计算机研究与发展
Journal of Computer Research and Development
DOI: 10.7544/issn1000-1239.202330691
61(5): 1325−1335, 2024

基于深度语义分析的警务卷宗知识抽取

马健伟　王铁鑫　江　宏　陈　涛　张　超　李博涵
（南京航空航天大学计算机科学与技术学院　南京　211106）
（jianweima@nuaa.edu.cn）

Knowledge Extraction Based on Deep Semantics Analysis Towards Police Dossier

Ma Jianwei, Wang Tiexin, Jiang Hong, Chen Tao, Zhang Chao, and Li Bohan
（*College of Computer Science and Technology*, *Nanjing University of Aeronautics and Astronautics*, *Nanjing* 211106）

Abstract　Police dossier, as one of the main records handled by the police department, contains massive and crucial policing information. As an important means, efficient information extraction from police dossier is helpful for case analysis, crime trend prediction, and the improvement of the public security management. However, the text of police dossier is written by police officers using natural language, which makes it difficult to extract crucial information. Traditional information extraction of police dossier heavily relies on manual effort and predefined templates, resulting in low efficiency and poor generality. Considering the particularity of police dossier, in this paper, a knowledge extraction method based on deep semantics analysis is proposed. This method consists of two core tasks: named entity recognition and relation extraction. Focusing on Chinese text, we propose a named entity recognition method that integrates structural and glyph features of Chinese characters. On the basis of entity recognition results, with the help of a specially constructed policing thesaurus, a relationship extraction method combining rule based and trigger word is proposed. Both on a publicly available Weibo dataset and a real dossier dataset provided by our partner a local police department, compared with several baseline named entity recognition models, our proposed method shows better performance in classifying exact entities and finding more potential entities. The automatically extracted relationships have also been verified and committed by the police department branch. A particular information system has been used in practice.

Key words　smart policing; police dossier; knowledge extraction; named entity recognition (NER); relation extraction

摘　要　卷宗作为公安机关办案、结案的主要记录，包含大量关键的警务信息.面向警务卷宗的信息抽取是分析案情、挖掘犯罪趋势、提高治安管理水平的重要手段.卷宗类文本多由基层警务人员采用自然语言书写，关键信息抽取难度大.传统的警务卷宗信息抽取，多依赖人工及预定义模板，效率低且通用性差.针对以上问题，参考卷宗的警务特征，提出了一种基于深度语义分析的卷宗知识抽取方法.该方法包含命名实体识别与关系抽取2个核心内容.提出的命名实体识别方法，融合了汉字结构特征和字形特征；提出的关系抽取方法建立在实体识别的基础上，实现基于触发规则和触发词的2种抽取模式.在公开的微博数据集、项目合作方**市**分局的真实卷宗集上，提出的命名实体识别方法对比基线方法，在实体识别精确率及召回率上综合表现优异；自动抽取的关系也得到**分局的认可.相关信息系统已在**分局部署使用.

关键词　智慧警务；警务卷宗；知识抽取；命名实体识别；关系抽取

收稿日期：2023-08-23；**修回日期**：2024-01-03
基金项目：国家自然科学基金面上项目 (61872182)
This work was supported by the General Program of the National Natural Science Foundation of China (61872182).
通信作者：王铁鑫（tiexin.wang@nuaa.edu.cn）

中图法分类号 TP391

近年来，随着经济与技术的快速发展，社会环境愈发复杂，治安案件类型与模式不断演化，如新型网络赌博、电信诈骗等. 各类案件的发生不仅严重破坏社会治安环境，也给公安机关的工作带来了极大的挑战. 2010 年以来，警务工作朝着网络化和合成化普及并快速发展，各地公安机关开始探索“智慧警务”建设. 然而，现阶段的“智慧警务”多针对结构化数据，鲜有针对文本类警务卷宗的处理方法. 作为公安办案结案的重要记录方式，警务卷宗中隐藏着大量的警务知识. 对这类隐藏知识的挖掘，能够揭示犯罪活动的形式和变化趋势，有助于公安机关及时调整工作重点，有效预防案件的发生. 然而，传统的卷宗处理多依赖人工审查，导致卷宗中隐藏的信息难以充分利用[1]. 同时，人工方式分析案件卷宗，消耗大量人力资源，加重基层民警压力，给其带来沉重的工作负担.

文本的自动知识抽取，可以有效代替人工审查，提高抽取效率的同时，减轻民警工作负担. 知识抽取包含 2 项核心任务[2]：命名实体识别 (named entity recognition, NER) 与关系抽取. 前者从文本中识别和提取具有特定含义的实体，可有效减少人工标注的工作量，降低标注错误的风险，亦是实现后者的基础. 本文针对中文警务卷宗，结合汉字的特点，利用卷积神经网络分别提取汉字的部首结构以及字形特征信息；采用 Co-Transformer 模型[3] 进行特征提取，并通过特征融合实现对文本的深度语义分析，进而提出一种新的命名实体识别方法. 同时，基于命名实体识别结果，结合警务领域知识，提出结合触发词与触发规则的特定关系抽取方法. 本文研究工作是与**市**公安分局①联合开展的，得到该局的大力支持，具体包括领域知识分享、提供真实案件卷宗作为验证数据集等. 本文提出的知识抽取方法，作为构件集成于该分局的“智慧警务建模与数据挖掘平台”.

1 相关工作

目前，智慧警务建设多面向并依赖于结构化数据，海量的历史卷宗没有得到充分利用. 本文针对自由文本类的警务卷宗，使用命名实体识别、关系抽取完成知识抽取任务.

1.1 命名实体识别与关系抽取

在过去的 10 余年间，人工智能技术，如长短期神经网络[4](long short-term memory, LSTM)、条件随机场[5](conditional random field, CRF)等，极大地提高了命名实体识别的性能[6-8]. 2018 年，谷歌发布了预训练 BERT 模型[9]. 该模型利用上下文语义信息，通过基于字符的训练方式(无须事先分词)，提高模型的泛化能力. BERT 模型在中文自然语言处理任务中表现出色. 针对中文词语之间没有空格分隔、命名实体识别划分任务难度大等问题，Zhang 等人[10] 提出一种利用汉字词典信息的 Lattice-LSTM 模型，将字符和词汇一起编码作为模型输入，结合调整后的 LSTM 作为上下文编码器，接受 Lattice 输入，以降低歧义的发生概率. 在此基础上，Li 等人[11] 在 Lattice 基础上，提出了一种新的位置编码方法 FLAT(flat Lattice Transformer)，通过使用相对位置更好地编码整个文本结构. FLAT 充分地利用 Lattice 信息，在命名实体识别任务上表现出色. 与此同时，一些研究者开始关注汉字本身所包含的信息，探索汉字字符的内在特征和语义信息，如汉字部首信息[12]、头尾信息[13]、笔画信息[14] 等.

文本的关系抽取方法，分为基于有监督学习和弱监督学习 2 种类型. 基于有监督学习的方法，依赖于大量人工标注的数据来训练模型，如文献 [15−17] 的研究工作. 相比之下，基于弱监督学习的方法只需要少量标记数据即可实现模型训练，包括利用句子级注意力的神经网络模型[18]、基于深度学习 TextRCNN 模型的短文本分类模型[19] 以及使用模式感知自注意机制的关系提取模型[20] 等. 这些方法通过有效利用已有数据及提取到的语义特征，有效地提升了关系抽取任务的性能. 然而弱监督学习在关系抽取任务中，缺乏有效应对文本噪声与歧义性的机制[21].

1.2 知识抽取实践应用

随着自然语言处理和人工智能的发展，知识抽取在警务、医疗、军事等众多领域中都有广泛的实践应用，为众多行业带来效率和智能化的提升.

针对警务应用，陈柱辉等人[22] 针对简要案情文本中的实体稠密分布、实体间相互嵌套以及实体简称问题，对字符向量生成方法进行了改进，提出了 RC-BiLSTM-CRF 网络架构，解决了预训练模型带来

① 考虑公安工作的敏感性与保密性，应合作分局要求，隐匿其名称并对其提供的数据进行修改后作展示使用.

的向量冗长问题，并通过修改参数提高收敛速度. 针对低频罪名数据量较少且易混淆罪名案情描述相似的问题，郭军军等人[23]提出了一种基于双向互注意力机制的案件辅助句融合方法，显著提高了低频及易混淆罪名的预测性能. 医疗领域中，罗凌等人[14]为解决中文电子病历标注数据稀缺、实体标注需要专业知识的问题，提出了一种基于笔画 ELMo 和多任务学习的中文电子病历识别方法. 针对军事领域应用，李健龙等人[24]为减少传统命名实体识别需要人工标注特征的工作量，通过无监督训练获取军事领域语料的分布式向量表示，使用双向 LSTM 递归神经网络模型来解决军事领域命名实体的识别问题.

综上，现有针对命名实体识别任务的研究大多集中于抽取单一维度的汉字特征以实现实体抽取任务，而忽略了句子的整体语义. 而在关系抽取任务中，主要采用基于深度学习的方法，针对特定测试集调整可变参数. 这类方法在通用性和准确性方面存在很大的局限性. 本文在 FLAT 位置编码结构的基础上，进一步考虑汉字的结构特征、字形特征，采用基于触发规则和触发词结合的关系抽取方法，能够以较高的精确率完成文本的命名实体识别、关系抽取任务.

2 警务卷宗知识抽取方法

2.1 方法框架

本文提出的文本知识抽取方法整体框架如图 1 所示. 以警务卷宗为输入：首先，使用改进的 VGG16 网络抽取汉字结构特征、字形特征，并依据 FLAT 位置编码构建汉字 Lattice 信息，实现命名实体识别并对识别出的实体进行分类；然后，在实体识别及分类的基础上，进行基于触发规则和触发词的关系抽取；最后，将抽取获得的知识（关系三元组）以结构化形式存储固化，以供进一步数据分析使用.

2.2 命名实体识别

命名实体识别针对卷宗中涉及的实体，如人、物、地点、事件等展开. 考虑中文汉字特征复杂、文本中词语之间没有空格分隔等特点，本文提出了一种结合汉字结构特征和字形特征的神经网络模型.

如图 2 所示，整个网络架构主要分为 3 层：输入表示层、语义编码层、标签预测层. 在输入表示层，为了将汉字图像转化为向量表示，使用卷积神经网络提取汉字的结构特征和字形特征，同时考虑上下文语义信息，使用 Flat-Lattice[13]嵌入，一起作为语义编码层的输入. 在语义编码层，使用 Co-Transformer 编码器来对句子的字、词、结构以及字形特征进行建模，将输入的汉字信息转换为更高级别的语义表示. 在标签预测层，使用 CRF 对句子进行标签预测，将经过编码的信息转换为最终的输出结果.

2.2.1 输入表示层

1）汉字结构特征提取

汉字作为象形文字，通常由较小的偏旁组成. 偏旁作为汉字的基本单元，给予汉字额外的语义信息，有助于识别中文命名实体. 表 1 展示了汉字偏旁及其含义的示例.

以偏旁“艹”为例，其通常表达草本植物相关的含义，比如花、草、菊等. 然而，在汉字中，偏旁仅能

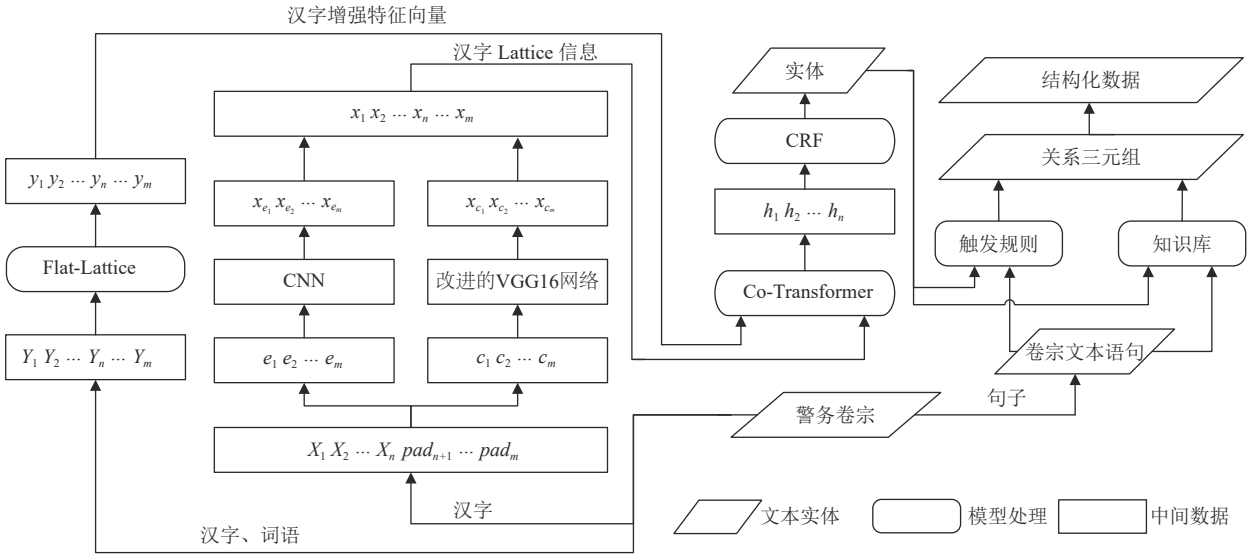

Fig. 1 The overall framework of our proposed method

图 1 本文提出方法的整体框架图

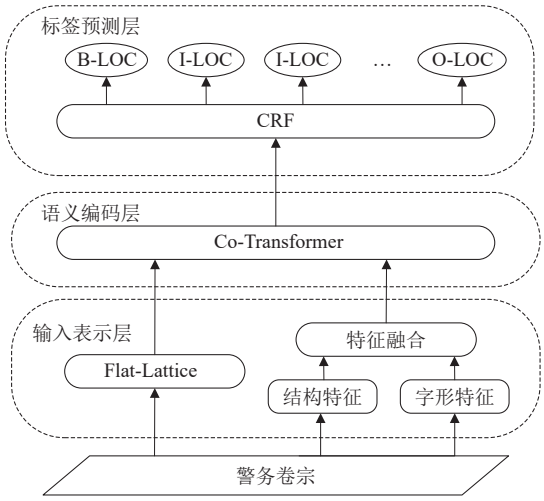

Fig. 2 The structure diagram of the improved NER network

图 2 改进后的命名实体识别网络结构图

Table 1 Example of Chinese Character Radicals

表 1 汉字偏旁示例

汉字偏旁	释义	示例
氵	水、液体	海、河、沟、深
艹	草本植物	花、草、菊、芳
木	树木、木制品	森、松、桥、根
月	身体	胶、胆、脉、脆

表达部分释义；偏旁之外，汉字其余结构(具有相同偏旁)包含汉字的不同特征，体现语义差异.

汉字由不同的部首组成. 例如，汉字“鹰”的结构由“广”“亻”“隹”“鸟”4 个部首构成，而汉字“榆”的结构则由“木”“人”“一”“月”“刂”5 个部首构成. 通过对汉字不同部首结构的分析，能够对汉字的深层语义进行提取、分析.

在自然语言处理领域，通常使用卷积神经网络处理序列数据. 由上述分析可知，每个汉字包含有限部首，因此本文使用卷积神经网络来提取汉字部首结构层面的特征嵌入. 具体的提取步骤为：首先，将汉字按部首结构拆分；再将拆分结果作为卷积神经网络的输入；最后使用最大池和全连接层获取汉字结构层面的特征嵌入.

2)汉字字形特征提取

汉字字形由图形演变为笔画形式，字形和释义有着密切的关系. 通过拆分汉字结构，并使用卷积神经网络，可提取汉字部首级别的特征. 然而，相同的部首构造也可能组成不同的汉字. 例如汉字古和叶都可以拆分为“十”和“口”，“木”和“几”也可以组成汉字朵和机. 此外，汉字形态和表意也有着紧密的关系，字形相近的汉字可能有相似的释义，如辩与辨、江与河、草与苗等. 因此，本文利用汉字的字形信息，采用卷积神经网络来提取汉字字形的特征，以获取字形中高维度的语义信息.

为获取汉字的字形特征，本文采用 VGG16 模型对汉字字形图像进行处理. VGG16 拥有 13 个卷积层和 3 个全连接层，相较于其他经典卷积神经网络模型，其拥有更多的可训练参数及网络深度，能够有效提取图像中的高维度特征[24]. 汉字图像大小及通道数不同于传统的 RGB 图像，本文对 VGG16 的输入层进行改进，以提高图像处理效率. 图 3 展示了改进后的 VGG16 网络架构.

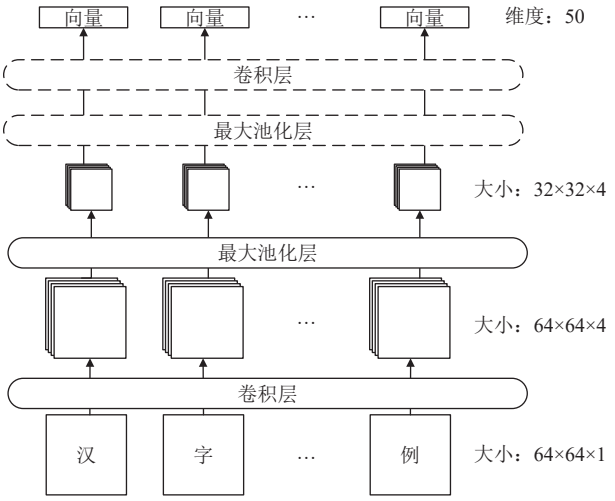

Fig. 3 The improved VGG16 network

图 3 改进的 VGG16 网络

参考《现代汉语常用字表》，中国常用汉字共 3 500 个；进一步，常用汉字可分为 2 500 个较常用汉字及 1 000 个次常用汉字. 本文从在线新华字典中收集了 4 702 张汉字图片，在包含常用汉字的基础上，对较常用、次常用的汉字实现覆盖. 利用改进后的 VGG16 网络的卷积层和池化层，将输入图片转换为 50 维向量表示. 50 维向量能够充分捕捉汉字特征并控制计算复杂度、节省存储空间.

3)特征融合

为了获得更全面的汉字特征，本文对 1)2)提取到的汉字结构特征和字形特征进行降维，并将降维后的 2 个特征向量进一步转换成相同维度的向量. 接着，将 2 个特征向量进行拼接，形成一个新的、维度更高的汉字增强特征向量，以更好地表征汉字所承载的信息.

2.2.2 语义编码层

Co-Transformer 是一种基于 Transformer 的神经网络结构，它采用了双向编码器-解码器的结构，用

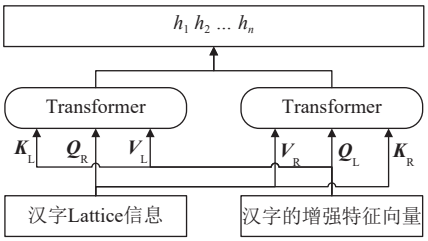

Fig. 4 The structure of Co-Transformer

图 4 Co-Transformer 结构

于对输入的多种特征向量进行建模和提取. Co-Transformer 结构如图 4 所示，左侧输入的是汉字的 Flat-Lattice 信息，即位置编码信息；右侧输入的是汉字的增强特征向量，即结构特征和字形特征.

图 4 中 $\boldsymbol{Q}_{\mathrm{L}}$, $\boldsymbol{K}_{\mathrm{L}}$, $\boldsymbol{V}_{\mathrm{L}}$ 表征 Flat-Lattice 信息，$\boldsymbol{Q}_{\mathrm{R}}$, $\boldsymbol{K}_{\mathrm{R}}$, $\boldsymbol{V}_{\mathrm{R}}$ 表征汉字特征信息，其计算如式(1)所示：

$$\begin{pmatrix} \boldsymbol{Q}_{(\mathrm{L,R}),i} \\ \boldsymbol{K}_{(\mathrm{L,R}),i} \\ \boldsymbol{V}_{(\mathrm{L,R}),i} \end{pmatrix}^{\mathrm{T}} = \boldsymbol{E}_{(\mathrm{L,R}),i} \begin{pmatrix} \boldsymbol{W}_{(\mathrm{L,R}),\boldsymbol{Q}} \\ \boldsymbol{I} \\ \boldsymbol{W}_{(\mathrm{L,R}),\boldsymbol{V}} \end{pmatrix}^{\mathrm{T}}. \tag{1}$$

$\boldsymbol{E}_{\mathrm{L}}$ 为 Flat-Lattice 嵌入，$\boldsymbol{E}_{\mathrm{R}}$ 为增强的汉字特征嵌入，$\boldsymbol{I}$ 是单位矩阵，$\boldsymbol{W}$ 是线性变换矩阵. 同时，本文使用式(2)对相对位置编码进行计算.

$$\boldsymbol{R}_{i,j} = ReLU(\boldsymbol{W}_{\mathrm{R}}(\boldsymbol{p}_{h_i-h_j} \oplus \boldsymbol{p}_{h_i-t_j} \oplus \boldsymbol{p}_{t_i-h_j} \oplus \boldsymbol{p}_{t_i-t_j})), \tag{2}$$

其中：$\boldsymbol{p}$ 是一个增强的位置编码向量，用于编码相对位置信息；h_i 和 t_i 分别表示第 i 个汉字或词的头部位置和尾部位置，⊕表示向量串联操作. 具体而言，$\boldsymbol{p}$ 的计算公式为：

$$\boldsymbol{p}_d^{2k} = \sin\left(\frac{d}{10\,000^{2k}/d_{\mathrm{model}}}\right), \tag{3}$$

$$\boldsymbol{p}_d^{2k+1} = \cos\left(\frac{d}{10\,000^{2k}/d_{\mathrm{model}}}\right), \tag{4}$$

其中 d 表示 h_i–h_j, h_i–t_j, t_i–h_j, t_i–t_j. 该相对位置编码可以有效地表示输入句子中汉字和词语之间的相对距离关系，以增强节点特征表示能力. 本文采用的相对位置编码方法可以将汉字或词语的位置信息转换为向量形式，得到表示汉字或词语之间相对位置的编码向量，准确地反映出不同汉字或词语之间的距离关系. 注意力计算采用式(5)(6)：

$$A_{\mathrm{L},ij} = \boldsymbol{W}_{\mathrm{L},\boldsymbol{Q}}^{\mathrm{T}}\boldsymbol{E}_{\mathrm{L},i}^{\mathrm{T}}\boldsymbol{E}_{\mathrm{L},j}\boldsymbol{W}_{\boldsymbol{K},\boldsymbol{E}} + \boldsymbol{W}_{\mathrm{L},\boldsymbol{Q}}^{\mathrm{T}}\boldsymbol{E}_{\mathrm{L},i}^{\mathrm{T}}\boldsymbol{R}_{i,j}\boldsymbol{W}_{\boldsymbol{K},\boldsymbol{R}} + \boldsymbol{u}^{\mathrm{T}}\boldsymbol{E}_{\mathrm{L},j}\boldsymbol{W}_{\boldsymbol{K},\boldsymbol{E}} + \boldsymbol{v}^{\mathrm{T}}\boldsymbol{R}_{i,j}\boldsymbol{W}_{\boldsymbol{K},\boldsymbol{R}}, \tag{5}$$

$$A_{\mathrm{R},ij} = \boldsymbol{W}_{\mathrm{R},\boldsymbol{Q}}^{\mathrm{T}}\boldsymbol{E}_{\mathrm{R},i}^{\mathrm{T}}\boldsymbol{E}_{\mathrm{R},j}\boldsymbol{W}_{\boldsymbol{K},\boldsymbol{E}} + \boldsymbol{W}_{\mathrm{R},\boldsymbol{Q}}^{\mathrm{T}}\boldsymbol{E}_{\mathrm{R},i}^{\mathrm{T}}\boldsymbol{R}_{i,j}\boldsymbol{W}_{\boldsymbol{K},\boldsymbol{R}} + \boldsymbol{u}^{\mathrm{T}}\boldsymbol{E}_{\mathrm{R},j}\boldsymbol{W}_{\boldsymbol{K},\boldsymbol{E}} + \boldsymbol{v}^{\mathrm{T}}\boldsymbol{R}_{i,j}\boldsymbol{W}_{\boldsymbol{K},\boldsymbol{R}}. \tag{6}$$

$\boldsymbol{W}$ 和 $\boldsymbol{E}$ 同式(1)中的定义，$\boldsymbol{A}_{\mathrm{L}}$ 为 Flat-Lattice 注意力机制，$\boldsymbol{A}_{\mathrm{R}}$ 为增强的字符信息注意力机制，$\boldsymbol{R}_{i,j}$ 为相对位置编码向量. 在该模型中通过式(7)计算 Flat-Lattice 嵌入和增强的字符信息嵌入的注意力得分：

$$Attention_{(\mathrm{L,R})}(\boldsymbol{A}_{(\mathrm{R,L})}, \boldsymbol{V}_{(\mathrm{L,R})}) = softmax(\boldsymbol{A}_{(\mathrm{R,L})})\boldsymbol{V}_{(\mathrm{L,R})}. \tag{7}$$

2.2.3 标签预测

本文利用 Co-Transformer 对特征进行提取后，采用 CRF 通过学习标签序列的联合概率分布来预测最优的标签序列.

在自然语言处理中，人工标注过程中可能会误差或错误导致模型预测不准. CRF 能够从全局的角度考虑标签之间的依赖关系，因而可以纠正局部错误，弥补人工标注的不足，提高标签序列的准确性.

对于给定的序列 H=($h_1\ h_2\ \cdots\ h_n$)，其包含 n 个字符，针对预测标签序列 Y=($y_1\ y_2\ \cdots\ y_n$)，可以通过式(8)计算具体得分：

$$S(H,Y) = \sum_{i=0}^{n} A_{y_i,y_{i+1}} + \sum_{i=0}^{n} P_{i,y_i}. \tag{8}$$

其中 $\boldsymbol{P}$ 是 Co-Transformer 输出的概率矩阵，P_{i,y_i} 表示第 i 个字符标签为 y_i 的概率得分，$\boldsymbol{A}$ 是 CRF 生成的转移矩阵，$A_{y_i,y_{i+1}}$ 表示从 y_i 到 y_{i+1} 的转移分数. 所有的标签归一化后得到 y 的最大概率，再使用式(9)计算得到最大分数，实现最优标签序列预测：

$$y^* = \arg\max(H, y'). \tag{9}$$

2.3 关系抽取

实体与关系相互关联，为了获得实体属性及实体之间的语义联系，本文通过定义并使用规则模板来识别满足特定触发规则的卷宗文本. 针对卷宗文本中无规则部分，本文通过构建案件知识库来完成对应的关系抽取任务.

2.3.1 基于触发规则的关系抽取

基于触发规则的关系抽取依赖于预定义的规则模板，实现“实体-关系-实体”三元组抽取. 通过归纳总结警务卷宗，本文发现人名实体与其特定意义属性关系词的跟随关系. 因此，定义对应的触发规则，实现这类句子中的实体-属性关系抽取.

上述抽取实体属性关系任务主要步骤为：1)使用 2.2 节提出的命名实体识别方法识别案件文本内实体；2)针对每个人名实体，寻找其后特定意义的属性关系词语，如“性别”“年龄”“住址”等，据此获得〈实体，属性名，属性〉三元组规则；3)对于包含属性关系词语的语句，使用预先设置的抽取规则抽取出其中的实体、属性和属性值，合成为三元组.

2.3.2 基于案件知识库的关系抽取

为正确识别实体间的关系类型，本文采用建立案件知识库的方法，该知识库内存储针对警务卷宗关系抽取的潜在触发词. 针对卷宗文本的特点及卷

Table 2 Example of Entity Relationship Types

表 2 实体关系类型示例

序号	实体关系类型	示例
1	人员类型	嫌疑人、报案人、被害人
2	案件类别	盗窃、诈骗、传销
3	具有	具有、享有、富有、具备
4	损失	损失、金额、破财、折损
5	协作	协作、搭档、搭伙、共同

宗的记录流程,通过与合作分局基层民警的多次深入讨论,确定了5种实体关系类型,如表2所示.

针对案件文本中的频发案件类型,本文选取了盗窃、强奸、侮辱等6个案件类型进行案件知识库的构建,并利用同义词词林扩展版[25]对其进行扩展.部分示例如表3所示.

Table 3 Example of Case Types

表 3 案件类型举例

序号	案件关系类型	示例
1	盗窃	盗窃、被盗、被偷、偷窃、行窃
2	强奸	奸污、强奸、糟踏、施暴、轮奸
3	侮辱	侮辱、羞辱、污辱、凌辱、折辱
4	诈骗	诈骗、骗子、欺骗、蒙骗、诓骗
5	传销	传销、传销商品
6	故意伤害	斗殴、打架、争斗、动手、打斗

对于无规则的案件文本,实体之间通常存在一定的特征词,因此本文利用构建的知识库实现关系抽取任务.基本思想是:将文本语句中的特征词与知识库进行匹配,以获得关系模型.通过对应的关系模型,可以形成相应的三元组表示.

主要实现过程为:1)从公安卷宗库中选取对应的案件文本,保留包含完整案件信息的文本数据;2)使用HanLP[26]进行分词、词性标注,再利用命名实体识别技术从案件文本数据中提取出与案件相关的实体;3)根据处理结果,利用TF-IDF(term frequency-inverse document frequency)算法计算某个词在案件知识库中的词频以及在整个语料库中的逆文档频率,以评估关键词在案件文本中的重要程度,从案件数据中抽取TF-IDF值最高的3个词作为特征词及实体信息,如果当前文本语句中没有与案件相关的信息,则处理下一条语句;4)如果当前文本语句中包含相应的信息,则将当前语句和知识库中的实体关系类型进行匹配,若存在相应关系,则返回对应实体以及关系信息三元组.

3 实 验

3.1 研究问题

为证明本文提出模型的有效性及可用性,实验设计2个研究问题:1)本文提出的基于汉字多特征融合的命名实体识别模型是否优于其他基线模型;2)本文提出的方法是否能够有效地提取警务卷宗中的实体.

3.2 实验设置

为了全面比较本文提出的命名实体识别方法与其他基线模型的性能表现,本文使用精确率、召回率、$F1$值作为模型的评价指标.

针对研究问题1,本文选取微博数据集作为输入,进行不同命名实体识别模型间的对比实验.该数据集包含约1 940条句子,均已完成标注.其中训练集中包含实体约1 890个,验证集包含实体约390个,测试集包含实体约490个.

针对研究问题2,本文采用公安真实卷宗作为测试数据集,该数据集包含1 000条公安案件数据,涵盖近1 400条语句.预处理阶段:使用3位序列标注BIO(begin, inside, outside)来实现数据集的实体边界划分和标注定义.针对案件文本数据特点,本文将实体类型定义为人名(PER)、地址名(LOC)、组织名(ORG)、案件类别名(CAS)、时间(TIM)和物品名(GOO)这6种类型.预处理后的实验数据集中共包含8 796个实体,具体的分类统计信息如表4所示.案件数据集随机划分为训练集、验证集和测试集,比例为8∶1∶1.

Table 4 Entity Type Statistics of Police Dossier Dataset

表 4 卷宗数据集中实体类型统计

序号	实体类型	数量	所占比例/%
1	人名 (PER)	3 686	41.91
2	地址名 (LOC)	1 288	14.64
3	组织名 (ORG)	368	4.18
4	案件类别名 (CAS)	1 319	15.00
5	时间名 (TIM)	1 689	19.20
6	物品名 (GOO)	446	5.07

3.3 结果分析

3.3.1 实体识别表现

针对微博数据集,基线模型多采用Name实体、

Nominal 实体及整个数据集上的 *F*1 值进行实验，本文沿用该对比策略，具体比较结果如表 5 所示. 对比所有基线模型，本文所提出模型在 Name 实体及整个数据集上取得的 *F*1 值均最高，在 Nominal 实体上表现略逊于 LR-CNN 模型. 由于该微博数据集为通用数据集，其包含的实体类型多元、复杂，数据噪声明显，因此，本实验中所有模型的性能表现均不够理想，且差别不大.

Table 5 Model Performance Comparative Results

表 5 模型性能对比结果 %

模型	*F*1 值		
	Name 实体	Nominal 实体	所有实体
Peng and Dredze	55.28	62.97	58.99
Lattice-LSTM	53.04	62.25	58.79
LR-CNN	57.14	**66.67**	59.92
PLT	53.55	64.90	59.76
FLAT			60.32
MECT	61.91	62.51	63.30
本文模型	**61.94**	64.98	**63.41**

注：黑体表示最优值.

表 5 的实验结果表明，本文所提出的基于汉字多特征融合的命名实体识别模型的表现略优于其他模型. 同时，本文模型是对 FLAT 和 MECT 这 2 种模型的整合及改进. 因此，与 FLAT 及 MECT 的对比可视为针对本文模型的消融实验. 对比 Flat 模型，在增加汉字字形特征的情况下，本文模型在 *F*1 值上提升了 3.11 个百分点；对比仅使用 Flat-Lattice 和汉字特征的 MECT 模型，本文模型 *F*1 值提升了 0.11 个百分点. 综上，汉字自身携带的结构及字型特征信息对于命名实体识别结果有潜在影响.

3.3.2 模型有效性

以真实案件文本为数据集，表 6 展示了 3 种命名实体识别方法的实验结果. 其中，FLAT 模型使用了字符级别的特征信息，MECT 模型进一步融合了汉字的基本结构信息.

Table 6 Model Performance Comparative Results of Police Dossier Dataset

表 6 警务卷宗数据集的模型性能对比结果 %

模型	精确率	召回率	*F*1 值
FLAT	89.81	90.11	89.96
MECT	91.72	90.62	91.17
本文模型	**92.89**	**90.74**	**91.80**

注：黑体表示最优值.

针对案件文本数据集，本文在模型配置中使用参数组合：*epoch*=50, *lr*=0.001 4, *radical_dropout*=0.1, *char_dropout*=0.2, *img_embed_lr_rate*=0.002 7. 本文模型的精确率、召回率和 *F*1 值均高于其他 2 种对比模型. 这表明本文所提出的增强的字符向量能够更加有效地提取出汉字的特征信息，从而提高命名实体识别的准确性. 此外，仅考虑了汉字字词特征的 FLAT 模型的效果最差，和 MECT 模型相比其 *F*1 值差 1.21 个百分点，和本文所提出的模型相比其 *F*1 值差 1.84 个百分点. 这说明在中文命名实体识别领域，仅使用字符级别的特征信息是不够的. 相比之下，融合 Flat-Lattice 及汉字字形信息的 MECT 模型的 *F*1 值和本文方法的 *F*1 值相差 0.63 个百分点，说明在本文模型设计中引入更多的语义信息能够对实体识别结果产生一定的积极影响. 该测试数据集中包含特定类型实体且数据噪声小，所有模型的命名实体识别表现明显优于其在通用数据集上的表现.

4 系统工具

4.1 系统分析

本节介绍面向警务卷宗的数据建模及挖掘原型系统. 该系统集成了前述章节的知识抽取技术，并已在**分局部署应用，显著提升了该局基层警务人员的工作效率.

该系统能够解决面向海量警务卷宗的数据建模及数据挖掘难题. 该系统支持将基层民警的办案经验通过拖拽数据表的方式形成特定数据研判模型并固化存储；同时，通过整合多个数据挖掘算法，该系统支持针对特定数据集的模式挖掘.

4.2 系统设计

4.2.1 系统总体设计

如图 5 所示，本文设计的原型系统可划分数据汇聚层、数据层、业务层、应用层.

数据汇聚层是原型系统的数据来源，位于系统的底层，主要功能为实现多源数据的整合. 本文提出的知识抽取方法集成于该层，以警务卷宗为数据源，通过应用知识抽取方法实现卷宗数据的结构化处理，并以三元组的形式存储，以满足进一步的数据处理需求.

数据层是警务结构化数据存储的支撑，能够保证警务数据的相对独立性，以提高数据的安全性. 该原型系统主要使用 PostgreSQL 关系型数据库作为数据层的存储平台.

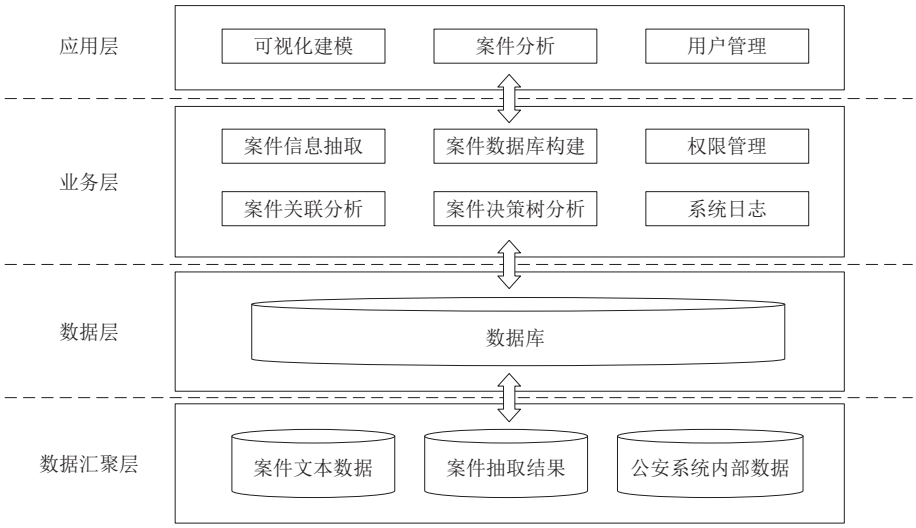

Fig. 5 System architecture diagram

图 5 系统架构图

业务层作为链接数据层及应用层的中间构件，由卷宗信息抽取、卷宗数据库构建、权限管理、案件关联分析、案件决策树分析、系统日志 6 个部分组成.

应用层作为与用户交互的界面，能够通过拖拽、点击的方式调用业务层的服务，使用透明传输的方式连接了应用层与数据层，保证了数据的安全性，实现了可视化建模、案件数据分析和面向用户服务的功能.

4.2.2 系统功能模块

警务人员通过浏览器访问相应服务器地址，通过身份验证进入原型系统. 警务人员通过使用自定义建模平台内的拖拽及配置功能，实现数据采集、数据建模及数据挖掘工作. 其中，数据挖掘模块提供的功能操作包括：数据探查、数据规则分析、决策树分析等.

系统通过预留功能接口的方式，支持系统进一步拓展，以实现挖掘算法实时更新、多背景拓展等目的.

4.3 系统实现

原型系统基于 Vue.js 框架，使用 JavaScript 语言及 Python 语言开发，采用 PostgreSQL 数据库存储警务数据作为输入来源.

4.3.1 自定义建模平台

进入自定义建模平台后，图 6 展示界面为拖拽式建模界面，依次为导航栏、资源库栏、建模画布、建模工具栏.

通过拖拽的方法可以将资源库栏内的数据表拖拽进入建模画布中，也可以将建模工具栏内的工具拖入建模画布中；节点化的数据表，可通过连线功能进行连接，实现数据表与建模工具的相关操作.

建模画布下方设置了对模型实现相关操作的功能键，依次为数据挖掘、模型训练、模型详情、保存模型、删除连线、删除节点及重置模型，能够对模型进行相关操作.

点击导航栏的查看数据表按钮，可以展示数据库内的数据表，包含数据库名、所属分类、数据库展示名等内容，能够对数据库及建模所用的数据表有进一步的理解.

点击导航栏的模型超市按钮，能够展示模型超市内存储的模型，通过双击保存的模型，能够将保存在数据库表内存储的模型加载进入建模画布中，能够实现模型管理、模型复用的效果.

4.3.2 数据挖掘平台

数据挖掘平台能够通过构件的方式对卷宗数据进行进一步规范化治理，现有原型系统汇聚了规则抽取技术、决策树可视化算法，能够对卷宗数据及警务数据库数据进一步地深度治理，实现对警务信息的全面补全、挖掘隐藏的知识和关联、提高警务数据的质量和可用性. 图 7 为平台中的一个数据挖掘构件实例.

5 结束语

针对警务卷宗信息提取问题，本文提出了基于深度学习的知识抽取方法，实现命名实体识别及领域特定关系的自动抽取. 利用 Co-Transformer 模型融

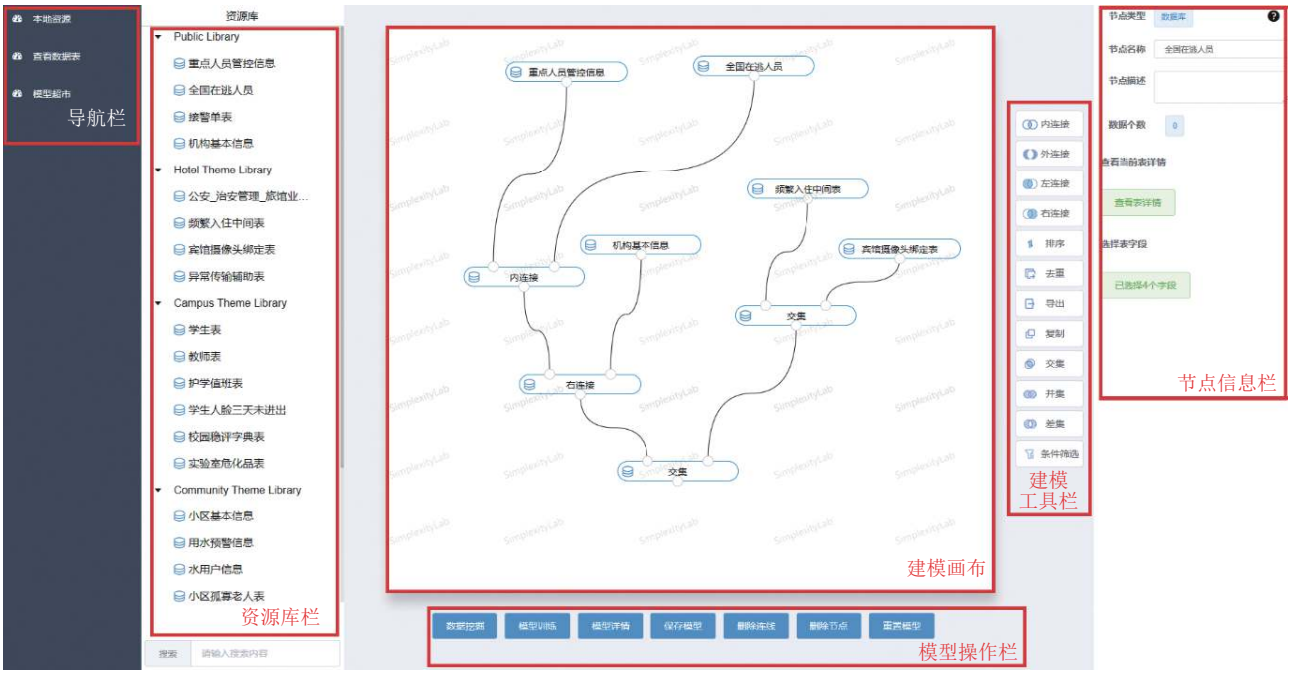

Fig. 6 An example of the modeling platform

图 6 建模平台示例

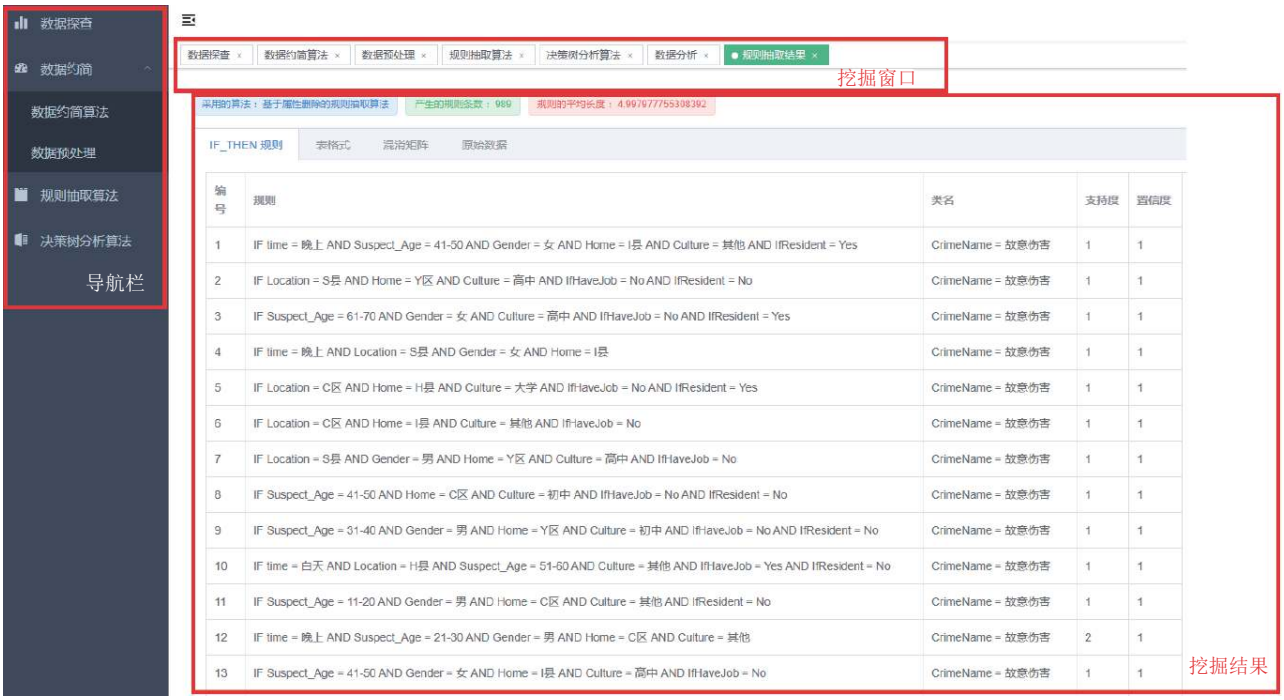

Fig. 7 An example of the mining platform

图 7 挖掘平台示例

合汉字字形特征、结构特征对汉字多个维度实现深度语义编码，进行实体类型标注. 使用微博数据集、真实卷宗数据集进行实验，本文提出的方法在实体识别精确率、召回率上综合表现优异，方法有效性和可用性得到了证明. 同时，针对公安系统的真实需求，本文构建了基于警务卷宗的建模及挖掘的原型系统，实现了预期的目标与功能.

未来工作中，考虑融合更多的汉字特征如发音等，以进一步提高命名实体识别的准确性. 同时，通过在更多公开数据集上进行模型训练，并针对不同

的领域采用不同的数据集进行训练,提高本文方法的通用性,使其可以迁移到其他应用领域.

作者贡献声明:马健伟提出实验设计,完成实验结果分析、论文撰写与修改;王铁鑫负责项目统筹、论文撰写与优化、实验设计;江宏提出核心算法思路并设计实验和实施;陈涛和张超参与技术路线实现并修改论文;李博涵参与论文写作指导及润色.

参 考 文 献

[1] Jin Xiaolong, Benjamin W, Cheng Xueqi, et al. Significance and challenges of big data research[J]. Big Data Research, 2015, 2(2): 59−64

[2] Deng Shumin, Ma Yubo, Zhang Ningyu, et al. Knowledge extraction in low-resource scenarios: Survey and perspective[J]. arXiv preprint, arXiv: 2202.08063, 2022

[3] Lu Jiasen, Dhruv B, Devi P, et al. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks[C]// Proc of the 33rd Int Conf on Neural Information Processing Systems. New York: ACM, 2019: 13−23

[4] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735−1780

[5] Sutton C, McCallum A. An introduction to conditional random fields[J]. Foundations and Trends® in Machine Learning, 2012, 4(4): 267−373

[6] Huang Zhiheng, Xu Wei, Yu Kai. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv preprint, arXiv: 1508.01991, 2015

[7] Peng Nanyun, Mark D. Named entity recognition for Chinese social media with jointly trained embeddings[C]//Proc of the 25th Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2015: 548−554

[8] Peng Nanyun, Mark D. Improving named entity recognition for Chinese social media with word segmentation representation learning[J]. arXiv preprint, arXiv: 1603.00786, 2016

[9] Jacob D, Chang Mingwei, Kenton L, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint, arXiv: 1810.04805, 2018

[10] Zhang Yue, Yang Jie. Chinese NER using lattice LSTM[J]. arXiv preprint, arXiv: 1805.02023, 2018

[11] Li Xiaonan, Yan Hang, Qiu Xiepeng, et al. FLAT: Chinese NER using flat-lattice transformer[C]//Proc of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2020: 6836−6842

[12] Xu Canwen, Wang Feiyang, Han Jialong, et al. Exploiting multiple embeddings for Chinese named entity recognition[C]//Proc of the 28th ACM Int Conf on Information and Knowledge Management. New York: ACM, 2019: 2269−2272

[13] Wu Shuang, Song Xiaoning, Feng Zhenhua. MECT: Multi-metadata embedding based cross-transformer for Chinese named entity recognition[J]. arXiv preprint, arXiv: 2107.05418, 2021

[14] Luo Ling, Yang Zhihao, Song Yawen, et al. Chinese clinical named entity recognition based on stroke ELMo and multi-task learning[J]. Chinese Journal of Computers, 2020, 43(10): 1943−1957(in Chinese)
(罗凌, 杨志豪, 宋雅文, 等. 基于笔画 ELMo 和多任务学习的中文电子病历命名实体识别研究[J]. 计算机学报, 2020, 43(10): 1943−1957)

[15] Sun Kai, Zhang Richong, Mao Yongyi, et al. Relation extraction with convolutional network over learnable syntax-transport graph[C]// Proc of the 34th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020: 8928−8935

[16] Guo Zhijiang, Zhang Yan, Lu Wei. Attention guided graph convolutional networks for relation extraction[C]//Proc of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2019: 241−251

[17] Christoph A, Marc H, Leonhard H. Improving relation extraction by pretrained language representations[J]. arXiv preprint, arXiv: 1906.03088, 2019

[18] Ji Guoliang, Liu Kang, He Shizhu, et al. Distant supervision for relation extraction with sentence-level attention and entity descriptions[C]// Proc of the 26th Int Joint Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2017: 3060−3066

[19] Li Bohan, Xiang Yuxuan, Feng Ding, et al. Short text classification model combining knowledge aware and dual attention[J]. Journal of Software, 2022, 33(10): 3565−3581(in Chinese)
(李博涵, 向宇轩, 封顶, 等. 融合知识感知与双重注意力的短文本分类模型[J]. 软件学报, 2022, 33(10): 3565−3581)

[20] Shang Yuming, Huang Heyan, Sun Xin, et al. A pattern-aware self-attention network for distant supervised relation extraction[J]. Information Sciences, 2022, 584: 269−279

[21] Phi V T, Santoso J, Shimbo M, et al. Ranking-based automatic seed selection and noise reduction for weakly supervised relation extraction[C]//Proc of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2018: 89−95

[22] Chen Zhuhui, Liu Xin, Zhang Mingjian, et al. Named entity recognition technology for brief case[J]. Computer Systems and Applications, 2022, 31(1): 47−54(in Chinese)
(陈柱辉, 刘新, 张明键, 等. 简要案情的命名实体识别技术[J]. 计算机系统应用, 2022, 31(1): 47−54)

[23] Guo Junjun, Liu Zhencheng, Yu Zhengtao, et al. Few shot and confusing charges prediction with the auxiliary sentences of case[J]. Journal of Software, 2021, 32(10): 3139−3150(in Chinese)
(郭军军, 刘真丞, 余正涛, 等. 融入案件辅助句的低频和易混淆罪名预测[J]. 软件学报, 2021, 32(10): 3139−3150)

[24] Li Jianlong, Wang Panqing, Han Qiyu. Military named entity recognition based on bidirectional LSTM[J]. Computer Engineering and Science, 2019, 41(4): 713−718(in Chinese)
(李健龙, 王盼卿, 韩琪羽. 基于双向 LSTM 的军事命名实体识别[J]. 计算机工程与科学, 2019, 41(4): 713−718)

[25] Che Wangxiang, Fengyunlong, Qin Libo, et al. N-LTP: An open-source neural language technology platform for Chinese[C]//Proc of the 2021 Conf on Empirical Methods in Natural Language Processing: System Demonstrations. Stroudsburg, PA: ACL, 2021: 42−49

[26] He Han, Jinho C. The stem cell hypothesis: Dilemma behind multi-task learning with transformer encoders [J]. arXiv preprint, arXiv: 2109.06939, 2021

Ma Jianwei, born in 1999. Master candidate. Student member of CCF. His main research interests include named entity recognition and autonomous driving testing.

马健伟, 1999 年生. 硕士研究生. CCF 学生会员. 主要研究方向为命名实体识别、自动驾驶测试.

Wang Tiexin, born in 1987. PhD, associate professor. Member of CCF. His main research interests include digital twins, model-based systems engineering, knowledge representation and modeling, and natural language processing.

王铁鑫, 1987 年生. 博士, 副教授. CCF 会员. 主要研究方向为数字孪生、基于模型的系统工程、知识表征与建模、自然语言处理.

Jiang Hong, born in 1997. Master. His main research interests include big data application and natural language processing.

江　宏, 1997 年生. 硕士. 主要研究方向为大数据应用、自然语言处理.

Chen Tao, born in 1997. Master candidate. Student member of CCF. His main research interests include knowledge representing and knowledge extraction.

陈　涛, 1997 年生. 硕士研究生. CCF 学生会员. 主要研究方向为知识表示、知识抽取.

Zhang Chao, born in 1999. Master candidate. Student member of CCF. His main research interests include knowledge extraction and entity alignment.

张　超, 1999 年生. 硕士研究生. CCF 学生会员. 主要研究方向为知识抽取、实体对齐.

Li Bohan, born in 1979. PhD, associate professor. Senior member of CCF. His main research interests include spatiotemporal databases, knowledge graphs, natural language processing, and recommendation systems.

李博涵, 1979 年生. 博士, 副教授. CCF 高级会员. 主要研究方向为时空数据库、知识图谱、自然语言处理、推荐系统.