electronics

MDPI

Article

An Ontology-Based and Deep Learning-Driven Method for Extracting Legal Facts from Chinese Legal Texts

Yong Ren [1], Jinfeng Han [1], Yingcheng Lin [1], Xiujiu Mei [1] and Ling Zhang [2,*]

[1] School of Microelectronics and Communication Engineering, Chongqing University, Chongqing 400044, China; renyong@cqu.edu.cn (Y.R.); hjf16@cqu.edu.cn (J.H.); linyc@cqu.edu.cn (Y.L.); mei_xj@cqu.edu.cn (X.M.)

[2] School of Public Policy and Administration, Chongqing University, Chongqing 400044, China

* Correspondence: zhangling1993@cqu.edu.cn

Abstract: The construction of smart courts promotes the in-deep integration of internet, big data, cloud computing and artificial intelligence with judicial trial work, which can both improve trials and ensure judicial justice with more efficiency. High-quality structured legal facts, obtained by extracting information from unstructured legal texts, are the foundation for the construction of smart courts. Based on the strong normative characteristics of Chinese legal text content and structure composition and the strong text feature learning ability of deep learning, this paper proposes an ontology-based and deep learning-driven method for extracting legal facts from Chinese legal texts. The proposed method utilizes rules and patterns generated in the process of knowledge modeling to extract simple entities, and then extracts complex entities hidden in legal text details with deep learning methods. Finally, the extracted entities are mapped into structured legal facts with clear logical relationships by the Chinese Legal Text Ontology. In the information extraction test of judicial datasets composed of Chinese legal texts on theft, the proposed method effectively extracts up to 38 categories of legal facts from legal texts and the number of categories extracted increases significantly. Among them, the rule-based extractor obtains an F1-score of 99.70%, and the deep learning-driven extractor obtains an F1-score of 91.43%. Compared with existing methods, the proposed method has great advantages in extracting the completeness and accuracy of legal facts.

Keywords: information extraction; ontology; BERT; Bi-LSTM; CRF; Chinese legal texts

Citation: Ren, Y.; Han, J.; Lin, Y.; Mei, X.; Zhang, L. An Ontology-Based and Deep Learning-Driven Method for Extracting Legal Facts from Chinese Legal Texts. *Electronics* **2022**, *11*, 1821. https://doi.org/10.3390/electronics11121821

Academic Editor: Arkaitz Zubiaga

Received: 10 May 2022
Accepted: 7 June 2022
Published: 8 June 2022

1. Introduction

In recent years, the number of unstructured Chinese legal texts has exploded with the continuous improvement of the legal system. A legal text is a legally binding written conclusion made by the court based on the facts of the case and the law, and represents the richest source of legal information [1]. The legal information contained in these texts can not only help judicial personnel and lawyers to understand the whole case, but also serve as the data basis for downstream legal applications such as knowledge graphs [2,3], case databases [4,5], information retrieval [6], and question answering systems [7]. However, due to the unstructured and noisy nature of legal texts, computers cannot directly obtain the legal information contained in them. In addition, there are currently more than 100 million legal texts on China Judgment Documents Online. This amount makes it difficult for humans, even legal professionals, to extract the legal information from legal texts quickly. In order to help judicial personnel to understand the whole case quickly and meet the needs of the downstream legal applications, it is crucial to study the information extraction method for automatically extracting legal information from Chinese legal texts.

Many studies have attempted to extract information from legal texts using a variety of techniques, including rule-based, ontology, machine learning, neural networks, and some linguistic methods [8–12]. However, applying these methods directly to Chinese legal texts

cannot achieve satisfactory results. First, most existing methods cannot be used to extract Chinese legal texts due to language limitations and the lack of datasets (the problem of language) [8,9,11,12]. Furthermore, the results of existing methods for extracting Chinese legal texts are all sentences, and the legal facts are still hidden in the sentence details [1,10]. Second, existing methods ignore the semantic relationship between legal facts because they do not model legal text knowledge. At the same time, existing methods tend to extract common entities in legal texts, such as person names, place names, and organization names, which are insufficient for downstream legal applications (the problem of incomplete information). Third, legal texts are usually written in a standard format, and each legal fact always exists in a fixed logical segment. However, existing methods directly regard the problem of information extraction as a named entity recognition task, ignoring the potential impact of legal text structural characteristics on entity labels (the problem of extraction accuracy).

To address the above issues, this paper proposes an ontology-based and deep learning-driven method, aiming at extracting legal facts from Chinese legal texts. First, an ontology known as the Chinese Legal Text Ontology (CLTO) is constructed by knowledge modeling of the Chinese legal text, including general ontology and special ontology. Ontology is an emerging research method in recent years which integrates the structural characteristics of a text and reduces semantic ambiguity. The CLTO is an improvement of Judicial Case Ontology [12], which is not suitable for Chinese legal texts. The entire legal text is then divided into several predefined logical segments using structural characteristics. Next, the corresponding entities are extracted from each logical segment using a hybrid method of rules and deep learning. For simple entities, the rules and patterns are generated in the process of knowledge modeling to extract entities. For complex entities, the pre-trained model Bidirectional Encoder Representations from Transformers (BERT) [13] is introduced into the task of legal text information extraction, and is combined with Bi-directional Long Short-Term Memory (Bi-LSTM) [14] and Conditional Random Field (CRF) [15] algorithms to extract entities. It is effective in solving the problem of polysemy in legal texts. Finally, the CLTO is used to map the extracted entities into structured legal facts with clear logical relationships. This study takes Chinese legal texts on theft from 2018 to 2021 as the corpus for knowledge modeling and evaluates the proposed method using manually annotated judicial datasets and the CAIL2021_IE dataset [16].

2. Related Work

This section presents related works on information extraction from legal texts. These works use a variety of methods, including rules, traditional machine learning such as CRF, ontology, deep learning such as LSTM, and hybrid methods.

2.1. Rule-Based Method

Earlier legal text information extraction systems mainly used rules and some language methods. The extraction performance of these systems is highly dependent on hand-crafted rules and patterns. Moens et al. [17] used paragraph classification and sentence analysis to extract information such as dates and case names from Belgian legal texts. Zhuang et al. [1] used regular expressions and feature dictionaries to extract basic case information from Chinese judgment documents. Solihin et al. [8] decomposed the problem of legal text information extraction into three sub-tasks of structure extraction, tokenization and entity extraction, each of which is implemented using a rule-based method, and finally successfully extracted a series of legal events from Indonesian judgment documents.

2.2. Traditional Machine Learning Method

The rule-based system has good extraction performance, but can only extract some information with fixed characteristics, and cannot extract information hidden in the details of legal text. The emergence of machine learning methods has solved this problem very well. Bach et al. [18] extracted key information from Vietnamese transport legal texts using

a CRF-based method. Iftikhar et al. [9] constructed an information extraction system called PULMS using CRF, Maximum Entropy (MaxEnt), and Trigram N Tag (TNT) algorithms, with the CRF algorithm achieving the best results. In addition, there are some studies that combined rules and traditional machine learning methods to get better extracted results. Dozier et al. [19] used a hybrid method of lookup, pattern rules, and statistics to extract legal facts such as judges, attorneys, and courts from U.S. case law. Andrew et al. [20] used a hybrid method of regular expressions and CRF to extract legal facts such as names, organizations, and personas from Luxembourg legal texts. Compared with the single method, the hybrid method improves both precision and recall.

2.3. Ontology-Based Method

Ontology-based methods have been proven to be the most suitable for extracting information from domain-specific texts [12]. Ontology makes domain knowledge explicit, and its result is more suitable for downstream legal applications. In recent years, ontology has also been used in the study of legal text information extraction. Buey et al. [21] used ontology and document cleaning methods to extract information such as document parameters and personnel parameters from Spanish notarization behavior. Araujo et al. [22] constructed a domain ontology called ODomJurBR, and used language rules to automatically extract information such as formal charges, convictions, and interrogations from Brazilian legal texts. Thomas et al. [12] proposed a knowledge-driven, semi-supervised pattern learning method to extract legal facts from judgement documents. This method used an ontology called Judicial Case Ontology to generate seed patterns to speed up the extraction process.

2.4. Deep Learning Method

In recent years, Deep Neural Networks (DNNs) have been widely used in a variety of Natural Language Processing (NLP) tasks. DNNs perform better than traditional machine learning in many fields, such as named entity recognition and text classification. There are two main types of DNNs, namely Convolutional Neural Networks (CNN) and Recursive Neural Networks (RNN). RNNs are an ideal choice for sequential items, such as text and speech [23]. At present, most studies of legal text information extraction use LSTM, a variant of RNN [24]. For example, Rao et al. [25] proposed a Multi-Bi-LSTM framework to extract legal facts such as parties' claims and judgments from Chinese civil legal texts, and the extracted results were sentences. Ji et al. [10] regarded the information extraction task as classification and extraction multi-task learning, and proposed an end-to-end joint model called JBLACN, which used a Bi-LSTM layer as the shared encoding layer for both tasks. Nuranti et al. [11] evaluated deep learning methods like CNN, LSTM+CRF, Bi-LSTM+CRF, etc., and machine learning methods like CRF, etc. methods on Indonesian legal texts, where the Bi-LSTM+CRF model achieved the highest accuracy. In addition, some studies [26] used Gated Recurrent Unit (GRU), another variant of RNN. However, in the task of legal text information extraction, the performance of GRU is not as good as that of LSTM.

According to the reviewed literature, it concludes that most studies focus on dealing with legal texts in other languages and rarely on Chinese legal texts. Furthermore, most studies only extract some generic entities, such as person names, place names, and organization names, which are insufficient for downstream legal applications. Therefore, this paper proposes a new information extraction method, which uses techniques such as rules, ontology, and deep learning, aiming at efficiently extracting more legal facts from Chinese legal texts.

3. Proposed Method

This section briefly introduces the proposed legal text information extraction method. Figure 1 shows the overall architecture of the proposed method, which is divided into four parts: knowledge modeling, preprocessing, paragraph classification, and fact extraction. The input is an unstructured legal text (DOC file) and the output is a structured legal

fact (JSON file). In this study, the legal text information extraction task is identified as a combined task of paragraph classification and fact extraction. The main idea of the proposed method is to use the strong normative characteristics of Chinese legal text content and structure composition to model their domain knowledge in order to obtain a knowledge model and corresponding extraction patterns to extract simple legal facts, and use deep learning methods to extract complex legal facts that cannot be extracted by rule-based methods. The following subsections describe the whole process in detail.

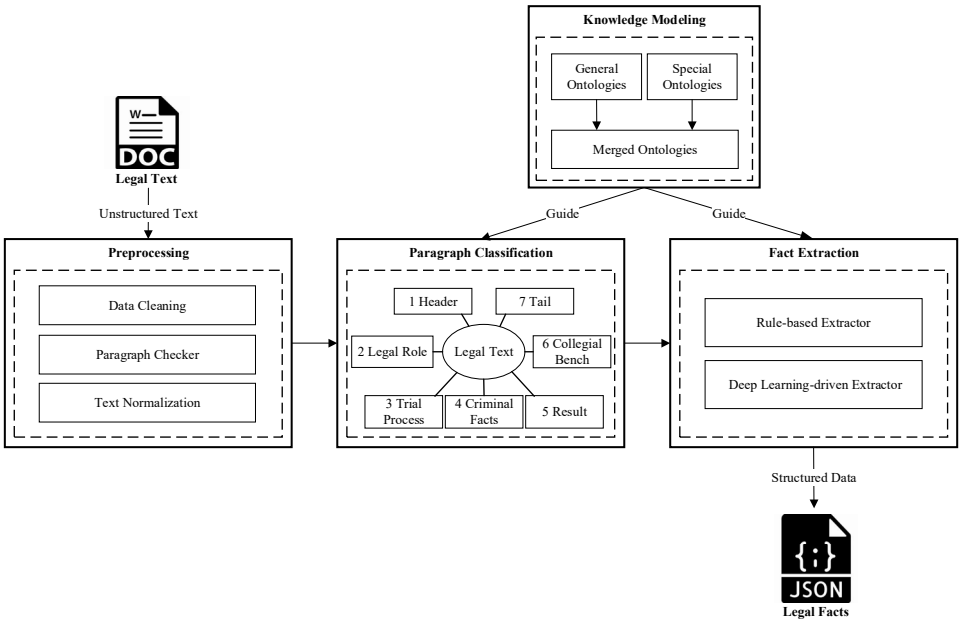

Figure 1. Overall architecture of the proposed method.

3.1. Knowledge Modeling

Because Chinese legal texts are long and complex and contain multiple legal facts, we first model the domain knowledge and structural characteristics in Chinese legal texts, and the modeling results are used to guide the process of paragraph classification and fact extraction. Knowledge modeling uses an ontology-based method that is widely used in the study of domain-specific information extraction. This study uses the Stanford Seven-step method to construct an ontology called CLTO. First, the domain category (Chinese legal text) of CLTO is determined. Then, the CLTO overloads existing ontology (Judicial Case Ontology [27]). Next, the important terms in CLTO are listed and the class hierarchy in CLTO is defined. Class hierarchy definition methods include top-down (constructed by ontology engineers), bottom-up (text extraction and semantic analysis), and intermediate methods (extension of a set of core concepts) [6]. This study uses the top-down method to define the class hierarchy of CLTO. Finally, the object properties and data properties of the CLTO are defined, and a set of individuals is created.

The CLTO includes general ontologies and special ontologies and is constructed by combining these two ontologies. General ontologies describe concepts and relationships common to legal texts. Special ontologies describe concepts and relationships specific to legal texts. For example, *defendant* exists in all types of legal texts (all types of legal texts have this concept), so it is divided into general ontologies. *Stolen item* only exists in the legal text on theft (only the legal text on theft has this concept), so it is divided into special ontologies. This classification is done to more quickly model knowledge of other types of legal texts. For example, Figure 2 models the domain knowledge of legal texts on *theft*. *Robbery* and *bribery* legal texts can be modeled quickly (only the concepts in the special ontology need to be modified). The CLTO is developed using Protege software [28]. Protege OWL content is automatically converted to XML files for ease of use during paragraph classification and fact extraction. Taking the Chinese legal text on theft as an example, Figure 2 shows a

snapshot of concepts in CLTO. The CLTO includes the concepts and relationships existing in Chinese legal texts on theft. This is similar to the Judicial Case Ontology defined in Reference 27, but the CLTO defined in this study is more applicable to Chinese legal texts. This is because CLTO is an improvement of Judicial Case Ontology [27], and the concepts and relationships defined in CLTO are better matched with Chinese legal texts. The CLTO can be obtained from https://github.com/HJF97/Chinese-Legal-Text-Ontology (accessed on 3 April 2022).

Figure 2. A snapshot of concepts in Chinese Legal Text Ontology.

3.2. Preprocessing

As shown in Figure 1, the input to the proposed method is an unstructured legal text. First, the input text needs to be preprocessed, including three sub-tasks of data cleaning, paragraph checking, and text normalization. The data cleaning removes non-ASCII word noise data contained in the text. The paragraph checker judges the completeness of a paragraph by the punctuation at the end of the paragraph, and corrects the truncated paragraphs. The text normalization replaces abbreviated forms with standard forms, such as replacing phrases describing dates (same day, next day) with standard forms (year-month-day). These three subtasks are necessary because noisy data, truncated paragraphs, and abbreviations all affect the process of paragraph classification and fact extraction.

3.3. Paragraph Classification

After preprocessing the legal text, the proposed method divides the preprocessed legal text into seven logical segments based on structural characteristics. In the Chinese

legal system, each legal text is written in a fixed order, and the extracted legal facts are always in a fixed logical segment [1]. Therefore, compared to directly extracting the whole legal text, the results of paragraph classification can effectively reduce the complexity and coupling of the fact extraction process. The paragraph classification is implemented using a rule-based method. First, the preprocessed legal text is stored in a string array according to the linefeed character. Then, corresponding rules and patterns are formulated according to the characteristics of each logical segment, such as location and keywords. Finally, the logical segment is matched from the string array using regular expressions. Table 1 describes the structural characteristics of each logical segment in Chinese legal texts.

Table 1. Structural characteristics of each logical segment in Chinese legal texts.

Logical Segment	Description	Characteristic
Header	Including trial court, document type and case number paragraphs	At the beginning of legal text No punctuation in the paragraph
Legal role	Including public prosecution organ, defendant and advocate paragraphs	After the case number paragraph Including keywords such as *public prosecution organ*, *defendant* and *advocate*
Trial process	Including participants, trial time and trial status information	Only one paragraph Including indictment number
Criminal fact	Including criminal fact paragraphs	After the trial process paragraph Before the result paragraph
Result	Including legal provision and judgment result paragraphs	Including keywords such as *the court considers* and *decides as follows* Before the collegial bench paragraph
Collegial bench	Including chief judge and judge paragraphs	Including keywords such as *chief judge* and *judge*
Tail	Including date of judgment, clerk and assistant judge paragraphs	At the end of legal text Including keywords such as *clerk* and *assistant judge*

3.4. Fact Extraction

After paragraph classification, the proposed method performs the fact extraction module. This module has two main components: rule-based and deep learning-driven extractors. The rule-based extractor is used to extract legal facts with specific trigger words. The deep learning-driven extractor is used to extract legal facts that cannot be extracted by the rule-based extractor. The following subsubsections describe these two extractors in detail.

3.4.1. Rule-Based Extractor

The rule-based extractor extracts legal facts from Chinese legal texts using rules and patterns generated in the process of knowledge modeling. Due to the normative nature of Chinese legal text writing, grammatical rules become a key processing resource for fact extraction tasks. In Chinese legal texts, most legal facts can be extracted by characteristics such as keywords, symbols, positions, and numbers. These characteristics are obtained from an in-depth analysis of legal texts in the process of knowledge modeling (see Section 3.1). The following describes the processing process of the rule-based extractor.

First, the relationship between legal facts is marked according to the subparagraph order. The logical paragraph is then divided further into subparagraphs. Next, legal facts are extracted from the subparagraphs using regular expressions. Finally, the extracted results are processed by length filtering, deduplication and digital conversion. The length filtering process removes unreasonable extracted results. The deduplication process removes the

same extracted results. The digital conversion process converts Chinese numerals into Arabic numerals. Figure 3 describes sample rules and sentences for extracting legal facts from each subparagraph.

Sub-paragraph	Legal Fact	Sample Rule	Sample Sentence
Trial court	1: Court	^[\u4e00-\u9fa5]{1,30}法院$	江苏省淮安市淮安区人民法院[1] **Translation:** Huaian District People's Court, Huaian City, Jiangsu Province[1]
Document type	1: Document type	^[\u4e00-\u9fa5]{1,10}(?=判决书)	刑事[1]判决书 **Translation:** Criminal[1] judgement
Case number	1: Case number	^（\d*）.*号$	（2021）苏0803刑初439号[1] **Translation:** （2021）Su 0803 XC No. 439[1]
Public prosecution organ	1: Public prosecution organ	(?<=公诉机关)([\u4e00-\u9fa5]{1,30}检察院)	公诉机关淮安市淮安区人民检察院[1]。 **Translation:** Public prosecution organ: Huaian District People's Procuratorate of Huaian City[1]
Defendant	1: Name of defendant 2: Gender of defendant 3: Birthday of defendant 4: Nation of defendant 5: Registered residence of defendant 6: Birthplace of defendant 7: Educational level of defendant 8: Current residence of defendant	(?<=被告人)(([\u4e00-\u9fa5]){2,8})(?=[，。]) (?<=[，。])([男女]) (?=[，。])(\d{4}年\d{1,2}月\d{1,2}日)(?=出生) (?<=[，。])([\u4e00-\u9fa5]{1,6}族)(?=[，。]) (?<=户籍地|所在地)([\u4e00-\u9fa5A-Za-z0-9]{1,20})(?=[，。]) (?<=出生地|出生于)([\u4e00-\u9fa5A-Za-z0-9]+)(?=[，。]) (?<=[，。])([\u4e00-\u9fa5]{1,10}(?=文化|毕业))(文盲(?=[，。])) (?<=住)([\u4e00-\u9fa5A-Za-z0-9]+)(?=[，。])	被告人石某某[1]，曾用名石XX，男[2]，1970年5月1日[3]出生于淮安市淮安区[6]，汉族[4]，小学[7]文化水平，无业，户籍地淮安市淮安区[5]，现住淮安市淮安区[8]。因涉嫌犯盗窃罪… **Translation:** Defendant XX Shi[1] , once used the name XX Shi, male[2], born on May 1, 1970[3] in Huaian District, Huaian City[6], Han[4] nationality, primary school[7] education level, unemployed, registered in Huai'an District, Huaian City[5], and now lives in Huaiyin District, Huaian City[8]. Suspected of committing theft... 被告人董某[1]，曾用名董X，女[2]，1974年11月20日[3]出生于淮安市淮阴区[6]，汉族[4]，初中[7]文化水平，农民，户籍地淮安市淮阴区[5]，现住淮安市淮阴区[8]。因涉嫌犯盗窃罪… **Translation:** Defendant X Dong[1], formerly known as X Dong, female, born on November 20, 1974[3] in Huaiyin District, Huaian City[6], Han[4] nationality, junior high school[7] education level, farmer, registered in Huaiyin District, Huaian City[5] , now lives in Huaiyin District[8], Huaian City. On suspicion of theft...
Advocate	1: Name of advocate 2: Work unit of advocate	(?<=辩护人)([\u4e00-\u9fa5]{2,8}(?=[，。]) (?<=[，。])([\u4e00-\u9fa5A-Za-z0-9]+(?=律师)	辩护人郎白[1]，浙江君安世纪律师事务所[2]律师。 **Translation:** Advocate Bai Lang[1], lawyer of Zhejiang Junan Century Law Firm[2]. 辩护人江耀乘[1]，广东德比律师事务所[2]律师。 **Translation:** Advocate Yaorong Jiang[1], lawyer of Guangdong Derby Law Firm[2].
Trial process	1: Indictment number 2: Date of public prosecution 3: Inquisitor	(?<=以|据)([\u4e00-\u9fa5]{1,5}[0-9（）]+号)(?=起诉书) (?<=于)(\d{4}年\d{1,2}月\d{1,2}日)(?=[\u4e00-\u9fa5]+公诉) (?<=检察[官长员])(([\u4e00-\u9fa5、])+)(?=出庭)	淮安市淮安区人民检察院以淮检刑诉（2021）87号[1]起诉书指控被告人石某某、董某犯盗窃罪，于2021年12月3日[2]向本院提起公诉…淮安市淮安区人民检察院指派检察员潘某[3]、杨某[3]出庭支持公诉… **Translation:** The People's Procuratorate of Huaian District, Huaian City[1] charged the defendants XX Shi and X Dong with the crime of theft with Huai Jian Criminal Prosecution (2021) No. 87, and filed a public prosecution with this court on December 3, 2021[2]...Assignment from the Huai'an District People's Procuratorate of Huai'an City Inquistors X Pan[3] and X Yang[3] appeared in court to support the prosecution...
Relevant legal provision	1: Legal provision name 2: Legal provision number	(《[\u4e00-\u9fa5]{1,20}》) ((第[\u4e00-\u9fa5]+条)(((第[\u4e00-\u9fa5]+款))*))	据此，依照《中华人民共和国刑法》[1]第二百六十四条[2]、第六十七条第三款[2]、第六十四条[2]和《中华人民共和国刑事诉讼法》[1]第十五条[2]之规定，判决如下： **Translation:** In accordance with the provisions of Article 264[2], Paragraph 3 of Article 67[2], Article 64[2] of the Criminal Law of the People's Republic of China[1] and Article 15[2] of the Criminal Procedure Law of the People's Republic of China[1], the judgment is as follows:
Judgment result	1: Name of sentenced 2: Charge 3: Prison term 4: Fine	(?<=被告人)([\u4e00-\u9fa5]{2,8})(?=犯) (?<=犯)([\u4e00-\u9fa5]+罪)(?=[，。]) (?<=判处|执行，)((无期徒刑|有期徒刑|拘役|管制|缓刑)(([\u4e00-\u9fa5]{1,4}年)?([\u4e00-\u9fa5]{1,4}月)?)) (?<=罚金)([\u4e00-\u9fa50-9]+元)	一、被告人石某某[1]犯盗窃罪[2]，判处有期徒刑十个月[3]，并处罚金人民币五千元[4]（刑期自判决执行之日起计算...）。 **Translation:** 1. Defendant XX Shi[1] committed the theft[2] and was sentenced to ten months in prison[3] and a fine of RMB 5,000[4] (the sentence shall be calculated from the date of execution of the judgment...). 二、被告人董某[1]犯盗窃罪[2]，判处有期徒刑八个月[3]，并处罚金人民币三千元[4]（刑期自判决执行之日起计算...）。 **Translation:** 2. Defendant X Dong[1] committed the theft[2] and was sentenced to eight months in prison[3] and a fine of RMB 3,000[4] (the sentence shall be calculated from the date of execution of the judgment...).
Collegial bench	1: Chief judge 2: Judge	(?<=^审判长)([\u4e00-\u9fa5]{2,8}) (?<=^审判员)([\u4e00-\u9fa5]{2,8})	审判长何某某[1] **Translation:** Chief judge: XX He[1] 审判员张某[2] **Translation:** Judge: X Zhang[2] 审判员陈某某[2] **Translation:** Judge: XX Chen[2]
Assistant judge	1: Assistant judge	(?<=法官助理)([\u4e00-\u9fa5]{2,8})	法官助理何某[1] **Translation:** Assistant judge: X He[1]
Date of judgment	1: Date of judgment	^(.{4})年(.{1,2})月(.{1,3})日$	二〇二一年十二月十五日[1] **Translation:** Dec. 15, 2021[1]
Clerk	1: Clerk	(?<=书记员)([\u4e00-\u9fa5]{2,8})	书记员王某某[1] **Translation:** Clerk: XX Wang[1]

Figure 3. Sample rules and sentences for extracting legal facts from each sub-paragraph. The extracted legal facts are distinguished by color and the superscript numbers.

Algorithm 1 is the pseudocode of the rule-based extractor. Due to space limitations, only the fact extraction process of the *legal role* logical segment is described here, and the extraction process of other logical segments is similar. The input to the algorithm is a matching pattern of logical segments and legal facts. Here, the matching pattern is stored in a *map*, because there are multiple legal facts in a paragraph. For example, the defendant paragraph contains legal facts such as name, gender, and birthday. The output to the algorithm is the extracted legal facts and relationships. Legal facts are obtained by scanning phrases in paragraphs and matching predefined patterns. Relationships are obtained through subparagraph order and predefined relationship maps in the process of knowledge modeling. The algorithm processes only one subparagraph per loop until the whole logical segment ends.

Algorithm 1. Rule-based Extractor.

Input:
LR_{LS}: Legal role logical segment
PPO_P: *Map* of patterns in the paragraph of public prosecution organ
D_P: *Map* of patterns in the paragraph of defendant
A_P: *Map* of patterns in the paragraph of advocate
Output:
D_R: *Map* of relationship in the paragraph of defendant
PPO_F: *Map* of legal facts in the paragraph of public prosecution organ
D_F: *Map* of legal facts in the paragraph of defendant
A_F: *Map* of legal facts in the paragraph of advocate
Initialize $D_R \leftarrow \emptyset$ $PPO_F \leftarrow \emptyset$, $D_F \leftarrow \emptyset$, $A_F \leftarrow \emptyset$;
for each paragraph $P \in LR_{LS}$ **do**
    i←Number of current paragraph in LR_{LS}
    if P_i is defendant paragraph and P_{i+1} is advocate paragraph **then**
        D_R←relationship mark between defendant and advocate
    end if
    if P_i is public prosecution organ paragraph **then**
        for all phrase$\in P_i$ that match each pattern$\in PPO_P$ **do**
            PPO_F←phrase
        end for
    else if P_i is defendant paragraph **then**
        for all phrase$\in$P that match each pattern$\in D_P$ **do**
            D_F←phrase
        end for
    else if P_i is advocate paragraph **then**
        for all phrase$\in$P that match each pattern$\in A_P$ **do**
            A_F←phrase
        end for
    else
        continue//There are no legal facts to extract from this paragraph.
    end if
end for
return D_R, PPO_F, D_F, A_F

However, the rule-based extractor is not sufficient to extract all categories of legal facts from Chinese legal texts. For example, in the sentence *Zhang broke the door lock of Li's house with a hammer*, the legal fact that needs to be extracted is the *hammer*. Without specific trigger words, the rule-based extractor cannot handle such legal facts. A deep learning-driven extractor is used to extract these legal facts. The following subsubsection describes how to use deep learning methods to extract legal facts hidden in textual details from Chinese legal texts.

3.4.2. Deep Learning-Driven Extractor

A logical segment of criminal facts in a Chinese legal text is represented as a sentence set $S = \{s_1, s_2, \cdots, s_m\}$, where s_m is the mth sentence in the logical segment. The goal is to extract legal facts from these sentences. For example, in sentence *Zhang broke the door lock of Li's house with a hammer*, the legal fact that needs to be extracted is the *hammer* (These legal facts have no specific trigger words). As in existing studies [10,11,26], the fact extraction problem is modeled as a sequence labeling task. The main idea of the model is to introduce the pre-trained model BERT [13] as an embedding layer to solve the polysemy problem in Chinese legal texts. At the same time, Bi-LSTM [14] and CRF [15] algorithms are incorporated into the model structure. As shown in Figure 4, the proposed extractor consists of an embedding layer, an encoding layer and an inference layer.

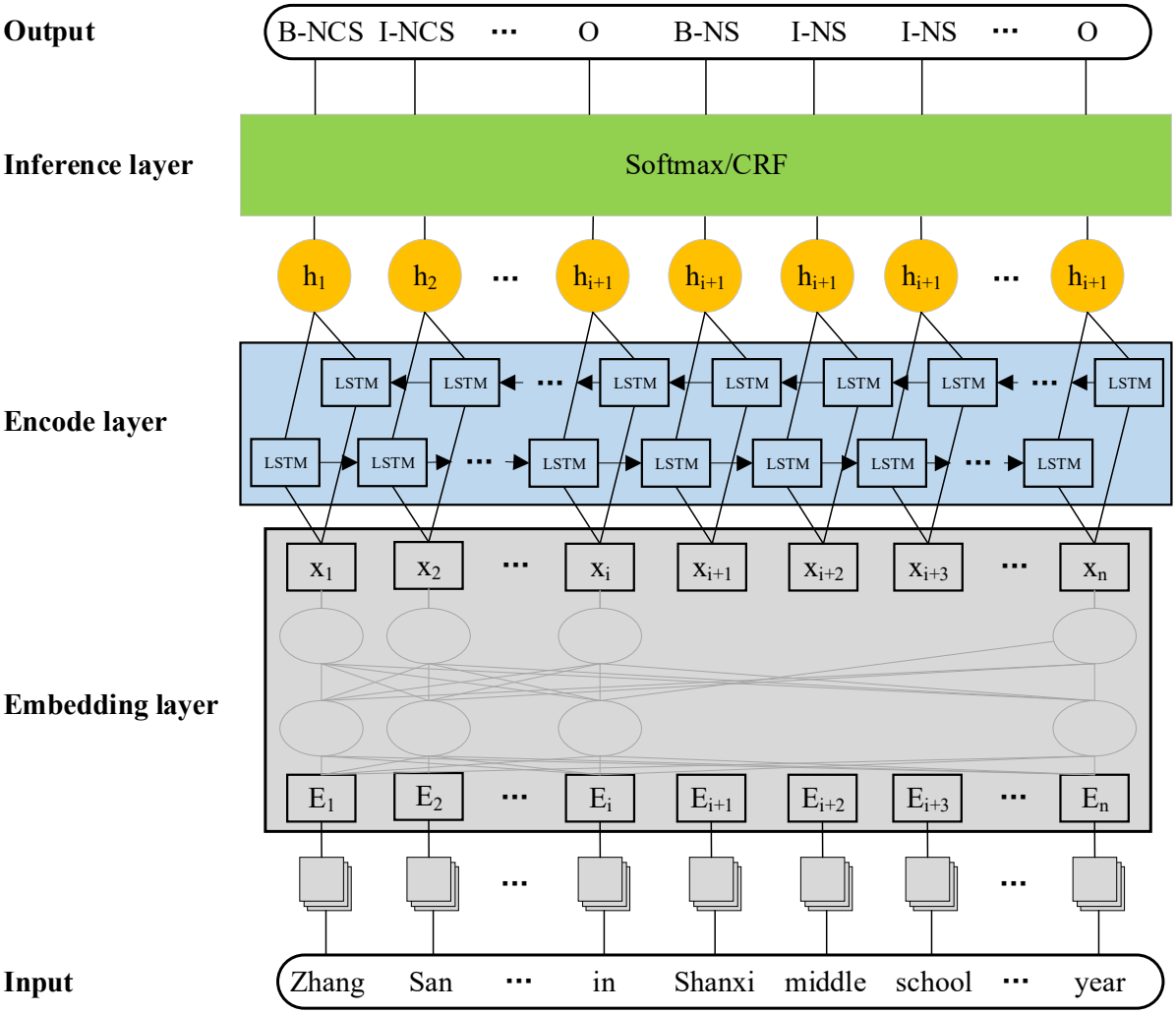

Figure 4. The architecture of the deep learning-driven extractor.

1. Embedding layer

The input to the model is a sentence in the sentence set $S = \{s_1, s_2, \cdots, s_m\}$. Sentence s_i containing n words is a word sequence $s = \{w_1, w_2, \cdots, w_n\}$. Then, each word w_i is represented by the input vector $\boldsymbol{E_i}$. The composition of $\boldsymbol{E_i}$ is as follows:

$$\boldsymbol{E_i} = \boldsymbol{E_t}(w_i) + \boldsymbol{E_s}(w_i) + \boldsymbol{E_p}(w_i) \tag{1}$$

where $\boldsymbol{E_t}(\cdot)$ is the token embeddings, $\boldsymbol{E_s}(\cdot)$ is the segmentation embeddings, and $\boldsymbol{E_p}(\cdot)$ is the position embeddings.

The embedding process of sentence s_i is expressed as:

$$\boldsymbol{X} = \mathrm{BERT}(\boldsymbol{E}, \theta_{bert}) \tag{2}$$

where $\boldsymbol{E} = \{\boldsymbol{E_1}, \boldsymbol{E_2}, \cdots, \boldsymbol{E_n}\}$ is the input vector representation, $\boldsymbol{X} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \cdots, \boldsymbol{x_n}\}$ is the output vector representation, and θ_{bert} is the relevant parameters.

2. Encode layer

In theory, RNN are ideal for processing sequential. But in practice, RNN suffers from the vanishing gradient problem [29]. At present, most studies of legal text information extraction use LSTM, a variant of RNN. Therefore, this study also uses LSTM to learn features from word sequences.

Given a vector sequence $\{\boldsymbol{x_1}, \boldsymbol{x_2}, \cdots, \boldsymbol{x_n}\}$, LSTM generates the corresponding vector representation $\{\boldsymbol{h_1}, \boldsymbol{h_2}, \cdots, \boldsymbol{h_m}\}$. The key equation of LSTM is shown below:

$$\begin{gathered} \boldsymbol{i_t} = \sigma(\boldsymbol{W_{ii}}\boldsymbol{x_t} + \boldsymbol{W_{hi}}\boldsymbol{h_{t-1}} + b_{ii} + b_{hi}) \\ \boldsymbol{f_t} = \sigma\Big(\boldsymbol{W_{if}}\boldsymbol{x_t} + \boldsymbol{W_{hf}}\boldsymbol{h_{t-1}} + b_{if} + b_{hf}\Big) \\ \boldsymbol{g_t} = tanh\Big(\boldsymbol{W_{ig}}\boldsymbol{x_t} + \boldsymbol{W_{hg}}\boldsymbol{h_{t-1}} + b_{ig} + b_{hg}\Big) \\ \boldsymbol{o_t} = \sigma(\boldsymbol{W_{io}}\boldsymbol{x_t} + \boldsymbol{W_{ho}}\boldsymbol{h_{t-1}} + b_{io} + b_{ho}) \\ \boldsymbol{c_t} = \boldsymbol{f_t}\boldsymbol{c_{t-1}} + \boldsymbol{i_t}\boldsymbol{g_t} \\ \boldsymbol{h_t} = \boldsymbol{o_t} tanh(\boldsymbol{c_t}) \end{gathered} \tag{3}$$

where $\sigma(\cdot)$ and $tanh(\cdot)$ are activation functions, $\boldsymbol{i_t}$ represents the input gate, $\boldsymbol{f_t}$ and $\boldsymbol{g_t}$ represent the forget gate, $\boldsymbol{o_t}$ represents the output gate, $\boldsymbol{c_t}$ represents long memory, and $\boldsymbol{h_t}$ represents short memory.

Bi-LSTM is composed of a forward LSTM and a backward LSTM, each LSTM has an output sequence. The process is represented as:

$$\begin{gathered} \overrightarrow{\boldsymbol{h_n}} = \overrightarrow{\mathrm{LSTM}}\Big(\overrightarrow{\boldsymbol{h_{n-1}}}, \boldsymbol{x_n}, \theta_{LSTM}\Big) \\ \overleftarrow{\boldsymbol{h_n}} = \overleftarrow{\mathrm{LSTM}}\Big(\overleftarrow{\boldsymbol{h_{n-1}}}, \boldsymbol{x_n}, \theta_{LSTM}\Big) \\ \boldsymbol{h_n} = \overrightarrow{\boldsymbol{h_n}} \oplus \overleftarrow{\boldsymbol{h_n}} \end{gathered} \tag{4}$$

where $\overrightarrow{\boldsymbol{h_n}}$ and $\overleftarrow{\boldsymbol{h_n}}$ is the output vector of the forward and backward LSTM at the nth word respectively, θ_{LSTM} is the training parameter of LSTM, $\oplus$ represents the splicing operation of $\overrightarrow{\boldsymbol{h_n}}$ and $\overleftarrow{\boldsymbol{h_n}}$, and $\boldsymbol{h_n}$ is the spliced vector.

Finally, the output sequence of sentence s_i is denoted as $\boldsymbol{H} = \{\boldsymbol{h_1}, \boldsymbol{h_2}, \cdots, \boldsymbol{h_n}\}$, which is the input of the inference layer.

3. Inference layer

The last layer uses the CRF algorithm to predict the label of each word due to the dependencies between labels. In the actual situation that there are a large number of referential nouns in Chinese legal texts, the CRF algorithm can use the adjacent labeling results to obtain the optimal label sequence. The algorithm is as follows:

The score of the embedding vector $\boldsymbol{X} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \cdots, \boldsymbol{x_n}\}$ of the input sentence and its predicted sequence $y = \{y_1, y_2, \cdots, y_n\}$ is defined as:

$$s(\boldsymbol{X}, y) = \sum_{i=0}^{n} \boldsymbol{T}_{y_i, y_{i+1}} + \sum_{i=1}^{n} \boldsymbol{E}_{i, y_i} \tag{5}$$

where $\boldsymbol{T}$ is the transition score matrix, e.g., $\boldsymbol{T}_{i,j}$ represents the transition score from label i to label j, and $\boldsymbol{E}$ is the encoding layer output score matrix, e.g., $\boldsymbol{E}_{i,j}$ represents the emission score from the i^{th} character to the j^{th} label.

All possible label sequences are passed through the SoftMax layer to obtain the probability distribution of the output sequence y as follows:

$$p(y|\boldsymbol{X}) = \frac{e^{s(\boldsymbol{X}, y)}}{\sum_{\widetilde{y} \in \boldsymbol{Y_X}} e^{s(\boldsymbol{X}, \widetilde{y})}} \tag{6}$$

where $\boldsymbol{Y_X}$ represents all possible label sequences of the input sequence $\boldsymbol{X}$. Finally, the output is the highest scoring label sequence, which is:

$$\mathrm{y}^* = \arg\max(\boldsymbol{X}, \widetilde{y}) \tag{7}$$

For a set of training samples $(\boldsymbol{X}, y)$, the loss function of the model is:

$$L = -\log(p(y|\boldsymbol{X})) = -s(\boldsymbol{X}, y) + \log\left(\sum_{\widetilde{y} \in Y_X} e^{s(\boldsymbol{X}, \widetilde{y})}\right) \tag{8}$$

4. Experimental Settings and Results

4.1. Dataset

The experimental dataset consists of the judicial dataset and the CAIL2021_IE dataset [16]. The judicial dataset is annotated by experts in the legal field and contains 500 Chinese legal texts on theft from several Chinese courts from 2018 to 2021. These texts are obtained on China Judgment Documents Online (https://wenshu.court.gov.cn/ (accessed on 22 December 2021)). The detailed statistics of the judicial dataset are shown in Table 2.

Table 2. Statistics of the judicial dataset.

No.	Legal Fact	Total	No.	Legal Fact	Total
1	Court	500	20	Name of sentenced	620
2	Document type	500	21	Charge	620
3	Case number	500	22	Prison term	584
4	Public prosecution organ	500	23	Fine	424
5	Name of defendant	620	24	Chief judge	227
6	Gender of defendant	605	25	Judge	668
7	Birthday of defendant	546	26	Assistant judge	177
8	Nation of defendant	485	27	Date of judgment	500
9	Registered residence of defendant	360	28	Clerk	468
10	Birthplace of defendant	232	29	Criminal suspect	1206
11	Educational level of defendant	550	30	Victim	560
12	Current residence of defendant	474	31	Time	526
13	Name of advocate	388	32	Spot	747
14	Work unit of advocate	386	33	Tools	148
15	Indictment number	488	34	Stolen money	182
16	Date of public prosecution	311	35	Stolen item	1134
17	Inquisitor	543	36	Organization	131
18	Legal provision name	921	37	Goods value	377
19	Legal provision number	921	38	Stolen profit	93

The CAIL2021_IE dataset [16], which consists of crime facts from Chinese legal texts on theft and contains 7500 sentences across 10 categories, is provided by China AI and Law Challenge (CAIL). Each word in the sentence is labeled by the BIO encoding format. The detailed statistics of the CAIL2021_IE dataset are shown in Table 3.

Table 3. Statistics of the CAIL2021_IE dataset.

No.	Legal Fact	Label	Total
1	Criminal suspect	B/I-NCS	6463
2	Victim	B/I-NVI	3108
3	Time	B/I-NT	2765
4	Spot	B/I-NS	3815
5	Tools	B/I-NTS	731
6	Stolen money	B/I-NSM	915
7	Stolen item	B/I-NSI	5884
8	Organization	B/I-NO	779
9	Goods value	B/I-NGV	2090
10	Stolen profit	B/I-NSP	481

4.2. Experimental Settings

Figure 5 illustrates the dataset used to evaluate each extractor. When the rule-based extractor is evaluated, it is tested on the judicial dataset. When the deep learning-driven extractor is evaluated, it is trained on the CAIL2021_IE dataset and tested on the judicial dataset.

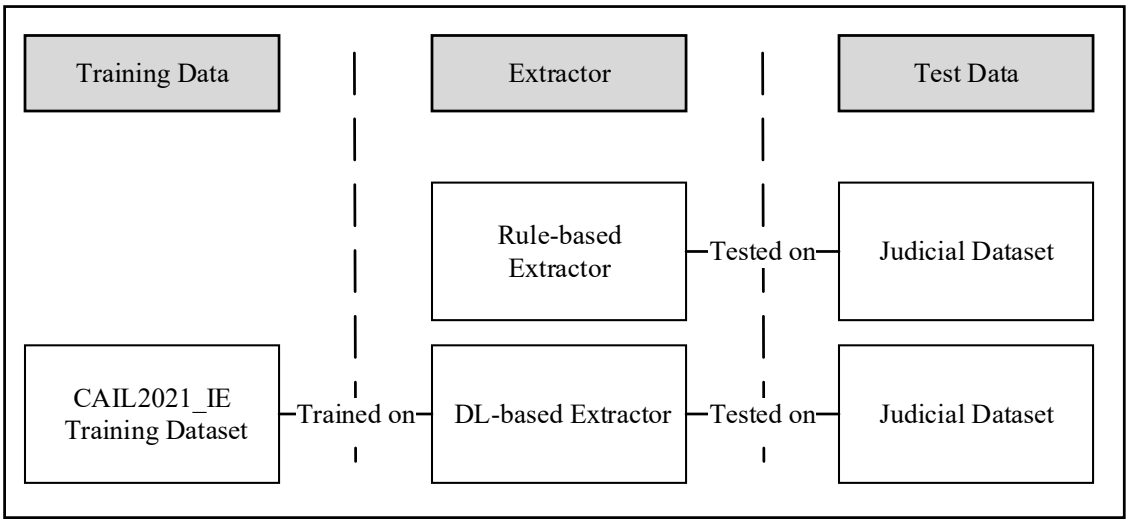

Figure 5. Setting of experimental datasets.

To demonstrate the effectiveness of the deep learning-driven extractor, it is compared with the following methods used in the study of legal text information extraction:

- CRF. CRF is a classic machine learning method that is often used for named entity recognition tasks. Ref. [9] used the CRF method in the task of legal text information extraction.
- Bi-LSTM and Bi-LSTM+CRF. LSTM is a variant of RNN and is often used for sequence labeling tasks. Refs. [10,26,30] used these methods, with Bi-LSTM+CRF performing the best.
- Bi-GRU and Bi-GRU+CRF. GRU is another variant of RNN and is also used for information extraction, such as Refs. [26,31].
- Multi-Bi-LSTM+CRF. Refs. [32,33] used this model. The model structure consists of multiple Bi-LSTM layers.

The evaluation of hyper-parameters uses a ten-fold cross validation method. The best hyper-parameters of the deep learning-driven extractor are shown in Table 4. In the embedding layer, the RoBERTa-wwm-ext model [34] is used. In the encoding layer, the LSTM dimension is set to 128 and only one Bi-LSTM sub-layer is used. The initial learning

rate is set to 3×10^{-5} with a decay rate 1×10^{-6}. The dropout rate is set to 0.5, the batch size is set to 16, and the maximum sequence length is 256. To provide fair comparisons, all of the compared methods set similar parameters.

Table 4. Hyper-parameters of the deep learning-driven extractor.

Parameter	Value	Parameter	Value
Pretrained language model	RoBERTa-wwm-ext	Batch size	16
LSTM dimension	128	Bidirectional	True
Maximum sequence length	256	LSTM layers	1
Learning rate	3×10^{-5}	Dropout rate	0.5
Decay rate	1×10^{-6}	Gradient clip	5

4.3. Evaluation Metrics

The proposed method is evaluated by using Precision, Recall, and F1-score combination metrics, as shown in Equation (9). An *exact match* strategy is used: the extracted legal facts are only correct when the boundaries are exactly aligned.

$$\begin{aligned} \text{Precision} &= \frac{\text{Number of legal facts correctly extracted}}{\text{Total number of legal facts extracted by system}} \\ \text{Recall} &= \frac{\text{Number of legal facts correctly extracted}}{\text{Total number of actual legal facts}} \\ \text{F1}_{\text{score}} &= \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \end{aligned} \tag{9}$$

4.4. Experimental Results and Discussion

The proposed method is tested on the judicial dataset consisting of 500 Chinese legal texts on theft (It took 17.186 s to extract these 500 legal texts). The proposed method can effectively extract up to 38 categories of legal facts and has excellent performance in Precision, Recall and F1-score. The experimental results of the proposed method and baselines are described below.

Table 5 describes the results of the rule-based extractor on the judicial dataset (sample rules and sentences are shown in Figure 3). These 28 legal facts are extracted with an average Precision of 99.85%, Recall of 99.55%, and F1-score of 99.70%. It is evident from Table 5 that the rule-based extractor can fully extract legal facts such as *court*, *document type*, *case number*, *judgment date*, and *clerk*. However, the extractor failed to fully extract legal facts such as *name of defendant*, *name of advocate*, *inquisitor*, and *charge*. Because of the *exact match* strategy, these legal facts that are not extracted correctly have always extra words or are missing some words. In addition, some logical segments are classified incorrectly in the process of paragraph classification, which also affects the extraction of legal facts.

Table 6 describes the comparison results of the deep learning-driven extractor and baseline on the judicial dataset, where the results are the average of all categories of legal facts. As can be seen from Table 6, the deep learning-driven extractor identifies the legal facts with an average Precision of 90.41%, Recall of 92.49%, and F1-score of 91.43%. These results show that the deep learning-driven extractor is more effective than existing methods in extracting legal facts from Chinese legal texts.

In addition, there are other observed results. First, the F1-score of the deep learning-driven extractor is 3.81% higher than that of Bi-LSTM+CRF, which shows that the dynamic word vector of the pre-trained model BERT has a greater improvement in the performance of Chinese legal text information extraction than the static word vector of Word2Vec. Furthermore, the F1-scores of Bi-LSTM(+CRF) are all higher than that of Bi-GRU(+CRF), which shows that LSTM network is more suitable for the information extraction of Chinese legal texts than the GRU network. In addition, the F1-scores of Bi-LSTM+CRF and Bi-GRU+CRF are significantly higher than those of Bi-LSTM and Bi-GRU, because CRF as an inference layer can capture the dependencies between each label. Finally, it can be seen that the F1-score of Multi-Bi-LSTM is lower than that of Bi-LSTM, which shows that

the multi-layer Bi-LSTM network cannot improve the extraction performance of Chinese legal texts.

Table 5. Results of the rule-based extractor on the judicial dataset.

No.	Legal Fact	P (%)	R (%)	F1 (%)	No.	Legal Fact	P (%)	R (%)	F1 (%)
1	Court	100	100	100	15	Indictment number	100	100	100
2	Document type	100	100	100	16	Date of public prosecution	100	100	100
3	Case number	100	100	100	17	Inquisitor	97.61	97.61	97.61
4	Public prosecution organ	100	100	100	18	Legal provision name	100	100	100
5	Name of defendant	99.68	99.52	99.60	19	Legal provision number	100	100	100
6	Gender of defendant	100	99.83	99.92	20	Name of sentenced	100	97.74	98.85
7	Birthday of defendant	100	100	100	21	Charge	99.70	97.42	98.53
8	Nation of defendant	100	100	100	22	Prison term	100	100	100
9	Registered residence of defendant	99.16	98.06	98.60	23	Fine	100	100	100
10	Birthplace of defendant	100	99.14	99.57	24	Chief judge	100	100	100
11	Educational level of defendant	100	100	100	25	Judge	100	100	100
12	Current residence of defendant	100	99.37	99.68	26	Assistant judge	100	100	100
13	Name of advocate	99.74	99.23	99.48	27	Date of judgment	100	100	100
14	Work unit of advocate	100	99.74	99.87	28	Clerk	100	100	100
						Average	99.85	99.55	99.70

Table 6. Comparison results of the deep learning-driven extractor and baselines.

Method	P (%)	R (%)	F1 (%)
CRF	86.97	85.85	86.34
Bi-GRU	84.14	81.33	82.50
Bi-GRU+CRF	88.76	86.46	87.56
Bi-LSTM	84.46	81.43	82.75
Bi-LSTM+CRF	88.04	87.32	87.62
Multi-Bi-LSTM+CRF	86.81	86.70	86.68
Proposed extractor	**90.41**	**92.49**	**91.43**

To better analyze the extracted results, Figures 6–8 describe the Precision, Recall and F1-score performance of the deep learning-driven extractor and baselines in each category (where the abscissa are: NCS-Criminal suspect, NVI-Victim, NT-Time, NS-Spot, NTS-Tools, NSM-Stolen money, NSI-Stolen item, NO-Organization, NGV-Goods value, NSP-Stolen profit). As can be seen from Figure 6, the proposed extractor achieves the highest average Precision (90.41%), and Multi-Bi-LSTM+CRF achieves the lowest average Precision (86.81%). It can also be observed that the proposed extractor outperforms baselines in eight legal fact categories (*Criminal suspect*, *Victim*, *Time*, *Spot*, *Stolen money*, *Stolen item*, *Organization*, *Goods value*). However, the proposed extractor obtains lower Precision in *Tools* (81.94%) and *Stolen item* (81.83%).

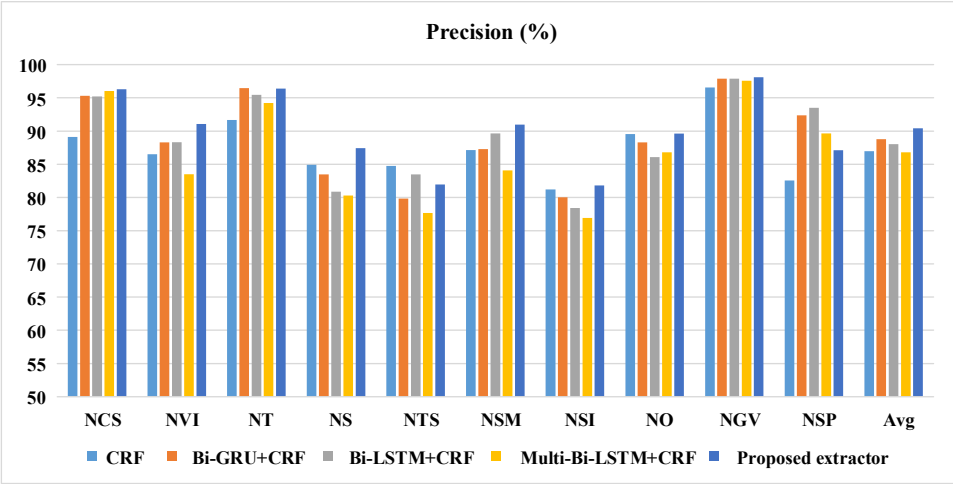

Figure 6. Precision performance of the deep learning-driven extractor and baselines in each category.

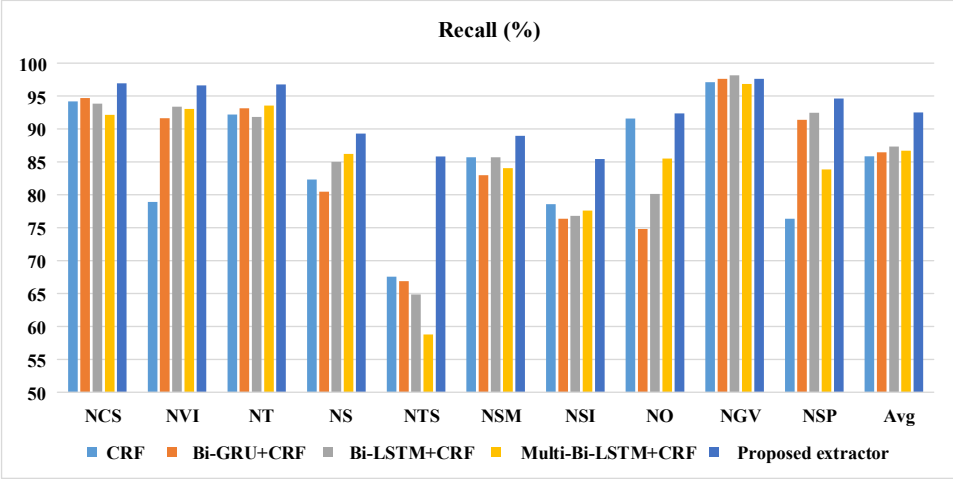

Figure 7. Recall performance of the deep learning-driven extractor and baselines in each category.

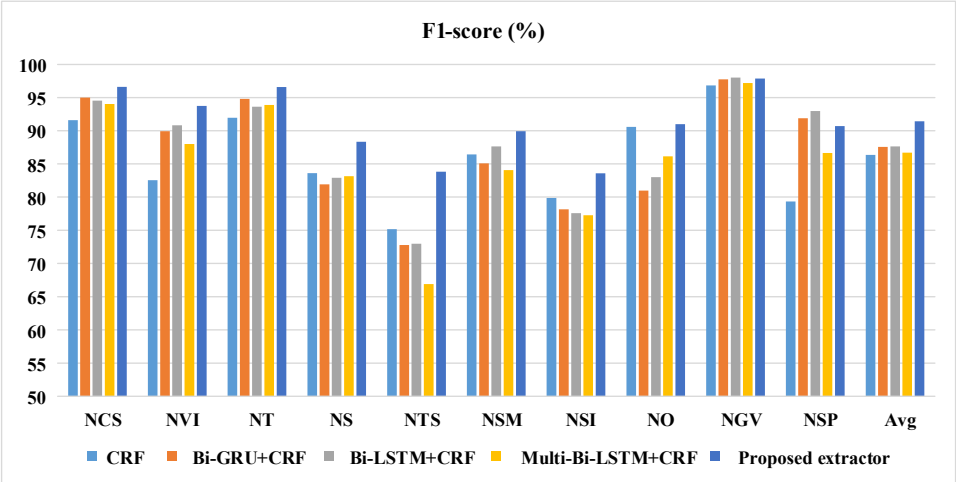

Figure 8. F1-score performance of the deep learning-driven extractor and baselines in each category.

As can be seen from Figure 7, the proposed extractor achieves the highest average Recall (92.49%), and CRF achieves the lowest average Recall (85.85%). It can also be observed that the proposed extractor outperforms baselines in 9 legal fact categories

(*Criminal suspect*, *Victim*, *Time*, *Spot*, *Tools*, *Stolen money*, *Stolen item*, *Organization*, *Stolen profit*). However, the proposed extractor obtains lower Recall in *Tools* (85.81%) and *Stolen item* (85.44%).

As can be seen from Figure 8, the proposed extractor achieves the highest average F1-score (91.43%), and CRF achieved the lowest average F1-score (86.34%). It can also be observed that the proposed extractor outperforms baselines in nine legal fact categories (*Criminal suspect*, *Victim*, *Time*, *Spot*, *Tools*, *Stolen money*, *Stolen item*, *Organization*, *Goods value*). However, the proposed extractor obtains a lower F1-score in *Tools* (83.83%), and *Stolen item* (83.59%).

As can be seen from Figures 6–8, the proposed extractor obtains the highest average Precision (90.41%), average Recall (92.49%), and average F1-score (91.43%). The proposed extractor outperforms baselines in most categories. However, the Spot, Tools, and Stolen item categories have poor extraction performance. The main reason is that there are ambiguous words and nested words (such as personal name and place name) in these legal facts. In addition, boundary recognition errors also resulted in poor extraction performance for these categories. As can be seen from Table 5, Figures 6–8, the proposed method effectively extracts up to 38 categories of legal facts from legal texts and the number of categories extracted increases significantly. Compared with existing methods, the proposed method has great advantages in extracting the completeness and accuracy of legal facts.

4.5. Comparison and Discussion with Other Related Works

Table 7 describes the comparison of the proposed method with other related legal text information extraction works.

Table 7. Comparison of the proposed method and other related works.

Work	Technique	Language	Number of Legal Fact Categories	Extract Hidden Legal Facts Support	Knowledge Modeling Support	Portability
Buey et al. (2016) [21]	Rule-based and ontology	Spanish	12	No	Yes	No
Zhuang et al. (2017) [1]	Rule-based	Chinese	7	No	No	No
Solihin et al. (2018) [8]	Rule-based	Indonesian	11	No	No	No
Iftikhar et al. (2019) [9]	Machine learning	English	9	Yes	No	No
Nuranti et al. (2020) [11]	Deep learning	Indonesian	10	Yes	No	No
Thomas et al. (2021) [12]	Semi-supervised learning and ontology	English	12	No	Yes	Yes
Proposed method	Rule-based, deep learning and ontology	Chinese	38	Yes	Yes	Yes

The proposed method is superior to existing works in the following ways: First, the proposed method is applicable to Chinese legal texts. Most of the existing works cannot be used to extract Chinese legal texts. Second, the proposed method uses both rule-based and deep learning-driven extractors, which can not only extract the legal facts with fixed linguistic rules, but also extract the legal facts hidden in legal texts. The proposed method extracts more kinds of legal facts and has a higher accuracy than the existing works. Third, the proposed method models the knowledge of Chinese legal texts. The knowledge modeling process makes the method compatible with other types of legal texts. Furthermore, the results of knowledge modeling also have a positive effect on the extraction performance.

The proposed method has some limitations as well. First, with no English annotated dataset, the extraction performance on English legal texts cannot be evaluated. Second, the

proposed method uses regular expressions for extracting a part of the legal facts. However, not all variants except the regular rules and patterns are considered, which results in some legal facts not being extracted. Furthermore, the proposed method is highly dependent on the structure of legal texts. If the legal text is poorly structured, it may affect the extraction performance and result in an increase in the number of false negatives.

5. Conclusions

This paper studies an ontology-based and deep learning-driven method for extracting legal facts from Chinese legal texts. The proposed method improves the performance of Chinese legal text information extraction through the strong normative characteristics of Chinese legal text content and structure composition and the strong text feature learning ability of deep learning. The experimental results show that the proposed method has excellent performance and is significantly superior to existing methods in extracting the completeness and accuracy of legal facts. Under the guidance of the knowledge model, the proposed method can be used to process various types of legal texts and can be better applied to the structured storage system of Chinese legal texts, which greatly improves the convenience of structured storage of legal texts and avoids a lot of manual labor by professionals in the judicial field.

In the future, we plan to improve our method in order to extract English legal texts. Second, we plan to incorporate a semi-supervised learning extractor into our method. In addition, we plan to focus on the construction of Chinese legal text ontology and construct the extracted legal facts into a knowledge graph.

Author Contributions: Conceptualization, Y.R., Y.L. and L.Z.; Data curation, J.H.; Formal analysis, J.H.; Funding acquisition, L.Z.; Investigation, J.H. and X.M.; Methodology, Y.R., J.H. and Y.L.; Project administration, L.Z.; Resources, Y.R., Y.L. and L.Z.; Software, J.H. and X.M.; Supervision, Y.R., Y.L. and L.Z.; Validation, J.H.; Visualization, J.H.; Writing—original draft, J.H.; Writing—review & editing, Y.R. and Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Program of China under grant 2020YFC0832700.

Data Availability Statement: The data presented in this study are available online.

Conflicts of Interest: The authors declare that they have no conflict of interest.

References

1. Zhuang, C.; Zhou, Y.; Ge, J.; Li, Z.; Li, C.; Zhou, X.; Luo, B. Information extraction from Chinese judgment documents. In Proceedings of the 2017 14th Web Information Systems and Applications Conference (WISA), Liuzhou, China, 11–12 November 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 240–244. [CrossRef]
2. Uyttendaele, C.; Moens, M.F.; Dumortier, J. Salomon: Automatic abstracting of legal cases for effective access to court decisions. *Artif. Intell. Law* **1998**, *6*, 59–79. [CrossRef]
3. Tiddi, I.; Schlobach, S. Knowledge graphs as tools for explainable machine learning: A survey. *Artif. Intell.* **2022**, *302*, 103627. [CrossRef]
4. Dozier, C.; Zielund, T. Cross document co-reference resolution applications for people in the legal domain. In Proceedings of the Conference on Reference Resolution and Its Applications, Barcelona, Spain, 25–26 July 2004; Association for Computational Linguistics: Stroudsburg, PA, USA, 2004; pp. 9–16.
5. Chaudhary, M.; Dozier, C.; Atkinson, G.; Berosik, G.; Guo, X.; Samler, S. Mining legal text to create a litigation history database. In Proceedings of the IASTED International Conference on Law and Technology, Cambridge, MA, USA, 9–11 October 2006.
6. Zhang, N.; Pu, Y.F.; Yang, S.Q.; Zhou, J.L.; Gao, J.K. An ontological Chinese legal consultation system. *IEEE Access* **2017**, *5*, 18250–18261. [CrossRef]
7. Khazaeli, S.; Punuru, J.; Morris, C.; Sharma, S.; Staub, B.; Cole, M.; Chiu-Webster, S.; Sakalley, D. A free format legal question answering system. In Proceedings of the Natural Legal Language Processing Workshop 2021, Punta Cana, Dominican Republic, 10 November 2021; pp. 107–113. [CrossRef]
8. Solihin, F.; Budi, I. Recording of law enforcement based on court decision document using rule-based information extraction. In Proceedings of the 2018 International Conference on Advanced Computer Science and Information Systems (ICACSIS), Yogyakarta, Indonesia, 27–28 October 2018; pp. 349–354. [CrossRef]

9. Iftikhar, A.; Jaffry, S.W.U.Q.; Malik, M.K. Information mining from criminal judgments of Lahore high court. *IEEE Access* **2019**, *7*, 59539–59547. [CrossRef]
10. Ji, D.; Tao, P.; Fei, H.; Ren, Y. An end-to-end joint model for evidence information extraction from court record document. *Inf. Process. Manag.* **2020**, *57*, 102305. [CrossRef]
11. Nuranti, E.Q.; Yulianti, E. Legal Entity Recognition in Indonesian Court Decision Documents Using Bi-LSTM and CRF Approaches. In Proceedings of the 2020 International Conference on Advanced Computer Science and Information Systems (ICACSIS), Depok, Indonesia, 17–18 October 2020; pp. 429–434. [CrossRef]
12. Thomas, A.; Sangeetha, S. Semi-supervised, knowledge-integrated pattern learning approach for fact extraction from judicial text. *Expert Syst.* **2021**, *38*, e12656. [CrossRef]
13. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
14. Huang, Z.; Xu, W.; Yu, K. Bidirectional LSTM-CRF models for sequence tagging. *arXiv* **2015**, arXiv:1508.01991.
15. Lafferty, J.; McCallum, A.; Pereira, F.C.N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the 18th International Conference on Machine Learning 2001 (ICML 2001), San Francisco, CA, USA, 28 June–1 July 2001; pp. 282–289.
16. China AI and Law Challenge. CAIL Information Extraction Dataset [Online]. Available online: http://cail.cipsc.org.cn/task9.html?raceID=7 (accessed on 21 December 2021).
17. Moens, M.F.; Uyttendaele, C.; Dumortier, J. Information extraction from legal texts: The potential of discourse analysis. *Int. J. Hum.-Comput. Stud.* **1999**, *51*, 1155–1171. [CrossRef]
18. Bach, N.X.; Thien, T.H.N.; Phuong, T.M. Question analysis for Vietnamese legal question answering. In Proceedings of the 2017 9th International Conference on Knowledge and Systems Engineering (KSE), Hue, Vietnam, 19–21 October 2017; pp. 154–159. [CrossRef]
19. Dozier, C.; Kondadadi, R.; Light, M.; Vachher, A.; Veeramachaneni, S.; Wudali, R. Named entity recognition and resolution in legal text. In *Semantic Processing of Legal Texts*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 27–43. [CrossRef]
20. Andrew, J.J. Automatic extraction of entities and relation from legal documents. In Proceedings of the Seventh Named Entities Workshop, Melbourne, Australia, 19 July 2018; pp. 1–8. [CrossRef]
21. Buey, M.G.; Garrido, A.L.; Bobed, C.; Ilarri, S. The AIS Project: Boosting Information Extraction from Legal Documents by using Ontologies. In Proceedings of the 8th International Conference on Agents and Artificial Intelligence, Rome, Italy, 24–26 February 2016; pp. 438–445. [CrossRef]
22. de Araujo, D.A.; Rigo, S.J.; Barbosa, J.L.V. Ontology-based information extraction for juridical events with case studies in Brazilian legal realm. *Artif. Intell. Law* **2017**, *25*, 379–396. [CrossRef]
23. Epelbaum, T. Deep learning: Technical introduction. *arXiv* **2017**, arXiv:1709.01412.
24. Staudemeyer, R.C.; Morris, E.R. Understanding LSTM—A tutorial into long short-term memory recurrent neural networks. *arXiv* **2019**, arXiv:1909.09586.
25. Rao, X.; Ke, Z. Hierarchical RNN for information extraction from lawsuit documents. *arXiv* **2018**, arXiv:1804.09321.
26. Fernandes, W.P.D.; Silva, L.J.S.; Frajhof, I.Z.; de Almeida, G.D.F.C.F.; Konder, C.N.; Nasser, R.B.; de Carvalho, G.R.; Barbosa, S.D.J.; Lopes, H.C.V. Appellate court modifications extraction for Portuguese. *Artif. Intell. Law* **2020**, *28*, 327–360. [CrossRef]
27. Thomas, A.; Sangeetha, S. A Legal Case Ontology for Extracting Domain-Specific Entity-Relationships from e-judgments. In Proceedings of the Sixth International Conference on Recent Trends in Information Processing & Computing (IPC), Bhopal, India, 27–28 October 2017.
28. Musen, M.A. The protégé project: A look back and a look forward. *AI Matters* **2015**, *1*, 4–12. [CrossRef] [PubMed]
29. Bengio, Y.; Simard, P.; Frasconi, P. Learning long-term dependencies with gradient descent is difficult. *IEEE Trans. Neural Netw.* **1994**, *5*, 157–166. [CrossRef] [PubMed]
30. Leitner, E.; Rehm, G.; Moreno-Schneider, J. Fine-grained named entity recognition in legal documents. In *International Conference on Semantic Systems*; Springer: Cham, Switzerland, 2019; pp. 272–287. [CrossRef]
31. Mandal, A.; Ghosh, K.; Ghosh, S.; Mandal, S. A sequence labeling model for catchphrase identification from legal case documents. *Artif. Intell. Law* **2021**, 1–34. [CrossRef]
32. Bach, N.X.; Thuy, N.T.T.; Chien, D.B.; Duy, T.K.; Hien, T.M.; Phuong, T.M. Reference extraction from Vietnamese legal documents. In Proceedings of the Tenth International Symposium on Information and Communication Technology, New York, NY, USA, 4–6 December 2019; pp. 486–493. [CrossRef]
33. Nguyen, T.S.; Nguyen, L.M.; Tojo, S.; Satoh, K.; Shimazu, A. Recurrent neural network-based models for recognizing requisite and effectuation parts in legal texts. *Artif. Intell. Law* **2018**, *26*, 169–199. [CrossRef]
34. Cui, Y.; Che, W.; Liu, T.; Qin, B.; Yang, Z. Pre-training with whole word masking for Chinese Bert. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 3504–3514. [CrossRef]