文章编号：2095-2163(2021)08-0083-05 中图分类号：TP391 文献标志码：A

基于深度学习的法律文本处理研究进展

李 尚，张宏莉，叶 麟，方滨兴
（哈尔滨工业大学 网络空间安全学院，哈尔滨 150001）

摘 要： 随着中国司法信息化建设的不断推进，以各类案件卷宗、裁判文书、法律法规以及司法解释为代表的法律文本数据量迅速增长，基于深度学习的法律文本处理研究已成为法律与人工智能这一交叉领域的热点问题。为了及时跟进该领域的最新研究成果，本文分别从法律文本表示、法律文本分类、法律文本挖掘与应用等 3 个方面梳理了该领域中的主要研究方向和国内外学者的代表性成果，并对该领域未来的发展趋势进行了分析和展望。

关键词： 深度学习；法律文本处理；文本表示；文本分类

Research advances of deep learning-based legal text processing

LI Shang, ZHANG Hongli, YE Lin, FANG Binxing
(School of Cyberspace Science, Harbin Institute of Technology, Harbin 150001, China)

【Abstract】 The persistent construction of China's judicial informatization has led to the rapid increase of legal text data represented by various case files, judgment documents, laws and regulations, and judicial interpretations. The research of legal text processing based on deep learning has become a hot issue in the intersection of law and artificial intelligence. In order to follow up the latest research results in this field, this paper compares the main research directions and representative achievements of domestic and foreign scholars from three aspects, including legal text representation, legal text classification, legal text mining and applications, and analyzes and outlooks the future development trend of this field.

【Key words】 deep learning; legal text processing; text representation; text classification

0 引 言

在司法领域，随着广大人民群众的法律意识不断增强，新案件的增长速度日益提高，再加上法律为了适应社会中层出不穷的新生事物而不断做出更新和完善，使得每天都有大量的新数据出现。这些数据来自于各类民事和刑事案件卷宗和判决文书，以及法律法规的补充扩展和司法解释。与此同时，中国司法信息化建设不断推进，这些数据经过筛选和清洗也更多地公开发布出来，中国裁判文书网是由最高人民法院主办的裁判文书发布网站，收录文书数量多达 1 亿余篇且仍在继续增长，目前已成为法律领域最大规模的数据资料库。

另一方面，随着数据量的不断增加，司法工作者的负担也日益繁重，法官和律师不仅需要查阅大量历史案例作为参考，还要对新的法律法规以及现有法律法规的补充扩展进行深入理解和研究。近年来，以深度学习和自然语言处理（natural language processing，NLP）为代表的人工智能技术不断取得新的突破，其研究成果已经推动了制造、医疗、教育等诸多领域的发展，提高了这些领域的生产效率，从而减轻了人们的劳动负担。而在司法领域，人工智能的相关研究总体上仍处于起步阶段。

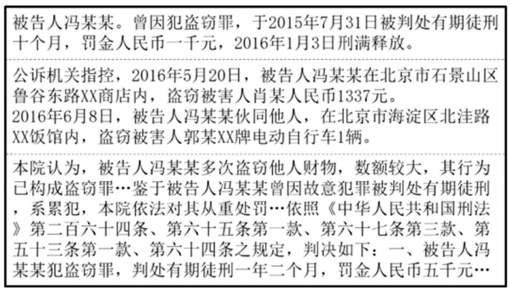
文本处理是传统机器学习和数据挖掘领域里相对基础但也非常重要的技术分支，包括文本表示、聚类、分类、检索等多个细分领域。而法律领域最主要的数据形式便是以裁判文书内容为代表的法律文本，如图 1 所示，其内容主要涉及对被告人信息、案件情节以及判决结果的描述。

为了充分挖掘法律文本数据的价值，减轻法律从业人员繁重的数据处理工作负担，近年来人工智能研究人员已经针对基于深度学习的法律文本处理技术开展了一系列工作，特别是在法律文本表示、法律文本分类以及几类典型的法律文本挖掘与应用方面，已产生一批代表性的成果。本文对这些研究工作和成果进行简要的梳理和分析。

基金项目： 国家重点研发计划课题(2018YFC0830902)。

作者简介： 李 尚(1989-)，男，博士研究生，主要研究方向：人工智能、信息安全；张宏莉(1973-)，女，博士，教授，博士生导师，主要研究方向：网络与信息安全、网络测量与建模、并行处理等；叶 麟(1982-)，男，博士，副教授，硕士生导师，主要研究方向：网络安全、网络测量、云计算等；方滨兴(1960-)，男，博士，教授，博士生导师，中国工程院院士，主要研究方向：网络与信息安全、并行计算等。

收稿日期： 2021-05-24


被告人冯某某。曾因犯盗窃罪，于2015年7月31日被判处有期徒刑十个月，罚金人民币一千元，2016年1月3日刑满释放。

公诉机关指控，2016年5月20日，被告人冯某某在北京市石景山区鲁谷东路XX商店内，盗窃被害人肖某人民币1337元。2016年6月8日，被告人冯某某伙同他人，在北京市海淀区北洼路XX饭馆内，盗窃被害人郭某XX牌电动自行车1辆。

本院认为，被告人冯某某多次盗窃他人财物，数额较大，其行为已构成盗窃罪…鉴于被告人冯某某曾因故意犯罪被判处有期徒刑，系累犯，本院依法对其从重处罚…依照《中华人民共和国刑法》第二百六十四条、第六十五条第一款、第六十七条第三款、第五十三条第一款、第六十四条之规定，判决如下：一、被告人冯某某犯盗窃罪，判处有期徒刑一年二个月，罚金人民币五千元…

图 1 法律文本(裁判文书)样例

Fig. 1 An example of legal text (judgment document)

1 基于深度学习的法律文本表示

文本表示是许多 NLP 应用中的基础性任务,对提升各类文本处理算法性能具有十分重要的作用。文本表示的目标是将非结构化的文本数据映射到低维向量空间中,进而可以用数学方法对文本进行计算和处理[1]。与通用领域的文本相比,法律文本具有领域性强、信息密集、结构特征相对明显等特点,更加有效的法律文本技术可以显著提升建模、分类、推理、挖掘等下游任务的性能,近年来已引起研究者的广泛兴趣。

1.1 基于嵌入的法律文本表示

字和词嵌入是对语言进行向量化表示的重要手段,但传统的嵌入方法(如 Word2Vec)对于法律文本中专业术语和领域知识的表达能力相对不足。Nay 通过在一个由案例法、成文法和行政法构成的法律语料库上应用 Word2Vec,训练得到了一个 Gov2Vec 的工具,可以有效地对语料中的法学概念进行编码,并能够学习到这些概念向量之间的隐含关系,成功运用在最高法院意见、总统行动和国会法案的摘要生成任务中[2];Chalkidis 和 Kampas 同样基于 Word2Vec 提出了 Law2Vec,通过包括英国、欧盟、加拿大、澳大利亚、美国和日本等国立法的大型语料库中训练法律词汇嵌入,并验证了法律词汇语义特征表示在文本分类、信息抽取和信息检索 3 个任务中的重要作用[3]。

自 2018 年以来,以 BERT 为代表的预训练语言模型已经形成了一种新的 NLP 范式[4]:首先使用大规模文本语料库进行预训练,再对特定任务的小数据集微调,从而降低单个 NLP 任务的难度。预训练语言模型的应用,大幅提升了命名实体识别、事件抽取、机器翻译、自动问答等多项 NLP 任务的性能,在法律文本处理领域也具有良好的应用前景。针对通用预训练语言模型对法律领域术语和知识表达能力较弱的问题,Zhong 等人提出了一个基于千万级法律文本(包括民事和刑事裁判文书)的中文预训练模型 OpenCLaP(Open Chinese Language Pre-trained Model Zoo),其支持最大 512 长度的文本输入以适配多种任务需求,经过微调使用后有效提升了案件要素抽取、判决结果预测、相似案例匹配等多个法律文本处理任务中基线模型的性能[5]。目前,如何将知识嵌入到预训练语言模型已成为该领域的研究热点,在法律文本表示领域开展此类研究同样有助于提升深度学习模型对于法律概念的理解和推理能力。

1.2 基于特征的法律文本表示

基于嵌入的法律文本表示方法充分发挥了深度神经网络在 NLP 任务中强大的潜在语义学习能力,但其产生的文本向量往往无法解释,这对于强调领域知识的法律文本是一个显著的缺陷。而传统的特征工程方法,由于需要大量人工标注工作,在大规模的法律语料库面前也显得捉襟见肘。因此,有研究者开始尝试这两种方法的结合,即在上层使用一定量的领域知识来定义法律文本表示的特征模式,然后在底层使用深度神经网络模型对这些特征进行学习和表示。

Li 等人根据中国刑法中对于盗窃罪的定义,归纳出与定罪量刑相关的 9 维特征(包括犯罪嫌疑人基本信息、是否累犯、是否携带武器、涉案物品价值等),然后使用长短期记忆(long short-term memory, LSTM)网络对法律文本进行编码,再根据生成的向量表示使用分类算法,判断是否符合某个特征,进而得到针对法律文本的 9 维向量表示,在实现了特征降维的同时,使得特征能够在法律知识框架下具备良好的解释性[6]。针对判决结果预测任务,Li 等人提出了一种基于注意力机制的法律文本表示模型,通过在涉及 10 类刑事罪名的裁判文书语料中进行训练,生成基于案件事实、被告人信息及相关刑法条文等多个层面的潜在语义特征表示向量,能够表示法律文本中人物、事件、法律条文 3 者之间的潜在逻辑关系,大幅提升了罪名、法律条文、刑期等预测任务的性能和预测结果的可解释性[7]。

2 基于深度学习的法律文本分类

文本分类是法律文本处理应用中的关键任务。不同的法律文本处理任务可以转化为不同类型的文本分类问题。例如:判断一个案件中的被告人是否有自首情节属于简单的二分类问题,分析案件类型

（涉嫌的主要罪名为互斥关系）属于多分类问题，判定被告人触犯了哪些法条则属于多标签分类问题。已有的研究工作也基本围绕这 3 类问题展开。

Aletras 等人使用多个支持向量机（Support Vector Machine，SVM）分类器对案件的若干语义学特征分别进行二分类，用于预测欧洲人权法院的判决[8]；Boella 等人使用词频-逆向文件频率（Term Frequency - Inverse Document Frequency，TF - IDF）算法和信息增益进行特征选择，然后训练 SVM 分类器，以识别法律文本所归属的领域[9]；Liu 等人在基于案例的推理系统中使用 K 最近邻（K - Nearest Neighbor，KNN）算法对 12 种常见的刑事罪名进行分类[10]；Katz 等人根据从案件概要中抽取的特征，构建了随机树模型以预测美国最高法院的决策[11]；Lin 等人首先根据人工定义的 21 类法律要素标签对案件描述的句子进行分类，再用于区分抢劫和恐吓罪名[12]；Liu 等人将多个法条的不同组合作为标签进行训练，将多标签分类问题简化为多分类问题[13-14]。这些早期的工作大多利用特征工程与统计机器学习模型的结合，使用有监督的学习方法训练分类器，模型分类性能和结果的可解释性都相对较好，但由于过度依赖特征设计和人工标注，在文本标签体系发生变化时可扩展性较差。

近年来，以各类神经网络为代表的深度学习模型凭借其强大的特征学习能力在多种 NLP 任务中发挥了重要作用，特别是针对大规模语料库的学习中，相比人工规则构造特征的方法更能够刻画数据丰富的语义信息。Wei 等人使用卷积神经网络（Convolution Neural Network，CNN）实现了一个法律文档分类器，其实验结果证明 CNN 模型在大规模训练集上取得的性能明显优于 SVM[15]；Chalkidis and Androutsopoulos 采用了完全不依赖人工标注的词语本身、词性标签和符号嵌入作为特征，使用双向 LSTM 网络完成了合同要素抽取任务[16]；Luo 等人提出了一个基于注意力机制的多标签神经网络分类器，通过将法律法规信息融入案件事实的向量表示，在提升案件罪名分类性能的同时使分类结果具备一定的可解释性[17]；Li 等人提出了一种多通道注意力神经网络框架，仅使用训练数据中罪名类型、适用法条、刑期 3 个极易获取的标签为监督对案情描述、被告人信息和法律条文进行联合编码，灵活的编码方式可以支持不同的多标签分类任务，均取得了较好的分类性能[7]；Wang 等人提出了一种层次化匹配神经网络，在构建案件罪名向量表示的过程中融入标签的层次信息，并借助语义匹配的方法完成罪名分类任务，取得了较高的准确率[18]。

3　法律文本挖掘与应用

随着法律文本表示和分类等法律文本处理技术的不断成熟，以及法律领域利用计算机和人工智能技术辅助业务开展的需求的快速增长，近年来涌现出一些代表性的法律文本挖掘方法及其应用。

3.1　法律判决预测

法律判决预测（Legal Judgment Prediction，LJP）是基于法律文本的最关键任务之一。在中国、德国、法国等采用大陆法系的国家中，判决结果是根据案件事实与成文法规决定的。在这一法律制度下，LJP 的任务就是通过案件事实描述文本与法律条文的匹配，来判断相关行为是否触犯某条法律，进而对应判罪名、适用法条以及刑期做出预测。

已有研究大多将罪名和法条预测任务用文本分类算法解决，包括早期使用统计机器学习模型，以及近期使用深度学习模型的方法。为了促进 LJP 的发展，Xiao 等人提出了一个大规模的中文裁判文书数据集 C-LJP，包含中国法院发布的 268 万件刑期案件文本[19]；在近期的一些工作中，Luo 和 Li 将研究重心放在如何使用基于注意力机制的神经网络去挖掘案件，描述不同部分之间的逻辑关系，为了更好的实现这一目的以及为后续预测结果提供更好的可解释性，引入了法律条文作为外部知识来引导神经网络的编码过程，在罪名和法条预测任务中取得了优异的性能[17,7]；Zhong 等人通过引入 LJP 各个子任务之间的拓扑关系，使得模型的预测过程更符合人类法官的判案逻辑，实验结果也证实了这一做法的有效性[20]。

在刑期预测方面，有部分工作通过将刑期划分为不同区间进而转化为分类问题解决，也有一些研究者按照更符合任务本身特性的回归问题去设计模型。Li 等人根据法律条文归纳出了盗窃案件除刑期外的 10 维特征，利用神经网络训练得到特征向量后再交由回归算法进行计算，取得了较高的准确率，但这一方法相对依赖人工引入外部知识和标注，无法高效地将预测模型扩展到支持更多类型的案件[21]；Chen 等人提出了一种采用门控机制的神经网络模型，以罪名为基础对案件进行刑期预测，有效提升了预测的准确率[22]。但总体而言，由于刑期这一数据类型连续性的特点，以及在现实中存在的法律之外的量刑因素，使得现有的模型性能都不理想。

3.2 相似案例检索

随着案件文档规模的日益增长,相似案例检索对于提高法律从业人员的工作效率具有重要意义,高质量的类案推送结果也有助于中国法律更加接近所追求的“类案类判”的目标。

在早期的研究工作中,Saravanan 和 Casanovas 提出了基于语义网和本体论的法律案例检索系统,在输入输出两端都比传统基于关键字的系统实用性更强,其缺点是严重依赖法律专家对于本体的编辑,而且以本体作为检索条件也无法满足当前“以案搜案”的业务需求[23-24]。

英美法系国家采用的是判例法,对一个案件作出判决时必须明确引用既往案件的判决,因此自然形成了一个案例引文网络,为引入图算法解决类案检索问题提供了基础。Wagh 等人基于案例引证网络节点的中心性和介数性提出了一种计算印度法院判决相似度的方法[25];Minocha 等人提出了一个法律离散度的概念,通过衡量两个案例的相邻节点集合的相似度,查找一个案例在引文网络中的相似案例[26]。针对引文网络通常非常稀疏的问题,有研究者开始引入机器学习算法对法律文本相似度进行计算,如基于段落相似度计算全文相似度、基于词频的贝叶斯统计方法、基于案件特征的最近邻算法,但这些基于统计特征的方法丢失了文本原有的语义信息。为了尽可能保留文本的语义信息,使用词嵌入和深度学习模型逐渐成为类案检索任务的主流方法。

4 结束语

针对法律文本处理问题,本文简要介绍了近年来以深度学习方法为主的相关研究成果,分别对法律文本表示、法律文本分类以及法律文本挖掘与应用领域的研究方向和进展进行了梳理和分析。除本文介绍的这些方向外,法律文本处理涉及到任务还包括法律问答、法律要素抽取、法律文本摘要等。

总体而言,传统的文本处理技术均可以在法律文本处理任务中发挥重要作用,而以词嵌入方法和神经网络为代表的深度学习模型的引入,更是能够充分学习海量法律文本中蕴含的庞大语义信息。但是,如何使深度学习模型更好地与法律专业知识进行融合,是目前众多研究工作面临的共同难题,如何兼顾模型性能和结果可解释性将成为该领域未来研究的焦点问题。

参考文献

[1] 李一鸣. 结合知识和神经网络的文本表示方法的研究[D]. 杭州: 浙江大学, 2019.

[2] NAY J J. Gov2Vec: Learning Distributed Representations of Institutions and Their Legal Text[C]//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 embeddings trained on large corpora[J]. 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[C]//In Proceedings of NAACL, 2019:4171-4186.

[5] ZHONG H, ZHANG Z, LIU Z, et al. Open Chinese Language Pre-trained Model Zoo[R/OL]. 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[C]//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[J]. IEEE Access, 2019, 7(1): 151144-151155.

[8] ALETRAS N, TSARAPATSANIS D, PREOIUC-PIETRO D, et al. Predicting judicial decisions of the European Court of Human Rights: A natural language processing perspective[J]. 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[C]//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[J]. Journal of Information Science and Engineering, 2004, 20(4): 783-800.

[11] KATZ D M, BOMMARITO II M J, BLACKMAN J. A general approach for predicting the behavior of the Supreme Court of the United States[J]. 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[J]. 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[C]// InProceedings of ISMIS, 2006: 681-690.

[14] LIU C, LIAO T. Classifying criminal charges in Chinese for web-based legal services[C]//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[C]//InProceedings of BigData, 2018: 3317-3320.

[16] CHALKIDIS I, ANDROUTSOPOULOS I. A deep learning approach to contract element extraction[C]//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[C]//InProceedings of EMNLP, 2017: 2727-2736.

(下转第 91 页)