How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence

Haoxi Zhong[1], **Chaojun Xiao**[1], **Cunchao Tu**[1], **Tianyang Zhang**[2],
Zhiyuan Liu[1]*, **Maosong Sun**[1]

[1]Department of Computer Science and Technology
Institute for Artificial Intelligence, Tsinghua University, Beijing, China
Beijing National Research Center for Information Science and Technology, China
[2]Beijing Powerlaw Intelligent Technology Co., Ltd., China
zhonghaoxi@yeah.net, {xcjthu,tucunchao}@gmail.com, zty@powerlaw.ai,
{lzy,sms}@tsinghua.edu.cn

Abstract

Legal Artificial Intelligence (LegalAI) focuses on applying the technology of artificial intelligence, especially natural language processing, to benefit tasks in the legal domain. In recent years, LegalAI has drawn increasing attention rapidly from both AI researchers and legal professionals, as LegalAI is beneficial to the legal system for liberating legal professionals from a maze of paperwork. Legal professionals often think about how to solve tasks from rule-based and symbol-based methods, while NLP researchers concentrate more on data-driven and embedding methods. In this paper, we describe the history, the current state, and the future directions of research in LegalAI. We illustrate the tasks from the perspectives of legal professionals and NLP researchers and show several representative applications in LegalAI. We conduct experiments and provide an in-depth analysis of the advantages and disadvantages of existing works to explore possible future directions. You can find the implementation of our work from https://github.com/thunlp/CLAIM.

1 Introduction

Legal Artificial Intelligence (LegalAI) mainly focuses on applying artificial intelligence technology to help legal tasks. The majority of the resources in this field are presented in text forms, such as judgment documents, contracts, and legal opinions. Therefore, most LegalAI tasks are based on Natural Language Processing (NLP) technologies.

LegalAI plays a significant role in the legal domain, as they can reduce heavy and redundant work for legal professionals. Many tasks in the legal domain require the expertise of legal practitioners and a thorough understanding of various legal documents. Retrieving and understanding legal documents take lots of time, even for legal professionals.

Therefore, a qualified system of LegalAI should reduce the time consumption of these tedious jobs and benefit the legal system. Besides, LegalAI can also provide a reliable reference to those who are not familiar with the legal domain, serving as an affordable form of legal aid.

In order to promote the development of LegalAI, many researchers have devoted considerable efforts over the past few decades. Early works (Kort, 1957; Ulmer, 1963; Nagel, 1963; Segal, 1984; Gardner, 1984) always use hand-crafted rules or features due to computational limitations at the time. In recent years, with rapid developments in deep learning, researchers begin to apply deep learning techniques to LegalAI. Several new LegalAI datasets have been proposed (Kano et al., 2018; Xiao et al., 2018; Duan et al., 2019; Chalkidis et al., 2019b,a), which can serve as benchmarks for research in the field. Based on these datasets, researchers began exploring NLP-based solutions to a variety of LegalAI tasks, such as Legal Judgment Prediction (Aletras et al., 2016; Luo et al., 2017; Zhong et al., 2018; Chen et al., 2019), Court View Generation (Ye et al., 2018), Legal Entity Recognition and Classification (Cardellino et al., 2017; ANGELIDIS et al., 2018), Legal Question Answering (Monroy et al., 2009; Taniguchi and Kano, 2016; Kim and Goebel, 2017), Legal Summarization (Hachey and Grover, 2006; Bhattacharya et al., 2019).

As previously mentioned, researchers' efforts over the years led to tremendous advances in LegalAI. To summarize, some efforts concentrate on symbol-based methods, which apply interpretable hand-crafted symbols to legal tasks (Ashley, 2017; Surden, 2018). Meanwhile, other efforts with embedding-based methods aim at designing efficient neural models to achieve better performance (Chalkidis and Kampas, 2019). More specifically, symbol-based methods concentrate more on utilizing interpretable legal knowledge to reason
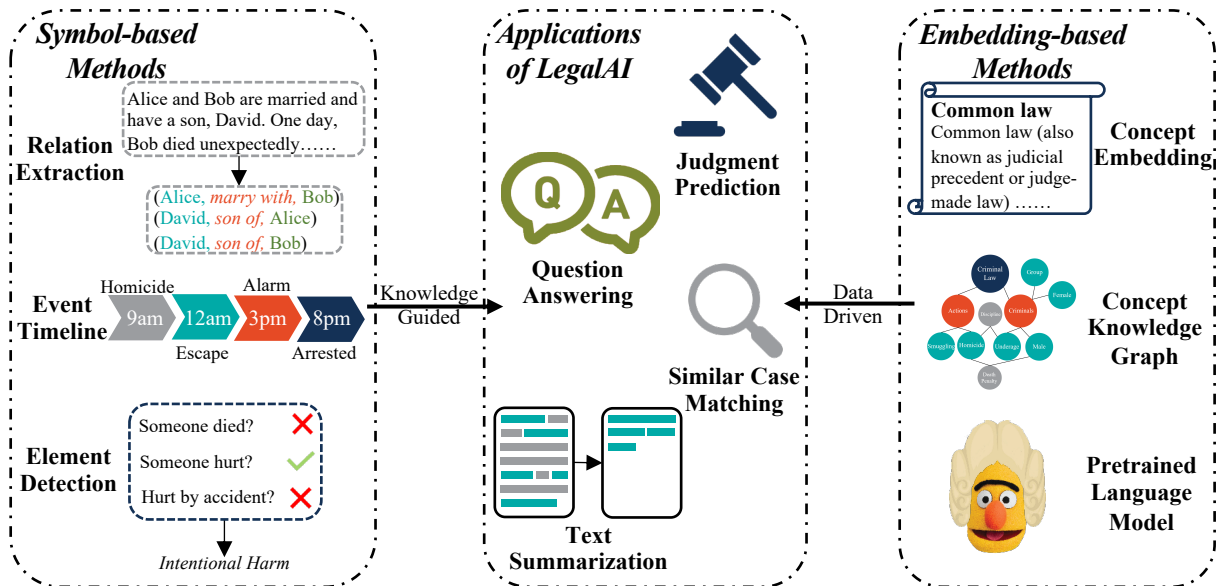
---

*Corresponding author.

Figure 1: An overview of tasks in LegalAI.

between symbols in legal documents, like events and relationships. Meanwhile, embedding-based methods try to learn latent features for prediction from large-scale data. The differences between these two methods have caused some problems in existing works of LegalAI. Interpretable symbolic models are not effective, and embedding-methods with better performance usually cannot be interpreted, which may bring ethical issues to the legal system such as gender bias and racial discrimination. The shortcomings make it difficult to apply existing methods to real-world legal systems.

We summarize three primary challenges for both embedding-based and symbol-based methods in LegalAI: (1) Knowledge Modelling. Legal texts are well formalized, and there are many domain knowledge and concepts in LegalAI. How to utilize the legal knowledge is of great significance. (2) Legal Reasoning. Although most tasks in NLP require reasoning, the LegalAI tasks are somehow different, as legal reasoning must strictly follow the rules well-defined in law. Thus combining predefined rules and AI technology is essential to legal reasoning. Besides, complex case scenarios and complex legal provisions may require more sophisticated reasoning for analyzing. (3) Interpretability. Decisions made in LegalAI usually should be interpretable to be applied to the real legal system. Otherwise, fairness may risk being compromised. Interpretability is as important as performance in LegalAI.

The main contributions of this work are con-

cluded as follows: (1) We describe existing works from the perspectives of both NLP researchers and legal professionals. Moreover, we illustrate several embedding-based and symbol-based methods and explore the future direction of LegalAI. (2) We describe three typical applications, including judgment prediction, similar case matching, and legal question answering in detail to emphasize why these two kinds of methods are essential to LegalAI. (3) We conduct exhaustive experiments on multiple datasets to explore how to utilize NLP technology and legal knowledge to overcome the challenges in LegalAI. You can find the implementation from github[1]. (4) We summarize LegalAI datasets, which can be regarded as the benchmark for related tasks. The details of these datasets can be found from github[2] with several legal papers worth reading.

2 Embedding-based Methods

First, we describe embedding-based methods in LegalAI, also named as representation learning. Embedding-based methods emphasize on representing legal facts and knowledge in embedding space, and they can utilize deep learning methods for corresponding tasks.

2.1 Character, Word, Concept Embeddings

Character and word embeddings play a significant role in NLP, as it can embed the discrete texts into

---

[1] https://github.com/thunlp/CLAIM
[2] https://github.com/thunlp/LegalPapers

continuous vector space. Many embedding methods have been proved effective (Mikolov et al., 2013; Joulin et al., 2016; Pennington et al., 2014; Peters et al., 2018; Yang et al., 2014; Bordes et al., 2013; Lin et al., 2015) and they are crucial for the effectiveness of the downstream tasks.

In LegalAI, embedding methods are also essential as they can bridge the gap between texts and vectors. However, it seems impossible to learn the meaning of a professional term directly from some legal factual description. Existing works (Chalkidis and Kampas, 2019; Nay, 2016) mainly revolve around applying existing embedding methods like Word2Vec to legal domain corpora. To overcome the difficulty of learning professional vocabulary representations, we can try to capture both grammatical information and legal knowledge in word embedding for corresponding tasks. Knowledge modelling is significant to LegalAI, as many results should be decided according to legal rules and knowledge.

Although knowledge graph methods in the legal domain are promising, there are still two major challenges before their practical usage. Firstly, the construction of the knowledge graph in LegalAI is complicated. In most scenarios, there are no ready-made legal knowledge graphs available, so researchers need to build from scratch. In addition, different legal concepts have different representations and meanings under legal systems in different countries, which also makes it challenging to construct a general legal knowledge graph. Some researchers tried to embed legal dictionaries (Cvrček et al., 2012), which can be regarded as an alternative method. Secondly, a generalized legal knowledge graph is different in the form with those commonly used in NLP. Existing knowledge graphs concern the relationship between entities and concepts, but LegalAI focuses more on the explanation of legal concepts. These two challenges make knowledge modelling via embedding in LegalAI non-trivial, and researchers can try to overcome the challenges in the future.

2.2 Pretrained Language Models

Pretrained language models (PLMs) such as BERT (Devlin et al., 2019) have been the recent focus in many fields in NLP (Radford et al., 2019; Yang et al., 2019; Liu et al., 2019a). Given the success of PLM, using PLM in LegalAI is also a very reasonable and direct choice. However, there are differences between the text used by existing PLMs and legal text, which also lead to unsatisfactory performances when directly applying PLMs to legal tasks. The differences stem from the terminology and knowledge involved in legal texts. To address this issue, Zhong et al. (2019b) propose a language model pretrained on Chinese legal documents, including civil and criminal case documents. Legal domain-specific PLMs provide a more qualified baseline system for the tasks of LegalAI. We will show several experiments comparing different BERT models in LegalAI tasks.

For the future exploration of PLMs in LegalAI, researchers can aim more at integrating knowledge into PLMs. Integrating knowledge into pretrained models can help the reasoning ability between legal concepts. Lots of work has been done on integrating knowledge from the general domain into models (Zhang et al., 2019; Peters et al., 2019; Hayashi et al., 2019). Such technology can also be considered for future application in LegalAI.

3 Symbol-based Methods

In this section, we describe symbol-based methods, also named as structured prediction methods. Symbol-based methods are involved in utilizing legal domain symbols and knowledge for the tasks of LegalAI. The symbolic legal knowledge, such as events and relationships, can provide interpretability. Deep learning methods can be employed for symbol-based methods for better performance.

3.1 Information Extraction

Information extraction (IE) has been widely studied in NLP. IE emphasizes on extracting valuable information from texts, and there are many NLP works which concentrate on IE, including name entity recognition (Lample et al., 2016; Kuru et al., 2016; Akbik et al., 2019), relation extraction (Zeng et al., 2015; Miwa and Bansal, 2016; Lin et al., 2016; Christopoulou et al., 2018), and event extraction (Chen et al., 2015; Nguyen et al., 2016; Nguyen and Grishman, 2018).

IE in LegalAI has also attracted the interests of many researchers. To make better use of the particularity of legal texts, researchers try to use ontology (Bruckschen et al., 2010; Cardellino et al., 2017; Lenci et al., 2009; Zhang et al., 2017) or global consistency (Yin et al., 2018) for named entity recognition in LegalAI. To extract relationship and events from legal documents, re-

searchers attempt to apply different NLP technologies, including hand-crafted rules (Bartolini et al., 2004; Truyens and Eecke, 2014), CRF (Vacek and Schilder, 2017), joint models like SVM, CNN, GRU (Vacek et al., 2019), or scale-free identifier network (Yan et al., 2017) for promising results.

Existing works have made lots of efforts to improve the effect of IE, but we need to pay more attention to the benefits of the extracted information. The extracted symbols have a legal basis and can provide interpretability to legal applications, so we cannot just aim at the performance of methods. Here, we show two examples of utilizing the extracted symbols for interpretability of LegalAI:

Relation Extraction and Inheritance Dispute. Inheritance dispute is a type of cases in Civil Law that focuses on the distribution of inheritance rights. Therefore, identifying the relationship between the parties is vital, as those who have the closest relationship with the deceased can get more assets. Towards this goal, relation extraction in inheritance dispute cases can provide the reason for judgment results and improve performance.

Event Timeline Extraction and Judgment Prediction of Criminal Case. In criminal cases, multiple parties are often involved in group crimes. To decide who should be primarily responsible for the crime, we need to determine what everyone has done throughout the case, and the order of these events is also essential. For example, in the case of crowd fighting, the person who fights first should bear the primary responsibility. As a result, a qualified event timeline extraction model is required for judgment prediction of criminal cases.

In future research, we need to concern more about applying extracted information to the tasks of LegalAI. The utilization of such information depends on the requirements of specific tasks, and the information can provide more interpretability.

3.2 Legal Element Extraction

In addition to those common symbols in general NLP, LegalAI also has its exclusive symbols, named legal elements. The extraction of legal elements focuses on extracting crucial elements like whether someone is killed or something is stolen. These elements are called constitutive elements of crime, and we can directly convict offenders based on the results of these elements. Utilizing these elements can not only bring intermediate supervision information to the judgment prediction task

but also make the prediction results of the model more interpretable.

Fact Description: One day, Bob used a fake reason for marriage decoration to borrow RMB 2k from Alice. After arrested, Bob has paid the money back to Alice.	
Whether did Bob sell something?	×
Whether did Bob make a fictional fact?	✓
Whether did Bob illegally possess the property of others?	✓
Judgment Results: Fraud.	

Table 1: An example of element detection from Zhong et al. (2020). From this example, we can see that the extracted elements can decide the judgment results. It shows that elements are useful for downstream tasks.

Towards a more in-depth analysis of element-based symbols, Shu et al. (2019) propose a dataset for extracting elements from three different kinds of cases, including divorce dispute, labor dispute, and loan dispute. The dataset requires us to detect whether the related elements are satisfied or not, and formalize the task as a multi-label classification problem. To show the performance of existing methods on element extraction, we have conducted experiments on the dataset, and the results can be found in Table 2.

Model	Divorce		Labor		Loan	
	MiF	MaF	MiF	MaF	MiF	MaF
TextCNN	78.7	65.9	76.4	54.4	80.3	60.6
DPCNN	81.3	64.0	79.8	47.4	81.4	42.5
LSTM	80.6	67.3	81.0	52.9	80.4	53.1
BiDAF	83.1	68.7	**81.5**	**59.4**	80.5	63.1
BERT	83.3	69.6	76.8	43.7	78.6	39.5
BERT-MS	**84.9**	**72.7**	79.7	54.5	**81.9**	**64.1**

Table 2: Experimental results on extracting elements. Here MiF and MaF denotes micro-F1 and macro-F1.

We have implemented several classical encoding models in NLP for element extraction, including TextCNN (Kim, 2014), DPCNN (Johnson and Zhang, 2017), LSTM (Hochreiter and Schmidhuber, 1997), BiDAF (Seo et al., 2016), and BERT (Devlin et al., 2019). We have tried two different versions of pretrained parameters of BERT, including the origin parameters (BERT) and the parameters pretrained on Chinese legal documents (BERT-MS) (Zhong et al., 2019b). From the results, we can see that the language model pretrained on the general domain performs worse

than domain-specific PLM, which proves the necessity of PLM in LegalAI. For the following parts of our paper, we will use BERT pretrained on legal documents for better performance.

From the results of element extraction, we can find that existing methods can reach a promising performance on element extraction, but are still not sufficient for corresponding applications. These elements can be regarded as pre-defined legal knowledge and help with downstream tasks. How to improve the performance of element extraction is valuable for further research.

4 Applications of LegalAI

In this section, we will describe several typical applications in LegalAI, including Legal Judgment Prediction, Similar Case Matching and Legal Question Answering. Legal Judgment Prediction and Similar Case Matching can be regarded as the core function of judgment in Civil Law and Common Law system, while Legal Question Answering can provide consultancy for those who are unfamiliar with the legal domain. Therefore, exploring these three tasks can cover most aspects of LegalAI.

4.1 Legal Judgment Prediction

Legal Judgment Prediction (LJP) is one of the most critical tasks in LegalAI, especially in the Civil Law system. In the Civil Law system, the judgment results are decided according to the facts and the statutory articles. One will receive legal sanctions only after he or she has violated the prohibited acts prescribed by law. The task LJP mainly concerns how to predict the judgment results from both the fact description of a case and the contents of the statutory articles in the Civil Law system.

As a result, LJP is an essential and representative task in countries with Civil Law system like France, Germany, Japan, and China. Besides, LJP has drawn lots of attention from both artificial intelligence researchers and legal professionals. In the following parts, we describe the research progress and explore the future direction of LJP.

Related Work

LJP has a long history. Early works revolve around analyzing existing legal cases in specific circumstances using mathematical or statistical methods (Kort, 1957; Ulmer, 1963; Nagel, 1963; Keown, 1980; Segal, 1984; Lauderdale and Clark, 2012). The combination of mathematical methods and legal rules makes the predicted results interpretable.

Fact Description:	One day, the defendant Bob stole cash 8500 yuan and T-shirts, jackets, pants, shoes, hats (identified a total value of 574.2 yuan) in Beijing Lining store.
Judgment Results	
Relevant Articles	Article 264 of Criminal Law.
Applicable Charges	Theft.
Term of Penalty	6 months.

Table 3: An example of legal judgment prediction from Zhong et al. (2018). In this example, the judgment results include relevant articles, applicable charges and the the term of penalty.

To promote the progress of LJP, Xiao et al. (2018) have proposed a large-scale Chinese criminal judgment prediction dataset, C-LJP. The dataset contains over 2.68 million legal documents published by the Chinese government, making C-LJP a qualified benchmark for LJP. C-LJP contains three subtasks, including relevant articles, applicable charges, and the term of penalty. The first two can be formalized as multi-label classification tasks, while the last one is a regression task. Besides, English LJP datasets also exist (Chalkidis et al., 2019a), but the size is limited.

With the development of the neural network, many researchers begin to explore LJP using deep learning technology (Hu et al., 2018; Wang et al., 2019; Li et al., 2019b; Liu et al., 2019b; Li et al., 2019a; Kang et al., 2019). These works can be divided into two primary directions. The first one is to use more novel models to improve performance. Chen et al. (2019) use the gating mechanism to enhance the performance of predicting the term of penalty. Pan et al. (2019) propose multi-scale attention to handle the cases with multiple defendants. Besides, other researchers explore how to utilize legal knowledge or the properties of LJP. Luo et al. (2017) use the attention mechanism between facts and law articles to help the prediction of applicable charges. Zhong et al. (2018) present a topological graph to utilize the relationship between different tasks of LJP. Besides, Hu et al. (2018) incorporate ten discriminative legal attributes to help predict low-frequency charges.

Experiments and Analysis

To better understand recent advances in LJP, we have conducted a series of experiments on C-LJP. Firstly, we implement several classical text classification models, including TextCNN (Kim, 2014), DPCNN (Johnson and Zhang, 2017),

5222

Task	Dev					Test				
	Charge		Article		Term	Charge		Article		Term
Metrics	MiF	MaF	MiF	MaF	Dis	MiF	MaF	MiF	MaF	Dis
TextCNN	93.8	74.6	92.8	**70.5**	1.586	93.9	72.2	93.5	67.0	1.539
DPCNN	94.7	72.2	93.9	68.8	1.448	94.9	72.1	94.6	69.4	1.390
LSTM	94.7	71.2	93.9	66.5	1.456	94.3	66.0	94.7	70.7	1.467
BERT	94.5	66.3	93.5	64.7	**1.421**	94.7	71.3	94.3	66.9	1.342
FactLaw	79.5	25.4	79.8	24.9	1.721	76.9	35.0	78.1	30.8	1.683
TopJudge	**94.8**	**76.3**	**94.0**	69.6	1.438	**97.6**	**76.8**	**96.9**	**70.9**	**1.335**
Gating Network	-	-	-	-	1.604	-	-	-	-	1.553

Table 4: Experimental results of judgment prediction on C-LJP. In this table, MiF and MaF denotes micro-F1 and macro-F1, and Dis denotes the log distance between prediction and ground truth.

LSTM (Hochreiter and Schmidhuber, 1997), and BERT (Devlin et al., 2019). For the parameters of BERT, we use the pretrained parameters on Chinese criminal cases (Zhong et al., 2019b). Secondly, we implement several models which are specially designed for LJP, including FactLaw (Luo et al., 2017), TopJudge (Zhong et al., 2018), and Gating Network (Chen et al., 2019). The results can be found in Table 4.

From the results, we can learn that most models can reach a promising performance in predicting high-frequency charges or articles. However, the models perform not well on low-frequency labels as there is a gap between micro-F1 and macro-F1. Hu et al. (2018) have explored few-shot learning for LJP. However, their model requires additional attribute information labelled manually, which is time-consuming and makes it hard to employ the model in other datasets. Besides, we can find that performance of BERT is not satisfactory, as it does not make much improvement from those models with fewer parameters. The main reason is that the length of the legal text is very long, but the maximum length that BERT can handle is 512. According to statistics, the maximum document length is 56, 694, and the length of 15% documents is over 512. Document understanding and reasoning techniques are required for LJP.

Although embedding-based methods can achieve promising performance, we still need to consider combining symbol-based with embedding-based methods in LJP. Take TopJudge as an example, this model formalizes topological order between the tasks in LJP (symbol-based part) and uses TextCNN for encoding the fact description. By combining symbol-based and embedding-based methods, TopJudge has achieved promising results on LJP. Comparing the results between TextCNN and TopJudge, we can find that just integrating the order of judgments into the model can lead to improvements, which proves the necessity of combining embedding-based and symbol-based methods.

For better LJP performance, some challenges require the future efforts of researchers: (1) **Document understanding and reasoning** techniques are required to obtain global information from extremely long legal texts. (2) **Few-shot learning**. Even low-frequency charges should not be ignored as they are part of legal integrity. Therefore, handling in-frequent labels is essential to LJP. (3) **Interpretability**. If we want to apply methods to real legal systems, we must understand how they make predictions. However, existing embedding-based methods work as a black box. What factors affected their predictions remain unknown, and this may introduce unfairness and ethical issues like gender bias to the legal systems. Introducing legal symbols and knowledge mentioned before will benefit the interpretability of LJP.

4.2 Similar Case Matching

In those countries with the Common Law system like the United States, Canada, and India, judicial decisions are made according to similar and representative cases in the past. As a result, how to identify the most similar case is the primary concern in the judgment of the Common Law system.

In order to better predict the judgment results in the Common Law system, Similar Case Matching (SCM) has become an essential topic of LegalAI. SCM concentrate on finding pairs of similar cases, and the definition of similarity can be various. SCM requires to model the relationship between cases from the information of different granularity, like fact level, event level and element level. In

other words, SCM is a particular form of semantic matching (Xiao et al., 2019), which can benefit the legal information retrieval.

Related Work

Traditional methods of Information Retrieve (IR) focus on term-level similarities with statistical models, including TF-IDF (Salton and Buckley, 1988) and BM25 (Robertson and Walker, 1994), which are widely applied in current search systems. In addition to these term matching methods, other researchers try to utilize meta-information (Medin, 2000; Gao et al., 2011; Wu et al., 2013) to capture semantic similarity. Many machine learning methods have also been applied for IR like SVD (Xu et al., 2010) or factorization (Rendle, 2010; Kabbur et al., 2013). With the rapid development of deep learning technology and NLP, many researchers apply neural models, including multi-layer perceptron (Huang et al., 2013), CNN (Shen et al., 2014; Hu et al., 2014; Qiu and Huang, 2015), and RNN (Palangi et al., 2016) to IR.

There are several LegalIR datasets, including COLIEE (Kano et al., 2018), CaseLaw (Locke and Zuccon, 2018), and CM (Xiao et al., 2019). Both COLIEE and CaseLaw are involved in retrieving most relevant articles from a large corpus, while data examples in CM give three legal documents for calculating similarity. These datasets provide benchmarks for the studies of LegalIR. Many researchers focus on building an easy-to-use legal search engine (Barmakian, 2000; Turtle, 1995). They also explore utilizing more information, including citations (Monroy et al., 2013; Geist, 2009; Raghav et al., 2016) and legal concepts (Maxwell and Schafer, 2008; Van Opijnen and Santos, 2017). Towards the goal of calculating similarity in semantic level, deep learning methods have also been applied to LegalIR. Tran et al. (2019) propose a CNN-based model with document and sentence level pooling which achieves the state-of-the-art results on COLIEE, while other researchers explore employing better embedding methods for LegalIR (Landthaler et al., 2016; Sugathadasa et al., 2018).

Experiments and Analysis

To get a better view of the current progress of LegalIR, we select CM (Xiao et al., 2019) for experiments. CM contains $8,964$ triples where each triple contains three legal documents (A, B, C). The task designed in CM is to determine whether

B or C is more similar to A. We have implemented four different types of baselines: (1) Term matching methods, TF-IDF (Salton and Buckley, 1988). (2) Siamese Network with two parameter-shared encoders, including TextCNN (Kim, 2014), BiDAF (Seo et al., 2016) and BERT (Devlin et al., 2019), and a distance function. (3) Semantic matching models in sentence level, ABCNN (Yin et al., 2016), and document level, SMASH-RNN (Jiang et al., 2019). The results can be found in Table 5.

Model	Dev	Test
TF-IDF	52.9	53.3
TextCNN	62.5	**69.9**
BiDAF	63.3	68.6
BERT	**64.3**	66.8
ABCNN	62.7	**69.9**
SMASH RNN	64.2	65.8

Table 5: Experimental results of SCM. The evaluation metric is accuracy.

From the results, we observe that existing neural models which are capable of capturing semantic information outperform TF-IDF, but the performance is still not enough for SCM. As Xiao et al. (2019) state, the main reason is that legal professionals think that elements in this dataset define the similarity of legal cases. Legal professionals will emphasize on whether two cases have similar elements. Only considering term-level and semantic-level similarity is insufficient for the task.

For the further study of SCM, there are two directions which need future effort: (1) **Elemental-based representation**. Researchers can focus more on symbols of legal documents, as the similarity of legal cases is related to these symbols like elements. (2) **Knowledge incorporation**. As semantic-level matching is insufficient for SCM, we need to consider about incorporating legal knowledge into models to improve the performance and provide interpretability.

4.3 Legal Question-Answering

Another typical application of LegalAI is Legal Question Answering (LQA) which aims at answering questions in the legal domain. One of the most important parts of legal professionals' work is to provide reliable and high-quality legal consulting services for non-professionals. However, due to the insufficient number of legal professionals, it is often challenging to ensure that non-professionals

	KD-Questions		CA-Questions		All	
	Single	All	Single	All	Single	All
Unskilled Humans	76.9	71.1	62.5	58.0	70.0	64.2
Skilled Humans	80.6	77.5	86.8	84.7	84.1	81.1
BiDAF	36.7	20.6	37.2	22.2	38.3	22.0
BERT	**38.0**	21.2	38.9	**23.7**	39.7	**22.3**
Co-matching	35.8	20.2	35.8	20.3	38.1	21.2
HAF	36.6	**21.4**	**42.5**	19.8	**42.6**	21.2

Table 6: Experimental results of JEC-QA. The evaluation metrics is accuracy. The performance of unskilled and skilled humans is collected from original paper.

Question: Which crimes did Alice and Bob commit if they transported more than 1.5 million yuan of counterfeit currency from abroad to China?

Direct Evidence

P1: Transportation of counterfeit money: ··· The defendants are sentenced to three years in prison.
P2: Smuggling counterfeit money: ··· The defendants are sentenced to seven years in prison.

Extra Evidence

P3: Motivational concurrence: The criminals carry out one behavior but commit several crimes.
P4: For motivational concurrence, the criminals should be convicted according to the more serious crime.

Comparison: seven years > three years

Answer: Smuggling counterfeit money.

Table 7: An example of LQA from Zhong et al. (2019a). In this example, direct evidence and extra evidence are both required for answering the question. The hard reasoning steps prove the difficulty of legal question answering.

can get enough and high-quality consulting services, and LQA is expected to address this issue.

In LQA, the form of questions varies as some questions will emphasize on the explanation of some legal concepts, while others may concern the analysis of specific cases. Besides, questions can also be expressed very differently between professionals and non-professionals, especially when describing domain-specific terms. These problems bring considerable challenges to LQA, and we conduct experiments to demonstrate the difficulties of LQA better in the following parts.

Related Work

In LegalAI, there are many datasets of question answering. Duan et al. (2019) propose CJRC, a legal reading comprehension dataset with the same format as SQUAD 2.0 (Rajpurkar et al., 2018), which includes span extraction, yes/no questions, and unanswerable questions. Besides, COLIEE (Kano et al., 2018) contains about 500 yes/no questions. Moreover, the bar exam is a professional qualification examination for lawyers, so bar exam datasets (Fawei et al., 2016; Zhong et al., 2019a) may be quite hard as they require professional legal knowledge and skills.

In addition to these datasets, researchers have also worked on lots of methods on LQA. The rule-based systems (Buscaldi et al., 2010; Kim et al., 2013; Kim and Goebel, 2017) are prevalent in early research. In order to reach better performance, researchers utilize more information like the explanation of concepts (Taniguchi and Kano, 2016; Fawei et al., 2015) or formalize relevant documents as graphs to help reasoning (Monroy et al., 2009, 2008; Tran et al., 2013). Machine learning and deep learning methods like CRF (Bach et al., 2017), SVM (Do et al., 2017), and CNN (Kim et al., 2015) have also been applied to LQA. However, most existing methods conduct experiments on small datasets, which makes them not necessarily applicable to massive datasets and real scenarios.

Experiments and Analysis

We select JEC-QA (Zhong et al., 2019a) as the dataset of the experiments, as it is the largest dataset collected from the bar exam, which guarantees its difficulty. JEC-QA contains 28, 641 multiple-choice and multiple-answer questions, together with 79, 433 relevant articles to help to answer the questions. JEC-QA classifies questions into knowledge-driven questions (KD-Questions) and case-analysis questions (CA-Questions) and reports the performances of humans. We implemented several representative question answering models, including BiDAF (Seo et al., 2016), BERT (Devlin et al., 2019), Co-matching (Wang et al., 2018), and HAF (Zhu et al., 2018). The experimental results can be found in Table 6.

From the experimental results, we can learn the

models cannot answer the legal questions well compared with their promising results in open-domain question answering and there is still a huge gap between existing models and humans in LQA.

For more qualified LQA methods, there are several significant difficulties to overcome: (1) **Legal multi-hop reasoning**. As Zhong et al. (2019a) state, existing models can perform inference but not multi-hop reasoning. However, legal cases are very complicated, which cannot be handled by single-step reasoning. (2) **Legal concepts understanding**. We can find that almost all models are better at case analyzing than knowledge understanding, which proves that knowledge modelling is still challenging for existing methods. How to model legal knowledge to LQA is essential as legal knowledge is the foundation of LQA.

5   Conclusion

In this paper, we describe the development status of various LegalAI tasks and discuss what we can do in the future. In addition to these applications and tasks we have mentioned, there are many other tasks in LegalAI like legal text summarization and information extraction from legal contracts. Nevertheless, no matter what kind application is, we can apply embedding-based methods for better performance, together with symbol-based methods for more interpretability.

Besides, the three main challenges of legal tasks remain to be solved. Knowledge modelling, legal reasoning, and interpretability are the foundations on which LegalAI can reliably serve the legal domain. Some existing methods are trying to solve these problems, but there is still a long way for researchers to go.

In the future, for these existing tasks, researchers can focus on solving the three most pressing challenges of LegalAI combining embedding-based and symbol-based methods. For tasks that do not yet have a dataset or the datasets are not large enough, we can try to build a large-scale and high-quality dataset or use few-shot or zero-shot methods to solve these problems.

Furthermore, we need to take the **ethical issues** of LegalAI seriously. Applying the technology of LegalAI directly to the legal system will bring ethical issues like gender bias and racial discrimination. The results given by these methods cannot convince people. To address this issue, we must note that the goal of LegalAI is not replacing the legal professionals but helping their work. As a result, we should regard the results of the models only as a reference. Otherwise, the legal system will no longer be reliable. For example, professionals can spend more time on complex cases and leave the simple cases for the model. However, for safety, these simple cases must still be reviewed. In general, LegalAI should play as a supporting role to help the legal system.

References

Alan Akbik, Tanja Bergmann, and Roland Vollgraf. 2019. Pooled contextualized embeddings for named entity recognition. In *Proceedings of NAACL*.

Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preotiuc-Pietro, and Vasileios Lampos. 2016. Predicting judicial decisions of the european court of human rights: A natural language processing perspective. *PeerJ Computer Science*, 2.

Iosif ANGELIDIS, Ilias CHALKIDIS, and Manolis KOUBARAKIS. 2018. Named entity recognition, linking and generation for greek legislation.

Kevin D Ashley. 2017. *Artificial intelligence and legal analytics: new tools for law practice in the digital age*. Cambridge University Press.

Ngo Xuan Bach, Tran Ha Ngoc Thien, Tu Minh Phuong, et al. 2017. Question analysis for vietnamese legal question answering. In *Proceedings of KSE*. IEEE.

Deanna Barmakian. 2000. Better search engines for law. *Law Libr. J.*, 92.

Roberto Bartolini, Alessandro Lenci, Simonetta Montemagni, Vito Pirrelli, and Claudia Soria. 2004. Semantic mark-up of Italian legal texts through NLP-based techniques. In *Proceedings of LREC*.

Paheli Bhattacharya, Kaustubh Hiware, Subham Rajgaria, Nilay Pochhi, Kripabandhu Ghosh, and Saptarshi Ghosh. 2019. A comparative study of summarization algorithms applied to legal case judgments. In *Proceedings of ECIR*. Springer.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko.

2013. Translating embeddings for modeling multi-relational data. In *Advances in neural information processing systems*, pages 2787–2795.

Mírian Bruckschen, Caio Northfleet, Paulo Bridi, Roger Granada, Renata Vieira, Prasad Rao, and Tomas Sander. 2010. Named entity recognition in the legal domain for ontology population. In *Workshop Programme*, page 16. Citeseer.

Davide Buscaldi, Paolo Rosso, José Manuel Gómez-Soriano, and Emilio Sanchis. 2010. Answering questions with an n-gram based passage retrieval engine. *Journal of Intelligent Information Systems*, 34(2):113–134.

Cristian Cardellino, Milagro Teruel, Laura Alonso Alemany, and Serena Villata. 2017. Legal NERC with ontologies, Wikipedia and curriculum learning. In *Proceedings of EACL*.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019a. Neural legal judgment prediction in English. In *Proceedings of ACL*.

Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2019b. Large-scale multi-label text classification on EU legislation. In *Proceedings of ACL*.

Ilias Chalkidis and Dimitrios Kampas. 2019. Deep learning in law: early adaptation and legal word embeddings trained on large corpora. *Artificial Intelligence and Law*, 27(2):171–198.

Huajie Chen, Deng Cai, Wei Dai, Zehui Dai, and Yadong Ding. 2019. Charge-based prison term prediction with deep gating network. In *Proceedings of EMNLP-IJCNLP*, pages 6363–6368.

Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of ACL*.

Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2018. A walk-based model on entity graphs for relation extraction. In *Proceedings of ACL*, pages 81–88.

František Cvrček, Karel Pala, and Pavel Rychlý. 2012. Legal electronic dictionary for Czech. In *Proceedings of LREC*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*.

Phong-Khac Do, Huy-Tien Nguyen, Chien-Xuan Tran, Minh-Tien Nguyen, and Minh-Le Nguyen. 2017. Legal question answering using ranking svm and deep convolutional neural network. *arXiv preprint arXiv:1703.05320*.

Xingyi Duan, Baoxin Wang, Ziyue Wang, Wentao Ma, Yiming Cui, Dayong Wu, Shijin Wang, Ting Liu, Tianxiang Huo, Zhen Hu, et al. 2019. Cjrc: A reliable human-annotated benchmark dataset for chinese judicial reading comprehension. In *Proceedings of CCL*. Springer.

Biralatei Fawei, Adam Wyner, and Jeff Pan. 2016. Passing a USA national bar exam: a first corpus for experimentation. In *Proceedings of LREC*.

Biralatei Fawei, Adam Wyner, Jeff Z Pan, and Martin Kollingbaum. 2015. Using legal ontologies with rules for legal textual entailment. In *AI Approaches to the Complexity of Legal Systems*, pages 317–324. Springer.

Jianfeng Gao, Kristina Toutanova, and Wen-tau Yih. 2011. Clickthrough-based latent semantic models for web search. In *Proceedings of SIGIR*. ACM.

Anne von der Lieth Gardner. 1984. An artificial intelligence approach to legal reasoning.

Anton Geist. 2009. Using citation analysis techniques for computer-assisted legal research in continental jurisdictions. *Available at SSRN 1397674*.

Ben Hachey and Claire Grover. 2006. Extractive summarisation of legal texts. *Artificial Intelligence and Law*, 14(4):305–345.

Hiroaki Hayashi, Zecong Hu, Chenyan Xiong, and Graham Neubig. 2019. Latent relation language models. *arXiv preprint arXiv:1908.07690*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8).

Baotian Hu, Zhengdong Lu, Hang Li, and Qingcai Chen. 2014. Convolutional neural network architectures for matching natural language sentences. In *Proceedings of NIPS*.

Zikun Hu, Xiang Li, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. 2018. Few-shot charge prediction with discriminative legal attributes. In *Proceedings of COLING*.

Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of CIKM*. ACM.

Jyun-Yu Jiang, Mingyang Zhang, Cheng Li, Michael Bendersky, Nadav Golbandi, and Marc Najork. 2019. Semantic text matching for long-form documents. In *Proceedings of WWW*. ACM.

Rie Johnson and Tong Zhang. 2017. Deep pyramid convolutional neural networks for text categorization. In *Proceedings of ACL*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Santosh Kabbur, Xia Ning, and George Karypis. 2013. Fism: factored item similarity models for top-n recommender systems. In *Proceedings of SIGKDD*. ACM.

Liangyi Kang, Jie Liu, Lingqiao Liu, Qinfeng Shi, and Dan Ye. 2019. Creating auxiliary representations from charge definitions for criminal charge prediction. *arXiv preprint arXiv:1911.05202*.

Yoshinobu Kano, Mi-Young Kim, Masaharu Yoshioka, Yao Lu, Juliano Rabelo, Naoki Kiyota, Randy Goebel, and Ken Satoh. 2018. Coliee-2018: Evaluation of the competition on legal information extraction and entailment. In *Proceedings of JSAI*, pages 177–192. Springer.

R Keown. 1980. Mathematical models for legal prediction. *Computer/LJ*, 2:829.

Mi-Young Kim and Randy Goebel. 2017. Two-step cascaded textual entailment for legal bar exam question answering. In *Proceedings of Articial Intelligence and Law*. ACM.

Mi-Young Kim, Ying Xu, and Randy Goebel. 2015. A convolutional neural network in legal question answering.

Mi-Young Kim, Ying Xu, Randy Goebel, and Ken Satoh. 2013. Answering yes/no questions in legal bar exams. In *Proceedings of JSAI*, pages 199–213. Springer.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of EMNLP*.

Fred Kort. 1957. Predicting supreme court decisions mathematically: A quantitative analysis of the "right to counsel" cases. *American Political Science Review*, 51(1):1–12.

Onur Kuru, Ozan Arkan Can, and Deniz Yuret. 2016. CharNER: Character-level named entity recognition. In *Proceedings of COLING*.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of NAACL*.

Jörg Landthaler, Bernhard Waltl, Patrick Holl, and Florian Matthes. 2016. Extending full text search for legal document collections using word embeddings. In *JURIX*, pages 73–82.

Benjamin E Lauderdale and Tom S Clark. 2012. The supreme court's many median justices. *American Political Science Review*, 106(4):847–866.

Alessandro Lenci, Simonetta Montemagni, Vito Pirrelli, and Giulia Venturi. 2009. Ontology learning from italian legal texts. *Law, Ontologies and the Semantic Web*, 188:75–94.

Shang Li, Hongli Zhang, Lin Ye, Xiaoding Guo, and Binxing Fang. 2019a. Mann: A multichannel attentive neural network for legal judgment prediction. *IEEE Access*.

Yu Li, Tieke He, Ge Yan, Shu Zhang, and Hui Wang. 2019b. Using case facts to predict penalty with deep learning. In *International Conference of Pioneering Computer Scientists, Engineers and Educators*, pages 610–617. Springer.

Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning entity and relation embeddings for knowledge graph completion. In *Proceedings of AAAI*.

Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. In *Proceedings of ACL*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019a. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zhiyuan Liu, Cunchao Tu, and Maosong Sun. 2019b. Legal cause prediction with inner descriptions and outer hierarchies. In *Proceedings of CCL*, pages 573–586. Springer.

Daniel Locke and Guido Zuccon. 2018. A test collection for evaluating legal case law search. In *Proceedings of SIGIR*. ACM.

Bingfeng Luo, Yansong Feng, Jianbo Xu, Xiang Zhang, and Dongyan Zhao. 2017. Learning to predict charges for criminal cases with legal basis. In *Proceedings of EMNLP*.

K Tamsin Maxwell and Burkhard Schafer. 2008. Concept and context in legal information retrieval. In *Proceedings of JURIX*.

Douglas L Medin. 2000. *Psychology of learning and motivation: advances in research and theory*. Elsevier.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Makoto Miwa and Mohit Bansal. 2016. End-to-end relation extraction using lstms on sequences and tree structures. In *Proceedings of ACL*, pages 1105–1116.

Alfredo Monroy, Hiram Calvo, and Alexander Gelbukh. 2008. Using graphs for shallow question answering on legal documents. In *Mexican International Conference on Artificial Intelligence*. Springer.

Alfredo Monroy, Hiram Calvo, and Alexander Gelbukh. 2009. Nlp for shallow question answering of legal documents using graphs. In *Proceedings of CICLing*. Springer.

Alfredo López Monroy, Hiram Calvo, Alexander Gelbukh, and Georgina García Pacheco. 2013. Link analysis for representing and retrieving legal information. In *Proceedings of CICLing*, pages 380–393. Springer.

Stuart S Nagel. 1963. Applying correlation analysis to case prediction. *Texas Law Review*, 42:1006.

John J. Nay. 2016. Gov2Vec: Learning distributed representations of institutions and their legal text. In *Proceedings of the First Workshop on NLP and Computational Social Science*.

Thien Huu Nguyen, Kyunghyun Cho, and Ralph Grishman. 2016. Joint event extraction via recurrent neural networks. In *Proceedings of NAACL*.

Thien Huu Nguyen and Ralph Grishman. 2018. Graph convolutional networks with argument-aware pooling for event detection. In *Proceedings of AAAI*.

Hamid Palangi, Li Deng, Yelong Shen, Jianfeng Gao, Xiaodong He, Jianshu Chen, Xinying Song, and Rabab Ward. 2016. Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 24(4).

Sicheng Pan, Tun Lu, Ning Gu, Huajuan Zhang, and Chunlin Xu. 2019. Charge prediction for multidefendant cases with multi-scale attention. In *CCF Conference on Computer Supported Cooperative Work and Social Computing*. Springer.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP*, pages 1532–1543.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Matthew E Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A Smith. 2019. Knowledge enhanced contextual word representations. In *Proceedings of EMNLP-IJCNLP*.

Xipeng Qiu and Xuanjing Huang. 2015. Convolutional neural tensor network architecture for community-based question answering. In *Proceedings of IJCAI*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

K Raghav, P Krishna Reddy, and V Balakista Reddy. 2016. Analyzing the extraction of relevant legal judgments using paragraph-level and citation information. *AI4JCArtificial Intelligence for Justice*, page 30.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of ACL*.

Steffen Rendle. 2010. Factorization machines. In *Proceedings of ICDM*. IEEE.

Stephen E Robertson and Steve Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of SIGIR*.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management*.

Jeffrey A Segal. 1984. Predicting supreme court cases probabilistically: The search and seizure cases, 1962-1981. *American Political Science Review*, 78(4):891–900.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*.

Yelong Shen, Xiaodong He, Jianfeng Gao, Li Deng, and Grégoire Mesnil. 2014. A latent semantic model with convolutional-pooling structure for information retrieval. In *Proceedings of CIKM*. ACM.

Yi Shu, Yao Zhao, Xianghui Zeng, and Qingli Ma. 2019. Cail2019-fe. Technical report, Gridsum.

Keet Sugathadasa, Buddhi Ayesha, Nisansa de Silva, Amal Shehan Perera, Vindula Jayawardana, Dimuthu Lakmal, and Madhavi Perera. 2018. Legal document retrieval using document vector embeddings and deep learning. In *Proceedings of SAI*. Springer.

Harry Surden. 2018. Artificial intelligence and law: An overview. *Ga. St. UL Rev.*

Ryosuke Taniguchi and Yoshinobu Kano. 2016. Legal yes/no question answering system using case-role analysis. In *Proceedings of JSAI*, pages 284–298. Springer.

Oanh Thi Tran, Bach Xuan Ngo, Minh Le Nguyen, and Akira Shimazu. 2013. Answering legal questions by mining reference information. In *Proceedings of JSAI*. Springer.

Vu Tran, Minh Le Nguyen, and Ken Satoh. 2019. Building legal case retrieval systems with lexical matching and summarization using a pre-trained phrase scoring model. In *Proceedings of Artificial Intelligence and Law*. ACM.

Maarten Truyens and Patrick Van Eecke. 2014. Legal aspects of text mining. In *Proceedings of LREC*.

Howard Turtle. 1995. Text retrieval in the legal world. *Artificial Intelligence and Law*, 3(1-2).

S Sidney Ulmer. 1963. Quantitative analysis of judicial processes: Some practical and theoretical applications. *Law and Contemporary Problems*, 28:164.

Thomas Vacek, Ronald Teo, Dezhao Song, Timothy Nugent, Conner Cowling, and Frank Schilder. 2019. Litigation analytics: Case outcomes extracted from US federal court dockets. In *Proceedings of NLLP Workshop*.

Tom Vacek and Frank Schilder. 2017. A sequence approach to case outcome detection. In *Proceedings of Articial Intelligence and Law*, pages 209–215. ACM.

Marc Van Opijnen and Cristiana Santos. 2017. On the concept of relevance in legal information retrieval. *Artificial Intelligence and Law*, 25(1).

Hui Wang, Tieke He, Zhipeng Zou, Siyuan Shen, and Yu Li. 2019. Using case facts to predict accusation based on deep learning. In *Proceedings of QRS-C*, pages 133–137. IEEE.

Shuohang Wang, Mo Yu, Jing Jiang, and Shiyu Chang. 2018. A co-matching model for multi-choice reading comprehension. In *Proceedings of ACL*.

Wei Wu, Hang Li, and Jun Xu. 2013. Learning query and document similarities from click-through bipartite graph with metadata. In *Proceedings of WSDM*. ACM.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, et al. 2018. Cail2018: A large-scale legal dataset for judgment prediction. *arXiv preprint arXiv:1807.02478*.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Tianyang Zhang, Xianpei Han, Heng Wang, Jianfeng Xu, et al. 2019. Cail2019-scm: A dataset of similar case matching in legal domain. *arXiv preprint arXiv:1911.08962*.

Jun Xu, Hang Li, and Chaoliang Zhong. 2010. Relevance ranking using kernels. In *Proceedings of AIRS*. Springer.

Yukun Yan, Daqi Zheng, Zhengdong Lu, and Sen Song. 2017. Event identification as a decision process with non-linear representation of text. *arXiv preprint arXiv:1710.00969*.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2014. Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Hai Ye, Xin Jiang, Zhunchen Luo, and Wenhan Chao. 2018. Interpretable charge predictions for criminal cases: Learning to generate court views from fact descriptions. In *Proceedings of NAACL*.

Wenpeng Yin, Hinrich Schütze, Bing Xiang, and Bowen Zhou. 2016. ABCNN: Attention-based convolutional neural network for modeling sentence pairs. *Transactions of the Association for Computational Linguistics*.

Xiaoxiao Yin, Daqi Zheng, Zhengdong Lu, and Ruifang Liu. 2018. Neural entity reasoner for global consistency in ner. *arXiv preprint arXiv:1810.00347*.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of EMNLP*.

Ni Zhang, Yi-Fei Pu, Sui-Quan Yang, Ji-Liu Zhou, and Jin-Kang Gao. 2017. An ontological chinese legal consultation system. *IEEE Access*, 5:18250–18261.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of ACL*.

Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun. 2018. Legal judgment prediction via topological learning. In *Proceedings of EMNLP*.

Haoxi Zhong, Yuzhong Wang, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. Iteratively questioning and answering for interpretable legal judgment prediction. In *Proceedings of AAAI*.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2019a. Jec-qa: A legal-domain question answering dataset. *arXiv preprint arXiv:1911.12011*.

Haoxi Zhong, Zhengyan Zhang, Zhiyuan Liu, and Maosong Sun. 2019b. Open chinese language pretrained model zoo. Technical report, Technical Report. Technical Report.

Haichao Zhu, Furu Wei, Bing Qin, and Ting Liu. 2018. Hierarchical attention flow for multiple-choice reading comprehension. In *Proceedings of AAAI*.