·开发研究与工程应用· 文章编号：1000-3428(2021)05-0277-08 文献标志码：A 中图分类号：TP391

面向法律文本的三元组抽取模型

陈彦光[1]，王 雷[2]，孙媛媛[1]，王治政[1]，张书晨[1]
（1.大连理工大学 计算机科学与技术学院，辽宁 大连 116024；2.辽宁省人民检察院第三检察部，沈阳 110033）

摘 要：在中国裁判文书网上的开源刑事判决文档中蕴藏着重要的法律信息，但刑事判决书文档通常以自然语言的形式进行记录，而机器难以直接理解文档中的内容。为使由自然语言记录的非结构化刑事判决书文本转化为结构化三元组形式，构建一种面向法律文本的司法三元组抽取模型。将三元组抽取过程看作二阶段流水线结构，利用预训练的基于Transformer的双向编码器表示模型先进行命名实体识别，再将识别结果应用于关系抽取阶段得到相应的三元组表示，从而实现对非结构化刑事判决书文本的信息提取。实验结果表明，在经过人工标注的刑事判决书数据集上，该模型相比基于循环神经网络的组合模型的F1值提高了28.1个百分点，具有更优的三元组抽取性能。

关键词：命名实体识别；关系抽取；预训练语言模型；Transformer编码器；流水线结构

开放科学（资源服务）标志码（OSID）：

中文引用格式：陈彦光，王雷，孙媛媛，等.面向法律文本的三元组抽取模型[J].计算机工程，2021，47(5)：277-284.
英文引用格式：CHEN Yanguang，WANG Lei，SUN Yuanyuan，et al.Triple extraction model for legal texts[J].Computer Engineering，2021，47(5)：277-284.

Triple Extraction Model for Legal Texts

CHEN Yanguang[1]，WANG Lei[2]，SUN Yuanyuan[1]，WANG Zhizheng[1]，ZHANG Shuchen[1]
（1.School of Computer Science and Technology，Dalian University of Technology，Dalian，Liaoning 116024，China；
2.The Third Procuratorial Department，People's Procuratorate of Liaoning Province，Shenyang 110033，China）

【Abstract】The open-source documents of criminal sentences on China judgments online contain important legal information.However，the documents are usually transcribed in the form of natural language and difficult for machines to understand.This paper proposes a triplet extraction model for legal texts to transform the unstructured texts recorded by natural language into structured triplets.In the construction of the model，the triplet extraction process is considered as a two-stage pipeline structure.The pretrained Bidirectional Encoder Representations from Transformer（BERT）model is used for Named Entity Recognition（NER），and the recognition results are applied to relation extraction to obtain the corresponding triplet representation，completing the information extraction for the unstructured legal texts of criminal senteces.Experimental results on the manually labeled dataset of criminal sentences show that the F1 score of the proposed model is 28.1 percentage points higher than that of combinational model based on recurrent neural network，demonstrating its excellent triplet extraction performance.

【Key words】Named Entity Recognition（NER）；relation extraction；pretrained language model；Transformer encoder；pipeline structure
DOI：10. 19678/j. issn. 1000-3428. 0057677

0 概述

随着中国司法信息的不断公开化，最高人民法院生效裁判文书全部在中国裁判文书网上公布，除法律有特殊规定的以外。在中国裁判文书网上的大量开源刑事判决书文档中蕴藏着重要的法律信息，但对于这些通过自然语言形式记录的刑事判决书文档，机器无法直接进行深层含义的理解，而自动化信息提取技术能将非结构化的自然语言文本转化为结构化的三元组形式，挖掘出文本中具有一定潜藏价

基金项目：国家重点研发计划（2018YFC0830603）。
作者简介：陈彦光（1995—），女，硕士研究生，主研方向为自然语言处理；王 雷，博士研究生；孙媛媛（通信作者），教授、博士生导师；王治政，博士研究生；张书晨，硕士研究生。
收稿日期：2020-03-11 **修回日期：**2020-05-08 **E-mail：**syuan@dlut.edu.cn

值的内容，并通过命名实体识别（Named Entity Recognition，NER）和关系抽取将非结构化的刑事判决书文本处理为结构化的三元组。刑事判决书中的案件事实描述文本s被表示为多个$<e_1,r,e_2>$三元组的形式，其中，e_1和e_2分别表示三元组的头实体和尾实体，r表示两个实体之间的关系类型[1]。

知识图谱以结构化的形式表示知识，通过对非结构化文本中难以理解的信息进行挖掘与分析，提高非结构化文本的查询性能及可解释性，通常作为搜索引擎、问答系统等实际应用中的底层支撑技术。目前，知识图谱的相关研究受到学术界和工业界的广泛关注，研究人员提出了许多知识图谱构建方法，但构建出的知识图谱多数面向通用领域，其中三元组抽取是知识图谱构建过程中的关键步骤。本文提出一个面向法律文本的三元组抽取模型，对非结构化的案件事实描述文本进行结构化表示。将三元组的抽取过程看作二阶段流水线结构，先进行命名实体识别，再将识别结果应用于关系抽取阶段得到相应的三元组表示。

1 相关工作

非结构化文本中的三元组抽取可分为命名实体识别和关系抽取两个阶段。命名实体识别用于提取文本中具有特定含义的实体短语，如人名、地名以及专有名词等。关系抽取对于文本中给定的实体对，通过上下文语义理解识别出实体之间的关系类型。

早期的命名实体识别工作主要包括基于规则和词典的命名实体识别方法与基于统计的命名实体识别方法。基于规则和词典的命名实体识别方法需要语言学专家和领域学者归纳规则模板和领域词典，通过匹配算法完成命名实体识别。基于统计的命名实体识别方法学习标注语料的训练过程并分析文本的语言特征，主要包括基于支持向量机（Support Vector Machine，SVM）的命名实体识别方法[2]、基于隐马尔科夫模型（Hidden Markov Model，HMM）的命名实体识别方法[3]以及基于条件随机场（Conditional Random Field，CRF）的命名实体识别方法[4]等。但这些早期工作对特征选择的要求较高，较大程度地依赖词典以及特征工程。随着深度学习技术的不断发展，使用神经网络进行命名实体识别的方法逐渐成为当前中文命名实体识别的主要研究方向[5-7]。由于基于神经网络的命名实体识别模型可以自动化地学习文本特征，从而减少对手工特征的依赖。目前主流的用于命名实体识别的神经网络模型为双向长短期记忆网络结合条件随机场（Bi-directional Long Short-Term Memory+Condition Random Field，BiLSTM+CRF）。近些年，在司法领域，许多学者对基于法律文书的命名实体识别方法开展了大量的相关研究工作[8-10]。

关系抽取工作一般可分为基于机器学习的关系抽取方法和基于深度学习的关系抽取方法。基于机器学习的关系抽取方法将关系抽取转化为分类任务，对两个实体之间的关系类型进行预测，该类方法先整合词性特征、实体类型、句法依存关系以及WordNet语义信息等语言学特征，再通过最大熵模型[11]、支持向量机模型[12-14]等基于统计模型的分类器对关系进行分类。随着深度学习技术的发展，研究人员提出了许多基于深度学习的关系抽取方法，通过对输入文本及实体位置信息等进行向量化表示，利用神经网络模型自动提取文本特征，预测实体对之间的关系类型，主要包括基于卷积神经网络的关系抽取方法[15-17]、基于循环神经网络的关系抽取方法[18-19]以及两者相结合的关系抽取方法[20]。随着自注意力机制研究的深入[21-22]，一些学者将Transformer架构[23]应用于关系抽取任务，利用基于Transformer的双向编码器表示（Bidirectional Encoder Representations from Transformer，BERT）[24]进行关系抽取[25]并取得了较好的效果。

近年来，预训练语言模型研究发展迅速，基于上下文信息捕捉单词的语义知识，通过在大规模语料上进行预训练，从而实现文本上下文相关特征的表示。在预训练语言模型研究中，一般通过特征集成和模型微调方式实现对预训练模型参数的迁移。特征集成方式将语言模型学习到的文本表示当作下游任务的输入特征进行应用，例如文献[26]提出的ELMo可在变化的语言语境下对词进行复杂特征建模。模型微调方式以整个预训练语言模型为基础，通过加入任务输出部分并对整个模型参数进行微调实现预训练模型的应用，例如：文献[24]提出的BERT模型通过Transformer编码器堆叠而成，实现对文本的双向特征表示，在11项自然语言处理任务中取得了最佳成绩；文献[27]提出的自回归预训练模型XLNet，在多项自然语言处理任务中取得了明显的性能提升。

2 司法三元组抽取模型

对于案件事实描述文本s，本文提出的司法三元组抽取模型能够将与其具有等价语义的三元组以$<e_1,r,e_2>$的形式进行预测。司法三元组抽取模型以BERT预训练语言模型为基础，搭建一个二阶段的流水线结构，主要包括实体识别模块和关系抽取模块两部分。实体识别模块用于对案件事实描述中具有特定含义的实体短语进行定位和分类，关系抽取模块旨在预测非结构化文本中每一对实体之间的关系类型。在关系抽取模块中，为强调给定实体对的位置和内容，借鉴文献[1]工作，在文本表示中加入实体信息的整合过程。针对流水线结构中的冗余实体

对信息所造成的影响，通过加入实体对筛选过程以减少无用信息的累积，并在关系抽取模块训练时，在训练集中适当增加负样本，以增强模型鲁棒性，本文提出两种策略来完善关系抽取模块的训练过程。此外，为进行有监督的模型训练以及验证模型在刑事判决书文本上的三元组抽取性能，本文以刑事判决书中的案件事实描述部分为数据基础，通过自然语言处理工具进行机器粗标与人工标注相结合的方式，构造一个面向涉毒类刑事案件的实体关系提取数据集。

司法三元组抽取模型的整体架构如图1所示，其中，$\boldsymbol{w}_i$表示输入文本的向量化表示，$\boldsymbol{h}_i$表示经过BERT模型编码得到的上下文语义向量，N表示输入序列长度，Trm表示BERT模型中的Transformer编码器单元。司法三元组抽取模型针对涉毒类案件刑事判决书文本进行研究，通过实体识别模块和关系抽取模块，实现对涉毒类刑事案件的结构化三元组抽取。

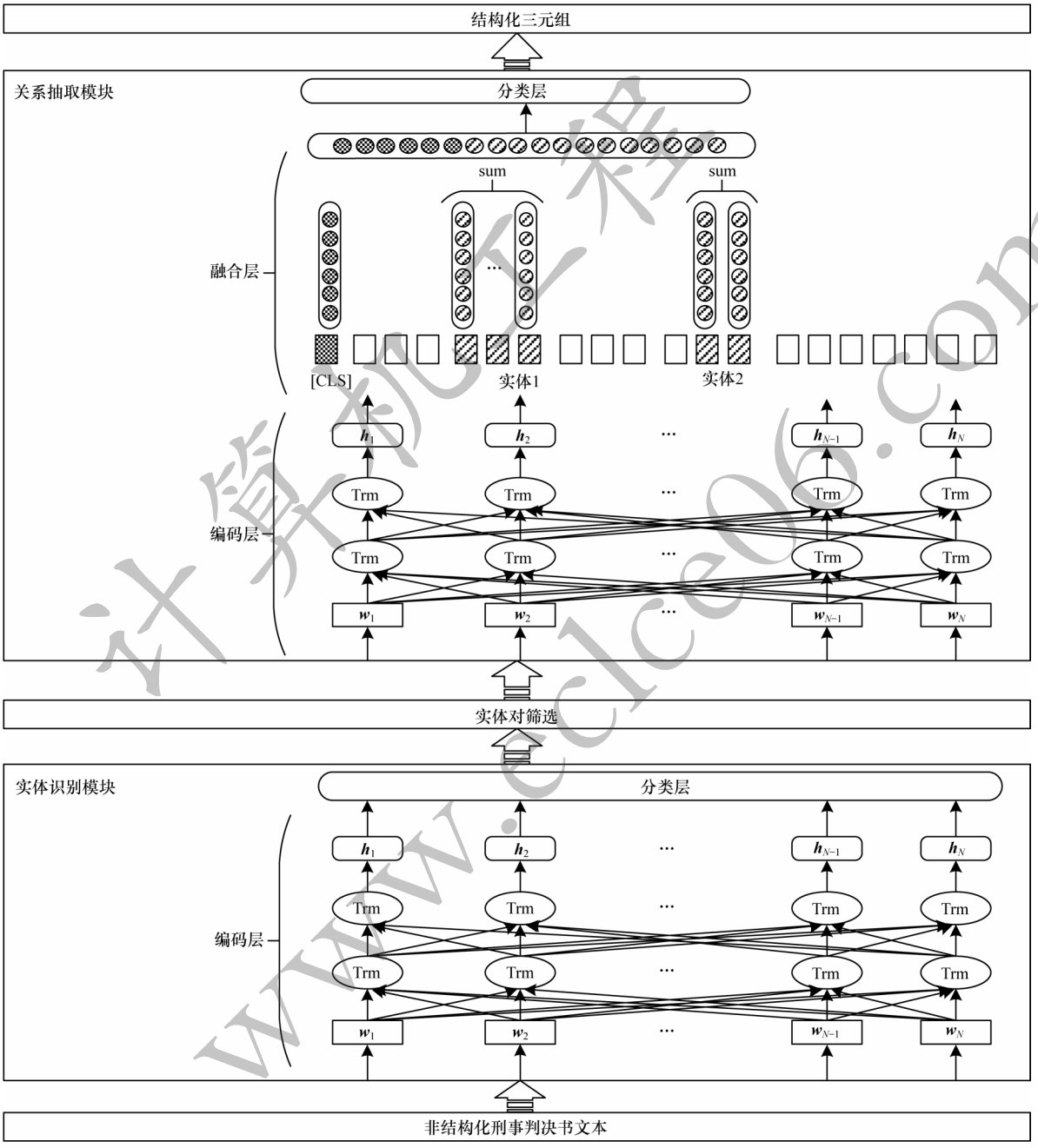

图1　司法三元组抽取模型的整体架构

Fig.1　The overall architecture of legal triplet extraction model

2.1　预训练语言模型

BERT模型由多层双向Transformer编码器堆叠而成，通过在大规模语料上进行无监督预训练获得文本的特征表示。BERT模型的输入部分可对单句及句子对进行表示，对于给定的字符，输入向量包括词嵌入信息、位置信息和分句信息3类信息表示，并且在BERT原始模型中具有‘[CLS]’、‘[SEP]’和‘[MASK]’3种特殊字符：‘[CLS]’符号置于每个输

入序列的首位，其对应的输出向量为该序列的向量表示，可直接用于分类任务；‘[SEP]’符号用于句子对作为输入时分隔序列中的两个句子，针对单句子作为输入的情况，将‘[SEP]’符号置于句子尾；‘[MASK]’符号应用在预训练阶段的覆盖语言模型中。

BERT模型通过覆盖语言模型（Masked Language Model，MLM）任务以及下一句预测（Next Sentence Prediction，NSP）任务完成对模型参数的预训练。在覆盖语言模型任务中，输入序列的部分字符通过‘[MASK]’符号被随机覆盖，该任务的目标是通过上下文文本预测被覆盖的字符，得到字符的双向上下文表示。下一句预测任务针对句子对输入，预测两句是否为文本中的连续语句，以此捕捉句子对之间的关系。在经过大规模语料预训练后，针对特定任务，还需使用任务相关的数据集对BERT模型进行微调，从而得到适用于具体任务的模型参数。

2.2 实体识别模块

实体识别模块是司法三元组抽取模型的主要模块之一，将刑事判决书案件事实描述部分中的命名实体全部标记处理，具体包括人名、地名、时间、毒品类型和毒品重量5类实体。针对输入文本中的每个字符，实体识别模块将预测该字符是否属于实体的一部分并给出实体类型，由此将实体识别过程转化为字符级的分类任务，预测指定字符的实体位置和实体类型，通过在以BERT模型为基础的编码层上添加一个多分类器进行实现。

按照BERT的输入格式，将案件事实描述文本处理为向量，作为实体识别模块的输入，该向量包含词嵌入、位置信息以及分句信息三部分。此外，在句首和句尾分别插入‘[CLS]’符号和‘[SEP]’符号。在模型微调过程中，使用编码层最后一层的隐层向量作为序列的特征表示，并通过多标签分类器对序列中的每个字符进行预测。标签序列x的分布可表示为：

$$p(x|s)=\mathrm{Softmax}(\boldsymbol{H}_{\mathrm{ER}}) \tag{1}$$

其中，$\boldsymbol{H}_{\mathrm{ER}}$为编码层最后一层的隐层向量表示。

在实体识别模块中，在BERT模型的基础上添加字符级多分类器形成实体识别模块的模型结构。为使实体识别模块可以利用BERT模型预训练阶段学习的文本特征，并学习下游的实体识别任务，还需对整个模型进行微调。首先通过载入预训练后的BERT模型权重对实体识别模块进行初始化；然后利用面向涉毒类刑事案件的实体识别数据集对实体识别模块进行有监督训练，完成相应参数的微调。由此得到的实体识别模块既包含预训练阶段的通用文本特征知识，又对法律实体识别任务进行了学习。对于训练样本$\{(s_i,x_i)\}|_{i=1}^{N}$，其中，s_i和x_i分别代表实体识别模块训练集中第i条样本的真实标签和预测标签，N为训练集中的样本数，使用交叉熵作为损失函数对实体识别模块的参数θ_{ER}进行学习：

$$L_{\mathrm{ER}}=\sum_{i=1}^{N}\ln p\left(x_i|s_i,\theta_{\mathrm{ER}}\right) \tag{2}$$

2.3 实体对筛选过程

实体对筛选过程的作用是减轻流水线结构中的冗余实体信息所造成的影响。该过程对实体识别模块的结果进行整合，选择可能具有关系的实体对并过滤不可能形成三元组的实体。在对司法三元组进行抽取的流水线中，实体对筛选过程置于实体识别模块后及关系抽取模块前。首先对文本中通过实体识别模块提取出的实体进行两两组合，形成实体对集合；然后通过关系类型分析，得出可能形成三元组的实体类型组合规则；最后依照这些规则对实体对集合进行筛选，得到可能存在关系的实体对，输入关系抽取模块中预测其关系。

2.4 关系抽取模块

关系抽取模块旨在通过上下文囊括的语义信息判断文本中给定的实体对存在的关系类型。为实现关系抽取模块的功能，给定一个描述文本s以及两个目标实体e_1和e_2，在文本中插入实体定位字符以供模型获取实体信息。实体定位字符分别为‘[E11]’、‘[E12]’、‘[E21]’和‘[E22]’4个字符。针对三元组的头实体e_1，将字符‘[E11]’和‘[E12]’分别置于e_1的首部和尾部，确定e_1的具体位置。针对三元组的尾实体e_2，按照相同的方式，在e_2首尾插入‘[E21]’和‘[E22]’字符进行定位。

关系抽取模块由编码层、融合层和分类层三部分组成，编码层用于提取文本特征及实体特征，融合层可将实体对的特征信息与上下文特征进行整合，分类层用于对文本中的每个实体对存在的关系类型进行预测。

2.4.1 编码层

编码层以BERT模型为基础对文本进行向量表示，分别对输入序列和实体对进行特征提取。将学习到的‘[CLS]’符号所对应的特征向量作为整个序列s的全局特征，通过$\boldsymbol{H}_{\mathrm{s}}$进行表示。将BERT模型最后一层的隐层向量看作是序列中每个字符的编码向量，以$\boldsymbol{h}$进行表示。为得到序列中的实体特征，对与头实体e_1和尾实体e_2相关的字符进行向量表示：

$$\boldsymbol{E}_1=\tanh\left(\sum_{i=m_1}^{m_2}\boldsymbol{h}_i\right) \tag{3}$$

$$\boldsymbol{E}_2=\tanh\left(\sum_{i=n_1}^{n_2}\boldsymbol{h}_i\right) \tag{4}$$

其中，$\boldsymbol{E}_1$和$\boldsymbol{E}_2$分别为实体e_1和e_2所对应的特征向量，m_1和m_2、n_1和n_2分别对应两个实体e_1、e_2在序列s中的

开始和结束位置。

2.4.2　融合层

融合层用于对编码层输出的序列特征$\boldsymbol{H}_s$和实体特征$\boldsymbol{E}_1$、$\boldsymbol{E}_2$进行整合，从而在序列特征中加入相应的实体对信息。为能够更好地学习各特征向量之间的关系，添加可训练的参数矩阵$\boldsymbol{W}_s$和$\boldsymbol{W}_e$，以对序列特征和实体特征所占的权重进行动态调整。在经过特征向量融合后，序列特征$\boldsymbol{H}_s$和实体特征$\boldsymbol{E}_1$、$\boldsymbol{E}_2$将整合为一个新的序列表示向量$\boldsymbol{S}$，其中包含序列s的全局文本信息以及其中的实体信息，具体表示为：

$$\boldsymbol{S}=\mathrm{concat}(\boldsymbol{W}_s\boldsymbol{H}_s,\boldsymbol{W}_e\boldsymbol{E}_1,\boldsymbol{W}_e\boldsymbol{E}_2) \tag{5}$$

2.4.3　分类层

分类层基于最终的序列表示$\boldsymbol{S}$对关系类型进行分类，通过Softmax分类器对文本中给定实体对存在的关系类型分布y进行预测：

$$p(y|s)=\mathrm{Softmax}(\boldsymbol{S}) \tag{6}$$

在关系抽取模块中，以BERT模型为基础，通过加入特征融合层和关系分类层形成关系抽取模块的模型结构。首先载入经过预训练的BERT模型权重作为关系抽取模型的初始权重，使得关系抽取模型具备预训练阶段学习的知识；然后通过在面向涉毒类刑事案件的关系抽取数据集上进行监督训练，并对模型参数进行微调，实现可用于法律文书关系抽取任务的模型。在训练过程中，通过交叉熵损失函数对关系抽取模块参数θ_{RE}进行学习：

$$L_{\mathrm{RE}}=\sum_{i=1}^{N}\ln p(y_i|s_i,\theta_{\mathrm{RE}}) \tag{7}$$

3　实验与结果分析

3.1　数据集构建

为实现中国司法领域的信息抽取，以涉毒类刑事判决书文本为基础，将其中的案件事实描述部分使用规则提取，在此基础上通过自然语言处理工具进行机器粗标与人工标注相结合的模式，标注出涉及到的法律实体及其之间的关系类型。选取涉毒类刑事案件中最具代表的贩卖毒品、非法持有毒品和容留他人吸毒3类案件作为研究主体，将1 750份刑事判决书中的案件事实描述文本作为原始语料，在此基础上进行标注形成数据集。

针对命名实体识别任务，使用BIO标注策略区分实体边界并预设人名、地名、时间、毒品类型和毒品重量5类实体。司法领域实体识别数据集中共包括19 321个实体。针对关系抽取任务，参考《中华人民共和国刑法》并结合3类涉毒类案件的判决依据，预定义持有(possess)、贩卖(给人)(sell_drug_to)、贩卖(毒品)(traffic_in)和非法容留(provide_shelter_for)4种关系类型，这4种关系涵盖了3类涉毒类案件中的犯罪行为。

将1 750条经过实体关系标注的案件事实描述文本以4∶1的比例进行随机划分，分别作为司法领域实体关系提取的训练集和测试集。训练集和测试集中实体与关系的统计情况分别如表1和表2所示。

表1　数据集中实体类型的统计情况

Table 1　Statistics of entity types in the dataset

实体类型	训练集数量	测试集数量	合计
人名	6 471	1 633	8 104
地名	1 779	462	2 241
时间	2 150	547	2 697
毒品类型	3 448	777	4 225
毒品重量	1 661	393	2 054
合计	15 509	3 812	19 321

表2　数据集中关系类型的统计情况

Table 2　Statistics of relation types in the dataset

关系类型	训练集数量	测试集数量	合计
possess	478	111	589
sell_drug_to	980	251	1 231
traffic_in	907	223	1 130
provide_shelter_for	842	201	1 043
合计	3 207	786	3 993

3.2　数据预处理与参数设置

由于本文三元组抽取模型采用流水线结构，因此会产生大量不存在关系类型的实体对，这些冗余的实体对将会对关系抽取模块的识别性能造成影响。为使关系抽取模块能更好地学习这种无关系类型的实体对特征，在训练过程中将不存在关系类型的实体组合作为负样本，以一定的比例添加到训练集中。

此外，本文还考虑关系方向性，即三元组$<e_1,r_a,e_2>$和$<e_2,r_b,e_1>$，这两个三元组的实体集合是一致的，但头尾实体位置互换，因此其存在的关系类型r_a和r_b是不同的，对于这一类头尾实体位置互换的三元组所存在的两个关系r_a和r_b，本文称其互为反向关系。关系的方向性对关系抽取模块的训练也有一定的影响，尤其在关系类型贩卖(给人)和非法容留中较为明显，由于在这两种关系中，头实体和尾实体对应的实体类型都为人名且表达形式相近，因此会对关系类型的预测产生影响。为使关系抽取模块能更好地学习关系的方向性，在训练过程中，将训练集中正样本所对应的反向关系作为负样本添加到训练集中。

在实验设置上，命名实体识别模块使用谷歌开源的中文BERT(BERT-Base，Chinese)模型，在此基础上进行微调完成对法律实体的识别，关系抽取模块分别使用中文BERT(BERT-Base，Chinese)模型和RoBERTa模型进行实验，其他参数设置如表3所示。

表3 实体识别模块与关系抽取模块的参数设置

Table 3 Parameters setting of entity recognition module and relation extraction module

参数	实体识别模块	关系抽取模块
批尺寸	32	8
学习率	5e-5	2e-5
训练轮数	3	5
序列长度	512	400

3.3 结果分析

在实验中,三元组抽取模型性能由精确率(P)、召回率(R)以及F1值(F)进行评估。评价指标的计算方式如下:

$$P=\frac{n_{\text{correct\_num}}}{n_{\text{predict\_num}}} \tag{8}$$

$$R=\frac{n_{\text{correct\_num}}}{n_{\text{true\_num}}} \tag{9}$$

$$F=\frac{2\times P\times R}{P+R} \tag{10}$$

其中,$n_{\text{correct_num}}$表示司法三元组抽取模型对所有实例抽取正确的三元组个数,$n_{\text{predict_num}}$表示司法三元组抽取模型预测出的三元组总数,$n_{\text{true_num}}$表示实际的三元组总数。其中,抽取出的三元组只有在两个实体e_1和e_2以及关系r都预测正确的情况下才被判定为正确的三元组。

实验采用3组不同的神经网络模型组合作为基线模型:组合模型1中实体识别使用双向长短期记忆网络结合条件随机场的模型(BiLSTM+CRF),关系抽取应用双向门循环单元结合注意力机制的模型(BiGRU+ATT);组合模型2中实体识别使用本文模型,关系抽取使用BiGRU+ATT;组合模型3中实体识别使用BiLSTM+CRF,关系抽取使用本文模型。不同的模型组合对三元组的抽取效果如表4所示,可以看出,本文提出的司法三元组抽取模型优于其他的组合模型,相比基于循环神经网络的组合模型1的F1值提高了28.1个百分点。由组合模型3的F1值高于组合模型2的F1值这一结果可以看出,本文关系抽取模块相比实体识别模块更有助于抽取性能的提升。

表4 组合模型与本文模型的三元组抽取结果对比

Table 4 Comparison of triplet extraction results of the combination models and the proposed model %

模型	精确率	召回率	F1值
组合模型1	42.3	59.3	49.4
组合模型2	45.3	62.7	52.6
组合模型3	69.6	81.0	74.9
本文模型	70.7	85.6	77.5

由于流水线结构中会产生大量不存在关系类型的实体对,因此为使关系抽取模块更加全面地学习这些无关系类型的实体对特征,在训练阶段通过添加负例样本完善关系抽取模型的训练过程。

3.3.1 正负样本比例对三元组抽取的影响

在实验中正负样本的比例对三元组的抽取效果产生了一定的影响。通过采用相同的随机种子,随机筛选不同比例的负样本添加到关系抽取模块的训练集中,确定用于训练关系抽取模块的最佳正负样本比例,分别选取正负样本比例为无负样本、1∶2、1∶3、1∶5和1∶7进行对比实验,结果如表5所示。

表5 基于不同正负样本比例的三元组抽取结果对比

Table 5 Comparison of triplet extraction results based on different positive/negative instance ratios %

正负样本比例	精确率	召回率	F1值
无负样本	22.7	91.3	36.3
1∶2	51.9	90.7	66.1
1∶3	67.5	87.4	76.2
1∶5	70.7	85.6	77.5
1∶7	71.8	85.0	77.8

随着关系抽取任务的训练集中负例样本占比逐渐增加,三元组抽取模型的整体抽取性能不断提升,F1值由无负样本的36.3%提升至正负样本比例为1∶7的77.8%,提高了41.5个百分点。这也证明了添加适当比例的负样本对关系抽取模块的训练过程具有积极作用,由实验结果中精确率的大幅提升也可看出,关系抽取模块通过负样本学习可更全面地学习不存在关系类型的实体对所具有的特征,能够更好地分辨出无关系类型的实体对。

3.3.2 反向关系对三元组抽取的增益效果

为验证反向关系对三元组抽取结果的影响,通过将正样本的反向关系作为负样本添加到训练集中,使关系抽取模块对关系方向性进行更好的学习,并选择具有不同正负样本比例的训练集分别进行实验,结果如表6所示,其中,"√"表示添加反向关系,"×"表示未添加反向关系。

表6 添加反向关系的三元组抽取结果对比

Table 6 Comparison of triplet extraction results of adding inverse relation %

正负样本比例	添加反向关系	精确率	召回率	F1值
无负样本	×	13.0	89.6	22.8
	√	22.7	91.3	36.3
1∶2	×	34.6	84.4	49.1
	√	51.9	90.7	66.1
1∶5	×	48.1	83.7	61.1
	√	70.7	85.6	77.5

由实验结果可以看出，关系方向性对关系抽取模块的训练过程十分重要，通过将正样本的反向关系添加到训练集中，使得本文模型对三元组抽取的精确率和召回率都有所提升，在无负样本、正负样本比例为1∶2和1∶5的条件下，F1值分别提高了13.5、17.0和16.4个百分点。由此说明将正样本的反向关系作为负样本进行模型训练这一策略能有效提升关系抽取模块的预测能力，有助于模型更好地区分具有相似头尾实体的实体对特征。

3.3.3　不同预训练语言模型对三元组抽取的影响

本文对关系抽取模块所使用的预训练语言模型进行对比实验，结果如表7所示，可以看出使用基于RoBERTa模型的关系抽取模块可更好地进行关系预测，在三元组抽取结果上达到79.6%的F1值。

表7　在1∶5正负样本比例下不同预训练语言模型的三元组抽取结果对比

Table 7　Comparison of triplet extraction results of different pretrained language models with the positive/negative instance ratio of 1∶5　%

预训练语言模型	精确率	召回率	F1值
BERT-Base Chinese	70.7	85.6	77.5
RoBERTa Chinese	72.6	88.1	79.6

4　结束语

本文建立一种将非结构化刑事判决书文本转化为结构化三元组形式的司法三元组抽取模型。该模型将预训练的BERT模型作为主体，在此基础上分别对实体识别模块和关系抽取模块进行微调，并搭建三元组抽取的流水线结构，实现对非结构化文本的信息提取。实验结果表明，该模型相比基于循环神经网络的组合模型的F1值提高了28.1个百分点，并通过加入两项针对关系抽取模块的训练策略能提升三元组抽取性能。下一步将继续优化本文模型的三元组抽取效果，并以此为基础构建司法知识图谱进行表示学习及知识推理等工作。

参考文献

[1] WU Shanchan, HE Yifan. Enriching pre-trained language model with entity information for relation classification[C]// Proceedings of the 28th ACM International Conference on Information and Knowledge Management. New York, USA: ACM Press, 2019: 2361-2364.

[2] ISOZAKI H, KAZAWA H. Efficient support vector classifiers for named entity recognition[C]//Proceedings of the 19th International Conference on Computational Linguistics. Philadelphia, USA: Association for Computational Linguistics, 2002: 1-7.

[3] BIKEL D M, MILLER S, SCHWARTZ R, et al. Nymble: a high-performance learning name-finder[C]//Proceedings of the 5th Conference on Applied Natural Language Processing. Philadelphia, USA: Association for Computational Linguistics, 1997: 194-201.

[4] LAFFERTY J D, MCCALLUM A, PEREIRA F C. Conditional random fields: probabilistic models for segmenting and labeling sequence data[C]//Proceedings of the 18th International Conference on Machine Learning. San Mateo, USA: Morgan Kaufmann Publishers Inc., 2001: 282-289.

[5] LAMPLE G, BALLESTEROS M, SUBRAMANIAN S, et al. Neural architectures for named entity recognition[C]// Proceedings of 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Philadelphia, USA: Association for Computational Linguistics, 2016: 260-270.

[6] ZHU Yuying, WANG Guoxin. CAN-NER: convolutional attention network for Chinese named entity recognition[C]// Proceedings of 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Philadelphia, USA: Association for Computational Linguistics, 2019: 3384-3393.

[7] ZHANG Yingcheng, YANG Yang, JIANG Rui, et al. Commercial intelligence entity recognition model based on BiLSTM-CRF[J]. Computer Engineering, 2019, 45(5): 308-314. (in Chinese)
张应成，杨洋，蒋瑞，等. 基于BiLSTM-CRF的商情实体识别模型[J]. 计算机工程，2019，45(5)：308-314.

[8] DOZIER C, KONDADADI R, LIGHT M, et al. Named entity recognition and resolution in legal text[M]. Berlin, Germany: Springer, 2010.

[9] QUARESMA P, GONCALVES T. Using linguistic information and machine learning techniques to identify entities from juridical documents[M]. Berlin, Germany: Springer, 2010.

[10] HAQ M I U, LI Q, HASSAN S. Text mining techniques to capture facts for cloud computing adoption and big data processing[J]. IEEE Access, 2019, 7: 162254-162267.

[11] KAMBHATLA N. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations[C]//Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. Philadelphia, USA: Association for Computational Linguistics, 2004: 22-29.

[12] ZHOU Guodong, SU Jie, ZHANG Jie, et al. Exploring various knowledge in relation extraction[C]//Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Philadelphia, USA: Association for Computational Linguistics, 2005: 427-434.

[13] CULOTTA A, SORENSEN J. Dependency tree kernels for relation extraction[C]//Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. Philadelphia, USA: Association for Computational Linguistics, 2004: 423-428.

[14] ZHOU Guodong, ZHANG Min, JI Donghong, et al. Tree kernel-based relation extraction with context-sensitive structured parse tree information[C]//Proceedings of 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Prague, Czech Republic: [s. n.], 2007: 728-736.
[15] ZENG Daojian, LIU Kang, LAI Siwei, et al. Relation classification via convolutional deep neural network[C]// Proceedings of the 25th International Conference on Computational Linguistics. Philadelphia, USA: Association for Computational Linguistics, 2014: 2335-2344.
[16] NGUYEN T H, GRISHMAN R. Relation extraction: perspective from convolutional neural networks [C]// Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing. Philadelphia, USA: Association for Computational Linguistics, 2015: 39-48.
[17] XIANG Bing, ZHOU Bowen. Classifying relations by ranking with convolutional neural networks [C]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Philadelphia, USA: Association for Computational Linguistics, 2015: 626-634.
[18] ZHANG Runyan, MENG Fanrong, ZHOU Yong, et al. Relation classification via recurrent neural network with attention and tensor layers [J]. Big Data Mining and Analytics, 2018, 1(3): 234-244.
[19] ZHOU Peng, SHI Wei, TIAN Jun, et al. Attention-based bidirectional long short-term memory networks for relation classification[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Philadelphia, USA: Association for Computational Linguistics, 2016: 207-212.
[20] SUN Ziyang, GU Junzhong, YANG Jing. Chinese entity relation extraction method based on deep learning [J]. Computer Engineering, 2018, 44(9): 164-170. (in Chinese)
孙紫阳,顾君忠,杨静. 基于深度学习的中文实体关系抽取方法[J]. 计算机工程,2018,44(9):164-170.
[21] VERGA P, STRUBELL E, MCCALLUM A. Simultaneously self-attending to all mentions for full-abstract biological relation extraction[C]//Proceedings of 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Philadelphia, USA: Association for Computational Linguistics, 2018: 872-884.
[22] WANG H, TAN M, YU M, et al. Extracting multiple-relations in one-pass with pre-trained transformers[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Philadelphia, USA: Association for Computational Linguistics, 2019: 1371-1377.
[23] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Neural Information Processing Systems Foundation, Inc. , 2017: 6000-6010.
[24] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional Transformers for language understanding[C]//Proceedings of 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Philadelphia, USA: Association for Computational Linguistics, 2019: 4171-4186.
[25] ALT C, HUBNER M, HENNIG L. Fine-tuning pre-trained transformer language models to distantly supervised relation extraction[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Philadelphia, USA: Association for Computational Linguistics, 2019: 1388-1398.
[26] PETERS M E, NEUMANN M, IYYER M, et al. Deep contextualized word representations[EB/OL]. [2020-02-01]. https://arxiv. org/abs/1802. 05365.
[27] YANG Zhilin, DAI Zihang, YANG Yiming, et al. XLNet: generalized autoregressive pretraining for language understanding [C]//Proceedings of Advances in Neural Information Processing Systems. Vancouver, Canada: Neural Information Processing Systems Foundation, Inc. , 2019: 5754-5764.

编辑 陆燕菲

(上接第276页)

[14] RICHMAN J S, MOORMAN J R. Physiological time-series analysis using approximate entropy and sample entropy [J]. American Journal of Physiology Heart & Circulatory Physiology, 2000, 278(6): 2039-2049.
[15] LOU Zhi, YAO Bo, YANG Jihai. Detection of gait activity segment in children with cerebral palsy based on surface electromyography [J]. Journal of Biomedical Engineering, 2017, 34(2): 342-349. (in Chinese)
娄智,姚博,杨基海. 基于表面肌电信号的小儿脑瘫步态活动段检测研究[J]. 生物医学工程学杂志,2017,34(2):342-349.
[16] MUKHOPADHYAY S, RAY G C. A new interpretation of nonlinear energy operator and its efficacy in spike detection [J]. IEEE Transactions on Biomedical Engineering, 1998, 45(2): 180-187.
[17] FOX E B, HUGHES M C, SUDDERTH E B, et al. Joint modeling of multiple time series via the beta process with application to motion capture segmentation [EB/OL]. [2020-02-10]. https://arxiv. org/abs/1308. 4747.
[18] JAIN S, NEALY R M. Splitting and merging components of a nonconjugate Dirichlet process mixture model[J]. Bayesian Analysis, 2007, 2(3): 445-472.
[19] TU Zhuowen, ZHU Songchuan. Image segmentation by data-driven Markov chain Monte Carlo[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence2002, 24(3): 657-673.
[20] HUANG Zhexue. Extensions to the k-means algorithm for clustering large data sets with categorical values[J]. Data Mining and Knowledge Discovery, 1998, 2(3): 283-304.

编辑 索书志

[14] ZHOU Guodong, ZHANG Min, JI Donghong, et al. Tree kernel-based relation extraction with context-sensitive structured parse tree information[C]//Proceedings of 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Prague, Czech Republic: [s. n.], 2007: 728-736.

[15] ZENG Daojian, LIU Kang, LAI Siwei, et al. Relation classification via convolutional deep neural network[C]// Proceedings of the 25th International Conference on Computational Linguistics. Philadelphia, USA: Association for Computational Linguistics, 2014: 2335-2344.

[16] NGUYEN T H, GRISHMAN R. Relation extraction: perspective from convolutional neural networks [C]// Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing. Philadelphia, USA: Association for Computational Linguistics, 2015: 39-48.

[17] XIANG Bing, ZHOU Bowen. Classifying relations by ranking with convolutional neural networks [C]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Philadelphia, USA: Association for Computational Linguistics, 2015: 626-634.

[18] ZHANG Runyan, MENG Fanrong, ZHOU Yong, et al. Relation classification via recurrent neural network with attention and tensor layers [J]. Big Data Mining and Analytics, 2018, 1(3): 234-244.

[19] ZHOU Peng, SHI Wei, TIAN Jun, et al. Attention-based bidirectional long short-term memory networks for relation classification[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Philadelphia, USA: Association for Computational Linguistics, 2016: 207-212.

[20] SUN Ziyang, GU Junzhong, YANG Jing. Chinese entity relation extraction method based on deep learning[J]. Computer Engineering, 2018, 44(9): 164-170. (in Chinese)
孙紫阳, 顾君忠, 杨静. 基于深度学习的中文实体关系抽取方法[J]. 计算机工程, 2018, 44(9): 164-170.

[21] VERGA P, STRUBELL E, MCCALLUM A. Simultaneously self-attending to all mentions for full-abstract biological relation extraction[C]//Proceedings of 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Philadelphia, USA: Association for Computational Linguistics, 2018: 872-884.

[22] WANG H, TAN M, YU M, et al. Extracting multiple-relations in one-pass with pre-trained transformers[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Philadelphia, USA: Association for Computational Linguistics, 2019: 1371-1377.

[23] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Neural Information Processing Systems Foundation, Inc., 2017: 6000-6010.

[24] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional Transformers for language understanding[C]//Proceedings of 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Philadelphia, USA: Association for Computational Linguistics, 2019: 4171-4186.

[25] ALT C, HUBNER M, HENNIG L. Fine-tuning pre-trained transformer language models to distantly supervised relation extraction[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Philadelphia, USA: Association for Computational Linguistics, 2019: 1388-1398.

[26] PETERS M E, NEUMANN M, IYYER M, et al. Deep contextualized word representations[EB/OL]. [2020-02-01]. https://arxiv. org/abs/1802. 05365.

[27] YANG Zhilin, DAI Zihang, YANG Yiming, et al. XLNet: generalized autoregressive pretraining for language understanding[C]//Proceedings of Advances in Neural Information Processing Systems. Vancouver, Canada: Neural Information Processing Systems Foundation, Inc., 2019: 5754-5764.

编辑 陆燕菲