Neural Computing and Applications
https://doi.org/10.1007/s00521-024-09907-4

ORIGINAL ARTICLE

DLEE: a dataset for Chinese document-level legal event extraction

Guochuan Xian[1] · Siyuan Du[1] · Xi Tang[1] · Yuan Shi[1] · Bofang Jia[1] · Banghao Tang[1] · Zhefu Leng[1] · Li Li[1]

Received: 22 October 2023 / Accepted: 23 April 2024

Abstract
Event extraction (EE) is capable of providing essential information to facilitate comprehension of legal cases by identifying event types and extracting corresponding arguments from legal case documents. In the legal field, events are often presented in the form of document, with arguments scattered across multiple sentences, which means that legal EE at the document level is needed to better capture the complete event. However, the existing legal EE datasets mainly focused on event extraction at the sentence level, with little attention given to the document level. Obviously, it put the development of document-level event extraction (DEE) in the legal field at a disadvantage. To address this challenge, we proposed DLEE, the first DEE dataset in the legal field with two distinctive features: (1) Document-level Semi-automated Annotation, ensuring effective annotation with high quality. (2) Large-scale and Fine-grained coverage, comprising 10,014 events and 99,423 arguments. Finally, we assessed the performance of commonly used DEE baseline models on DLEE. It revealed that the DLEE is an open question, and further attention is needed for the improvement of the models' performance.

Keywords Dataset · Document-level event extraction · Legal · Semi-automatic annotation

1 Introduction

Event, as a special form of information, refers to something that one or more participants participate in at a specific time and a specific location [1]. Event extraction (EE) is an important research area in the field of information extraction and has been widely used in recommendation [2], cyber security [3], knowledge graph construction [4], knowledge question answering [5], and other fields [6, 7].

EE aims to extract key information from unstructured text to eventually form a structured event, including time, place, characters, actions, etc. It is usually divided into two sub-tasks [8]: 1) Event detection. Extracting event triggers from the text and identifying event types. 2) Event argument extraction. Extracting arguments of the event and assigning the corresponding argument roles.

✉ Li Li
lily@swu.edu.cn

Guochuan Xian
guocxian@email.swu.edu.cn

[1] School of Computer and Information Science, Southwest University, Chongqing 400715, China

In recent years, there has been a growing research interest in utilizing EE technology to enhance Legal Artificial Intelligence (LegalAI) [9], aiming to mitigate the burden on humans to understand and retrieve legal documents. Inspired by the previous success of Legal Event Detection (LED) [10], Similar Case Matching (SCM) [9], and Legal Judgment Prediction (LJP) [11], we recognize the pivotal role of arguments in these tasks. Specifically, the downstream applications of event arguments are depicted in Fig. 1. In the LJP, *Wang severely injured* the victim, *Ren*, using a *sharp knife*, therefore committed an offense under *Criminal Law Article 234* and was convicted of the *crime of intentional injury*, resulting in a prison sentence of *56 months*. In the SCM, Case 2 was matched to Case 1.

Despite the promising prospects of EE in the legal domain, there are few EE datasets specifically tailored for legal applications. Moreover, the existing datasets for legal EE exhibit limitations in terms of scale and coverage. The following are some specific issues associated with current legal EE datasets:

- **Constraints of sentence level**: Existing legal datasets [12–14] are primarily the sentence-level EE (SEE)

Published online: 16 May 2024

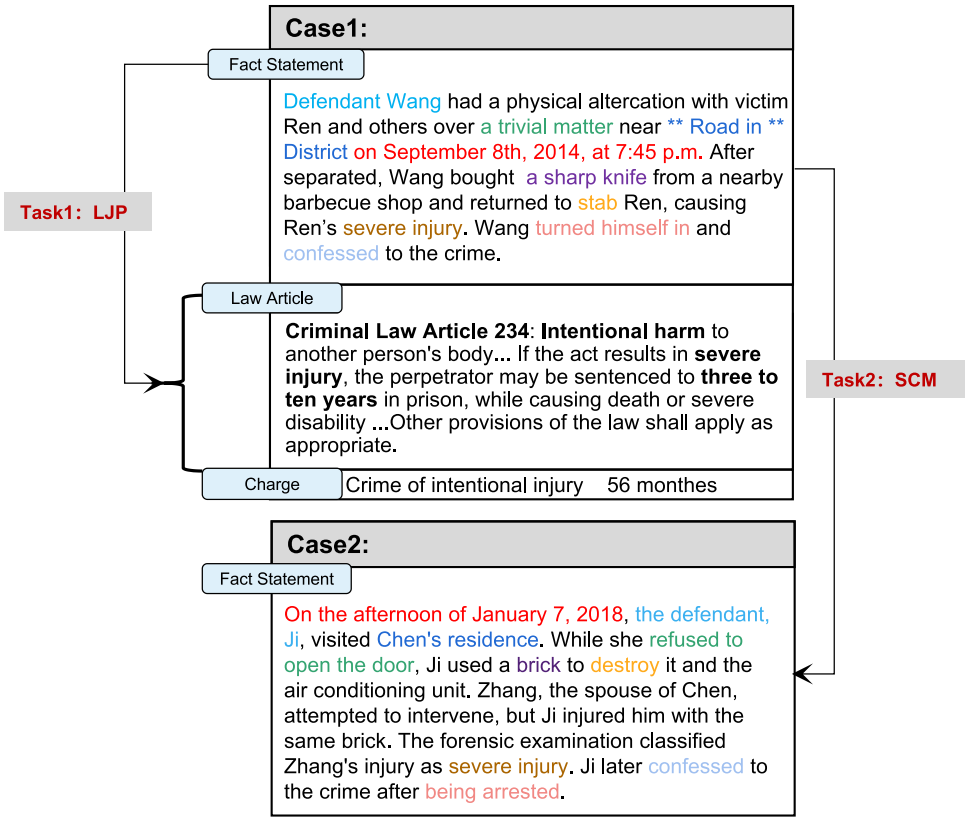

Fig. 1 An example of different kinds of (different colors) arguments within two cases, which play an important role in legal downstream tasks: Legal Judgment Prediction (LJP) and Similar Case Matching (SCM). The key information in Law Article, which is the applicable terms to Case 1, is represented in bold. The Charge is the corresponding verdict for Case 1. Additionally, Case2 is a similar case to Case 1

datasets. However, in legal cases, arguments are often distributed across multiple sentences, and events are often presented in the form of documents, making it challenging to capture complete argument roles within a single sentence.

- **Limited data with coarse granularity**: Although LEVEN [10] offers an extensive and comprehensive dataset on LED, it does not include argument information. Additionally, the existing SEE datasets have limitations in terms of a restricted number of events, low coverage of event types, incomplete argument roles, and unavailable data. As a result, these constraints impede the possibility of conducting in-depth and comprehensive analyses of legal cases.

In order to alleviate the above two problems, we propose a large-scale and fine-grained Chinese legal DEE dataset. Our data source consists of more than 63,000 criminal judgment documents, with the majority sourced from a corpus [15] comprising over 43,000 Chinese criminal judgment documents. The remaining 20,000 cases were collected from the Chinese Judgment Document Online[1]. The criminal judgment documents underwent standardized processing, with data not meeting the criteria being filtered out. Subsequently, 10,014 records were preserved for further annotation, ensuring the dataset's high quality. The dataset primarily exhibits the following characteristics:

- **Document-level Semi-automatic Annotation**: To ensure the quality of DLEE, we manually selected documents from the provided corpus. Furthermore, we have invited legal experts and law students, aiming to enhance the dataset's overall quality. They were involved in designing an event schema and utilizing our semi-automated platform to extract triggers and arguments from the fact statement section of each criminal judgment document.
- **Large-scale and Fine-grained**: DLEE encompasses a comprehensive event schema, consisting of 7 top-level event types, 49 sub-level event types, and a total of 378 corresponding argument roles. Each criminal judgment document is classified into one of 49 event types. Furthermore, we have annotated 10,014 criminal judgment documents along with nearly 100,000 associated arguments, averaging 9.9 event arguments per document.

To sum up, we make the following contributions:

- We proposed DLEE, the first DEE dataset in the Chinese legal field, aiming to alleviate the scarcity of such datasets. Meanwhile, we explored the potential use of DLEE for downstream applications such as SCM and

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

LJP. DLEE is anticipated to attract more research on LegalAI in Chinese.

- We explored a semi-automated process for constructing a DEE dataset and developed a comprehensive annotation system to achieve efficient and accurate data annotation in the Chinese legal domain. This approach serves as a valuable reference for future dataset creation efforts.
- We conducted a thorough analysis of the dataset. The results illustrate that the DLEE dataset is qualitatively competitive. We also implemented three DEE baseline models to assess challenges faced by those models. The results show these models on DLEE need further improvement to achieve practical application.

Following is the organization of the rest of the paper. Section 2 provides the related work for the paper. Section 3 presents the data collection process. Section 4 analyzes the data from different aspects. Section 5 presents the experiment performance on the DLEE dataset. Section 6 presents the conclusion and future work.

2 Related work

2.1 Document-level event extraction

EE, as a crucial subtask of information extraction (IE), has witnessed progress through multiple assessment tasks such as Message Understanding Conference (MUC) [16], Automatic Content Extraction Conference (ACE) [1], and Text Analysis Conference Knowledge Base Population (TAC-KBP) [17] evaluation task. But most of the research has predominantly focused on SEE [8, 18, 19] or Event Detection(ED) [20, 21]. Actually, downstream applications require information about event arguments, which are often distributed across multiple sentences. Therefore, it is critical to explore DEE methods that can both identify corresponding event types and capture scattered arguments.

In the field of DEE, earlier works mainly depended on manually designed characteristics [22, 23]. Recently, neural networks and pre-trained language models have gained popularity. Yang et al. [24] developed DCFEE, a system for DEE in the Chinese Financial domain. This system automatically generates a large-scale labeled dataset and extracts events from entire documents. To identify trigger words and candidate arguments, the system treats DEE as a sequence labeling task. It employs an event detection model to identify events in the document and applies an argument-filling strategy to fill in missing arguments. However, this approach is limited in its ability to handle dispersed arguments. Zheng et al. [25] addressed this issue by directly filling out a form based on the document, dividing the task into three sub-tasks: entity extraction to extract candidate elements, event detection to identify triggered events, and event table filling to populate the table with relevant elements. Xu et al. [26] proposed Git, which utilizes a graph interaction network to capture global interactions among entity mentions and different sentences, enabling more effective extraction of scattered elements and the capture of event relationships. For further excavating multi-scale and multi-amount argument relations, Liang et al. [27] proposed the RAAT model, which achieves state-of-the-art performance on two public datasets in DEE tasks.

In summary, although existing works have demonstrated promising results in addressing the cross-sentence and multi-event issues from different perspectives. The lack of a comprehensive and publicly available DEE dataset in the Chinese legal field is a significant challenge. Constructing such a dataset is of paramount importance to enable the utilization of advanced models in promoting the development of LegalAI.

2.2 Related datasets

Due to the high cost of annotating data and time-consuming treatment processes, the DEE dataset is primarily limited to a few specific domains. The following are several domains with corresponding datasets:

Finance field. The application of EE in the finance domain is always a hot research direction. Yang et al. [24] proposed a DCFEE system as early as 2018, which labeled 2976 announcements with 4 types and 3,044 sentences. This approach employed key sentences to determine the event type of each article and used an argument-filling strategy to extract the arguments of both the key sentence and surrounding sentences. To make document-level event annotating easier, Zheng et al. [25] reformalized the DEE task with the no-trigger-words design. They labeled 32, 040 financial documents with 5 types and 35 argument roles, in which a few documents may contain multiple event types. Han et al. [28] came up with a large-scale benchmark, DuEE-Fin, which consists of more than 15,000 events categorized into 13 event types, and more than 81,000 event arguments mapped in 92 argument roles. Notably, this benchmark allows one document to contain several events, multiple arguments to share the same argument role, and one argument to play different roles in different events.

News field. MUC-4 [29] consists of 1700 news reports and covers 4 event types and 5 argument roles, some documents in the dataset describe multiple events. On average, each document in the dataset contains 403.27 tokens and 7.12 paragraphs. Roles Across Multiple Sentences (RAMS) [30] contains 9,124 annotated events

across 139 types, 65 argument roles, and 3,993 news articles with an average length of approximately 40 sentences. However, it may result in the loss of many arguments using the method of the 5-sentence window around the sentence containing the event trigger for EE. WikiEvents [31] comprises only 246 documents, covering 50 event types and 59 argument roles. It provides comprehensive event and coreference annotations, but the cross-sentence argument annotations are kept to a minimum. DocEE [32], a document-level news event extraction dataset including 27,485 documents with 59 event types and 356 argument roles. Each document is only annotated with one event, similar to DLEE.

Chinese legal field. Although there are currently no dedicated datasets for DEE in the legal field, there are several datasets for SEE, including the following: Li et al. [12] applied EE to the legal field for the first time and annotated 3,100 case materials, constructing a sentence-level dataset that has 13 event types and 40 argument roles. Shen et al. [13] labeled a dataset, which contains 2,380 instances with 11 predefined event types and 26 argument roles. Li et al. [14] selected 3000 larceny judgment documents, and divided the fact statement section of each document into individual sentences, resulting in 6538 sentences. These sentences were then categorized into one of the 5 event types.

More details of the above datasets are listed in Table 1. The critical information of a case is typically scattered across multiple sentences, necessitating the expansion of the dataset from sentence level to document level in order to achieve a more fine-grained extraction. Thus, in the Chinese legal field, constructing a DEE dataset is an urgent imperative.

3 Dataset construction

In this section, we will present a comprehensive description of the DLEE annotation process. The construction of DLEE involved three main steps: event schema construction, candidate selection, and semi-automatic annotation.

3.1 Event schema construction

Event schema plays a crucial role in the event extraction task as the quality of definition directly impacts its ability to obtain the desired results from the model, especially the key information of the event. Recognizing our limited expertise in the legal field, we have sought assistance from 3 legal experts to design the schema. According to the criminal case statistics published on the Chinese Judgment Document Online, we counted the number of criminal cases corresponding to each charge and designated the top 50 criminal charges as our event types. However, as one of the criminal charges involved privacy and confidentiality, accessing the case data without jeopardizing the case's confidentiality was not feasible. Therefore, we selected the 49 criminal charges remaining as our event types. In order to better distinguish between different types of criminal charges, we classified them into 7 top levels based on the category of *Chinese Criminal Law*. For instance, the charge of *dangerous driving* belongs to *Endangering public safety*. Table 2 shows the detailed distribution of event type in DLEE.

Following the work [32], to present the information of a case in the form of an event, we extract a comprehensive event from the entire document, mapping each criminal charge to a sub-level event type. From the perspective of judicial decision-making, key elements of criminal cases have the greatest impact on the outcome of judgment.

Table 1 Statistics of widely-adopted EE datasets (*isDoc*: whether annotated materials are at document-level

Dataset	#isDoc	#Sents or Docs	#EventTypes	#ArgRoles	#Events	Language	Domain
MUC-4	√	1700	5	4	–	English	News
RAMS	√	3993	139	65	9124	English	News
WikiEvents	√	246	50	59	3951	English	News
DocEE	√	27,485	59	356	27,485	English	News
DCFEE	√	2976	4	–	3044	Chinese	Finance
ChFinAnn	√	32,040	5	35	–	Chinese	Finance
Duee-fin	√	11,699	13	92	15,850	Chinese	Finance
DivorceEE*	×	–	13	40	–	Chinese	Legal
CLEE*	×	6538	5	20	6538	Chinese	Legal
DyHiLED*	×	–	11	26	2380	Chinese	Legal
DLEE	√	10,014	49	378	10,014	Chinese	Legal

Sents or Docs: the number of sentences or documents. *EventTypes*: event types. *ArgRoles*: argument roles. *Events*: event instances. Unavailable datasets are denoted by *; while, - indicates unaccessible values.)

Table 2 Top-level event types and sub-level event types, with the number of annotated events

Disrupting market economic order	**Events**	***Infringing property***	**Events**
Illegal absorption of public deposits	176	*Larceny*	1,123
Falsification of VAT invoices	136	*Robbery*	476
Production, sale of toxic and harmful foods	123	*Fraud*	324
Manufacture and sale of counterfeit drugs	112	*Intentional destruction of property*	146
Organize and lead MLM campaigns	105	*Embezzlement*	145
Sales of counterfeit registered trademarks	95	*Misappropriation of funds*	96
Obstruction of Credit card management	94	*Refusal to pay labour remuneration*	94
Misconduct	**Events**	***Bribery and corruption***	**Events**
Dereliction	99	*Taking bribes*	281
Abuse of power	89	*State Assets Offences*	67
Endangering public safety	**Events**	***Disturbing public order***	**Events**
Dangerous driving	853	*Smuggling, trafficking, transport, manufacture of drugs*	584
Traffic accident	376	*Accommodating other people's drug use*	247
Illegal possession of firearms, ammunition	165	*Opening of casinos*	251
Arson	111	*Affray*	188
Major liability accident offence	104	*Obstruction of public duties*	181
Offences against power equipment	104	*Covering up or concealing the proceeds of crime*	159
Damage to flammable explosive equipment	95	*Deforestation*	152
Violating democratic rights	**Events**	*Refusal to enforce sentences or convict*	148
Intentional injury	588	*Poaching*	133
Voluntary manslaughter	268	Alteration, trade in documents and seals of State organs	119
Rape	255	*Environmental pollution*	110
Illegal detention	162	*Illegal mining*	103
Kidnapping	130	*Illegal fishing of aquatic products*	103
Child molestation	109	*Illegal occupation of agricultural land*	96
Trafficking in children and women	105	*Harbouring and concealing criminals*	95
Bigamy	98	*Seduction, retention, introduction to prostitution*	41

Taking *intentional injury* as an example, the *severity of the injury* is a critical element because it directly relates to the judge's decision. However, these key elements often manifest in different forms in different cases. Therefore, we selected 20 documents for each criminal charge to enhance the inclusiveness of key elements. These key elements were then summarized and categorized by three legal experts using the majority voting process, based on the content of the documents. Finally, we determined the argument roles corresponding to each event type by using these key elements. Table 3 displays four examples of event schema in DLEE.

In total, we have defined 49 event types and 378 argument roles. On average, each event type has 7.7 argument roles. Figure 2 presents an instance of annotation using the schema we have established.

3.2 Candidate selection

Candidate document selection. Among the 63,255 cases in the original corpus, more than 49 criminal charges specified in our schema were identified. Consequently, cases with lower frequencies of criminal charge occurrences (i.e., those not falling within our schema) were excluded, retaining only those associated with the 49 criminal charges defined in our schema. Based on the distinct distributions of 49 criminal charges, we initially selected 52,688 documents from a total of 63,255 cases in the original corpus. Criminal judgment documents typically consist of multiple sections, such as the facts of the crime, the analysis of the judgment, and the outcome of sentences. The crucial information in a criminal judgment document is primarily found in the fact statement section. Hence, we utilize it as our annotation text. As revealed by our statistics, the fact statement section in a criminal judgment document typically ranges from 50 to 1,000 words in length. In order to ensure the quality of the dataset, we eliminated any data that fall outside of this range. A total of 31,256 documents that satisfied this constraint were preserved for the next steps of processing. Our primary objective is to extract comprehensive events. Therefore, it is most appropriate to map each individual criminal charge to an event type. However, the presence of multiple charges in a criminal judgment document results in the fragmentation of the fact statement section. For instance, in a criminal judgment document involving both

Table 3 Four examples of sub-level event types and corresponding arguments in DLEE

Affray	Arson
Time	Time
Place	Place
Severity of injury	Reason for arson
Number of brawlers	Method of arson
Crime tool	Arson tool
Motive of the crime	Severity of injury
Compensation amount	Burned item
Attack action	Property damage
Injured area	Post-incident behavior
Method of apprehension	Method of apprehension
Attitude upon apprehension	Attitude upon apprehension
Victim's response	Victim's response
Larceny	**Dangerous driving**
Time	Time
Place	Place
Number of thefts	Vehicle type
Involved amount	Reason for dangerous driving
Stolen item	Alcohol level
Quantity of stolen items	Driver qualification
Crime tool	Property damage
Post-incident behavior	Severity of injury
Method of apprehension	Post-incident behavior
Attitude upon apprehension	Attitude upon apprehension
Victim's response	

robbery and *intentional injury*, the first half may exclusively describe the *robbery*; while, the second half describes the*intentional injury*, which contradicts our initial definition of extracting a comprehensive event from a document. Therefore, we excluded all the criminal judgment documents that involved multiple crimes. After this step, we were left with 18,723 documents. Finally, we filtered out some low-quality data with unclear descriptions of crime fact statements and obtained 10,014 candidate data.

Candidate trigger selection. Trigger candidates provided convenience for trigger annotation. Initially, we engaged three legal experts to review a subset of the selected documents from the previous step. The experts compiled a set of candidate triggers for each type of event by considering the document's content and drawing on their legal expertise. Due to the possibility of incomplete coverage in the trigger candidates, we augment it with NTU Chinese FrameNet Lexicon (CFN-Lex) [33]. CFN-Lex is a total of 36K lexical units that cover 779 frames for FrameNet [34] in Chinese. This resource is extracted from a large-scale bilingual corpus to achieve higher coverage in terms of lexical units, which is helpful in annotation campaigns. Lexical units within the same frame exhibit the characteristics of semantic relevance. Therefore, we retrieve the triggers from CFN-Lex and incorporate the lexical units from the corresponding frame into the trigger candidates. For instance, in Fig. 3, the trigger candidates of *larceny* event include *take away* and *theft*. Therefore, we should include lexical units from the corresponding frame in our trigger candidates. Overall, we gathered 787 candidate triggers across 49 event types. On average, each sub-level event type has 16.1 candidate triggers.

3.3 Semi-automatic annotation

We have devised an efficient and straightforward annotation workflow to ensure the accuracy and quality of data annotation. A total of 30 law students were divided into 10 groups, with each group consisting of one annotator and two independent validators. Prior to the formal annotation, both annotators and validators underwent comprehensive training to increase their familiarity with the entire data annotation process.

Furthermore, we developed a customized annotation system named DLEE Annotation Platform so as to optimize the annotation process and increase efficiency. The workflow of annotation using the DLEE Annotation Platform is depicted in Fig. 4.

DLEE annotation platform. DLEE Annotation Platform plays a vital role in promoting efficient and accurate data annotation. Figure 5 is a screenshot of the DLEE Annotation Platform used for the annotation of a contract fraud case. Its user-friendly interface and convenient features have made it highly popular among annotators. The platform comprises three main modules: 1) **Case previewing module:** We provide annotators and validators with a window to preview the original data. Annotators can select the triggers and argument candidates by simply clicking on the previewed content. 2) **Trigger scanning module:** By combining pre-defined trigger candidates, this module automatically scans the text for trigger words, allowing annotators to select the most appropriate one. 3) **Argument extraction module:** To extract arguments automatically, we initially manually annotated 3000 data samples for training a BERT_question-answering (QA) [18] model. This model has been proven to yield excellent results in argument extraction tasks. We deployed the trained model on the DLEE Annotation Platform to perform preliminary argument extraction on candidate data. Subsequently, in order to continuously improve the performance of the model, the newly annotated data will be used for further training.

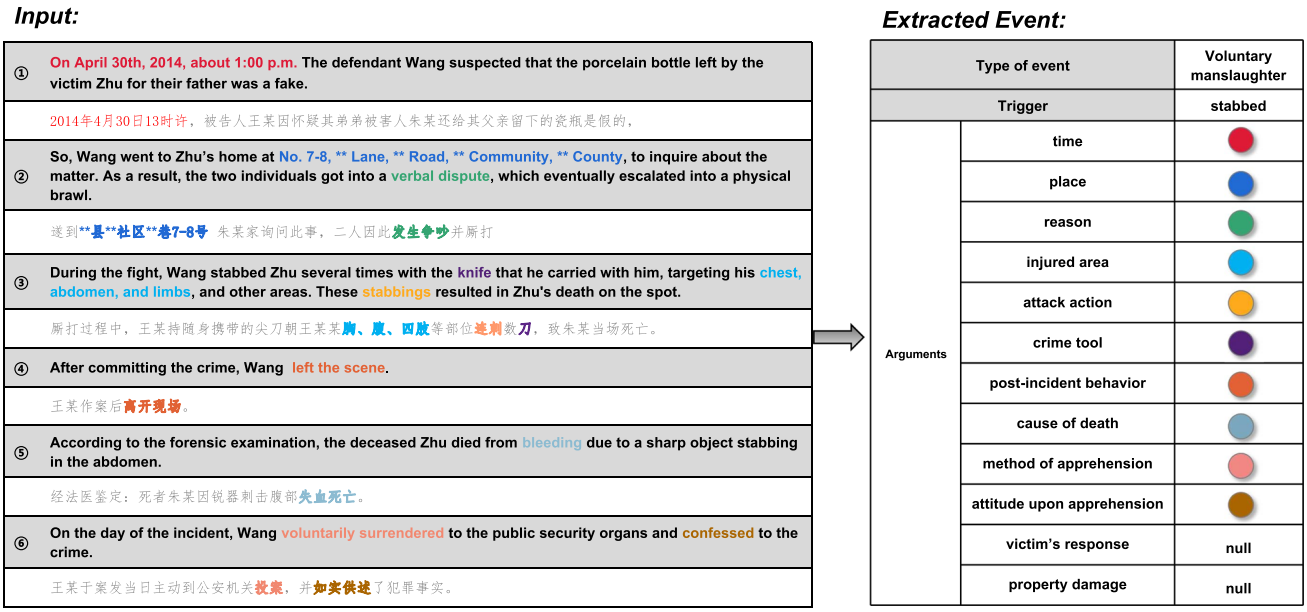

Fig. 2 An instance of the annotation, showing the annotated case on the left and the corresponding annotated result on the right. Distinct colors are used to indicate different argument roles. *victim's response* and *property damage* are not found in this document, so we use *null* to denote. To demonstrate the annotated arguments and real scenarios clearly, we divided the original document into six sentences. Moreover, the corresponding Chinese translation is beneath each English sentence

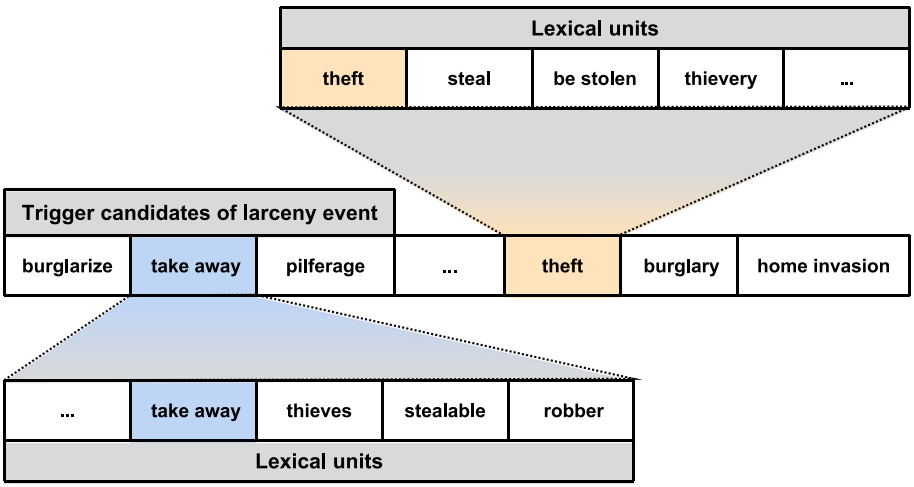

Fig. 3 An example of augmentation for the trigger candidates of *larceny* event utilizing CFN-Lex. Lexical units with the same semantics are included in the trigger candidates

Initial annotation. To familiarize annotators with event schemas, we assigned each group five specific event types, enabling them to focus on annotating and validating candidate data of their respective event types, without the need to handle others. Notably, criminal judgment documents contain the charges of the cases corresponding to event types. Therefore, the annotators are able to identify the event type directly by leveraging the functionality of case previewing. Then, annotators retrieve trigger words in annotation texts utilizing the functionality of trigger scanning. If no match is found in the trigger candidates, annotators will manually select the most appropriate trigger based on the content of the case. The annotators make manual modifications based on the results obtained from the argument extraction module to complete the annotation. The modifications were implemented based on the following key points: 1) If multiple different arguments appear for the same argument role, each of them needs to be annotated individually. 2) If an argument is repeated multiple times in the text, it is annotated only once. 3) If an argument spans an excessively long range, attempts should be made to minimize its span while preserving its original semantics. 4) If an argument is incorrectly assigned to an argument role, it should be manually reassigned.

Multi-round validation. Due to the various limitations of annotators, especially their subjectivity, there will

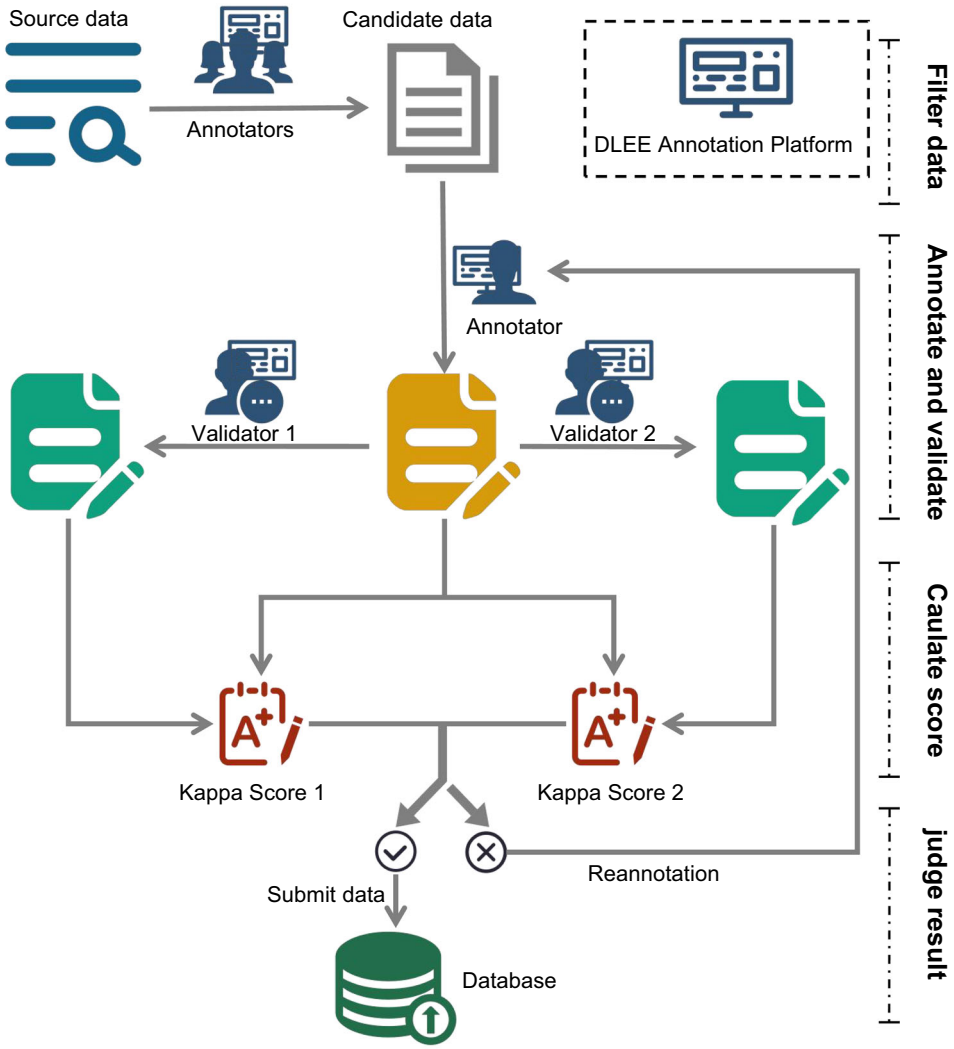

Fig. 4 Semi-automatic annotation process of the DLEE dataset. Initially, candidate data is annotated by an annotator by the DLEE annotation platform. Subsequently, two validators independently examine the annotated data. If *kappa score 1* and *kappa score 2* both exceed a predefined threshold, the annotated case is accepted and stored in the database. Otherwise, the data are returned to the annotator for reannotation

inevitably be irregularities and omissions in the annotation results. Therefore, to ensure the quality of the annotation, we adopted the approach from previous studies [35] by calculating the kappa coefficient [36] as a measure of inter-annotator agreement (IAA). Each case undergoes independent validation by two validators following the initial annotation, which involves modifying any irregular annotations and supplementing any missing arguments. If the initial annotation has been performed effectively, no modifications are necessary during the validation phase. Correspondingly, this annotation will also achieve a higher kappa score. We calculate the IAA between each validation result and the initial annotation result individually, yielding two kappa scores k_1 and k_2. The annotated data will be submitted to our final dataset only if both k_1 and k_2 exceed the predetermined threshold σ ($\sigma \geq 0.7$). Otherwise, a new round of the annotation process will be conducted until the data satisfies our requirements. To enhance flexibility and robustness, we decided to use a dynamic threshold to control the entire annotation process in an adaptive way. The dynamic threshold is calculated based on the continuously updated mean and standard deviation. After manual annotation and validation by law students, the average kappa score of the submitted data reached 0.869, which demonstrates the satisfactory quality of the dataset we constructed. Whenever a new data sample is submitted to the dataset, we utilize the kappa scores obtained during the validation phase of the data to calculate the mean of the kappa scores k by formula (1):

$$\bar{k} = \frac{k_1 + k_2 + \cdots + k_n}{n} \tag{1}$$

and the standard deviation of the kappa scores by formula (2):

$$s_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(k_i - \bar{k})^2} \tag{2}$$

Then, the updated parameters are utilized to calculate the dynamic threshold by formula (3):

Fig. 5 Interface during the utilization of the DLEE annotation platform is depicted in this screenshot. The left-hand box displays the specific details of the case; while, the right-hand box serves as the operational workspace. Annotators utilize this area for selecting charges, as well as annotating trigger words and arguments. The preview of the annotation results is presented at the bottom of the left-hand box, facilitating real-time visibility of the annotation

$$\sigma = \bar{k} + 2s_n \qquad (\sigma \geq 0.7) \tag{3}$$

4 Data analysis

In this section, we present a comprehensive analysis of DLEE. Firstly, we introduce specific parameters of the dataset and compare them with other datasets. Subsequently, we provide a detailed description of the distribution characteristics of the data.

4.1 Statistics of data size

Statistics. We annotated 10,014 distinct criminal judgment documents, with each document corresponding to a specific event type. Our event schema consists of 7 top-level event types, 49 distinct sub-level event types, and 378 argument roles. Overall, we annotated 10,014 events and 99,423 arguments. On average, each event contains 9.9 arguments. All statistical data are presented in Table 1.

Comparisons to related datasets. We compare DLEE with other DEE datasets and SEE legal datasets. 1) **SEE legal datasets**. Compared with DLEE's diverse event type annotations, DivorceEE [12] only covers divorce cases, making it less applicable to the downstream task. CLEE [14] contains a considerable number of annotated larceny cases. However, the scope of event types within the dataset is excessively limited, consisting of only five event types. In comparison, the DLEE dataset encompasses a wider range of event types, totaling up to 49. In DyHiLED [13], the amount of labeled data are relatively small, with 2,380 instances and 11 pre-defined event types; while, DLEE is five times larger. 2) **DEE datasets in other domains**. DocEE [32] is the largest news DEE dataset collected from Wikipedia. The limited 59 event types are inadequate to capture the extensive range of news occurrences in real-life scenarios. DocEE-fin [28] represents the largest compilation of Chinese financial events, playing a significant role in propelling the development of financial EE. However, its applicability in the legal domain is limited.

4.2 Data distribution

Event types and events. Similar to previous DEE datasets [30–32], DLEE conforms to the long-tail distribution, which is the distribution of legal cases in real-world scenarios, as shown in Fig. 6. To preserve the original distribution of data in criminal judgment documents, we annotated different numbers of events corresponding to each event type. The number of annotated event instances for each event type can be found in Table 2.

To further analyze the distribution of event types and events among different top-level event types, we conducted a statistical analysis as shown in Table 4. According to statistics, *Disturbing public order* exhibits the highest proportion, characterized by the largest number of annotated events and diverse event types. The occurrence of the

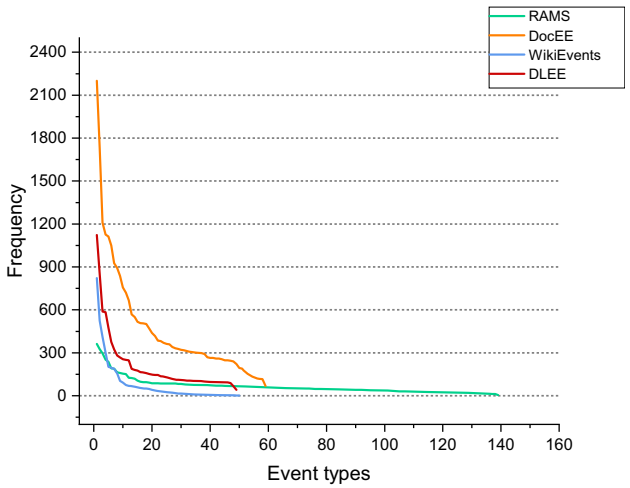

Fig. 6 Distribution of events in four widely used DEE datasets: RAMS, DocEE, WikiEvents, and DLEE

Table 4 Distribution of sub-level event types and events in the top-level event type, with the percentage of events

Top-level event type	#Types	#Events	Percent (%)
Disrupting market economic order	7	841	8.4
Misconduct	2	188	1.9
Infringing property	7	2404	24
Endangering public safety	7	1808	18.1
Disturbing public order	16	2710	27.1
Bribery and corruption	2	348	3.5
Violating democratic rights	8	1715	17.1

Misconduct events is the lowest, representing only 1.9% of the total, which closely matches the actual distribution of charges. The consistent distribution with real scenarios ensures the quality of DLEE.

Argument roles and arguments. DLEE includes up to 99,423 arguments, which provide detailed information for legal events. These arguments play a vital role in the judge's decision-making and the propulsion of similar case retrieval. We conducted a comprehensive analysis of all the arguments and their corresponding argument roles. Then we found that 90 percent of argument roles had multiple arguments, which posed a great challenge to argument extraction. For example, *knife, hammer, brick* could be taken as the value for the argument role of *crime tool* at the same time. Additionally, the arguments range in length from 1 to 33 characters. Based on our annotation result, we found it is sufficient for capturing key information in legal cases. Notably, approximately 70% of the arguments are fewer than 10 characters.

5 Experiments

In this section, we conducted experiments to evaluate the performance of baseline models on DLEE and conducted another additional experiment to analyze the influence of arguments on models' performance. Furthermore, we explored the possibility of explainable legal case matching [37] on DLEE.

Benchmark settings. Following the work [32], we presented similar settings to evaluate the performance of these baseline models. Due to the inability to evenly divide events of certain types. The annotated data for each event type is randomly divided into training, test, and development sets in an approximate ratio of 8:1:1, and the development set contains slightly more events. Table 5 presents three specific divisions for overall performance.

To assess model performance on various top-level event types, the events under the top-level event types are divided into training, validation, and test sets in the same ratio, with the number of events for each top-level event type shown in Table 4.

Baseline models and metrics. Similar to other tasks in the field of IE, in recent years, there has been a proliferation of noteworthy information extraction models in DEE. We selected the following baseline models widely used in other DEE tasks for comparison: 1) **BERT** _QA [38] transforms the EE task into a QA task to mitigate the error propagation associated with entity recognition in EE. It leverages trigger question templates and argument question templates to extract corresponding triggers and arguments in an end-to-end fashion. 2) **PTPCG** [39] involves constructing pruned complete graphs to derive various combinations of event arguments. Subsequently, the results of event types and event argument combinations are integrated to generate the final set of event records. 3) **LIC_-DEE** [40] was presented by the 2021 Language and Intelligence Challenge (LIC2021), and it utilizes ERNIE [41] as an embedding layer and adopts a pipeline mode, which initially detects the event type and extracts the corresponding arguments. The hyper-parameters of models are listed in Appendix.

Table 5 Detailed statistics of the Training set, Development set, and Test set for overall performance

	Train	Dev	Test	Total
#Documents	7982	1025	1007	10,014
#Events	7982	1025	1007	10,014
#Arguments	79,242	10,216	9965	99,423

Table 6 Overall performance on event extraction

Models	Event Classification			Argument Extraction		
	P.	R.	F1.	P.	R.	F1.
BERT_QA	92.36	96.39	94.33	56.00	61.85	58.78
PTPCG	91.20	95.32	93.21	59.33	53.15	56.07
LIC_DEE (ERNIE)	59.69	67.29	63.26	61.61	68.87	65.04
BERT_QA (legal)	94.44	97.34	95.87	57.98	62.06	59.95
PTPCG (legal)	93.28	96.41	94.82	60.72	54.91	57.67
LIC_DEE (legal)	64.53	65.53	65.02	61.43	68.77	64.89

To ensure a fair comparison among different baselines, we employ Precision, Recall, and F1 scores as evaluation metrics for both event classification and argument extraction. Similar to previous work [8, 26, 32], event classification, used for event detection, aims to extract trigger and determine the event type; while, argument extraction extracts the mentioned entity and identifies the corresponding role.

5.1 Overall performance

Firstly, we implemented the aforementioned three baseline models on DLEE. Additionally, in order to further evaluate the impact of different encoders on these models, we changed these models by replacing the embedding layers. Specifically, the baselines without *legal* and *ERNIE* indicate encoding the sentences with Bert-Base-Chinese [42]. LIC_DEE (ERNIE) uses *ERNIE* as the encoder; while, these models with *legal* employ Criminal Document Bert [43], which was trained by 6.63 million criminal judgment documents. Table 6 presents the results on all event types.

We can observe that: For BERT_QA and PTPCG, the model with legal performs better. This indicates the encoder trained with criminal judgment documents learned legal knowledge, resulting in the model's better performance on event classification and argument extraction. BERT_QA (legal) demonstrated decent performance on DLEE, outperforming other models with the F1 score of 95.87% in event classification and 59.95% in argument extraction. Its success can be attributed to its approach of formulating event extraction as a QA task and leveraging question templates to provide valuable information about event types and argument roles. LIC_DEE (ERNIE) demonstrates superior performance in argument extraction, potentially attributed to the incorporation of new phrases in its training corpus and entity-level masking design, as well as the utilization of high-quality Chinese corpora.

However, both LIC_DEE (ERNIE) and LIC_DEE (legal) perform significantly worse in event classification compared to the other two models. This could be attributed to the model design that heavily relies on trigger word identification for event type recognition, which potentially leading to error accumulation. Moreover, this approach overlooks the impact of semantic relationships within the text on classification.

5.2 Top-level performance

To explore the performance of models on different top-level event types, we conducted an experiment using two models: BERT_QA and BERT_QA (legal), which have been demonstrated to have relatively better performance in the aforementioned experiments.

The final results are presented in Table 7, indicating: Criminal Document Bert encoder significantly enhances effectiveness. Except for a 6.91% decrease in argument extraction for *misconduct*. (The Argument Extraction F1

Table 7 Top-level performance on event extraction

Top-level event types	BERT_QA						BERT_QA (legal)					
	Event Classification			Argument Extraction			Event Classification			Argument Extraction		
	P.	R.	F1.	P.	R.	F1.	P.	R.	F1.	P.	R.	F1.
Disrupting market economic order	96.16	97.73	96.94	54.96	61.69	58.13	96.83	98.55	97.68	56.06	61.43	58.62
Misconduct	88.00	92.44	90.17	68.27	65.95	**67.09**	89.30	96.43	92.73	60.86	59.51	**60.18**
Infringing property	94.89	97.74	96.29	50.12	59.40	54.37	96.58	97.55	97.06	52.59	59.92	56.01
Endangering public safety	94.13	96.57	95.34	57.19	64.28	60.53	95.13	98.05	96.57	61.30	65.45	63.31
Disturbing public order	96.19	98.03	97.10	56.09	58.83	57.43	96.47	98.63	97.54	58.03	59.52	58.77
Bribery and corruption	85.50	96.08	90.48	52.74	63.87	57.78	92.64	96.08	94.33	60.52	68.13	64.10
Violating democratic rights	91.89	96.15	93.97	53.55	58.93	56.11	94.25	96.09	95.16	56.95	60.45	58.65

Table 8 Adjusting the ratios of argument roles and their corresponding arguments in the training and development sets, while maintaining the test set unchanged, to evaluate their influence on the model's argument extraction performance

Models	100%			50%			30%		
	P.	R.	F1.	P.	R.	F1.	P.	R.	F1.
BERT_QA	56.00	61.85	58.78	52.67	48.58	50.54	34.36	45.82	39.27
PTPCG	59.33	53.15	56.07	48.68	46.19	47.40	35.61	35.41	35.51
LIC_DEE (ERNIE)	61.61	68.87	65.04	55.01	52.83	53.90	44.03	43.26	43.64
BERT_QA (legal)	57.98	62.06	59.95	50.26	52.13	51.18	42.07	37.17	39.47
PTPCG (legal)	60.72	54.91	57.67	51.43	46.61	48.90	37.59	39.94	38.73
LIC_DEE (legal)	61.43	68.77	64.89	55.35	60.59	57.85	46.50	47.45	46.97

scores for the top-level event type *Misconduct* in the BERT_QA and BERT_QA (legal) models are bolded to indicate the greatest difference between the two F1 scores), the remaining show improvement. This may be attributed to the relatively small proportion of *Misconduct* events in the original training corpus of Criminal Document Bert, resulting in insufficient information learned by the encoder. Top-level event types with fewer events have worse classification performance. *Misconduct* and *Bribery and corruption* exhibit relatively low scores of event classification on both models, which indicates that the number of events greatly influences the performance for event classification. The performance of argument extraction is not obviously related to the number of events. In the BERT_QA, although the number of events for Misconduct is the smallest, it achieves the highest argument extraction score, whereas Disturbing public order, despite having the most events, has a modest argument extraction score. We also found that some event types have more events, but fewer arguments. It implies that the aforementioned phenomenon may be caused by the difference in the number of arguments for the top-level event types.

5.3 Analysis of the argument extraction of DLEE

In order to conduct a more in-depth investigation into the influence of annotating argument granularity on the same baseline models, we adjusted the proportions of argument roles and their corresponding arguments within the training and validation sets, while keeping the test set unchanged. We established three different ratios for comparative analysis. It is worth noting that 100% signifies that no modifications were made to the training, development, and test datasets, and the specific division is the same with Table 5.

As shown in Table 8, we can see that: 1) As we reduced the proportions of arguments and argument roles within the training and development sets, the performance of all baseline models in argument extraction gradually declined. This outcome underscores the significance of fine-grained annotation in the context of argument extraction. 2) When the ratio was adjusted to 50%, the model's performance dropped to approximately 50%, with the most significant decrease being 11.14%. This indicates that the model's generalization capability is relatively weak. Therefore, we believe it is necessary to further enhance the model's generalization ability in future work.

5.4 Discussion

In this section, we mainly discuss the phenomena of model misclassification and extraction errors when facing some complex cases. For an ideal event extraction process, the model first performs event classification based on the input text, followed by the extraction of corresponding arguments according to the event types. Therefore, the prerequisite for obtaining accurate event extraction results is that the model correctly identifies the event type. In most cases, the model is able to accurately obtain the correct event type as expected. When faced with two highly similar event types, the model frequently struggles to discern the distinctions between them, resulting in the acquisition of incorrect event types and extraction of undesired argument information. Figure 7 illustrates an example of misclassification and extraction errors by the model, where the model incorrectly classifies a case belonging to the offense of dangerous driving as an event type of traffic accident and extracts unexpected argument information. We are planning to capture the subtlest feature differences to address the above challenge effectively in the future.

5.5 Case study

Similar case matching. In order to promote the development of LegalAI, we conducted a case study to explore the approach of explainable SCM. It is worth noting that whether or not the two criminal cases are similar was heavily dependent on the annotation at the sentence or document level by legal professionals, while more detailed information about the similarities remains unknown. Fortunately, the application of the DEE model on DLEE provides another feasible method for explainable SCM. The extracted arguments provided detailed information about cases which makes them play a bigger role in

On February 12, 2015, around 3 p.m., the defendant, Mr. Zhang, drove an unlicensed ordinary two-wheeled motorcycle while intoxicated. While traveling from south to north along the *** Highway in *** District, Mr. Zhang collided with an unlicensed electric bicycle driven by Mr. Yuan at the *** section, causing injuries to *** and varying degrees of damage to both vehicles. Upon examination, Mr. Zhang's blood alcohol content was found to be 95.64 milligrams per 100 milliliters.

Dangerous driving		**Traffic accident**	
time	February 12, 2015, around 3 p.m.	time	February 12, 2015, around 3 p.m.
place	the *** Highway in *** District	place	the *** Highway in *** District
vehicle type	electric bicycle	vehicle type	electric bicycle
vehicle type	two-wheeled motorcycle	vehicle type	two-wheeled motorcycle
alcohol level	95.64 milligrams per 100 milliliters	the cause of the accident	unlicensed
severity of injury	causing injuries	severity of injury	causing injuries
reason for dangerous driving	intoxicated		

Fig. 7 A comparison chart between the ground truth and the predicted results of a case. The upper box contains the basic facts of the case, the left box contains the event type and argument information of the ground truth, and the right box contains the event type and argument information predicted by the model

Case1	**S1:** On the evening of June 20, 2019, the defendant Li and his friends were eating supper at ** Night Food Street, during which they drank approximately two cups of Liquor. After, Li drove an ordinary passenger car, passing through ** Road and ** Road. **S2:** At around 23:41 that night, he was caught by the police officers at the intersection of ** Road and ** Road. **S3:** The on-site breath alcohol test showed a result of 103 mg/100 ml. ... **S6:** According to the identification by ** Public Security Bureau's Physical Evidence Appraisal Institute, Li 's blood ethanol content was determined to be 103.2 mg/100 ml. **S7:** Li has no objections to the prosecutor's facts, charges, and sentencing recommendations, and signed the statement and had no objections in court.
Case2	**S1:** The public prosecution alleges that on June 2, 2017, at around 23:26, the defendant Zang was caught by on-duty police officers at the intersection of ** Road and ** Road in ** District, ** City, while driving a car under the influence of alcohol. **S2:** The blood alcohol concentration in Zang's blood was determined to be 161.5 milligrams per 100 milliliters. **S3:** Zang does not dispute the facts, charges, and sentencing recommendations put forward by the public prosecution, and has signed the record of the court session without objection. **S4:** The police officers conducted an on-site breath alcohol test, which showed an alcohol content of 146 milligrams per 100 milliliters. ...

Fig. 8 An example of argument extraction for two similar cases. Sn represents different sentences of the cases

Case1	**S1:**The ** District People's Procuratorate accuses the defendant Yang1 of being overbearing and eager to win. **S2:** ...on August 9, 2013, around 22:00, at the entrance of ** Hotel on ** Street. **S3:**They engaged in mutual fighting with Feng×× and a group assembled by him, including Liu×, Zhang××, Han×, Miao×, etc., wielding machetes, bayonets, and other tools. **S4:**The injuries to Feng××'s head and left neck constitute level two minor injuries, while the injury to his right hand constitutes minor injury; the injury to Zhang××'s left shoulder constitutes minor injury. The defendant Yang1 was later apprehended. **S5:**The prosecution believes that the actions of the defendant Yang1 constitute a certain crime.
Case2	**S1:**The People's Procuratorate of ** District, ** City, ** Province, accuses that at around 3:00 p.m. on May 18, 2014, there was a conflict between Mr. Cao and Mr. Kang, who then agreed to fight. **S2:** At around 6:00 p.m. on the same day, Mr. Cao gathered Mr. Cui (both already sentenced) and the defendant Mr. Liu, along with more than 10 others, wielding steel pipes, machetes, and other deadly weapons, to engage in a group fight with the group gathered by Mr. Kang in ** District, ** City. **S3:** Mr. Kang was injured. **S4:** According to forensic examination, Mr. Kang suffered a fracture of the left ulna and avulsion of the beak bone, affecting the joint surface, classified as a level one minor injury.

Fig. 9 A more complex example of argument extraction for two similar cases, where the case includes additional and more intricate argument information. Sn represents different sentences of the cases

determining the similarity between two cases. Therefore, we employed a BERT_QA model trained on DLEE to extract the arguments from two similar cases in LeCaRD [15], and demonstrate the key elements in these cases.

Arguments in different colors represent the extracted information from two cases shown in Figs. 8 and 9. Figure 9 is a more complex example than Figure 8, as it involves a greater number of arguments, some of which belong to the same argument role. For instance, in Case 1 of Figure 9, *head and left neck*, *right hand*, and *left shoulder* all belong to the argument role of *injured body*

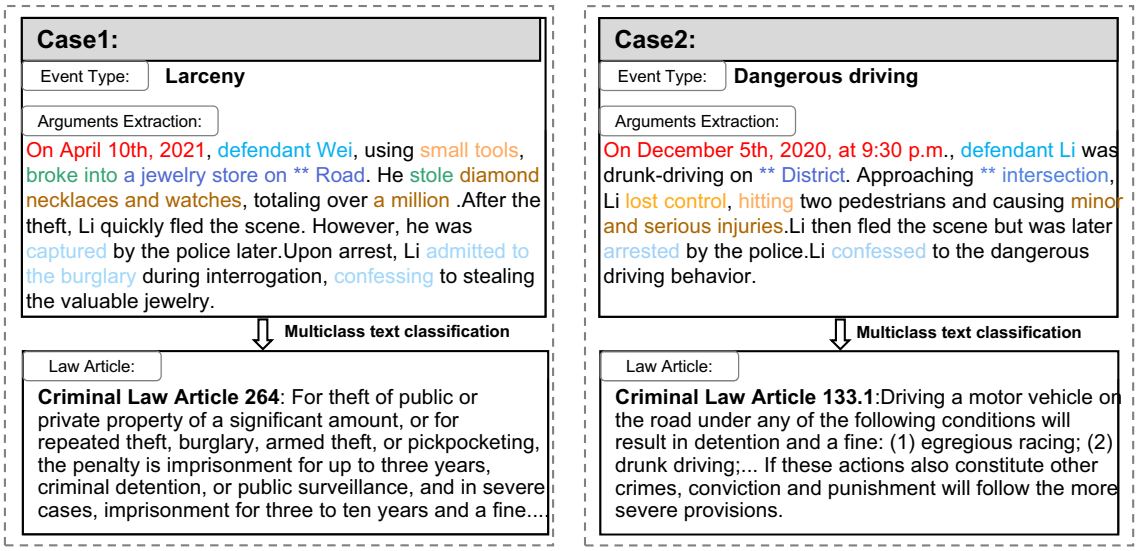

Fig. 10 Two examples of using event extraction to obtain key elements from case facts for predicting law articles

parts. We collected 100 similar cases from LeCaRD and extracted their arguments. In which 68 cases, legal experts confirmed that the extracted arguments are helpful in adjudicating similar cases. Hence, arguments provide specific support for determining the similarity between cases. However, as the model determines the quality of argument extraction, it needs further improvement.

Law article prediction. In order to further clarify the practical significance of extracting event from the basic facts of judgment document, we take the law article prediction task for example. The specific approach is combining previous research and our own work to further explore the facilitating role of key elements extracted from basic case facts to law article prediction.

For the task of law article prediction, in conjunction with the work from [11], the significance of our DLEE dataset lies in its ability to facilitate large-scale event extraction model training (including 10,014 events). Moreover, it allows for the fine-grained extraction of arguments from the case fact, with an average of 9.9 arguments per case fact. Consequently, DLEE can be employed to preprocess the law article prediction data from CAIL2018 [44] by training an event extraction model, thereby enhancing the efficiency of the law article prediction task. Specifically, we utilize an event extraction model to extract information from the case fact, obtaining both its event type (charge) and arguments (key elements). Subsequently, a multi-classification model is employed to achieve the goal of predicting the relevant law article.

Figure 10 illustrates two instances of law article prediction. Case 1 and Case 2 represent events with different topics. Event extraction models are applied first, with the event classification module identifying their respective themes (event types). They include *Larceny* and *Dangerous Driving*. The event arguments extraction module identifies key elements, such as *stole diamond necklaces and watches, valued at a million, minor and serious injuries, confessed*. Building upon this foundation, a multi-classification model is further utilized to accomplish law article prediction, they are *Criminal Law Article 264 and 133.1*.

6 Conclusion and future work

In this paper, we propose the DLEE, the first DEE dataset in the legal area featuring a comprehensive event schema and a substantial number of annotated event instances. DLEE provides significant advantages compared with the existing SEE legal datasets. It comprises 10,014 events and 99,423 arguments with 378 argument roles. In order to improve annotation efficiency and the quality of the dataset, we built a DLEE annotation platform, which is efficient in our work for both semi-automated annotation and multi-round validation. We conducted several experiments to evaluate the performance of baseline models on DLEE. The performance of models is unsatisfactory on DLEE, which requires more effort. In the future, we will be committed to continuously improving the capabilities of legal DEE models on LegalAI tasks.

Appendix: Hyper-parameters of the models

See Table 9.

Table 9 Hyper-parameters of baseline models

BERT_QA		PTPCG		LIC_DEE	
batch_size	8	batch_size	96	batch_size	4
learning_rate	0.001	learning_rate	0.001	learning_rate	0.0005
optimizer	Adam	dropout	0.1	optimizer	Adam
gradient_accumulation_steps	5	optimizer	Adam	warmup_proportion	0.1
warmup_steps	0	hidden_size	768	hidden size	768
hidden_size	768	gradient_accumulation_steps	8	gradient accumulation steps	8

Funding National Natural Science Foundation of China(No.61877051)

Data availability Details of our dataset can be found online at https://anonymous.4open.science/r/DLEE-DATA/README.md. The dataset is available on request.

Declarations

Conflict of interest The authors declared no potential Conflict of interest with respect to the research, authorship, and publication of this article.

References

1. Doddington G, Mitchell A, Przybocki M, Ramshaw L, Strassel S, Weischedel R (2004) The automatic content extraction (ACE) program—tasks, data, and evaluation. In: Proceedings of the fourth international conference on language resources and evaluation (LREC'04). European Language Resources Association (ELRA), Lisbon, Portugal. http://www.lrec-conf.org/proceedings/lrec2004/pdf/5.pdf
2. Gao L, Wu J, Qiao Z, Zhou C, Yang H, Hu Y (2016) Collaborative social group influence for event recommendation. In: Proceedings of the 25th ACM international on conference on information and knowledge management. CIKM '16, pp 1941–1944. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/2983323.2983879
3. Man Duc Trong H, Trong Le D, Pouran Ben Veyseh A, Nguyen T, Nguyen TH (2020) Introducing a new dataset for event detection in cybersecurity texts. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), pp 5381–5390. Association for Computational Linguistics, Online. https://aclanthology.org/2020.emnlp-main.433
4. Du L, Ding X, Xiong K, Liu T, Qin B (2021) ExCAR: Event graph knowledge enhanced explainable causal reasoning. In: Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (Volume 1: Long Papers), pp 2354–2363. Association for Computational Linguistics, Online. https://aclanthology.org/2021.acl-long.183
5. Souza Costa T, Gottschalk S, Demidova E (2020) Event-qa: A dataset for event-centric question answering over knowledge graphs. In: Proceedings of the 29th ACM international conference on information & knowledge management. CIKM '20, pp 3157–3164. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3340531.3412760
6. Sims M, Park JH, Bamman D (2019) Literary event detection. In: Proceedings of the 57th annual meeting of the association for computational linguistics, pp 3623–3634. Association for Computational Linguistics, Florence, Italy. https://aclanthology.org/P19-1353
7. Lai VD, Nguyen MV, Kaufman H, Nguyen TH (2021) Event extraction from historical texts: a new dataset for black rebellions. In: Findings of the association for computational linguistics: ACL-IJCNLP 2021, pp. 2390–2400. Association for Computational Linguistics, Online. https://aclanthology.org/2021.findings-acl.211
8. Chen Y, Xu L, Liu K, Zeng D, Zhao J (2015) Event extraction via dynamic multi-pooling convolutional neural networks. In: Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (Volume 1: Long Papers), pp 167–176. Association for Computational Linguistics, Beijing, China. https://aclanthology.org/P15-1017
9. Zhong H, Xiao C, Tu C, Zhang T, Liu Z, Sun M (2020) How does NLP benefit legal system: a summary of legal artificial intelligence. In: Proceedings of the 58th annual meeting of the association for computational linguistics, pp 5218–5230. Association for Computational Linguistics, Online. https://aclanthology.org/2020.acl-main.466
10. Yao F, Xiao C, Wang X, Liu Z, Hou L, Tu C, Li J, Liu Y, Shen W, Sun M (2022) LEVEN: A large-scale Chinese legal event detection dataset. In: Findings of the association for computational linguistics: ACL 2022, pp. 183–201. Association for Computational Linguistics, Dublin, Ireland. https://aclanthology.org/2022.findings-acl.17
11. Feng Y, Li C, Ng V (2022) Legal judgment prediction via event extraction with constraints. In: Proceedings of the 60th annual meeting of the association for computational linguistics (Volume 1: Long Papers), pp 648–664. Association for Computational Linguistics, Dublin, Ireland. https://aclanthology.org/2022.acl-long.48
12. Li C, Sheng Y, Ge J, Luo B (2019) Apply event extraction techniques to the judicial field. In: Adjunct proceedings of the 2019 ACM international joint conference on pervasive and ubiquitous computing and proceedings of the 2019 ACM international symposium on wearable computers. UbiComp/ISWC '19 Adjunct, pp. 492–497. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3341162.3345608

13. Shen S, Qi G, Li Z, Bi S, Wang L (2020) Hierarchical Chinese legal event extraction via pedal attention mechanism. In: Proceedings of the 28th international conference on computational linguistics, pp 100–113. International Committee on Computational Linguistics, Barcelona, Spain (Online). https://aclanthology.org/2020.coling-main.9
14. Li Q, Zhang Q, Yao J, Zhang Y (2020) Event extraction for criminal legal text. In: 2020 IEEE international conference on knowledge graph (ICKG), pp 573–580. https://doi.org/10.1109/ICBK50248.2020.00086
15. Ma Y, Shao Y, Wu Y, Liu Y, Zhang R, Zhang M, Ma S (2021) Lecard: a legal case retrieval dataset for Chinese law system. In: Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval. SIGIR '21, pp 2342–2348. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3404835.3463250
16. Grishman R, Sundheim B (1996) Message Understanding Conference- 6: a brief history. In: COLING 1996 Volume 1: The 16th international conference on computational linguistics. https://aclanthology.org/C96-1079
17. Mitamura T, Liu Z, Hovy EH (2015) Overview of tac kbp 2015 event nugget track. Theory and Applications of Categories
18. Liu J, Chen Y, Liu K, Bi W, Liu X (2020) Event extraction as machine reading comprehension. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), pp 1641–1651. Association for Computational Linguistics, Online. https://aclanthology.org/2020.emnlp-main.128
19. Wang S, Yu M, Chang S, Sun L, Huang L (2022) Query and extract: refining event extraction as type-oriented binary decoding. In: Findings of the association for computational linguistics: ACL 2022, pp 169–182. Association for Computational Linguistics, Dublin, Ireland. https://aclanthology.org/2022.findings-acl.16
20. Liu S, Li Y, Zhang F, Yang T, Zhou X (2019) Event detection without triggers. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, Volume 1 (Long and Short Papers), pp 735–744. Association for Computational Linguistics, Minneapolis, Minnesota. https://aclanthology.org/N19-1080
21. Tong M, Xu B, Wang S, Cao Y, Hou L, Li J, Xie J (2020) Improving event detection via open-domain trigger knowledge. In: Proceedings of the 58th annual meeting of the association for computational linguistics, pp 5887–5897. Association for Computational Linguistics, Online. https://aclanthology.org/2020.acl-main.522
22. Ahn D (2006) The stages of event extraction. In: Proceedings of the workshop on annotating and reasoning about time and events, pp 1–8. Association for Computational Linguistics, Sydney, Australia. https://aclanthology.org/W06-0901
23. Gupta P, Ji H (2009) Predicting unknown time arguments based on cross-event propagation. In: Proceedings of the ACL-IJCNLP 2009 conference short papers, pp 369–372. Association for Computational Linguistics, Suntec, Singapore. https://aclanthology.org/P09-2093
24. Yang H, Chen Y, Liu K, Xiao Y, Zhao J (2018) DCFEE: a document-level Chinese financial event extraction system based on automatically labeled training data. In: Proceedings of ACL 2018, system demonstrations, pp 50–55. Association for Computational Linguistics, Melbourne, Australia. https://aclanthology.org/P18-4009
25. Zheng S, Cao W, Xu W, Bian J (2019) Doc2EDAG: an end-to-end document-level framework for Chinese financial event extraction. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), pp 337–346. Association for Computational Linguistics, Hong Kong, China https://aclanthology.org/D19-1032
26. Xu R, Liu T, Li L, Chang B (2021) Document-level event extraction via heterogeneous graph-based interaction model with a tracker. In: Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (Volume 1: Long Papers), pp 3533–3546. Association for Computational Linguistics, Online. https://aclanthology.org/2021.acl-long.274
27. Liang Y, Jiang Z, Yin D, Ren B (2022) RAAT: relation-augmented attention transformer for relation modeling in document-level event extraction. In: Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: human language technologies, pp 4985–4997. Association for Computational Linguistics, Seattle, United States. https://aclanthology.org/2022.naacl-main.367
28. Han C, Zhang J, Li X, Xu G, Peng W, Zeng Z (2022) Duee-fin: a large-scale dataset for document-level event extraction. In: Natural language processing and Chinese computing: 11th CCF international conference, NLPCC 2022, Guilin, China, September 24–25, 2022, proceedings, Part I, pp 172–183. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-031-17120-8_14
29. McLean V (1992) Fourth message understanding conference (MUC-4). https://aclanthology.org/M92-1000
30. Ebner S, Xia P, Culkin R, Rawlins K, Van Durme B (2020) Multi-sentence argument linking. In: Proceedings of the 58th annual meeting of the association for computational linguistics, pp 8057–8077. Association for Computational Linguistics, Online. https://aclanthology.org/2020.acl-main.718
31. Li S, Ji H, Han J (2021) Document-level event argument extraction by conditional generation. In: Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies, pp 894–908. Association for Computational Linguistics, Online. https://aclanthology.org/2021.naacl-main.69
32. Tong M, Xu B, Wang S, Han M, Cao Y, Zhu J, Chen S, Hou L, Li J (2022) DocEE: a large-scale and fine-grained benchmark for document-level event extraction. In: Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: human language technologies, pp 3970–3982. Association for Computational Linguistics, Seattle, United States. https://aclanthology.org/2022.naacl-main.291
33. Yang T-H, Huang H-H, Yen A-Z, Chen H-H (2018) Transfer of frames from English FrameNet to construct Chinese FrameNet: a bilingual corpus-based approach. In: Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan. https://aclanthology.org/L18-1139
34. Baker CF, Fillmore CJ, Lowe JB (1998) The Berkeley FrameNet project. In: 36th Annual Meeting of the Association for computational linguistics and 17th international conference on computational linguistics, Volume 1, pp 86–90. Association for Computational Linguistics, Montreal, Quebec, Canada. https://aclanthology.org/P98-1013
35. Artstein R, Poesio M (2008) Survey article: inter-coder agreement for computational linguistics. Comput Linguist 34(4):555–596. https://doi.org/10.1162/coli.07-034-R2
36. Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Measur 20(1):37–46. https://doi.org/10.1177/001316446002000104
37. Yu W, Sun Z, Xu J, Dong Z, Chen X, Xu H, Wen J-R (2022) Explainable legal case matching via inverse optimal transport-based rationale extraction. In: Proceedings of the 45th international acm sigir conference on research and development in information retrieval. SIGIR '22, pp. 657–668. Association for

Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3477495.3531974
38. Du X, Cardie C (2020) Event extraction by answering (almost) natural questions. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), pp 671–683. Association for Computational Linguistics, Online. https://aclanthology.org/2020.emnlp-main.49
39. Zhu T, Qu X, Chen W, Wang Z, Huai B, Yuan N, Zhang M (2022) Efficient document-level event extraction via pseudo-trigger-aware pruned complete graph. In: Proceedings of the thirty-first international joint conference on artificial intelligence, IJCAI-22, pp 4552–4558. International Joint Conferences on Artificial Intelligence Organization, Vienna. Main Track. https://doi.org/10.24963/ijcai.2022/632
40. Contributors P (2021) PaddleNLP: an easy-to-use and high performance NLP library. https://github.com/PaddlePaddle/PaddleNLP
41. Sun Y, Wang S, Li Y, Feng S, Chen X, Zhang H, Tian X, Zhu D, Tian H, Wu H (2019) ERNIE: enhanced representation through knowledge integration. CoRR arXiv:1904.09223
42. Devlin J, Chang M-W, Lee K, Toutanova K (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, Volume 1 (Long and Short Papers), pp 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota. https://aclanthology.org/N19-1423
43. Zhong H, Zhang Z, Liu Z, Sun M (2019) Open Chinese language pre-trained model zoo. Technical Report. https://github.com/thunlp/openclap
44. 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