Trends in Social Sciences and Humanities Research
ISSN: 2959-9938
DOI:10.61784/tsshr231229

RESEARCH PROGRESS IN LEGAL TEXT PROCESSING BASED ON DEEP LEARNING

K. Smith Leiren
CICERO - Center for International Climate Research, Norway.

Abstract: With the continuous advancement of China's judicial informatization construction, the amount of legal text data represented by various case files, judgment documents, laws and regulations, and judicial interpretations Rapidly growing, research on legal text processing based on deep learning has become a hot issue at the intersection of law and artificial intelligence. In order to promptly follow up on the latest developments in this field, New research results, this article covers legs and countries in this field. The representative achievements of domestic and foreign scholars are analyzed and the future development trends of this field are analyzed and prospected.
Keywords: Deep learning; Legal text processing; Textual representation; Text Categorization

1. INTRODUCTION

In the judicial field, with the increasing legal awareness of the broad masses of the people, continues to strengthen, the growth rate of new cases is increasing day by day, coupled with the legal In order to adapt to the endless new things emerging in society, we constantly make updates. and improvement, so that a large amount of new data appears every day. These numbers The data comes from various civil and criminal case files and judgment documents, and and supplementary expansion and judicial interpretation of laws and regulations. At the same time, China As China's judicial informatization construction continues to advance, these data have been screened and Purges are also being published more publicly. The China Judgment Documents Network is run by Judgment document release website hosted by the Supreme People's Court, including documents quantity up to More than 100 million articles and still growing, it has become a legal The largest data database in the legal field.

On the other hand, as the amount of data continues to increase, judicial workers The burden is also becoming increasingly heavy. Judges and lawyers not only need to consult a large number of historical cases as a reference, as well as new laws and regulations as well as existing Supplementary expansion of laws and regulations for in-depth understanding and research. in recent years Come, with deep learning and natural language processing (natural language processing, NLP) The artificial intelligence technology represented by AI continues to make new breakthroughs, and its research results have promoted manufacturing, medical care, education The development of many other fields has improved the production efficiency of these fields, thus And reduce people's labor burden. In the judicial field, artificial intelligence Energy-related research is generally still in its infancy.

Text processing is a related field in traditional machine learning and data mining fields. Basic but also very important technical branches, including text representation, aggregation Category, classification, retrieval and other subdivisions. The legal field is the most important The data form is the legal document represented by the content of the judgment document. This book, as shown in the picture As shown in 1, its content mainly involves information about the defendant and the case. A description of the circumstances of the incident and the outcome of the verdict.

In order to fully tap the value of legal text data and alleviate legal Practitioners are burdened with heavy data processing work.

2. LEGAL TEXT REPRESENTATION BASED ON DEEP LEARNING

The text representation is many NLP Basic tasks in applications play a very important role in improving the performance of various text processing algorithms. The goal of text representation is to map unstructured text data into a low-dimensional vector space, so that the text can be calculated and processed using mathematical methods [1] . Compared with texts in general fields, legal texts have the characteristics of strong domain, dense information, and relatively obvious structural features. More effective legal text technology can significantly improve modeling, classification, The performance of downstream tasks such as reasoning and mining has attracted widespread interest among researchers in recent years.

2.1 Embedding-Based Legal Text Representation

Character and word embeddings are an important means of vectorized representation of language, but traditional embedding methods (like Word2Vec) has relatively insufficient ability to express professional terms and domain knowledge in legal texts. Nay By applying Word 2 Vec on a legal corpus consisting of case law, statutory law and administrative law, a Gov 2 Vec is obtained by training This tool can effectively encode legal concepts in the corpus

and learn the implicit relationships between these concept vectors. It has been successfully used in the task of generating summaries of Supreme Court opinions, presidential actions, and congressional bills [2] ; Chalkidis and Kam pas Also based on Word2Vec proposed Law2Vec trains legal vocabulary embeddings in a large corpus including the legislation of the United Kingdom, the European Union, Canada, Australia, the United States, and Japan, and verifies the use of legal vocabulary semantic feature representation in text classification, information extraction, and information retrieval. important role in three tasks [3] .

since Since 2018, with Pre-training language represented by BERT model has formed a new NLP Paradigm [4] : First use the big pre-training on large-scale text corpora, and then on decimals for specific tasks Data set fine-tuning to reduce individual NLP The difficulty of the task. The application of pre-trained language models has greatly improved named entity recognition and event extraction. retrieval, machine translation, automatic question and answer, etc. NLP task performance, in The field of legal text processing also has good application prospects. Targeted at pass Use the pre-trained language model to express terms and knowledge in the legal field to improve the understanding and reasoning capabilities of the deep learning model for legal concepts.

2.2 Feature-Based Legal Text Representation

Embedding-based legal text representation approaches take full advantage of depth neural network in NLP Powerful latent semantic learning ability in tasks, However, the text vectors it generates are often unexplainable, which is important for emphasizing the collar Domain knowledge of legal texts is a significant shortcoming. And the traditional The engineering method requires a lot of manual annotation work and is difficult to implement on a large scale. The legal corpus is also stretched thin. Therefore, there are studies Researchers began to try a combination of these two methods, that is, using a certain Amount of domain knowledge is used to define the characteristic patterns of legal text representation, and then Finally, a deep neural network model is used at the bottom layer to learn these features. Practice and expression.

Li According to the definition of the crime of theft in Chinese criminal law, Retrieve information related to conviction and sentencing 9- dimensional characteristics (including basic information of criminal suspects, whether they are repeat offenders, whether they carry weapons, and the value of items involved in the case) wait), and then use long short-term memory (long short - term memory, LSTM) network encodes legal text and then generates Quantitative representation uses a classification algorithm to determine whether it meets a certain feature, and then get legal text 9- dimensional vector representation, implementing features While reducing dimensionality, the features can have good interpretability under the legal knowledge framework [6] . For the judgment result prediction task, Li et al. A legal text representation model based on attention mechanism is proposed. by involving The system is trained on the corpus of judgment documents for 10 types of criminal crimes, and generates data based on case facts, defendant information and relevant criminal law provisions. latent semantic feature representation vectors at multiple levels such as text, which can represent Characters, events and legal provisions in legal texts The potential logical relationship between the three greatly improves the prediction of charges, legal provisions, prison terms, etc. service performance and the interpretability of prediction results [7] .

3. LEGAL TEXT CLASSIFICATION BASED ON DEEP LEARNING

Text classification is a critical task in legal text processing applications. Different legal text processing tasks can be transformed into different types of text classification problems. For example: Determining whether the defendant in a case has the circumstances to surrender is a simple two-classification problem, analyzing the type of case (the main alleged crimes are mutually exclusive) It is a multi-classification problem, and determining which laws the defendant has violated is a multi-label classification problem. Existing research work basically focuses on this 3 types of questions are expanded.

Aletras wait people make use Multiple support vector machines (Support Vector Machine, SVM) Classifier for several semantics of cases The features are classified into two categories respectively and used to predict the judgment of the European Court of Human Rights. Judgment [8] ; Boella et al. used word frequency - inverse file frequency (Term Frequency – Inverse Document Frequency, TF – IDF) algorithm and information gain for feature selection and then training SVM classifier to identify the field to which legal texts belong [9] ; Liu et al. Using K in case-based reasoning systems nearest neighbor (K – Nearest Neighbor, KNN) Algorithm pair 12 common criminal offense Classify by name [10] ; Katz et al. built a random tree model to predict the decision-making of the U.S. Supreme Court based on features extracted from case summaries [11] ; Lin et al. first based on the artificial definition of 21 types of legal elements Tags classify sentences describing cases and are used to distinguish robberies and intimidation charges [12] ; Liu and others combined different combinations of multiple legal provisions into Train for labels, reducing multi-label classification problems to multi-classification Question [13-14] . Most of these early works utilized feature engineering and Incorporation of statistical machine learning models using supervised learning methods When training a classifier, both model classification performance and result interpretability are comparable. is relatively good, but due to over-reliance on feature design and manual annotation, the paper This labeling system has poor scalability when changes occur.

In recent years, deep learning models represented by various types of neural networks have With its powerful feature learning ability, the model can be used in a variety of NLP On task Played an important role, especially for large-scale corpus learning, compared with the method of constructing features using artificial rules, it is better able to describe the

data Rich semantic information. Wei Using convolutional neural networks, et al. (Convolution Neural Network, CNN) Implemented a legal document classifier whose experimental results demonstrate The CNN model achieves significantly better performance on large-scale training sets than SVM [15] ; Chalkid is and Androutsopoulos uses words that do not rely on human annotation at all. itself, part-of-speech tags and symbol embeddings as features, using double Towards LSTM The network completed the contract element extraction task [16] ; Luo et al. Proposed a multi-label neural network classification based on attention mechanism device, by integrating legal and regulatory information into the vector representation of case facts, While improving the classification performance of case crimes, the classification results have a certain Certain interpretability [17] ; Li proposed a multi-channel attention neural network framework that only uses the crime types and applicable laws in the training data. article, sentence The three easily accessible labels are the supervisor's description of the case, the The defendant information and legal provisions are jointly coded, and the coding method is flexible. The formula can support different multi-label classification tasks and has achieved good results. classification performance [7] ; Wang proposed a hierarchical matching Neural network integrates the hierarchical information of labels in the process of constructing the vector representation of case charges, and completes the charges with the help of semantic matching method. Classification tasks have achieved high accuracy [18] .

4. LEGAL TEXT MINING AND APPLICATION

With legal text processing technologies such as legal text representation and classification the continued maturation of the legal sector, and the use of computers and artificial intelligence in the legal field The rapid growth in demand for technology-assisted business development has emerged in recent years Some representative legal text mining methods and their applications are proposed.

4.1 Legal Judgment Prediction

Legal Judgment Prediction (Legal Judgment Prediction (LJP) is one of the most critical tasks based on legal text. In China, Germany, In countries such as France that adopt civil law systems, the verdict is determined based on the facts of the case and statutory regulations. Under this legal regime, LJP The task is to determine whether the relevant behavior violates a certain law by matching the case fact description text with legal provisions, and then make predictions about the corresponding charges, applicable legal provisions, and prison terms.

Most of the existing studies use text segmentation to predict crimes and legal provisions. algorithm-like solving, including early use of statistical machine learning models, and Recent approaches using deep learning models. in order to promote LJP hair Zhan, Xiao et al proposed a large-scale database of Chinese judgment documents data set C- LJP, including the Chinese court issued 2.68 million prison cases File text [19] ; In some recent work, Luo and Li will study The focus is on how to use neural networks based on attention mechanisms to mine Dig cases and describe the logical relationships between different parts for better To achieve this purpose and provide better interpretable results for subsequent predictions Interpretability, legal provisions are introduced as external knowledge to guide the neural network The coding process of the network has achieved excellent results in the task of predicting crimes and legal provisions. Different properties [17, 7] ; Zhong by introducing LJP Each sub-responsibility The topological relationship between services makes the prediction process of the model more consistent with human Similar to the judge's judgment logic, the experimental results also confirmed the effectiveness of this approach Effectiveness [20] .

In terms of sentence prediction, some work divides the sentence into different intervals and then transforms it into a classification problem. There are also some researchers who design models based on regression problems that are more in line with the characteristics of the task itself. Li et al. summarized the punishment for theft cases in addition to the prison term based on legal provisions. 10- dimensional features, using neural network training to obtain feature vectors and then submitting them to the regression algorithm for calculation, achieving a high accuracy. However, this method relies relatively on manually introducing external knowledge and annotations, and cannot efficiently expand the prediction model to support more types of cases. Part [21] ; Chen proposed a neural network model using a gating mechanism to predict the sentence of the case based on the crime, which effectively improved the accuracy of the prediction [22] . But in general, due to the continuous characteristics of the sentence data type and the existence of sentencing factors outside the law in reality, the performance of the existing models is not ideal.

4.2 Similar Case Search

As the size of case documents increases, similar case retrieval It is of great significance to improve the work efficiency of legal practitioners. High-quality push results of similar cases also help Chinese law get closer The goal pursued is " si milar cases and similar judgments ".

At early stages research work In, Saravanan and Casanovas Proposed a legal case retrieval system based on semantic web and ontology, It is more practical than traditional keyword-based systems at both input and output ends. Its disadvantage is that it relies heavily on legal experts to edit the ontology. Moreover, using ontology as a search condition cannot satisfy the current " case-based search". "Project " business needs [23-24] .

Common law countries use case law to decide a case. Judgments in previous cases must be clearly cited when issuing a judgment, so naturally A case citation network was formed to introduce graph algorithms to solve similar cases. Retrieval questions provide the basis. Wagh based on case citation network proposed a method to calculate the

centrality and betweenness of network nodes. Indian Court Method to determine similarity [25] ; Minocha et al proposed a method The concept of law dispersion, by measuring the adjacent node sets of two cases The similarity of the combination is used to find similar cases of a case in the citation network [26] . In response to the problem that citation networks are usually very sparse, some researchers have begun to introduce machine learning algorithms to calculate the similarity of legal texts. Calculation, such as calculating full-text similarity based on paragraph similarity, calculating based on word frequency Bayesian statistical methods and nearest neighbor algorithms based on case characteristics, but this Some methods based on statistical features lose the original semantic information of the text. interest. In order to preserve the semantic information of the text as much as possible, word embeddings are used And deep learning models have gradually become the mainstream methods for similar case retrieval tasks.

5. CONCLUSION

Aiming at the problem of legal text processing, this article briefly introduces the Related research results based on deep learning methods in the past, respectively Legal text representation, legal text classification, and legal text mining and application The research directions and progress in the field were sorted out and analyzed. Eliminate capital In addition to these directions introduced in this article, legal text processing also involves tasks Including legal questions and answers, legal element extraction, legal text abstracts, etc.

Overall, traditional text processing techniques can be used in legal play an important role in text processing tasks, and word embedding methods and The introduction of deep learning models represented by neural networks can even Fully learn the huge semantic information contained in massive legal texts. but is, how to make deep learning models better match legal expertise The integration of Taking into account model performance and result interpretability will become a future research topic in this field. the focus of research.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Li Yiming. Research on text representation method combining knowledge and neural network. Hangzhou: Zhejiang University, 2019.

[2] NAY J J. Gov2Vec: Learning Distributed Representations of Institutions and Their Legal Text. In Proceedings of the First Workshop on NLP and Computational Social Science, 2016: 49-54.

[3] CHALKIDIS I, KAMPAS D. Deep learning in law: early adaptation and legal word embedding s trained on large corpora. Artificial Intelligence and Law, 2019, 27(2): 171-198.

[4] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL, 2019: 4171-4186.

[5] ZHONG H, ZHANG Z, LIU Z, et al. Open Chinese Language Pre -trained Model Zoo. 2019. https: / /github. com /thunlp /opendap.

[6] LI S, ZHANG H, YE L, et al. Evaluating the Rationality of Judicial Decision with LSTM -based Case Modeling. In Proceedings of ICDSC, 2018: 392-397.

[7] LI S, ZHANG H, YE L, et al. MANN: A Multichannel Attentive Neural Network for Legal Judgment Prediction. IEEE Access, 2019, 7(1): 151144-151155.

[8] LETTERS N, TSARAPATSANIS D, PREOI UC:PIETRO D, et al.Predicting judicial decisions of the European Court of Human Rights Rights: A natural language processing perspective 1 J 1 PeerJ Computer Science, 2016, 2 (10): 93.

[9] BOELLA G, CARO L D, HUMPHREYS L. Using classification to support legal knowledge engineers in the eunomos legal document management system. In Proceedings of Juris - Information, 2011: 1-12.

[10] LIU C, CHANG C, HO J. Case instance generation and refinement for case-based criminal summary judgments in Chinese. Journal of Information Science and Engineering, 2004, 20 (4): 783-800.

[11] KATZ D M, BOM MARITO II M J, BLACKMAN J. A general approach for predicting the behavior of the Supreme Court of the United States. PLoS One, 2017, 12(4): 12.

[12] LIN W, KUO T, CHANG T, et al. Exploiting machine learning models for Chinese legal documents labeling, case classification, and sentencing prediction. Computational Linguistics and Chinese Language Processing, 2012, 17(4): 49-68.

[13] LIU C, HSIEH C. Exploring phrase -based classification of judicial documents for criminal charges in Chinese . InProceedings of ISMIS, 2006: 681-690.

[14] LIU C, LIAO T. Classifying criminal charges in Chinese for web- based legal services. InProceedings of APWC, 2005: 64-75.

[15] WEI F, QIN H, YE S, et al. Empirical study of deep learning for text classification in legal document review. InProceedings of BigData, 2018: 3317-3320.

[16] CHALKIDIS I, ANDROUTSOPOULOS I. A deep learning approach to contract element extraction. In Proceedings of JURIX, 2017: 155-164.

[17] LUO B, FENG Y, XU J, et al. Learning to predict charges for criminal cases with legal basis. InProceedings of EM NLP, 2017: 2727-2736.
[18] WANG P, FAN Y, NIU S, et al. Hierarchical matching network for crime classification. In Proceedings of SIGIR, 2019: 325-334.
[19] XIAO C, ZHONG H, GUO Z, et al. CAIL2018: A large-scale legal dataset for judgment prediction. https://arxiv.org/abs/1807.02478v 1
[20] ZHONG H, GUO Z, TU C, et al. Legal judgment prediction via topological learning. In Proceedings of EM NLP, 2018: 3540-3549.
[21] LI S, ZHANG H, YE L, et al. Prison Term Prediction on Criminal Case Description with Deep Learning. CM C - Computers Materials & Continua, 2020, 62(3): 1217-1231.
[22] CHEN H, CAI D, DAI W, et al. Charge-based prison term prediction with deep gating network. In Proceedings of EM NLP-IJCNLP, 2019: 6363-6368.
[23] SARAVANAN M, RAVINDRAN B, RAMAN S. Improving legal information retrieval using an ontological framework. Artificial Intelligence & Law, 2009, 17(2): 101-124.
[24] CASANOVAS P, CASANOVAS P, PALM IRANI M, et al. Semantic web for the legal domain: the next step. Semantic Web, 2016, 7(3): 213-227.
[25] WAGH R S, ANAND D. Application of citation network analysis for improved similarity index estimation of legal case documents: a study. In Proceedings of ICCTAC, 2017: 1-5.
[26] MINOCHA A, SINGH N, SRIVASTAVA A, et al. Finding Relevant Indian Judgments using Dispersion of Citation Network. In Proceedings of the Web Conference, 2015: 1085-1088.