计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(12):211−219 [doi: 10.15888/j.cnki.csa.008860]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

融合字位置特征的铁路事故命名实体识别①

陈业明, 戴　齐, 刘　捷

(西南交通大学 计算机与人工智能学院, 成都 611756)
通信作者: 戴　齐, E-mail: qdai@swjtu.edu.cn

摘　要: 铁路事故的相关信息以事故概况文本的形式存在, 对于铁路安全工作有重要意义. 但由于缺乏有效的信息抽取手段, 导致分散在文本中的铁路事故知识没有得到充分的利用. 命名实体识别是信息抽取的重要子任务, 目前关于事故领域的命名实体识别问题研究较少. 针对铁路事故命名实体识别问题, 提出一种融合字位置特征的命名实体识别模型, 该模型通过全连接神经网络获取字的位置特征, 并与语义层面的字向量合并作为字的最终向量表示输入 BiLSTM-CRF 模型获取最优标签序列. 实验结果表明, 模型在铁路事故文本命名实体识别问题上的准确率、召回率和 *F*1 值分别为 93.29%、94.77% 和 94.02%, 相比于传统模型, 取得了更好的效果, 为铁路事故知识图谱的构建奠定基础.

关键词: 命名实体识别; 铁路事故; 字位置特征; 双向长短期记忆网络 (BiLSTM); 条件随机场; 知识图谱; 自然语言处理

引用格式: 陈业明,戴齐,刘捷.融合字位置特征的铁路事故命名实体识别.计算机系统应用,2022,31(12):211–219. http://www.c-s-a.org.cn/1003-3254/8860.html

Named Entity Recognition of Railway Accident Texts with Character Position Features

CHEN Ye-Ming, DAI Qi, LIU Jie

(School of Computing and Artificial Intelligence, Southwest Jiaotong University, Chengdu 611756, China)

Abstract: Relevant information of railway accidents, existing in the form of accident overview texts, is of great significance to railway safety work. However, due to the lack of effective information extraction methods, the knowledge of railway accidents scattered in the texts has not been fully utilized. Named entity recognition is an important subtask of information extraction, and there are few studies on named entity recognition of accidents. A named entity recognition model fused with character position features is proposed for the named entity recognition of railway accidents. The model obtains the character position features through a fully connected neural network. It merges them with the character vectors at the semantic level as the final vector representation of the characters, which is then input to the BiLSTM-CRF model to obtain the optimal label sequence. The experimental results show that the accuracy, recall, and *F*1 value of the model on the named entity recognition of railway accident texts are 93.29%, 94.77%, and 94.02% respectively. This model yields better effects than traditional models and lays a foundation for the construction of a railway accident knowledge graph.

Key words: named entity recognition; railway accident; character position features; bidirectional long short-term memory (BiLSTM); conditional random field; knowledge graph; natural language processing (NLP)

1　引言

铁路是国家大力发展的重要基础设施、大众化的交通工具, 在国家综合交通运输体系中有不可替代的地位. 在铁路行业中, 铁路安全占据了极其重要的位置.

① 基金项目: 国家铁路集团有限公司科技研究开发重点课题 (N2020S009)
收稿时间: 2022-04-14; 修改时间: 2022-05-22; 采用时间: 2022-05-28; csa 在线出版时间: 2022-07-28

安全是铁路运输的基本要求, 是铁路运输永恒的主题. 国家铁路局公布的铁路安全情况公告显示: 近年来, 我国铁路安全总体上呈现稳定、有序、可控的发展态势, 但每年仍有数百人因铁路事故丧生. 目前, 铁路事故的相关信息以事故概况文本的形式存在, 且以纸质方式存档. 这些文本记录了事故的发生过程, 包含大量关于事故时间, 事故地点, 事故类型等的铁路事故知识. 但由于缺乏有效的信息抽取与存储手段, 导致在传统的铁路安全管理与铁路事故分析中并不能充分利用这些信息. 知识图谱是一种描述现实世界中概念、实体及其关系的管理海量知识的工具. 构建铁路事故知识图谱可实现铁路事故知识的有效集成与持续积累、事故案列快速检索, 同时可统计事故时间、地点、列车等数据的分布情况完成事故分析, 为事故预防工作提供知识支持. 铁路事故知识图谱对于提升铁路安全工作效率、铁路运输安全性具有重要的意义, 而铁路事故命名实体识别是构建铁路事故知识图谱的基础.

命名实体识别 (named entity recognition, NER), 也称为实体抽取, 其目的是识别出文本中表示命名实体的成分, 并对其进行分类[1]. 实体是知识图谱中最基本的元素, 命名实体识别是知识图谱构建过程中基础与关键的一步. 铁路事故文本的命名实体识别属于特定领域的实体抽取问题, 主要任务是从非结构化铁路事故文本数据中识别出事故列车、地点、时间等不同类型的实体. 相较于其他领域的实体抽取问题, 铁路事故的命名实体识别问题拥有专有名词多、实体边界模糊、实体表述方式多的特点. 因此, 针对铁路事故命名实体识别问题, 本文提出了一种融合字位置特征的铁路事故文本命名实体识别方法, 并采用人工标注的方法获取实验语料, 通过实验验证了方法的有效性.

本文的组织结构如下: 第 2 节分别介绍通用领域与事故领域命名实体识别的相关工作; 第 3 节介绍本文提出的融合字位置特征的铁路事故文本命名实体识别模型; 第 4 节介绍实验语料, 对本文提出的模型进行实验, 并对实验结果进行分析; 第 5 节为结束语.

2 相关工作

命名实体识别的研究开展得较早且根据应用场景的不同可以分为通用领域的命名实体识别与特定领域的命名实体识别, 发展到现在主要有基于规则的方法、基于统计模型的方法和基于深度学习的方法.

早期的命名实体识别主要采用人工编写规则的方法, 规则的制定依赖于专家与特定领域, 导致工作量巨大且可移植性差.

基于统计模型的方法利用标注语料进行模型训练, 常用于命名实体识别的统计学习模型包括隐马尔可夫模型 (hidden Markov models, HMM)[2]、条件随机场模型 (conditional random field, CRF)[3] 等. 与传统基于规则的方法相比, 基于统计模型的方法不用定义繁琐的规则, 但良好的识别效果往往需要人工定义特征, 导致人力资源消耗较大.

近年来, 随着深度学习技术被广泛应用于各类自然语言处理问题并展现出良好的效果, 基于深度学习的命名实体识别方法应运而生并成为时下命名实体识别研究的热点. 相较于传统的命名实体识别方法, 基于深度学习的方法通过神经网络模型自动提取特征, 实现端到端的命名实体识别, 避免了人工对规则或特征的定义.

学者们对于基于深度学习的命名实体识别方法在通用领域进行了大量的研究, 例如, Huang 等人[4] 首次提出使用 BiLSTM-CRF 模型进行命名实体识别, 在通用领域的命名实体识别语料 CoNLL-2003 数据集上获得了 88.83% 的 $F1$ 值. Ma 等人[5] 在文献 [4] 的基础上使用 CNN 模型获取单词的字符级向量表示, 提出了 LSTM-CNNs-CRF 模型, 获得了 2.38% 的 $F1$ 值的提升. Strubell 等人[6] 将 IDCNNs 模型应用于命名实体识别任务, 在与 BiLSTM-CRF 模型的识别效果相当的情况下极大地缩短了模型的训练时间. 申晖等人[7] 提出 BSTTC 模型进行中文命名实体识别, 采用星型的 Transformer 结构提取句子特征, 提高识别效果的同时还减少了模型训练时间.

目前, 关于事故文本的命名实体识别研究还处在起步阶段. 宋建炜等人[8] 采用改进的预训练语言模型结合 BiLSTM-CRF 模型在建筑施工安全事故文本的命名实体识别问题上取得了 95.18% 的 $F1$ 值. 刘鹏等人[9] 基于 Lattice-LSTM 模型[10] 识别煤矿事故命名实体, 并以此为基础构建了煤矿安全知识图谱. 具体到铁路事故文本命名实体识别的研究, Hua 等人[11] 使用 CRF 模型识别我国铁路事故的事故原因与事故结果实体, $F1$ 值均到达了 80% 以上. Li 等人[12] 使用 BiLSTM-CRF 模型对我国铁路事故及故障分析报告中的事故故障时间、事故故障等级等 8 类命名实体进行了命名实体识别并通过实验验证了该模型在理论和实践上的适用性. 上述研究直接将通用领域的实体抽取模型应用于事故

文本, 而没有考虑到事故文本本身的特点, 限制了识别效果的提升.

相较于一般领域的文本, 铁路事故文本有以下特点: (1) 事故文本包含大量铁路领域的专有名词和专业术语, 一些通用领域常用的自然语言处理技术, 如词性标注、文本分词等难以应用于铁路事故文本中; (2) 对事故的描述遵循一定的规律, 导致实体在文本中出现的位置较为固定, 例如事故时间通常处在句首, 事故类型通常处在句尾等; (3) 待识别的实体存在省略、嵌套的情况, 导致实体的边界模糊. 基于上述特点, 本文提出一种融合字位置特征的铁路事故文本命名实体识别方法, 该方法通过 Word2Vec 模型获取字的语义层面向量表示, 并采用全连接的神经网络获取字在句中的位置特征, 将字的语义层面向量表示与字的位置特征合并作为字最终的向量表示, 最后通过 BiLSTM-CRF 模型获取字的最优标签.

3 融合字位置特征的实体识别模型

3.1 模型整体结构

融合字位置特征的命名实体识别模型由输入层、嵌入层、双向 LSTM 层、CRF 层与输出层组成, 模型的整体结构如图 1 所示.

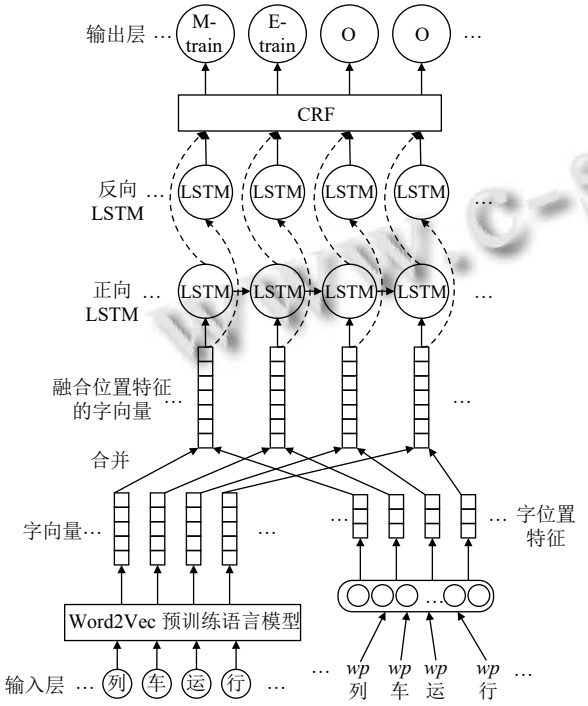

图 1 融合字位置特征的铁路事故命名实体识别模型

模型首先接收字序列与字位置序列作为输入, 然后在嵌入层分别使用 Word2Vec 预训练语言模型与全连接的神经网络获取语义层面字向量与字位置特征并将字向量与字位置特征进行合并得到字的最终向量表示, 将字的向量表示输入 BiLSTM 模型自动提取特征, 获取上下文信息完成实体的初识别, 最后使用 CRF 模型输出文本的最优标签序列, 实现铁路事故文本的命名实体识别.

3.2 Word2Vec 预训练语言模型

中文的命名实体识别可以在字层面进行, 也可以在词层面进行. 词层面的命名实体识别需要首先对文本进行分词, 但铁路事故文本包含大量铁路领域的专有名词和专业术语, 直接使用现有的分词工具对事故文本进行分词会出现分词错误影响命名实体识别结果的情况. 故本文在字层面进行命名实体识别研究.

深度学习模型无法理解符号化的文本, 而只能接收数值型输入. 因此, 基于深度学习的命名实体识别需要将输入的字表示成向量的形式. 在自然语言处理领域, 通过使用大规模无标注的文本语料来训练深层网络结构, 从而得到字向量, 这种深层网络结构通常被称为“预训练模型”[13]. 目前最常用的预训练语言模型是 2013 年 Mikolov 发布的 Word2Vec 模型[14,15].

Word2Vec 模型的实质是一个浅层的神经网络, 旨在通过字的 one-hot 编码训练神经网络, 使得该神经网络能预测给定字的上下文或通过给定的上下文预测字本身, 二者分别被称为 Skip-gram 模型与 CBOW 模型, 而训练完成后神经网络的权重便可作为字的向量表示. 本文使用 Word2Vec 的 Skip-gram 模型获取铁路事故文本语义层面的字向量表示, 模型的整体架构如图 2.

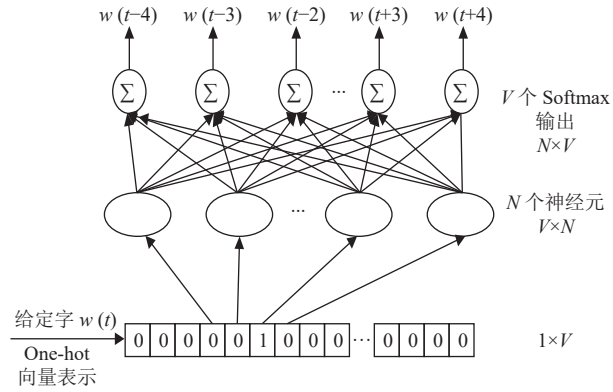

图 2 文本字向量表示 Skip_gram 神经网络模型

图 2 中, Skip_gram 模型以预测给定字的上下文为目标学习隐藏层中神经元的权重, 学习完成后, 通过 $V\times N$ (V 为语料中字的总数, N 为嵌入维度) 大小的权重矩阵得到字的 N 维向量表示.

3.3 字位置特征获取

在通过预训练语言模型获得语义层面字向量的基础上, 一些研究还将通过传统特征工程得到的向量融入字的向量表示中, 且在一些特定领域得到了比融合前更好的识别效果. 例如雷树杰等人[16] 以英文武器装备名识别为背景, 证明了基于深度学习的命名实体识别问题中语言学特征与领域特征存在的有效性与必要性, 董瑞等人[17] 设计了一组维吾尔语语言学特征并为每个特征分配一个向量, 将语言学特征向量、字符特征向量与词向量合并输入 BiLSTM-CRF 模型中, 获得了更高的识别精度.

铁路事故领域, 由于记录人员对事故的描述遵循一定的规律, 不同概念的实体在句中出现的位置都各不相同且较为固定. 该特点使字在句中的位置可以成为除语义层面字向量之外判断该字标签的另一个重要特征. 由于语料中句子的长度各不相同, 本文通过式 (1) 对字的位置进行归一化处理.

$$wp_c = \frac{i}{len(sentence)},\ 1 \leqslant i \leqslant len(sentence) \tag{1}$$

其中, i 表示字 c 是句中的第 i 个字, $len(sentence)$ 表示句子的长度, wp_c 表示字 c 在句中的位置. 计算得到字在句中的位置后, 本文使用一个全连接的神经网络提取字的位置特征, 模型架构如图 3 所示.

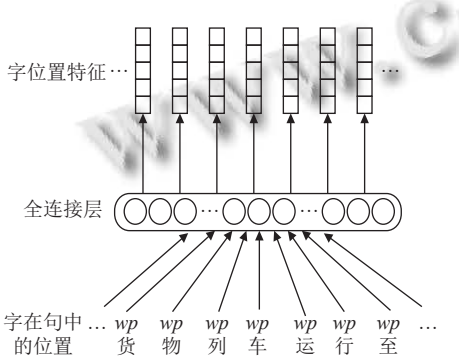

图 3 字位置特征提取模型

3.4 BiLSTM 模型

长短期记忆 (long short-term memory, LSTM) 单元[18] 是带有门控机制的循环神经网络 (recurrent neural network, RNN) 模型, 门控机制实现了对长距离信息的有效利用, 有效解决了传统 RNN 模型的短期记忆问题与梯度消失问题. LSTM 单元的具体计算过程如式 (2)–式 (7) 所示:

$$i_{(t)} = \sigma(W_{xi}x_{(t)} + W_{hi}h_{(t-1)} + b_i) \tag{2}$$

$$f_{(t)} = \sigma\left(W_{xf}x_{(t)} + W_{hf}h_{(t-1)} + b_f\right) \tag{3}$$

$$o_{(t)} = \sigma(W_{xo}x_{(t)} + W_{ho}h_{(t-1)} + b_o) \tag{4}$$

$$g_{(t)} = \tanh\left(W_{xg}x_{(t)} + W_{hg}h_{(t-1)} + b_g\right) \tag{5}$$

$$c_{(t)} = f_{(t)} \otimes c_{(t-1)} + i_{(t)} \otimes g_{(t)} \tag{6}$$

$$h_t = o_{(t)} \otimes \tanh(c_{(t)}) \tag{7}$$

其中, W 与 b 分别表示输入门、遗忘门、输出门的权重矩阵与偏置向量, $x_{(t)}$ 表示单元输入, $c_{(t)}$ 代表记忆单元状态, $i_{(t)}$、$f_{(t)}$、$o_{(t)}$ 分别表示输入门、遗忘门、输出门. $g_{(t)}$ 是仅由当前输入得到的中间状态, 用于更新记忆单元的信息, $h_{(t)}$ 是单元的输出. σ表示 *Sigmoid* 激活函数, tanh 表示双曲正切激活函数, $\otimes$为点乘运算. 单向的 LSTM 只利用了字的上文信息, 而对于命名实体识别任务, 字的下文信息同样重要, 故本文使用双向长短时记忆 (bidirectional long short -term memory, BiLSTM) 结构[19] 进行模型训练, 其结构如图 4 所示.

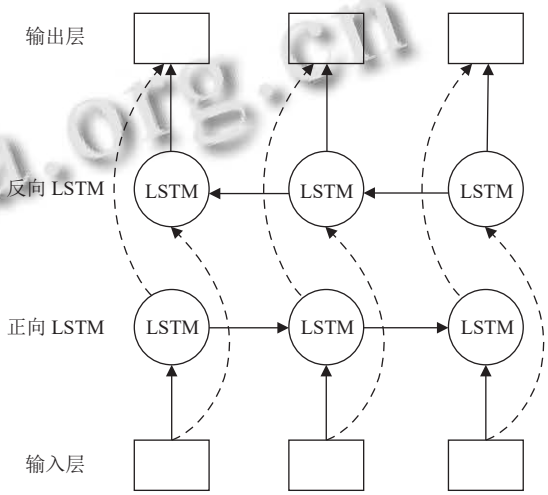

图 4 BiLSTM 结构

由图 4 中可以看出, 输入层将字的向量表示分别输入正向与反向 LSTM 单元, 在输出层将两个方向的计算结果进行拼接构成最终的输出, 因此 BiLSTM 结构可以综合考虑上下文中的信息, 获得更好的识别效果.

3.5 CRF 模型

通过 BiLSTM 模型获取字属于每个标签的概率仅在字的层面进行预测, 忽略了命名实体识别问题中标签之间的约束关系, 例如“B-Time”标签后面不可能出现“M-Train”标签, 文本的开头不可能是“E-”标签等, 从而仅使用 BiLSTM 模型进行命名实体识别会出现预测错误的情况. 针对上述不足, 本文将 CRF 模型[3] 加入最终的模型中, 对 BiLSTM 模型的输出进行处理, 获得全局最优的标记序列.

定义输入句子$S=\{w_1,w_2,\cdots,w_n\}$, BiLSTM 模型的输出结果$P_{n\times m}$ (n 表示句子中字的个数, m 表示标签种类, P_{ij} 表示字 i 属于标签 j 的概率), 标签转移概率矩阵$A_{(m+2)\times(m+2)}$. 则对于一个预测序列$y=\{y_1,y_2,\cdots,y_n\}$, 它的概率可以表示为:

$$K(X,y)=\sum_{i=0}^{n}A_{y_i,y_{i+1}}+\sum_{i=1}^{n}P_{i,y_i} \tag{8}$$

其中, A_{ij} 表示由标签转移到标签的概率, A_{y_0,y_1}与$A_{y_n,y_{n+1}}$分别表示句子起始标签与结尾标签的概率. 对所有可能的序列进行归一化得到在句子 S 条件下序列 y 的概率为:

$$p(y\mid S)=\frac{\mathrm{e}^{K(X,y)}}{\sum_{\tilde{y}\in Y_x}K(X,\tilde{y})} \tag{9}$$

其中, Y_x 表示所有可能的标记集合, 包括不符合标记规则的标记序列.

在训练过程中标记序列的似然函数:

$$\log(p(y\mid S))=K(X,y)-\log\left(\sum_{\tilde{y}\in Y_x}K(X,\tilde{y})\right) \tag{10}$$

预测时, 由式 (11) 得到概率最大的一个标签序列:

$$y^*=\underset{\tilde{y}\in Y_x}{\operatorname{argmax}}K(X,\tilde{y}) \tag{11}$$

4 实验结果及分析

4.1 实验语料

本文从各铁路局编制的典型事故案列汇编中收集整理了 1 000 份铁路事故案例文本获得了共计 72 174 字的实验语料并根据铁路事故知识图谱的应用需要定义了事故时间、事故列车、事故地点、经济损失、事故类型和伤亡情况 6 类实体类型. 实体类型其对应的实体举例如表 1 所示.

表 1 铁路事故领域实体类型

概念	实体举例
事故时间 (Time)	2015年4月17日16时50分
事故列车 (Train)	xxxxx次列车
事故地点 (Location)	xx线/xx站
经济损失 (Loss)	5.96万元
事故类型 (Type)	铁路交通一般A类事故
伤亡情况 (Casualties)	2人轻伤

实体的标注采用“BMEO”标注体系, 其中“B-概念”表示当前字为该概念实体的起始字; “M-概念”表示当前字为该概念实体的内部字; “E-概念”表示当前字为该概念实体的结尾字; “O”表示当前字为非实体字. 根据上述实体标注规范与表 1 中列出的 6 类概念, 可以得到 19 类标签, 如表 2 所示.

表 2 标签类别

序号	标签	标签含义
1	B-Time	事故时间实体起始字
2	M-Time	事故时间实体内部字
3	E-Time	事故时间实体结尾字
4	B-Train	事故列车实体起始字
5	M-Train	事故列车实体内部字
6	E-Train	事故列车实体结尾字
7	B-Location	事故地点实体起始字
8	M-Location	事故地点实体内部字
9	E-Location	事故地点实体结尾字
10	B-Loss	经济损失实体起始字
11	M-Loss	经济损失实体内部字
12	E-Loss	经济损失实体结尾字
13	B-Type	事故类型实体起始字
14	M-Type	事故类型实体内部字
15	E-Type	事故类型实体结尾字
16	B-Casualties	伤亡情况实体起始字
17	M-Casualties	伤亡情况实体内部字
18	E-Casualties	伤亡情况实体结尾字
19	O	非实体

使用命名实体语料标注工具 YEDDA[20] 对实验语料中出现的命名实体进行人工标注, 一个标注示例如图 5 所示.

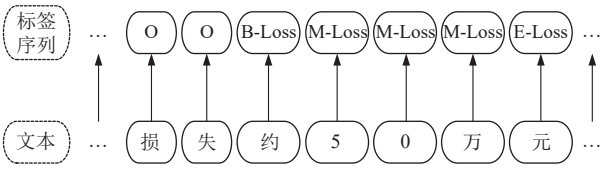

图 5 实体标注示例

标注完成后, 总计获得 4 201 个命名实体. 本文以 8:1:1 的比例将标注的实验语料划分为训练集、验证集与测试集, 各个数据集中命名实体的具体数量如表 3 所示.

表 3 铁路事故数据集命名实体数量

实体类别	训练集	验证集	测试集	合计
事故时间	800	100	100	1 000
事故列车	799	100	100	999
事故地点	782	98	98	978
经济损失	48	3	19	70
事故类型	790	99	100	989
伤亡情况	126	16	23	165
合计	3 345	416	440	4 201

为验证融合字位置特征的命名实体识别模型在铁路事故领域的有效性, 本文在上述数据集上进行对比实验与消融实验. 此外, 本文还在 Zhang 等人[10] 标注的中文简历数据集上进行实验验证模型在不同领域的适用性.

4.2 评价指标

本实验使用准确率 P, 召回率 R 与调和平均数 $F1$ 作为模型的评价指标, 假设实体概念总数为 N, 对于第 i 个实体类别:

$$P_i = \frac{TP_i}{TP_i + FP_i} \times 100\% \tag{12}$$

$$R_i = \frac{TP_i}{TP_i + FN_i} \times 100\% \tag{13}$$

$$F1_i = \frac{2 \times P_i \times R_i}{P_i + R_i} \times 100\% \tag{14}$$

其中, TP_i 为模型识别出第 i 个实体类别的正确实体个数; FP_i 为模型识别出第 i 个实体类别的错误实体个数; FN_i 为模型未识别出第 i 个实体类别的正确实体个数, $F1_i$ 为 P_i 与 R_i 的调和平均数. 根据各类别的实体数量在总实体数量中的占比对各类别实体的识别效果进行加权平均, 可以得到模型的总体准确率 P、总体召回率 R 和总体调和平均数 $F1$:

$$P = \frac{\sum_{i=1}^{N} TP_i}{\sum_{i=1}^{N} TP_i + \sum_{i=1}^{N} FP_i} \times 100\% \tag{15}$$

$$R = \frac{\sum_{i=1}^{N} TP_i}{\sum_{i=1}^{N} TP_i + \sum_{i=1}^{N} FN_i} \times 100\% \tag{16}$$

$$F1 = \frac{2 \times P \times R}{P + R} \times 100\% \tag{17}$$

4.3 模型搭建与参数设置

本文采用谷歌人工智能团队开发的深度学习框架 TensorFlow 搭建命名实体识别模型, 预训练语言模型使用第三方开源库 Gensim 中自带的 Word2Vec 模块对实验语料进行训练, 验证集与测试集中若有训练集中未出现的字时, 对该字进行随机初始化获取字向量. 模型参数设置如下: 预训练字向量维度为 128, 获取字位置特征的全连接神经网络神经元个数为 16, BiLSTM 模型神经元个数为 128, 使用 dropout 技术[21] 防止模型过拟合, dropout 值为 0.2, CRF 模块中全连接层参数为字标签类别数 19, 优化器采用可加速收敛的 Adam[22], 学习率为 0.002, 训练轮数为 20, 批处理句子数量设为 16.

4.4 实验结果及分析

为验证本文提出的融合字位置特征的铁路事故文本命名实体识别模型的有效性, 在铁路事故文本语料上共进行了 4 组对比实验, 使用的模型分别为 CRF、BiGRU-CRF、IDCNNs 与 BiLSTM-CRF. 其中 CRF 模型使用开源的条件随机场工具包 CRF++进行训练, IDCNNs 模型由 4 层一维卷积操作与 1 层空洞卷积操作组成, BiGRU-CRF 与 BiLSTM-CRF 的模型参数与第 4.3 节中的参数设置相同, 实验结果如表 4 所示.

表 4 对比实验结果 (%)

模型	P	R	$F1$
CRF	92.49	89.55	90.99
BiGRU-CRF	90.81	92.05	91.42
IDCNNs	91.35	93.64	92.48
BiLSTM-CRF	92.57	93.41	92.99
本文	93.29	94.77	94.02

从表 4 可知, 本文提出的融合字位置特征的命名实体识别模型的准确率为 93.29%, 召回率为 94.77%, $F1$ 值为 94.02%, 3 项指标均为最优, 验证了本文提出的模型的有效性. 相比于 BiLSTM-CRF 模型, 模型的准确率、召回率、$F1$ 值分别提高了 0.72%、1.36%、1.03%, 说明了全连接神经网络获取字的位置特征的有效性. 由于本文将字的位置特征加入字的向量表示, 所以能在一定程度上通过字在句中的位置判断字对应的标签, 正确识别训练集中未出现过的事故实体. 例如在 “…57011 次接触网检修车申请区间返回明光站…”句中, “明光站”这一事故地点实体在训练集中没有出现

过, BiLSTM-CRF 模型没有识别出该实体, 但融合字位置特征的命名实体识别模型判断出这些字处在事故地点实体经常出现在句中的位置, 再结合语义信息与上下文信息, 成功识别出了该实体.

为说明本文提出的融合字位置特征的铁路事故文本命名实体识别方法中各模块在命名实体识别问题中的作用, 使用 TensorFlow 自带的 Embedding 层替代 Word2Vec 模块, 移除 CRF 模块与字位置特征获取模块进行消融实验, 实验结果如表 5 所示.

表 5 消融实验结果 (%)

模型	P	R	$F1$
Embedding+PF+BiLSTM+CRF	91.86	92.27	92.06
Word2Vec+PF+BiLSTM	86.51	91.82	89.08
Word2Vec+BiLSTM+CRF	92.57	93.41	92.99
Word2Vec+PF+BiLSTM+CRF	93.29	94.77	94.02

由表 5 可见, 使用 Embedding 层替代 Word2Vec 模块后, 准确率降低了 1.43%, 召回率降低了 2.5%, $F1$ 值降低了 1.96%, 说明 Word2Vec 预训练语言模型能获取语义层面的字向量, 提升模型识别效果. 移除 CRF 模块后, 准确率降低了 6.78%, 召回率降低了 2.95%, $F1$ 值降低了 4.94%, 说明 CRF 模块能考虑标签之间的约束关系, 得到全局最优的标注序列, 对模型识别效果的提升较大. 移除字位置特征获取模块造成模型识别效果的差异与实验结果的分析如上文所述, 说明对于铁路事故文本, 将字的位置特征加入字的向量表示中能进一步提升模型的识别能力.

模型对于铁路事故各个类别实体的识别效果如表 6.

表 6 各类别识别效果对比 (%)

实体类别	P_i	R_i	$F1_i$
事故时间	99.01	100.00	99.50
事故列车	95.05	96.00	95.52
事故地点	84.16	86.73	85.43
经济损失	89.47	89.47	89.47
事故类型	97.03	98.00	97.51
伤亡情况	87.50	91.30	89.36

从表 6 可见, 事故时间、事故类型与事故列车这 3 类实体的 $F1$ 值均达到了 95% 以上, 主要是因为这 3 类实体都具有较明显的边界特征且表述方式相对固定, 如事故时间多以“日、分、左右、许”等字结尾, 事故类型多以“事故”两字结尾, 事故列车常以字母或数字开头, 以“列车”两字结尾. 同时, 这 3 类实体在文本中所处的位置也较为固定, 事故时间常出现在句子开头, 事故类型常出现在句子结尾, 事故列车多出现在句中靠前的位置. 而事故地点、经济损失与伤亡情况 3 类实体的 $F1$ 值略有下降, 分别为 85.43%、89.47% 与 89.36%, 其中, 经济损失与伤亡情况两类实体的训练数据较少造成识别效果相对较差, 事故地点类实体识别效果较差的主要原因是我国铁路车站及线路名称众多, 导致实体边界特征不明显.

在标注错误的实例中, 较为典型的有以下几类: (1) 事故列车实体“51677B 次”中的“次”字被错误标注为“M-Train”, 这是因为该实体省略了“列车”两字, 使得实体边界识别错误. (2) “精河站检车员”中“精河站”作为“检车员”的定语出现, 不代表事故地点, 但由于本文提出的方法未考虑句子的语法特征, 导致该词被错误标注为事故地点类实体. (3) 事故地点实体“西固城站”只识别出了“固城站”, “津霸联络线”只识别出了“联络线”, 造成这类错误的原因是事故地点实体的边界特征不明显.

简历文本与事故文本类似, 拥有实体出现位置较为固定的特点, 故本文在中文简历数据集[10] 上进行实验验证模型在不同领域的适用性, 实验结果如表 7 所示.

表 7 中文简历数据集实验结果 (%)

实体类别	P_i	R_i	$F1_i$
国籍	96.43	96.43	96.43
学历	93.04	95.54	94.27
籍贯	100.00	83.33	90.91
姓名	95.37	91.96	93.64
单位	84.55	88.07	86.27
专业	77.78	84.85	81.16
民族	100.00	100.00	100.00
职位	92.33	91.97	92.15

从表 7 可以看出, 本文提出的融合字位置特征的命名实体识别方法在中文简历数据集上的 $F1$ 值均高于 80%, 在多数实体类别上 $F1$ 值高于 90%, 实验结果说明融合字位置特征的命名实体识别方法适用于不同领域的实体识别研究.

在提取相应实体后, 将铁路事故知识以三元组的形式加载到 Neo4j 图数据库中进行知识的整合与存储. 结构化铁路事故知识局部效果如图 6 所示 (由事故时间实体、事故列车实体与事故类型实体拼接作为铁路事故的唯一标识实体). 由此可完成铁路事故快速检索、事故相关信息的统计分析工作.

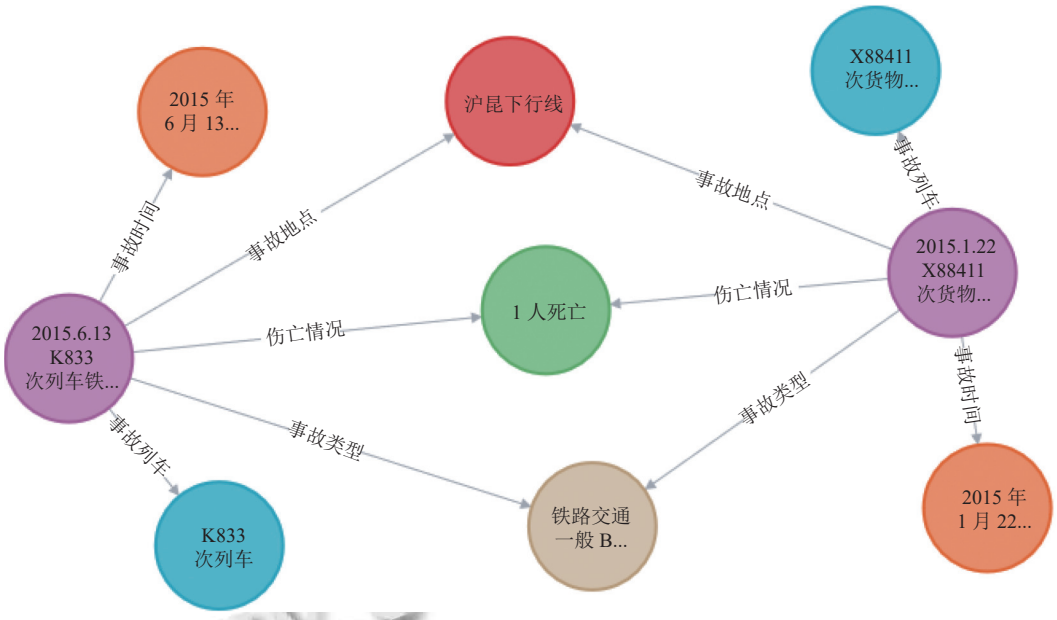

图 6 结构化铁路事故知识局部效果

5 结论与展望

本文根据铁路事故领域知识图谱的应用需求, 对铁路事故文本的命名实体识别问题进行了研究, 定义了该领域 6 类概念. 在此基础上, 收集整理并标注了 1 000 篇铁路事故案例文本作为实验语料, 提出了融合字位置特征的命名实体识别方法. 该方法关注到事故文本本身的特点, 将字的位置特征加入字的向量表示中. 实验表明, 相比于传统模型, 本文提出的方法具有更优的识别效果且适用于实体在文本中出现位置较为固定的其他特殊领域的实体识别研究.

铁路事故领域没有大规模的标注数据, 采用人工标注的方法费时费力. 在未来的研究工作中, 将基于所提的方法引入迁移学习, 使用少量标注数据进行铁路事故命名实体识别研究. 另外, 为构建铁路事故领域知识图谱, 后续会开展事故原因的实体关系联合抽取研究, 从而更好地满足铁路事故领域知识图谱的应用需求.

参考文献

1 刘浏, 王东波. 命名实体识别研究综述. 情报学报, 2018, 37(3): 329–340. [doi: 10.3772/j.issn.1000-0135.2018.03.010]

2 Bikel DM, Miller S, Schwartz R, *et al*. Nymble: A high-performance learning name-finder. Proceedings of the 5th Conference on Applied Natural Language Processing. Washington: ACM, 1997. 194–201.

3 何炎祥, 罗楚威, 胡彬尧. 基于 CRF 和规则相结合的地理命名实体识别方法. 计算机应用与软件, 2015, 32(1): 179–185, 202. [doi: 10.3969/j.issn.1000-386x.2015.01.046]

4 Huang ZH, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging. arXiv:1508.01991, 2015.

5 Ma XZ, Hovy E. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin: Association for Computational Linguistics, 2016. 1064–1074.

6 Strubell E, Verga P, Belanger D, *et al*. Fast and accurate entity recognition with iterated dilated convolutions. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen: Association for Computational Linguistics , 2017. 2670–2680.

7 申晖, 张英俊, 谢斌红, 等. 基于 BSTTC 模型的中文命名实体识别. 计算机系统应用, 2021, 30(6): 262–270. [doi: 10.15888/j.cnki.csa.007935]

8 宋建炜, 邓逸川, 苏成. 基于预训练语言模型的建筑施工安全事故文本的命名实体识别研究. 图学学报, 2021, 42(2): 307–315.

9 刘鹏, 叶帅, 舒雅, 等. 煤矿安全知识图谱构建及智能查询方法研究. 中文信息学报, 2020, 34(11): 49–59. [doi: 10.3969/j.issn.1003-0077.2020.11.007]

10 Zhang Y, Yang J. Chinese NER using lattice LSTM. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne: Association for Computational Linguistics, 2018. 1554–1564.

11 Hua LL, Zheng W, Gao SG. Extraction and analysis of risk

factors from Chinese railway accident reports. 2019 IEEE Intelligent Transportation Systems Conference (ITSC). Auckland: IEEE, 2019. 869–874.

12 Li XQ, Shi TY, Li P, *et al*. BiLSTM-CRF model for named entity recognition in railway accident and fault analysis report. Proceedings of the Asia-Pacific Conference on Intelligent Medical 2018 & International Conference on Transportation and Traffic Engineering 2018. Beijing: ACM, 2018. 1–5.

13 李舟军, 范宇, 吴贤杰. 面向自然语言处理的预训练技术研究综述. 计算机科学, 2020, 47(3): 162–173. [doi: 10.11896/jsjkx.191000167]

14 Mikolov T, Chen K, Corrado G, *et al*. Efficient estimation of word representations in vector space. arXiv:1301.3781, 2013.

15 Mikolov T, Sutskever I, Chen K, *et al*. Distributed representations of words and phrases and their compositionality. Proceedings of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe: ACM, 2013. 3111–3119.

16 雷树杰, 邢富坤, 王闻慧. 融合多类型特征的特定领域实体识别研究. 计算机应用与软件, 2019, 36(11): 210–217. [doi: 10.3969/j.issn.1000-386x.2019.11.034]

17 董瑞, 杨雅婷, 蒋同海. 融合多种语言学特征的维吾尔语神经网络命名实体识别. 计算机应用与软件, 2020, 37(5): 183–188.

18 Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735–1780. [doi: 10.1162/neco.1997.9.8.1735]

19 Graves A, Jaitly N, Mohamed A. Hybrid speech recognition with deep bidirectional LSTM. 2013 IEEE Workshop on Automatic Speech Recognition and Understanding. Olomouc: IEEE, 2013. 273–278.

20 Yang J, Zhang Y, Li LW, *et al*. YEDDA: A lightweight collaborative text span annotation tool. Proceedings of ACL 2018, System Demonstrations. Melbourne: Association for Computational Linguistics, 2017. 31–36.

21 Hinton GE, Srivastava N, Krizhevsky A, *et al*. Improving neural networks by preventing co-adaptation of feature detectors. arXiv:1207.0580, 2012.

22 Kingma DP, Ba J. Adam: A method for stochastic optimization. Proceedings of the 3rd International Conference on Learning Representations. arXiv:1412.6980, 2015.

(校对责编: 孙君艳)