计算机科学
COMPUTER SCIENCE

基于Transformer的司法文书命名实体识别方法

王颖洁，张程烨，白凤波，汪祖民

引用本文

王颖洁，张程烨，白凤波，汪祖民. 基于Transformer的司法文书命名实体识别方法[J]. 计算机科学, 2024, 51(6A): 230500164-9.

WANG Yingjie, ZHANG Chengye, BAI Fengbo, WANG Zumin. Named Entity Recognition Approach of Judicial Documents Based on Transformer [J]. Computer Science, 2024, 51(6A): 230500164-9.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230500164

基于 Transformer 的司法文书命名实体识别方法

王颖洁[1]　张程烨[1]　白凤波[2]　汪祖民[1]
1 大连大学信息工程学院　大连 116622
2 广西民族大学人工智能学院　南宁 530006
(wb@hongstech.com)

摘　要　命名实体识别是自然语言处理领域的关键任务之一,是实现下游任务的基础。目前针对司法领域的相关研究相对较少,司法系统的信息化和智能化转型仍有许多问题亟需解决。相比其他领域的文本,司法文书存在专业性强、语料资源少等局限,导致现有的司法文书识别结果较低。因此,从以下 3 方面开展研究:首先,提出了一种多标签层级迭代的文本标注方式,可以对原始司法文书文本进行自动化标注,同时有效地提升司法文书命名实体识别任务的实体识别效果;其次,提出了一种交融式的 Transformer 神经网络模型,对汉字固有属性的深层特征进行了充分利用,用于对司法文书进行命名实体识别;最后,对所提出的标注方法和模型与其他神经网络模型进行了对比实验。所提出的文本标注方式可以较为准确地实现司法文书的标注任务;同时,所提出的模型在通用数据集中相对于对照模型有较大的提高,并在司法领域数据集中取得了良好的效果。

关键词:自然语言处理;数据标注;Transformer 模型;深度学习;司法信息化
中图分类号　TP391

Named Entity Recognition Approach of Judicial Documents Based on Transformer

WANG Yingjie[1],ZHANG Chengye[1],BAI Fengbo[2] and WANG Zumin[1]
1 College of Information Engineering,Dalian University,Dalian 116622,China
2 School of Artificial Intelligence,Guangxi Minzu University,Nanning 530006,China

Abstract　Named entity recognition is one of the key tasks in the field of natural language processing,and it is the foundation of downstream tasks. At present,there are relatively few research results on the judicial field,and there are still many problems need to be solved in the informatization and intelligent transformation of the judicial system. Compared with texts in other fields,judicial documents have limitations such as strong professionalism and few corpus resources,leading to low recognition results of existing judicial documents. Therefore,the research is carried out from the following three aspects. Firstly,a multi-label hierarchical iterative annotation method(ML-HIA) is proposed,which can automatically annotate the original judicial documents and effectively improve the effect of the entity recognition task of judicial documents. Secondly,an feature mixed Transformer(FM-Transformer) neural network model,which makes full use of the deep features of the inherent attributes of Chinese characters,is proposed to identify named entities of judicial documents. Finally,the proposed method and model are compared with other neural network models. The proposed method of text annotation can realize the task of judicial document annotation accurately. At the same time,compared with other models,the proposed model has a great improvement in the general dataset,and has achieved good results in the judicial datasets.

Keywords　Natural language processing,Data annotation,Transformer model,Deep learning,Judicial informatization

1　引言

当前,人工智能技术逐渐被应用于生活的各个领域,为日常的生活和工作提供了便利。命名实体识别(Named Entity Recognition,NER)是自然语言处理(Natural Language Processing,NLP)领域中的一项基础任务,主要用于识别文本中各类实体的边界和类别。作为自然语言处理的第一个环节,NER 的识别效果可以直接影响到语义识别、情感分析、文本摘要和知识图谱等后续下游任务的准确率。因此,对于 NER 方法的优化可以极大地改善 NLP 整体流程的效果。

通用领域的命名实体识别主要针对人名、地名、组织名和国家名 4 类实体,然而,在司法领域的命名实体识别任务中,增加了需要识别的司法类型的实体,具有较强的领域针对性。其中,司法命名实体指包括司法机构、司法章程、法律条文等一系列领域专有名词组成的集合。当前,司法文书命名实体识别主要存在以下 3 方面的挑战:1)当前国内外针对司法领域的命名实体识别研究相对较少,且大多为基于统计学习和机器学习的方法;2)中文司法文书的现有标注数据集数量较少,且大多存在标注不规范的现象;3)司法实体具有较强的多义性,从而影响识别的准确性。

目前主流的机器学习方式是以有监督的深度学习方式为主,对于标注数据有着强依赖性需求。然而,当前在司法领域中缺乏命名实体识别(Named Entity Recognition,NER)研究所需要的领域数据集,仅能依赖通用数据集进行模型的训练,在司法领域识别效果较差。当前对于该类文本的标注大多采用基于人工标注的方式,相关的研究集中于开发可用性高、操

通信作者:白凤波(baif@gxun.edu.cn)

作便捷的标注平台。在完成目标文本的标注后，可以使用BLEU算法[1]、METHOR算法[2]、ROUGE算法[3]、SPICE算法[4]、CIDEr算法[5]和ZenCrowd算法[6]对当前标注的质量进行评定。

为了改善自然语言处理任务在司法领域中的效果，我们通过对司法领域专业书籍进行实体爬取，构建了司法领域的实体词典。同时，为了提高标注数据在司法领域的泛用性，提出了一种多标签层级迭代的方式，从而为文本添加司法标注信息。并且，针对汉字丰富的特征信息，提出了一种汉字的多特征融合方法。最后，提出了一种交融式的Transformer神经网络模型结构(Feature-Mixed Transformer，FM Transformer)，不仅可以较为准确地对常规实体进行识别，而且可以较为准确地识别司法领域实体。

本文的主要贡献包括4个方面：

1)构建了一种司法领域的实体词典语料库；

2)提出了一种多标签层级迭代标注(Multi-Label Hierarchical Iterative Annotation，ML-HIA)算法，能够为原始司法文书文本自动添加司法类型的标签；

3)提出了一种汉字的多特征融合方法，对汉字的偏旁特征、读音特征和字形特征进行了提取和融合；

4)提出了一种交融式的Transformer神经网络模型结构FM-Transformer，并在不同的数据集中对所提出的模型进行了对照实验。实验结果表明，所提出的模型在司法领域数据集和通用数据集中取得了良好的实体识别效果。

本文第2章介绍了NER领域的相关研究方法；第3章介绍了所采用的词典构建方法和多标签层级迭代的方法；第4章介绍了所使用的汉字多特征融合方法及所提出的交融式的Transformer神经网络模型结构；第5章介绍了所使用的数据集和相关的验证实验；最后总结全文并展望未来。

2 相关工作

2.1 命名实体识别

命名实体识别属于信息提取任务的子任务，其目的是对文本中预定义类型的实体进行提取和分类。常见的预定义实体包括专有名词(如人名、地名、组织名等)、数字和日期3类，同时在面对不同的应用场景时，也会定义相应的专有名词实体。NER技术大致可分为基于规则的方法、基于统计模型的方法和基于深度学习的方法3类。

在早期的命名实体识别研究中，基于规则的方法由于其易于实现且无需训练的特点，在规模较小的数据集中取得了较好的效果。Pan[7]根据特定语料制定了相应的识别规则，在结合统计特征后能够减少对大规模语料库的依赖。然而，所制定的规则无法对应所有文本类型，一旦文本发生变化就需要重新制定。Feng等[8]为了增强模型的领域泛化性和实体相关性，在CRF模型的基础上引入正则表达式和特征模板，通过结合上下文信息，能够更精确地匹配文本中的词序列与特征模板，实现了更好的识别性能。基于统计模型的方法的关键在于针对特定的使用场景来选择合适的模型进行训练。

Zhao等[9]提出了一种多标签卷积神经网络方法，将字符级向量、词语级向量和词典相结合，以较少的特征工程在NCBI和CDR语料库上实现了最优的性能。Wang等[10]针对医疗领域标注数据不足的问题，提出了一种基于序列生成对抗网络的模型，可以在不使用外部资源的情况下自动生成标注数据，并且具有一定的泛用性。为了解决文本数据中固有噪声的干扰，Aguilar等[11]提出了一种多任务神经网络架构，通过将CNN和BiLSTM并行使用，能够从字词序列、词性标签和地名词典中学习到文本的深层特征。该方法在处理噪声较大的文本时鲁棒性较好，但对于实体边界的处理效果仍有待改进。为了解决这个问题，Guo等[12]在模型中引入注意力机制，同时针对中文语料，采用CNN提取部首嵌入特征，以丰富语义信息。该方法在农业和医疗领域的语料集中均取得了较好的效果，具有一定的泛化能力。

在特定的领域场景中，语料库的规模相对较小，实体的名称也呈现出较大的多样性，导致神经网络训练的效果显著降低。针对这个问题，Zhang等[13]提出了基于GAN的共同特征学习方法，采用带有注意力机制的BiLSTM模型作为生成模型，并采用CNN作为判别模型，在信息安全领域的数据集中取得了较高的标注效果。同时，基于光滑近似逼近思想对离散类型的文本数据进行处理，解决了标注数据缺乏和同一实体标注不一致的问题。Das等[14]基于图聚类算法，采用无监督方法提取语料库中的实体关系，极大提高了模型的领域泛用性，避免了模型的再训练过程。由于实体抽取的效果对分词的效果有较强的依赖性，因此有学者提出在文档级别开展文本的实体抽取工作。Zhao等[15]通过预训练的方式获得实体的分布式表示，减少了分词错误导致的性能影响，保证了实体标签的一致性。

2.2 文本自动标注

数据标注在机器学习的过程中，扮演着举足轻重的作用。在自然语言处理领域，需要大量预标注的文本信息，因此如何对文本进行经济高效的标注，成为了当前研究的重点之一。目前，对文本标注的研究总体可以分为两类。第一类是开发新型的标注工具和标注平台。Yu等[16]使用MARKUS标注工具对数字古籍文本进行了标注，证明了该系统处理中文文言文的有效性。当前现有的标注平台界面设计对用户较为友好，但在实际使用的过程中仍然存在不足之处。1)数据在标注之后，无法对标注的结果进行质量分析，标注的可信度无法得到保障。2)不同标注平台所依赖的环境存在较大差异，在多设备或服务器协同工作的情况下配置较繁琐，从而降低了用户的使用体验和标注速度。3)对于大型的标注任务，人工标注的难度较大，缺乏在标注过程中的辅助工具。

针对现有标注平台的问题，Zhang等[17]在医疗领域构建了一种具有标注数据分析模块的半自动标注平台。同时，该平台融合了多种实体自动识别的算法，极大地降低了标注人员的工作难度。Lawson等[18]从激励标注人员的角度出发，在亚马逊的Mechanical Turk(MTurk)标注系统中引入了竞争性支付机制，避免了标注人员为了追求标注数量而主动降低标注质量的行为。在文献[19-21]的基础上，Zhang等[22]提出了一种基于数据标签的标注系统OneLabeler。相比标注算法和标注界面的改进，该系统更强调对不同场景的适应性，降低了用户配置使用环境的难度。

另一类针对文本标注的研究是改进并优化标注算法，从而降低人工标注的数量，减轻标注人员的工作负担。传统的命名实体标签仅包含一些常见类型，例如人名、地名、组织名和日期等[23]。这些通用的命名实体类型具有领域无关性，因此具有极强的泛用性。然而，在处理特定领域的实体识别任务时，采用通用的标注往往不能实现令人满意的效果。由于

领域相关的标注数据较少，因此有学者尝试采用半监督学习的方式解决样本不足的问题。对于半监督学习而言，主要依赖于少量带标注类别的样本和大量无类别标注的普通数据。Zhu 等[24]在半监督学习的基础上，利用信息熵理论的特点，在候选标注样本中选择信息量最大的样本进行标注。Sun 等[25]在贝叶斯(Bayes)算法的基础上，对不同学科的知识点进行自动标注，并使用 AdaBoost 算法辅助训练参数。这种方法极大提高了多标签分类的精度，并且比朴素贝叶斯算法更具通用性和稳定性。在仅有 Marmo 作为数学实体公开数据集的背景下，Beyette 等[26]提出了一种半自动标记算法，并采用唯一字符串技术来对齐文档。

对于命名实体识别任务所需要的标注文本，存在着以下 3 点要求[27]。首先，实体的标注涉及类别的选择和所包含内容的确定。除了 MUC 会议定义的标准标签类别之外，带有领域特征的类型很难相互统一。同时，非人工标注的方式很难确定标注范围和划定实体识别边界。最后，一词多义的现象在文本中普遍存在，如何使机器正确理解其中的含义是一项亟需解决的问题。

裁判文书是记载司法审判活动过程，明确当事人权利义务的司法产品[28]。由于缺乏司法领域的标注数据集，Weng 等[29]对数据集中的裁判文书进行人工标注，得到了对应的结构化标注数据集。然而，这种方法对于大规模的裁判文书并不适用，需要寻找一种更加便捷快速的标注方法。

2.3 中文特征融合

早期的中文命名实体识别方法根据预先定义的汉字转换编码，将每个汉字转换为特征向量输入到神经网络中。随着中文 NER 技术的不断发展，这种方法逐渐显露出一定的局限性。首先，该方法仅利用了汉字自身的特征，并没有结合当前字在文章中的位置信息，从而会出现上下文语义缺失的问题。同时，与英文单词不同，汉字自身蕴含了丰富的象形特征，而这种固有的特征信息无法在单一编码向量中充分利用。因此，如何对汉字的各类特征信息进行充分利用成为了当前研究的重点。目前的研究按照特征的类别进行划分，可以分为词语特征融合和汉字特征融合两大类。

词语特征融合的方法是通过使用栅格结构[30]对单个汉字和词语共同处理，结合其所在的起止位置汉字来确定词的位置。该方法有效弥补了分词错误导致的误差，在中文 NER 任务中取得了良好的效果。然而，这种结构仅可在 LSTM 模型中实现，并且无法使用 GPU 并行计算，网络的训练时间较长。为了解决这些问题，Sui 等[31]构建了一种结合字词特征的联合图神经网络，分别建立了字符和自匹配词的联系、字符和上下文的联系、自匹配词和上下文的局部联系，能够消除依赖树解析时的误差和信息损失。Gui 等[32]为了尽可能消除中文文本中的歧义现象，采用设置一个全局中继节点的方式对长距离的字词特征信息进行处理。同时，采用了一种图聚合循环机制来处理文中词边界模糊的问题，极大地提高了实体识别的准确率。Ma 等[33]通过将不同长度的单词输入对应的处理层，并加入整个句子的长依赖信息，不仅能够减少词表冲突和歧义问题，而且可以使用 GPU 进行并行计算，缩短了模型的训练时间。Kong 等[34]将每个字匹配到的标签汇成一个标签嵌入向量，并对融合词典的嵌入向量与字向量直接拼接，可以在保证模型识别效果的情况下极大提高训练速度。

自 Transformer 模型于 2017 年应用于 NLP 任务以来，其对长距离序列的高效处理能力得到了研究者的广泛关注。相比 CNN 和 RNN 而言，Transformer 具有时间复杂度更低、模型可解释性更强、处理长序列能力更强等优势。Li 等[35]根据 Transformer 中注意力机制的无偏性，将文献[30]中的栅格结构迁移到 Transformer 中。为了定位词的位置，对每个汉字和词语都引入了对应起止位置的位置向量。因此，FLAT(Flat LAttice Transformer)模型可以建立每个汉字和全部匹配词语的联系。

当前，汉字特征融合主要包括汉字的字形特征融合、笔画特征融合、偏旁特征融合和读音特征融合。首先，汉字的字形特征即为其固有的形态特征，也就是将汉字视作图像进行处理[36]。作为一种象形文字，Su 等[37]直接使用 CNN 从汉字的位图中提取特征，并根据提取的图向量特征进行特征融合。Meng 等[38]针对传统计算机视觉模型对汉字字形识别效果欠佳的问题，采用了古汉语文字对汉字字符集进行了扩充，并使用了一种改进的 CNN 模型用于汉字图像处理，有效地提高了模型的泛化性。

作为汉字最基本的语义单位，笔画对于汉字的语义有较大的影响。因此，为了在汉字级别进行语义增强，需要对笔画特征进行深入的分析和研究。Cao 等[39]首次提出了使用汉字的笔画特征信息进行语义增强，将笔画的信息采用 n-gram 的方式进行提取，并为不同笔画分配相应的 ID 值作为标识。相比词语增强的方法，采用笔画特征信息代替词语进行训练可以实现更高的实体识别准确率。

在中文文本中，汉字的偏旁在一定程度上可以反映汉字所属的类别。因此，汉字的偏旁特征可以在分类任务中有效地提升模型的性能。Sun 等[40]对汉字的偏旁特征进行了提取和分析，对 CRF 模型进行了增强，实验证明在中文相似性判断和分词两方面都实现了较明显的提升。同时，Hardmeier 等[41]也通过实验证明，对词根和偏旁等汉字的固有特征进行提取可以有效地提升模型的效果。

在语言学的研究中，无论以何种语言书写的文本，只有当其作为口语的记录时才具有实际意义。因此，作为汉字的重要语言学特征之一，汉字的读音也蕴含着大量的深层特征。在文献[39]的研究基础上，Zhang 等[42]引入了汉字的拼音特征，证明了融合拼音、字形和偏旁特征的中文实体识别效果高于仅融合字形和偏旁特征。Zhu 等[43]在中文文本中引入汉字的拼音特征，并采用结构相同的模型进行对比，结果证明拼音特征的引入对实体的识别准确率有较好的提升。Chaudhary 等[44]将多种语言的读音特征融入网络模型中，在提高模型识别效果的同时提升了模型的泛用性。

3 司法文书数据预处理

目前，针对司法文书的 NER 数据集较少，因此需要对原始的司法文书数据进行预处理，从而得到足够的语料训练数据。本次使用的司法文书数据来源为中国裁判文书网。该网站于 2013 年 7 月 1 日由最高人民法院开通，截至到 2023 年 3 月 12 日，共涵盖司法文书 139 490 527 篇，其中民事文书的数量已超过 8000 万篇。因此，可以从该网站中获得大量实时更新的司法文书数据。本次选择民事文书和刑事文书作为语料库的资源，形成约 130 万字的司法文书语料库。

3.1 司法文书数据清洗

在爬取到文书数据后，需要对其进行预处理操作，具体包

括数据清洗、数据对齐和结构化存储3部分。其中，对文本数据的清洗采用正则表达式的方法，删除文本中的无意义信息，如网页标签、特殊符号等。为了方便文本数据的使用，将一篇文章以句子为单位存储。同时，删除其中过长的句子，以降低训练的开销。最后，选取相应的格式进行结构化存储。

3.2 司法领域实体词典构建

本文在《中国百科大词典-法学》中对司法实体进行了抽取，并将得到的实体处理为词典形式。同时，将词典中的法律实体按照词长降序排列，并删除重复出现的实体。所构建的词典包括2889个法律实体，其词长分布如图1所示。从图中可以看出，大多数法律实体的长度集中在2～5个汉字之间。

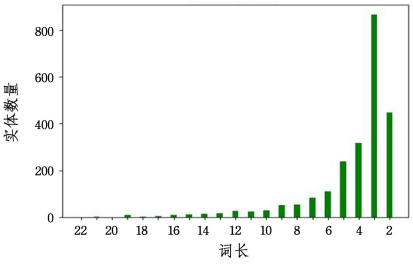


图1 法律实体词长分布情况

Fig. 1 Legal entity word's length distribution

3.3 多标签层级迭代标注算法

本节将对所提出的多标签层级迭代标注方法进行详细说明。ML-HIA方法首先对初始文本进行标注初始化，也即将标注均记为“O”(Outside)。接着，计算初始文本的LR值，其中LR代表文本的司法相关度(Law Relation，LR)。若LR值大于分类阈值，则将司法标注设置为首层，也即设置该类标注的优先级为最高级。最后，描述了多标签层级迭代标注算法的整体实现流程。下文将对每个模块的内容和算法进行简要描述。

3.3.1 标签类型标注

在命名实体识别任务中，标注是由标签头部和标签类型两个部分组成。标签头部常用的标注方法包括BIO，BIOES和BME等。本次选择采用基于BIO的标注方式进行标注，其中B代表实体的开始位置(Beginning)，I代表实体不包括首部的内部位置(Inside)，O代表实体之外的区域。在标签头部之后，即为所代表的标签类型，如人名(PER)、地名(LOC)、机构名(ORG)等。本次在原有标签类型的基础上，增加了司法标签类型(LAW)，图2给出了带有司法标签类型的标注示例。在标注结束后，由具有多年司法领域工作经验的中国政法大学的专家团队进行人工核对，从而得到相对准确的标注结果。

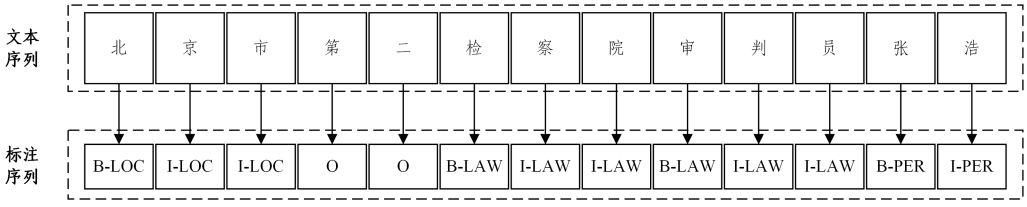


图2 司法标签标注示例

Fig. 2 Examples of judicial labeling

3.3.2 司法相关度计算

为了区分文本是否属于司法文书文本，本次采用计算文本司法相关度的方法进行分类。首先计算司法词典中的实体出现在待标注文本中的TF-IDF值。TF-IDF是词频(Term Frequency，TF)和逆向文档频率(Inverse Document Frequency，IDF)的乘积，其中TF和IDF的计算式如下：

$$TF_k=\frac{a_{k,m}}{\sum_i a_{i,m}} \tag{1}$$

$$IDF_k=\log\left(\frac{n_{\text{all}}}{n_k+1}\right) \tag{2}$$

$$TFIDF_k=TF_k * IDF_k \tag{3}$$

其中，$a_{k,m}$表示第k个实体在第m句中出现的次数，$\sum_i a_{i,m}$表示在第m句中所有类型实体出现的次数之和；n_{all}表示语料库中所有句子的总数，n_k表示包含实体k的句子数。在得到所有司法实体的TF-IDF值后，即可进行司法相关度LR的计算，LR的计算式如下：

$$LR=\frac{1}{T}\sum_{i=1}^{T} TFIDF_i \tag{4}$$

其中，$\sum_{i=1}^{T} TFIDF_i$表示对词典中TF-IDF值最大的前T个实体进行累加。经测试，$T=10$时可以证明其领域代表性，并选择$LR=0.4$作为临界阈值，可以对司法文书文本和普通文本进行有效的区分。所提出的多标签层级迭代标注算法的流程如下。通过层级迭代的方式，可以有效地处理在标注过程中出现的同一实体具有多类标签的问题，即实体重叠问题。同时，通过调整不同层级间的迭代顺序，可以灵活地在不同领域的数据集中进行标注工作，得到更符合当前领域的标注数据。在司法文书命名实体识别任务中，即可将司法实体的迭代顺序置于最优先。

算法1 多标签层级迭代标注算法

输入：司法相关度LR、输入的文本T={s_1，s_2，s_3，…，s_n}，文本中的第i个句子 s_i={[c_1，“O”]，[c_2，“O”]，[c_3，“O”]，…，[c_m，“O”]}，司法领域词典dict，相关性阈值horizon

输出：带有标注的司法语料集Set

```
1. if LR > horizon:
2.     for i=1 to len(T)
3.     for j=1 to len(dict)
4.       if len(dict[j]) > len(s_i)
5.       continue
6.       else
7.       if dict[j] in s_i:
8.         head=s_i.find(dict[j])
9.         tail=head+len(dict[j])-1
10.      if s_i[head][1]! ="O" or s_i[tail][1]!="O"
11.        break
12.     for k to len(dict[j])
13.        temp=head+k
14.        if s_i[temp][1] =="O" and k ==0
```

```
        newstr=list(s_i[temp][1])
        newstr[0]="B-LAW"
        s_i[temp][1]="".join(newstr)
        if s_i[temp][1] =="O" and k ! =0
        newstr=list(s_i[temp][1])
        newstr[0]="I-LAW"
        s_i[temp][1]="".join(newstr)
      end for
    end for
    end for
Set=T
return Set
```

4 FM-Transformer 神经网络模型

在得到司法文书的标注数据后，便可以对相应的神经网络进行训练。为了充分利用汉字自身的固有特征，从而提高中文司法文书的实体识别效果，提出了一种交融式的 Transformer 神经网络模型 FM-Transformer（Feature-Mixed Transformer），并在公开数据集和司法领域数据集上进行了对照实验。

所提出的 FM-Transformer 模型首先对输入的文本进行预训练，得到各个汉字对应的嵌入向量。同时，提取汉字的偏旁特征和读音特征，并将其作为汉字的特征向量输入到卷积神经网络（Convolution Neural Network，CNN）模型中，并对两者的输出向量进行池化操作，从而得到汉字的特征向量。最后，将嵌入向量和特征向量共同输入到 FM-Transformer 模型中，得到模型预测的标签和识别的实体。4.1 节和 4.2 节分别对模型的各个部分进行了介绍和说明。

4.1 汉字特征处理

作为一种象形文字，汉字自身存在着更多的固有特征，如偏旁特征、读音特征和字形特征等。汉字的偏旁特征可以体现其表达属性的大致类别。例如，汉字中的“江”“河”“湖”“海”等字均带有偏旁“氵”，共同表达了“水流”的含义；汉字中的“林”“树”“枝”“桠”等字均带有偏旁“木”，共同表达了“植物”的含义。同时，汉字的读音特征也可以反映出使用者希望表达的内容。例如，对于相同的汉字“鲜”，读作“xian1”时表达的含义为新鲜、鲜明；在读作“xian3”时，则表达出稀有、稀少的含义。汉字的字形特征同样可以一定程度上体现汉字的含义，例如“国”“园”等字可以视作一处被围墙所包裹的地点。在对司法文书文本进行分析后，本次选择采用汉字的偏旁特征和读音特征两类固有属性作为汉字的特征信息，并将其输入 CNN 进行处理。

本次所使用的 CNN 模型结构如图 3 所示。对于句子中第 i 个字 c_i，对其偏旁和拼音两项特征进行处理，得到偏旁的嵌入向量 $\boldsymbol{r}_i$ 和拼音的嵌入向量 $\boldsymbol{s}_i$。接着，将 $\boldsymbol{r}_i$ 和 $\boldsymbol{s}_i$ 分别输入到 CNN 模型中，经过最大池化和正则化处理后，得到两者的特征向量 $\boldsymbol{r}_i^c$ 和 $\boldsymbol{s}_i^c$，并将其输入到全连接层得到 $\boldsymbol{r}_i^f$ 和 $\boldsymbol{s}_i^f$，其数学描述如下：

$$r_i^c = CNNLayer(r_i) \tag{5}$$

$$s_i^c = CNNLayer(s_i) \tag{6}$$

$$r_i^f = FullyConnected(r_i^c) \tag{7}$$

$$s_i^f = FullyConnected(s_i^c) \tag{8}$$

最后，将 r_i^f 和 s_i^f 进行融合，得到 c_i 的特征向量 w_i^f。特征融合的数学描述如下：

$$w_i^f = Concat(r_i^f, s_i^f) \tag{9}$$

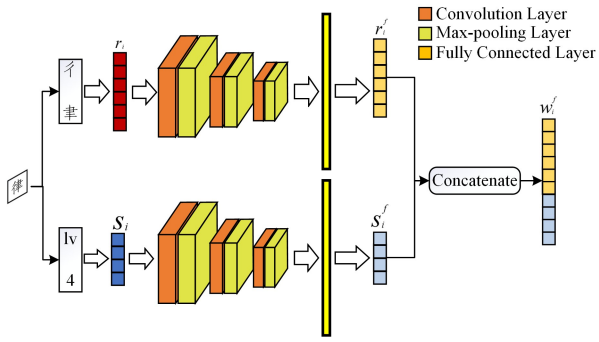


图 3 CNN 处理汉字特征流程

Fig. 3 Features of Chinese characters processed by CNN model

4.2 交融式 Transformer

FM-Transformer 模型的结构如图 4 所示。首先，将向量化的汉字和对应的特征向量共同输入到 Transformer 模型中。为了充分利用汉字的特征信息，编码器的部分采用了交替式的结构。

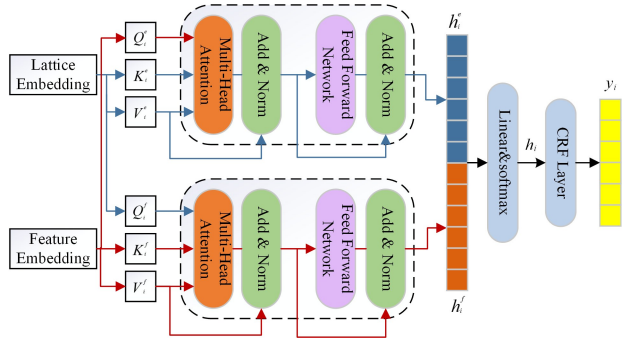


图 4 FM-Transformer 模型的结构

Fig. 4 Structure of FM-Transformer model

对于句子中第 i 个字 c_i，在经过词嵌入处理后得到其嵌入向量 $\boldsymbol{w}_i^e$。接着，在 Transformer 的编码器部分，计算其相对位置编码 R_{ij}，从而确定词的起始和结束位置，R_{ij} 的计算式如下：

$$R_{ij} = ReLU(pos) = ReLU(W_r Concat(span)) \tag{10}$$

$$span = span(h_i - h_j), span(t_i - t_j) \tag{11}$$

其中，h_i 和 t_i 为 c_i 的起始位置和结束位置，$span(x)$ 为相对位置信息计算函数，其表达式如下：

$$span_x^{2k} = sin\left(\frac{x}{10\,000^{2k/d_{\text{model}}}}\right) \tag{12}$$

$$span_x^{2k+1} = \cos\left(\frac{x}{10\,000^{2k/d_{\text{model}}}}\right) \tag{13}$$

为了结合汉字的嵌入特征和固有特征，我们在计算注意力时将两者进行了融合。对于输入的 w_i^e 和 w_i^f，首先利用线性变换得到汉字嵌入特征对应的 Q_i^e，K_i^e 和 V_i^e 以及汉字固有特征对应的 Q_i^f，K_i^f 和 V_i^f。其计算式如下：

$$\begin{bmatrix} Q_i^e \\ K_i^e \\ V_i^e \end{bmatrix} = \begin{bmatrix} W_i^K \\ W_i^Q \\ W_i^V \end{bmatrix} w_i^e \tag{14}$$

$$\begin{bmatrix} Q_i^f \\ K_i^f \\ V_i^f \end{bmatrix} = \begin{bmatrix} W_i^K \\ W_i^Q \\ W_i^V \end{bmatrix} w_i^f \tag{15}$$

接着，在计算注意力得分时，采用了下述方法进行实现：

$$\alpha_{ij}^{e}=(Q_i^{e}+p^{e})^{\mathrm{T}}w_j^{e}+(Q_i^{e}+q^{e})^{\mathrm{T}}R_{ij}W_R \tag{16}$$

$$\alpha_{ij}^{f}=(Q_i^{f}+p^{f})^{\mathrm{T}}w_j^{f}+(Q_i^{f}+q^{f})^{\mathrm{T}}R_{ij}W_R \tag{17}$$

其中，α_{ij}^{e}为嵌入特征的注意力得分，α_{ij}^{f}为固有特征的注意力得分。最后，将两个 Transformer 模型输出的结果h_i^{e}和h_i^{f}进行融合得到h_i，并在通过 CRF 层后得到最终的预测向量y_i。其中 h_i和y_i的数学描述如下：

$$h_i=W_i Concat(h_i^{e},h_i^{f})+b \tag{18}$$

$$y_i=CRF(\mathrm{Softmax}(h_i)) \tag{19}$$

由于司法文书中的司法实体所占比例相对较少，因此会出现类别不平衡影响模型性能的问题。为了解决这个问题，本次使用的损失函数为：

$$Loss=\log\left(1+\sum_{i\in NEG}e^{s_i}\sum_{j\in POS}e^{-s_j}\right) \tag{20}$$

其中，POS 和 NEG 为样本的正负类别集合；s_i和s_j为目标分类的得分。

5 实验分析

为了测试本文提出的 ML-HIA 标注算法，在民事案件判决书、刑事案件判决书、微博数据集和 Resume 简历数据集 4 个语料库之间进行了对比实验。同时，我们将所标注的数据集用于命名实体识别任务，从而验证是否可以有效地对法律实体进行识别。最后，为了验证多层机制的有效性，以中国裁判文书网的文书为原始语料，进行了相应的消融实验。

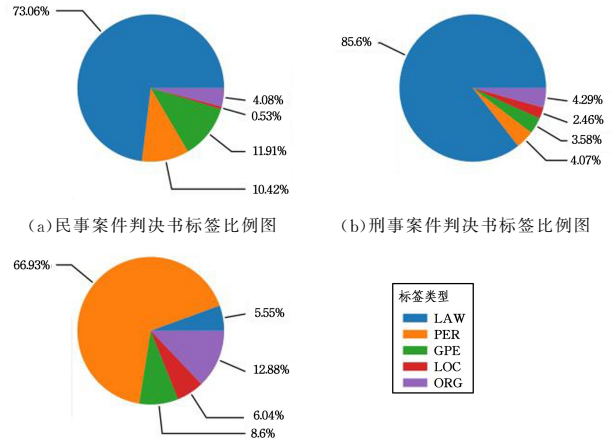
5.1 实验数据集及对照模型

本次所采用的数据集分为司法领域数据集和通用数据集两类。其中司法领域数据集为中国裁判文书网中公开的司法文书数据，其中以民法、刑法两类居多，因此可以在对其进行爬取和标注后，作为司法领域数据集进行使用。测试算法所使用的文本数据包括民事案件判决书、刑事案件判决书和微博数据集 3 个语料库。其中用于算法验证的民事案件判决书数据集共 979 283 字，刑事案件判决书数据集共 365 318 字，微博数据集共 105 016 字。

本次实验使用的通用数据集采用微博数据集[1)]和 Resume 简历数据集，两种数据集均为公开数据集，其中微博数据集包括约 1500 句微博文本数据，可以在简单处理后直接使用。本文选用了 4 种近年来表现相对优异的方法来对所提出的模型进行对比分析，包括 CRF，Lattice-LSTM，LRCNN 和 Soft-lexicon 模型 4 类。

5.2 ML-HIA 算法在司法文书标注任务中的效果

为了验证所提出的 ML-HIA 算法的可靠性，本次选择了对中国裁判文书网中的民事案件判决书和刑事案件判决书进行标注测试。如图 5 所示，ML-HIA 算法可以在司法文书中对司法实体实现良好的标注，在通用数据集中也可实现较为理想的标注效果。



(a)民事案件判决书标签比例图 (b)刑事案件判决书标签比例图

(c)微博数据集标签比例图

图 5 不同数据集中标注标签分布

Fig. 5 Distribution of labeled tags in different datasets

同时，为了确定各类标签的层级迭代顺序，也即各类标签的标注优先级，统计了各类标签优先级最高时的实体标注数量，如图 6 所示。从图中可以得出，当优先对司法实体(LAW)进行标注时，可以达到更高的标注数量。

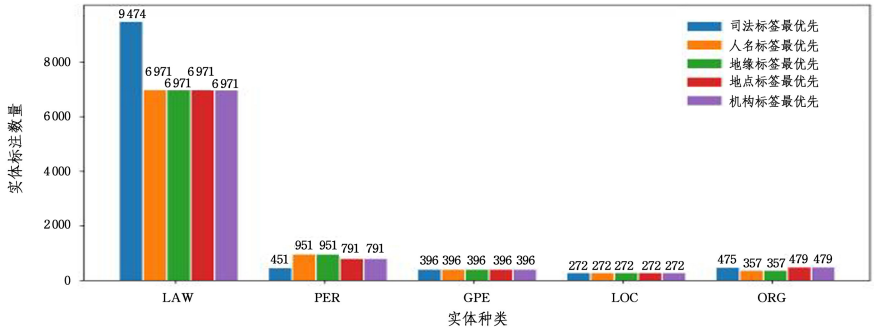


图 6 各类标签最高优先级时标注实体数量

Fig. 6 Labeled entities with the highest priority on each label

5.3 FM-Transformer 模型实体识别效果

为了验证 FM-Transformer 模型在 NER 任务中的效果，首先将 Peng 等[45-46]、He 等[47-48]和 Cao 等[49]在微博数据集中的实验结果作为基准进行对照。同时，在 Resume 数据集上将 FM-Transformer 模型与近年来提出的 Lattice，TENER，FLAT 和 Lexicon 模型进行对照，两个数据集上的对照结果如表 1 和表 2 所列。所提出的 FM-Transformer 模型在微博数据集中的 F1 值相比基准模型均有较大的提升，在 Resume 数据集中也取得了较好的效果。

由表 1 和表 2 中的结果可以发现，融合词组特征的模型识别效果均优于单独使用字符特征的模型，且融合更多汉字特征的 FM-Transformer 模型可以在训练中学习到更深层级

1) https://gitcode.net/mirrors/hltcoe/gold-en-horse/-/tree/master/data

的特征表示，从而达到更好的识别效果。

表 1　各模型在微博数据集中的识别效果

Table 1　F1-score of each model in Weibo dataset

（%）

Compared Model	F1-score
Peng andDredze(2015)	56.05
Peng andDredze(2016)	58.99
He and Sun(2017a)	54.82
He and Sun(2017b)	58.23
Cao(2018)	58.70
FLAT	60.32
Proposed Model	**64.82**

表 2　各模型在 Resume 数据集中的识别效果

Table 2　F1-score of each model in Resume dataset

（%）

Compared Model	F1-score
Lattice	94.46
TENER	95.00
FLAT	95.45
Lexicon	95.59
Proposed Model	**95.80**

5.4　FM-Transformer 模型在司法 NER 中的效果

由于司法文书存在着司法实体定义灵活、司法文书语料库较少的问题，限制了司法领域的命名实体识别准确率。本次将 FM-Transformer 模型应用于民事案件判决书和刑事案件判决书两类司法文书中，其识别效果如表 3 所列。从表中可以看出，FM-Transformer 模型对民事案件类型的判决文书有较高的识别能力，对刑事案件类型的判决文书识别效果欠佳。其原因在于，刑事案件类型的司法实体所占比例较少，导致其 F1 值相对民事案件较低。

表 3　FM-Transformer 模型在司法命名实体识别中的效果

Table 3　Effect of FM-Transformer model on judicial NER

（%）

司法文书类别	Accuracy	Precision	Recall	F1
民事案件判决书	99.14	87.56	92.28	89.86
刑事案件判决书	64.82	50.14	67.70	57.62

表 4 和图 7 为各类模型在民事案件判决书数据集中对各类实体的识别效果。

表 4　各模型在民事案件判决书数据集中的识别效果

Table 4　Recognition effect of each model in civil dataset

（%）

Compared Model	Accuracy	F1
CRF	97.29	85.36
Lattice-LSTM	98.74	83.81
LRCNN	99.03	87.73
Soft-Lexicon	99.12	88.45
Proposed Model	**99.14**	**89.86**

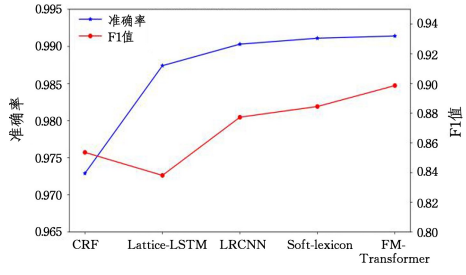


图 7　不同模型在民事案件数据集下的效果

Fig. 7　Effect of different models in civil dataset

由图 7 可以得到，在语料较大的民事案件判决书数据集中，融合汉字词语信息的 Lattice-LSTM，LRCNN，Soft-lexicon 和所提出的 MF-Transformer 模型均可以取得更好的识别效果。同时，相比 Lattice-LSTM 使用单一类型的分词，保留更多分词结果的 Soft-lexicon 模型在准确率和 F1 值上可以取得更为良好的命名实体识别效果。

表 5 和图 8 为各类模型在刑事案件判决书数据集中对各类实体的识别效果。图 8 中的对照结果显示，在处理诸如刑事案件判决书数据集等语料较少的数据集时，Lattice-LSTM 和 LRCNN 的 F1 值明显高于 CRF 模型，Soft-lexicon 的 F1 值也高于 Lattice-LSTM 和 LRCNN，而 MF-Transformer 达到了最高的 F1 值 57.62%。

表 5　各模型在民事案件判决书数据集中的识别效果

Table 5　Recognition effect of each model in criminal dataset

（%）

Compared Model	Accuracy	F1
CRF	96.95	41.23
Lattice-LSTM	97.07	50.82
LRCNN	97.04	53.87
Soft-Lexicon	96.64	56.03
Proposed Model	**97.16**	**57.62**

由此可见，融合汉字的词语级别特征和汉字的深层特征均能够对模型的识别效果有所改善。本文所提出的 MF-Transformer 模型在民事案件判决书数据集和刑事案件判决书数据集中，准确率和 F1 值两项评估指标均取得了最优的效果，表明了该模型对于司法领域的实体识别准确性较高。

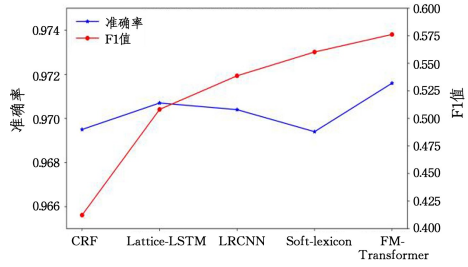


图 8　不同模型在刑事案件数据集下的效果

Fig. 8　Effect of different modelsin criminal dataset

结束语　针对司法领域语料库较少、命名实体识别效果较低的问题，本文提出了一种多标签层级迭代的文本数据标注方法。首先，通过爬取司法领域的资料，构建了司法领域的实体词典。接着，通过计算基于 TF-IDF 算法的文本司法相关度 LR，可以有效地区分普通文本和司法文本。同时，提出了一种交融式的 Transformer 神经网络模型。该模型可以更加充分地利用汉字的固有特征，从而实现更好的实体识别效果。

实验结果表明，所提出的 ML-HIA 算法可以有效地减少人工标注的工作量，并保证数据标注的准确性。该算法可以在司法领域中实现良好的效果，具有一定的参考价值。同时，所提出的 MF-Transformer 神经网络模型可以有效地识别民事案件判决文书，并在通用数据集中实现了良好的效果。然而，该模型在司法领域内的泛化性仍有不足，有待进行进一步的研究。

参 考 文 献

[1]　PAPINENI K，ROUKOS S，WARD T，et al. Bleu：A method for

automatic evaluation of machine translation[C]// Proceedings of the Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2002: 311-318.

[2] LIN C Y. ROUGE: A package for automatic evaluation of summaries[C]// Proceedings of the Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2004: 74-81.

[3] LAVIE A, AGARWAL A. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments[C]// Proceedings of the Workshop on Statistical Machine Translation. Stroudsburg, PA: ACL, 2007: 228-231.

[4] ANDERSON P, FERNANDO B, JOHNSON M, et al. SPICE: Semantic propositional image caption evaluation[C]// Proceedings of the 14th European Conference on ComputerVision(ECCV 2016). Amsterdam, The Netherlands, 2016: 382-398.

[5] VEDANTAM R, ZITNICK C L, PARIKH D, et al. CID Er: Consensus-based image description evaluation[C]// Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Piscataway, NJ: IEEE, 2015: 4566-4575.

[6] DEMARTINI G, DIFALLAH D E, CUDREMAUROUX P. ZenCrowd: Leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking[C]// Proceedings of the 21st international conference on World Wide Web. New York: ACM, 2012: 469-478.

[7] PAN Z G. Research on the recognition of Chinese named entity based on rules and statistics[J]. Information Science, 2012, 30(5): 708-712.

[8] FENG Y, JIANG B, WANG L, et al. Cybersecurity named entity recognition using multi-modal ensemble learning[J]. IEEE Access, 2020, 8: 63214-63224.

[9] ZHAO Z H, YANG Z B, LUO L, et al. Disease named entity recognition from biomedical literature using a novel convolutional neural network[J]. BMC Medical Genomics, 2017, 10(S5): 75-83.

[10] WANG P H, LI M Z, LI S. Data augmentation for Chinese clinical named entity recognition[J]. Journal of Beijing University of Posts and Telecommunications, 2020, 43(5): 84-90.

[11] AGUILAR G, MAHARJAN S, SOLORIO T, et al. A multi-task approach for named entity recognition in social media data[J]. arXiv: 1906.04135, 2019.

[12] GUO X C, TANG Z, DIAO L, et al. Recognition of Chinese agricultural diseases and pests named entity with joint radical-embedding and self-attention mechanism[J]. Transactions of the Chinese Society for Agricultural Machinery, 2020, 51(S2): 335-343.

[13] ZHANG H, GUO Y B, LI T. Domain named entity recognition combining GAN and BiLSTM-attention-CRF[J]. Journal of Computer Research and Development, 2019, 56(9): 1851-1858.

[14] DAS P, DAS K A, NAYAK J, et al. A graph based clustering approach for relation extraction from crime data[J]. IEEE Access, 2019, 7, 101269-101282.

[15] ZHAO P F, ZHAO C J, WU H R, et al. Named entity recognition of Chinese agricultural text based on attention mechanism[J]. Transactions of the Chinese Society for Agricultural Machinery, 2021, 52(1): 185-192.

[16] YU Y X, LI X. Research on text annotation method of ancient works from the perspective of digital humanities: a case study on MARKUS[J]. Big Data Research, 2022, 8(6): 15-25.

[17] ZHANG K L, ZHAO X, GUAN T F, et al. A platform for entity and entity relationship labeling in medical texts[J]. Journal of Chinese Information Processing, 2020, 34(6), 36-44.

[18] LAWSON N, EUSTICE K, PERKOWITZ M, et al. Annotating large email datasets for named entity recognition with Mechanical Turk[C]// Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk. Stroudsburg: ACL, 2010: 71-79.

[19] BOSTOCK M, OGIEVETSKY V, HEER J. D3: Data-Driven Documents[J]. IEEE Transactions on Visualization and Computer Graphics, 2011, 17(12): 2301-2309.

[20] MENDEZ G G, NACENTA M A. iVoLVER: Interactive Visual Language for Visualization Extraction and Reconstruction[C]// Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. New York: ACM, 2017: 4073-4085.

[21] REN D H, HOLLERER T, YUAN X R. iVisDesigner: Expressive Interactive Design of Information Visualizations[J]. IEEE Transactions on Visualization and Computer Graphics, 2014, 20(12): 2092-2101.

[22] ZHANG Y, WANG Y, ZHANG H D, et al. OneLabeler: A Flexible System for Building Data Labeling Tools[C]// Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. Stroudsburg: ACL, 2022: 1-22.

[23] FAN H, HUANG H C, WANG X, et al. Research on Knowledge Extraction Technology of Power Grid Text Data based on Semantic annotation[C]// Proceedings of the Third Smart Grid Conference. 2018: 146-150.

[24] ZHU Y, JING L P, YU J. An Active Labeling Method for Text Data Based on Nearest Neighbor and Information Entropy[J]. Journal of Computer Research and Development, 2012, 49(6): 1306-1312.

[25] SUN Q, HE W X, CHEN L Y, et al. Multi-Label Automatic Labeling for Question Attributes Based on Adaboost and Bayes Algorithms[C]// Proceedings of 2018 Chinese Automation Congress. Piscataway, NJ: IEEE, 2018, 2955-2960.

[26] BEYETTE D, WANG Z L, LIN J, et al. semi-automatic LaTeX-based labeling of mathematical objects in PDF documents: MOP data set[C]// Proceedings of the ACM Symposium on Document Engineering 2019. Stroudsburg: ACL, 2019: 1-4.

[27] FORT K, EHRMANN M, NAZARENKO A, et al. Towards a methodology for named entities annotation[C]// Proceedings of the Third Linguistic Annotation Workshop. Stroudsburg: ACL, 2009: 142-145.

[28] CAI Y B. AI Assistance: How to Handle Civil and Commercial Cases[J]. Oriental Law, 2018, 18(3): 131-139.

[29] WENG Y, GU S Y, LI J, et al. Paragraph Context-Based Text Classification Approach for Large-Scale Judgment Text Structuring[J]. Journal of Tianjin University (Science and Technology), 2021, 54(4): 418-425.

[30] ZHANG Y, YANG J. Chinese NER using lattice LSTM[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2018: 1554-1564.

[31] SUI D B, CHEN Y B, LIU K, et al. Leverage lexical knowledge for Chinese named entity recognition via collaborative graph network[C]// Proceedings of the 2019 Conference on Empirical

Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Stroudsburg：ACL，2019：3830-3840.

[32] GUI T，ZOU Y C，ZHANG Q，et al. A lexicon-based graph neural network for Chinese NER[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Stroudsburg：ACL，2019：1040-1050.

[33] MA R T，GUI T，ZHANG Q，et al. CNN-based Chinese NER with lexicon rethinking[C]// Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence. Menlo Park，CA：AAAI，2019：4982-4988.

[34] KONG B，LIU S Q，WEI F Y，et al. Chinese relation extraction using extend softword[J]. IEEE Access，2021，9：110299-110308.

[35] LI X N，YAN H，QIU X P，et al. FLAT：Chinese NER using flat-lattice transformer[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg：ACL，2020：6836-6842.

[36] LIU F，LU H，LO C，et al. Learning character-level compositionality withvisual features[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Stroudsburg：ACL，2017：2059-2068.

[37] SU T R，LEE H Y. Learning Chinese word representations from glyphs of characters[C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Stroudsburg：ACL，2017：264-273.

[38] MENG Y X，WU W，LI X Y，et al. Glyce：Glyph-vectors for Chinese character representations[C]// Proceedings of the 33rd International Conference on Neural Information Processing Systems. New York：ACM，2019：2746-2757.

[39] CAO S H，LU W，ZHOU J，et al. Cw2vec：Learning Chinese wordembeddings with stroke n-gram information[C]// Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence，the 30th innovative Applications of Artificial Intelligence，and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence. Menlo Park：AAAI，2018：5053-5061.

[40] SUN Y M，LIN L，YANG N，et al. Radical-enhanced Chinese character embedding[C]// Proceedings of the 21st International Conference on Neural Information Processing. Berlin，Heidelberg：Springer，2014：279-286.

[41] SHAO Y，HARDMEIER C，TIEDEMANN J，et al. Character-based joint segmentation andpos tagging for Chinese using bidirectional RNN-CRF[C]// Proceedings of the Eighth International Joint Conference on Natural Language Processing. Stroudsburg：ACL，2017：173-183.

[42] ZHANG Y，LIU Y G，ZHU J J，et al. Learning Chinese word embeddings from stroke，structure and pinyin of characters[C]// Proceedings of the 28th ACM International Conference on Information and Knowledge Management. New York：ACM，2019：1011-1020.

[43] ZHU W H，JIN X，NI J Y，et al. Improve word embedding using both writing and pronunciation[J]. PLoS One，2018，13(12)：1-13.

[44] CHAUDHARY A，ZHOU C，LEVIN L，et al. Adapting word embeddings to new languages with morphological and phonological subword representations[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg：ACL，2018：3285-3295.

[45] PENG N Y，DREDZE M. Named entity recognition for Chinese social media with jointly trained embeddings[C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Stroudsburg：ACL，2015：548-554.

[46] PENG N Y，DREDZE M. Improving named entity recognition for Chinese social media with word segmentation representation learning[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics(Volume 2：Short Papers). Stroudsburg：ACL，2016：149-155.

[47] HE H F，SUN X. F-score driven max margin neural network for named entity recognition in Chinese social media[C]// Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics：Volume 2. Stroudsburg：ACL，2017：713-718.

[48] HE H F，SUN X. A unified model for cross-domain and semi-supervised named entity recognition in Chinese social media[C]// Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. Menlo Park：AAAI，2017：3216-3222.

[49] CAO P F，CHEN Y B，LIU K，et al. Adversarial transfer learning for Chinese named entity recognition with self-attention mechanism[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg：ACL，2018：182-192.

WANG Yingjie，born in 1977，Ph.D，associate professor，is a member of CCF (No. 39234M). Her main research interests include software engineering and trustworthy software.

BAI Fengbo，born in 1978，Ph.D，senior software engineer，is a member of CCF (No. F6846M). His main research interests include natural language processing，data science，evidence science，etc.