sutskever et al., 2014): a sequence-to-sequence model where attention is used to assign weights to context.
- **PGN** (See et al., 2017): a model that employs the pointer generator network. During decoding, it expands the context distribution to the dynamic vocabulary, which solves the out-of-vocabulary problem.
- **HRED** (Serban et al., 2016): a hierarchical long short-term memory network structure which can encode multiple sentences hierarchically.
- **Transformer** (Vaswani et al., 2017): a network architecture using self-attention mechanism and positional encoding.
- **LLaMA** (Touvron et al., 2023): a large language model based on transformer architecture.
- **LLaMA+SFT** (Ouyang et al., 2022): a model which employs Supervised Fine-Tuning on the basis of large language model LLaMA.

For text classification tasks, the following models are tested:

- **BiLSTM** (Klein et al., 2017): a bidirectional encoding structure that solves the problem of RNN's difficulty to memorize long sequences.

Table 5: Fact Finding and Dialogue Generation Experimental Results.

model	Fact Finding			Dialogue Generation		
	R-1	R-2	R-L	R-1	R-2	R-L
S2S+attention	40.71	22.11	33.95	27.69	16.29	22.63
PGN	41.28	22.35	34.63	28.48	17.91	23.97
HEAD	44.02	24.21	37.73	28.59	19.03	24.13
LLaMA	52.85	42.76	54.91	48.43	47.28	53.65
LLaMA+SFT	54.43	44.29	57.61	48.35	49.79	54.84

Table 6: Feature Recognition, Elements Identification and Role Recognition Experimental Results.

model	FR		EI		RR	
	Mic	Mac	Mic	Mac	Mic	Mac
BiLSTM	72.51	31.92	69.26	27.62	83.69	40.03
BERT	74.63	34.58	73.53	32.87	85.16	41.29
LLaMA	82.71	75.43	83.84	71.29	89.72	76.81
LLaMA+SFT	85.64	78.39	88.43	74.59	90.07	77.20

- **Bert** (Devlin et al., 2019): a pre-trained language model using mask mechanism, which can be applied to a variety of downstream tasks.
- **LLaMA** (Touvron et al., 2023): a large language model based on transformer architecture.
- **LLaMA+SFT** (Ouyang et al., 2022): a model which add SFT fine-tuning technology on the basis of large language model LLaMA.

4.2 Evaluation

We use two types of evaluation metrics: for natural language generation tasks, we use ROUGE (Lin, 2004), and report ROUGE-1, ROUGE-2 and ROUGE-L scores, while for classification tasks, we use micro-average F1 scores (mic) and macro-average F1 scores (mac).

5 Result discussion

5.1 Text Summarization

Table 5 (columns 2-4) shows the results of the Fact Finding task over different tested baselines.

For traditional models, compared to **S2S+attention**, **PGN** shows better performance, mainly because the fact entities usually appear in the dialogue, so copying the entities from the dialogue into the generated fact is an efficient solution. **HRAD** achieves better results, mainly because the input of text is a dialogue where the hierarchical information is essential for representation, and hierarchical coding is more conducive to obtaining semantic information. The large language models (LLMs) show superior performance, especially the model after SFT fine-tuning achieves a new performance level. Pre-training a large language model on massive amounts of data is a major advance in NLP.

5.2 Dialogue Generation

Table 5 (columns 5-7) shows the results of the Dialogue Generation. Similar to Fact Finding, the dialogue generation task is also conducted with the mainstream generation models. There are certain similarities between the generation of dialogue and the generation of facts. The goal of those two tasks is to obtain concrete factual information from the dialogue.

From the results in Table 5, it can be seen that the **LLaMA+SFT** model achieves here the best results, too. The dialogue generation task aims to generate the judge's utterance through the analysis of the previous part of dialogue between the judge, the plaintiff and the defendant. Compared to the model **S2S+attention**, **PGN** produces better results. Usually, the judge's utterances are in the form of questions with the objective to find out the truth of the matter. The judge will continuously ask questions to the plaintiff and the defendant, and will further investigate the content mentioned in their replies. For example, the plaintiff said ”He signed an IOU”, and next, the judge will further investigate the fact of the ”IOU”. Therefore getting the contextual key words and phrases makes sense for generating judges' utterance generation. In addition, a copy mechanism in **PGN** contributes to the better performance of the generation model.

5.3 Text Classification

Table 6 shows the results of the Feature Recognition, Elements Identification and Role Recognition. They all use the same classification baseline models. The difference is that Role Recognition is a three-class classification of a single sentence, while feature recognition and elements recognition are multi-label classification tasks for entire dialogues.

From the experimental results, it can be concluded that **LLaMA+SFT** achieves the best classification results. It outperforms **BiLSTM** and **Bert** models by a large margin, not only in single sentence classification but also in long text classification. Hence, it is promising to do classification using pre-trained large language models.

5.4 Practical significance

Nowadays, a large number of judges are under a high workload. In addition to adjudicating cases in court, judges also undertake a large number of transactional tasks such as litigation guidance, post-judgment questions and answers, law popularization, investigation and research. If AI technology can effectively support the administrative work of judicial personnel, its application in the judicial field would save effort and costs.

The five tasks proposed in this paper have important practical applications. Studying Fact Finding and Dialogue Generation can be of great significance in the research of judicial assistants. For example, judge's utterances could be generated to let the judge use it as a prompt when questioning the plaintiff and the defendant, or to simulate actual trial debate for educational or preparatory purposes. Generating corresponding facts or judgments after the trial could support the task of summarizing the case. The research on Element Identification and Feature Recognition could help judges quickly overview and understand the elements of a case, which are of great significance for case filing. Finally, the task of Role Recognition could lead to providing sufficient support or refutation depending on speaker's role, and could form a part of multi-tasking approaches to automatic court debate analysis/simulation.

5.5 Ethics Statement

Finally, we would like to briefly reflect on ethical issues. The dataset is created on the basis of real cases, and should ensure the fairness and impartiality of court judgments (Pitoura et al., 2018; Mahoney, 2015; Lim et al., 2020). Unbalanced dataset distribution and social bias could lead to potential risks of machine learning, and researchers should be aware of such risks. To address those issues, we have carefully removed sensitive data (eg, name, gender, race, etc.). We have also adopted a cross-training approach to ensure a more balanced dataset.

6 Conclusions

We proposed a large-scale judicial dataset, Court Debate Dataset (CDD) which contains real judicial debates and is annotated by experienced judges and students of law schools. CDD can be applied in academic research on a variety of downstream tasks, including Fact Finding, Dialogue Generation, Feature Recognition, Element Identification and Role Recognition. Academic research results could be then put into practice in real-world applications leading to the interplay of theory and practice, and promoting the process of Legal Intelligence.

In the future, we will continue to develop new models based on the provided dataset to improve results across diverse sub-tasks.

References

Paheli Bhattacharya, Kripabandhu Ghosh, Arindam Pal, and Saptarshi Ghosh. 2020. Hier-spcnet: A legal statute hierarchy-based heterogeneous network for computing legal case document similarity. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1657–1660. ACM.

Asia J Biega, Peter Potash, Hal Daumé, Fernando Diaz, and Michèle Finck. 2020. Operationalizing the legal principle of data minimization for personalization. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 399–408. ACM.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural legal judgment prediction in English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4317–4323, Florence, Italy. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Qian Dong and Shuzi Niu. 2021. Legal judgment prediction via relational learning. In *Proceedings of the 44rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 983–992. ACM.

X Duan, B. Wang, Z. Wang, W. Ma, Y. Cui, D. Wu, S. Wang, T. Liu, T. Huo, and Z. Hu. 2019. Cjrc: A reliable human-annotated benchmark dataset for chinese judicial reading comprehension.

P. N. Gray. 1997. *Artificial legal intelligence*. Artificial Legal Intelligence.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Lin Li, Lingyun Zhao, Peiran Nai, and Xiaohui Tao. 2022. Charge prediction modeling with interpretation enhancement driven by double-layer criminal system. *World Wide Web*, 25(1):381–400.

Sora Lim, Adam Jatowt, Michael Färber, and Masatoshi Yoshikawa. 2020. Annotating and analyzing biased sentences in news articles using crowdsourcing. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1478–1484, Marseille, France. European Language Resources Association.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Bingfeng Luo, Yansong Feng, Jianbo Xu, Xiang Zhang, and Dongyan Zhao. 2017. Learning to predict charges for criminal cases with legal basis. pages 2727–2736.

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 *Proceedings of the 44rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 993–1002. ACM.

Kathleen Mahoney. 2015. Judicial bias: The ongoing challenge. *Journal of Dispute Resolution*, 2015(1):4.

V. Malik, R. Sanjay, S. K. Nigam, K. Ghosh, and A. Modi. 2021. Ildc for cjpe: Indian legal documents corpus for court judgment prediction and explanation.

Gayle McElvain, George Sanchez, Sean Matthews, Don Teo, Filippo Pompili, and Tonya Custis. 2019. Westsearch plus: A non-factoid question-answering system for the legal domain. In *Proceedings of the 42rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1361–1364. ACM.

Stuart Nagel. 1960. Using simple calculations to predict judicial decisions. *American Behavioral Scientist*, 4(4):24–28.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Evaggelia Pitoura, Panayiotis Tsaparas, Giorgos Flouris, Irini Fundulaki, Panagiotis Papadakos, Serge Abiteboul, and Gerhard Weikum. 2018. On measuring bias in online information. *ACM SIGMOD Record*, pages 16–21.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Iulian Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3776–3783. AAAI.

Yunqiu Shao. 2020. Towards legal case retrieval. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2485–2485. ACM.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Conference on Neural Information Processing Systems*, pages 3104–3112. MIT Press.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Conference on Neural Information Processing Systems*, pages 5998–6008. MIT Press.

Pengfei Wang, Yu Fan, Shuzi Niu, Ze Yang, Yongfeng Zhang, and Jiafeng Guo. 2019. Hierarchical matching network for crime classification. In *Proceedings of the 42rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 325–334. ACM.

Chaojun Xiao, Haoxiang Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, and Jianfeng Xu. 2018. Cail2018: A large-scale legal dataset for judgment prediction. *ArXiv*, abs/1807.02478.

Chaojun Xiao, Haoxiang Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Tianyang Zhang, Xianpei Han, Heng Wang, and Jianfeng Xu. 2019. Cail2019-scm: A dataset of similar case matching in legal domain. *ArXiv*.

Nuo Xu, Pinghui Wang, Long Chen, Li Pan, Xiaoyan Wang, and Junzhou Zhao. 2020. Distinguish confusing law articles for legal judgment prediction. pages 3086–3095, Online. Association for Computational Linguistics.

Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun. 2018. Legal judgment prediction via topological learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3540–3549, Brussels, Belgium. Association for Computational Linguistics.

Haoxi Zhong, Yuzhong Wang, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. Iteratively questioning and answering for interpretable legal judgment prediction. 01, pages 1250–1257.

Xin Zhou, Yating Zhang, Xiaozhong Liu, Changlong Sun, and Luo Si. 2019. Legal intelligence for e-commerce: Multi-task learning by leveraging multiview dispute representation. In *Proceedings of the 42rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 315–324. ACM.

A Appendices

A.1 Features

The 12 features mentioned in Section 3.2 are:

1. Whether the litigation period has expired,
2. Whether to demand repayment,
3. whether there is a written loan agreement,
4. whether the loan is a private loan,
5. whether the guarantor provides a guarantee,
6. whether the interest rate is agreed on,
7. whether repayment period is agreed upon,
8. whether the loan period is agreed upon,
9. whether the default clause is agreed upon,
10. whether there is a repayment action,
11. whether the borrower provides the loan as
12. whether the principal and interest are still owed.

A.2 Element

The 14 element tags mentioned in Section 3.2 include:

1. Loan amount,
2. Loan period,
3. Loan start time,
4. Loan end time,
5. Repayment time,
6. Principal payment,
7. Interest payment,
8. Liquidated damages,
9. Outstanding principal balance,
10. Delivery Date,
11. Delivery Amount,
12. Annual Interest rate,
13. Monthly interest rate,
14. Overdue interest rate.