doi:10.12068/j.issn.1005－3026.2026.20240234

航空事故领域的知识抽取方法研究与实现

刘 军[1]，曹 悦[1]，刘向军[2]，王宏艳[1]
（1.东北大学 计算机科学与工程学院，辽宁 沈阳 110169；2.中软信息系统工程有限公司，北京 100081）

摘 要：随着航空运输业与信息技术的快速发展，航空应急管理给海量、异构的航空安全数据高效利用带来了挑战.本文针对航空事故知识图谱的知识抽取问题，即命名实体识别与关系抽取问题，提出以下方法：1）提出基于BERT（bidirectional encoder representations from Transformers）的改进BiGRU－IDCNN－CRF模型，实现94.69%的命名实体识别精确率；2）构建基于强化学习的聚类远程监督关系抽取模型，结合改进K均值聚类与远程监督标注降低数据噪声，并通过强化学习优化去噪过程，最终结合分段卷积神经网络（PCNN）与注意力机制，实现84.16%的关系抽取精确率.实验结果表明，本文方法有效提升了航空事故知识图谱的信息提取质量，为航空安全管理提供了精准的信息支撑.

关 键 词：航空事故；知识抽取；命名实体识别；关系抽取；远程监督；强化学习
中图分类号：TP 391 **文献标志码**：A **文章编号**：1005－3026(2026)01－0089－10

Research and Implementation of Knowledge Extraction in Aviation Accident Domain

LIU Jun[1], *CAO Yue*[1], *LIU Xiang-jun*[2], *WANG Hong-yan*[1]
（1. School of Computer Science & Engineering, Northeastern University, Shenyang 110169, China; 2. China Software Information System Engineering Co. Ltd., Beijing 100081, China. Corresponding author: LIU Jun, E－mail: liujun@cse.neu.edu.cn）

Abstract: In light of the rapid development of air transportation and information technology, the efficient utilization of massive and heterogeneous aviation safety data in aviation emergency management faces challenges. The problem of knowledge extraction for an aviation accident knowledge graph was studied, specifically named entity recognition and relation extraction, and the following methods were proposed: 1) An improved BiGRU－IDCNN－CRF model based on bidirectional encoder representations from Transformers (BERT) was presented, achieving a named entity recognition accuracy of 94.69%; 2) A reinforcement learning-based clustering distant supervision relation extraction model was constructed, in which data noise was reduced by integrating improved K－means clustering with distant supervision labeling, and the denoising process was optimized via reinforcement learning; a combination of piecewise convolutional neural network (PCNN) and an attention mechanism was applied to achieve a relation extraction accuracy of 84.16%. Experimental results indicate that the quality of information extraction for the aviation accident knowledge graph is effectively improved, providing accurate information support for aviation safety management.

Key words: aviation accident; knowledge extraction; named entity recognition; relation extraction; distant supervision; reinforcement learning

在大数据与互联网技术的推动下，航空领域知识呈指数级增长，面对多源化的航空数据以及不断出现的航空事故新知识，传统的知识表示工具已不能满足航空事故应急管理的需求.因此，

收稿日期：2024－12－24
基金项目：国家自然科学基金青年基金资助项目（62501133）.
作者简介：刘 军（1969—），男，辽宁沈阳人，东北大学教授，博士生导师.
通信作者：刘 军，E－mail：liujun@cse.neu.edu.cn.

对于航空事故来说如何更有效利用大量的历史积累记录,提高安全规则提供的辅助分析支持,以及如何避免类似的事故发生已经成为迫切需要解决的关键问题.航空事故的成因复杂多样,多种因素构成复杂的网络关系,无法用二维关系显示.航空事故文本中包含大量数据,可能蕴含大量显性与隐性知识,挖掘这些数据之间的联系能够辅助快速决策.因此,知识图谱[1]作为实体关系的直接表征,在航空领域具有广阔的应用潜力和前景.

航空事故知识图谱的构建过程主要包括:数据获取、知识抽取、知识融合和知识加工.知识抽取是航空事故知识图谱构建的关键性工作,分为命名实体识别和关系抽取两大过程.命名实体识别技术最初是通过人工编写规则进行实体抽取的,随后发展出基于统计机器学习系统和特征工程系统.

随着深度学习方法的兴起,卷积神经网络、循环神经网络(RNN)、长短期记忆网络(LSTM)等模型被引入命名实体识别领域,并取得了显著的效果[2-4].随着BERT预训练语言模型在自然语言领域的多项任务中取得出色成绩,许多学者开始将其应用于命名实体识别任务.Qin等[5]提出基于BERT-BiGRU-CRF模型的中文电子病历命名实体识别方法;Li等[6]提出师生蒸馏学习模型,使学生模型能够融合源语言在实体识别与实体相似性评估方面的优势;Boudjellal等[7]在阿拉伯语的小型生物医学数据集中训练单语言的BERT模型,并取得较好的成果;Zhou等[8]结合对比学习和基于原型的伪标签提高跨语言命名实体识别的准确率;Ma等[9]采用双塔BERT模型分别对文本字符和对应标签进行编码,通过将二者进行点积运算实现分类.

关系抽取是知识图谱构建中的重要子任务,早期方法包括基于模板匹配和基于监督学习2种方法,但存在模板构建困难、耗时和可移植性差的问题.随着深度学习的发展,深度学习方法被引入关系抽取领域[10],包括使用新型深度神经网络(deep neural network,DNN)[11]和LSTM建立的全局优化的神经模型[12].然而,这些方法需要大量训练语料库,对于某些领域语料库不足的问题,提出了远程监督来解决[13],且随着技术的发展,图神经网络[14]、注意力机制[15]和未标记的远程监督[16]也被应用于基于远程监督的关系抽取任务.Ji等[17]提出的远程监督结合句子级注意力和实体定义的关系抽取模型能够提高抽取的精确率,但存在数据噪声的问题.Chen等[18]提出一种新的对比学习框架,旨在提高远程监督关系抽取的性能,但由于依赖于自动生成的标签,仍可能存在标签不一致和噪声问题.Luo等[19]利用动态转移矩阵来处理数据中的噪声,增强远程监督关系抽取,但转移矩阵的构建存在挑战,可能影响模型对噪声的敏感性.Zhou等[20]提出基于自选择注意力机制的远程监督关系提取方法,但在处理包含噪声的实例时表现欠佳.

目前针对航空事故领域文本的关系抽取方法研究较少,航空事故关系抽取的效果仍有待提升,需要进一步研究和改进.针对航空事故文本中实体边界模糊、结构复杂且实体长度较长、含有大量数字与字母组合词、内部相关关系较为密切、文本专业性强、文本之间存在长距离依赖且文本训练语料库少等问题,本文重点研究知识抽取中的命名实体识别与关系抽取方法,以用于航空事故知识图谱的构建.首先识别航空事故句子级文本中的实体,然后抽取实体对之间的关系,进而对抽取结果进行存储.

1 基于改进BiGRU-IDCNN-CRF的命名实体识别模型

结合航空事故文本数据的实体成分复杂、结构嵌套、含有大量数字与字母组合词以及实体内部相关关系较为密切等特点,提出改进的BiGRU-IDCNN-CRF模型来进行实体识别.

1.1 改进BiGRU

在BiGRU网络前向传播和后向传播的过程中,因为2个过程是相互独立且提取的信息量存在差异,对航空事故文本实体之间复杂度较大的文本抽取效果欠佳.因此为了确保2个过程之间紧密联系、相互制约以获得更好的结果,本文提出了贡献因子α,并将其加入到前向传播和后向传播中,使用α调整GRU(gated recurrent unit)的前向传播和后向传播对后续数据的影响,输出层通过这两层的权重以及偏置得到和标签集维度一样的向量作为最终输出,计算公式如式(1)所示:

$$h_t^{\mathrm{B}}=\alpha W_h\overrightarrow{h}_t+(1-\alpha)W_h\overleftarrow{h}_t+b_t. \tag{1}$$

其中:h_t^{B}为BiGRU的最终输出;t为时间步;W_h为输出层的权重;b_t为输出层的偏置;在提取航空事故文本特征过程中,$\overrightarrow{h}_t$为前向传输的输出状态,$\overleftarrow{h}_t$

为后向传输的输出状态.贡献因子α的值将通过多次实验来确定.

1.2　实体识别模型的设计

为了解决传统字向量嵌入方法无法表征字多义性问题,引入BERT模型将航空事故文本向量化.航空事故文本通过BERT模型将语义充分表示,将得到的高质量语义向量输入到改进的BiGRU-IDCNN-CRF模型中,对航空事故知识图谱的实体进行识别,具体的模型如图1所示.

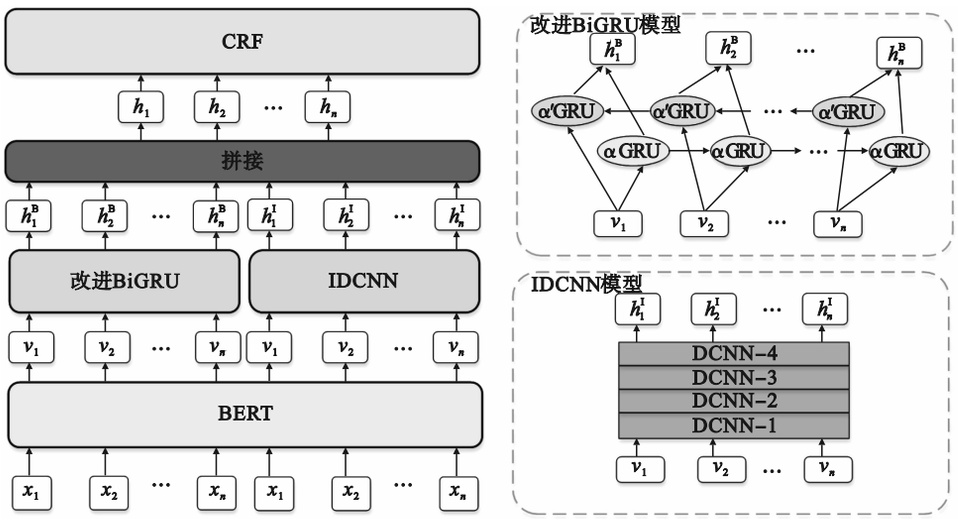

图1　基于BERT嵌入的改进BiGRU-IDCNN-CRF航空事故识别模型

Fig. 1　Improved BiGRU-IDCNN-CRF aviation accident recognition model based on BERT embedding

1）BERT:航空事故的句子级文本m_i,用$X^{m_i}=(x_1,x_2,\cdots,x_{x_n})$作为模型的输入,其中$x_i$表示句子文本的第$i$个字,将全部的文本依次按字的形式输入预训练语言模型BERT中.BERT使用字向量、位置向量和句子级向量进行嵌入,然后将它们拼接相加以产生最终的字向量输出$\boldsymbol{V}=(v_1,v_2,\cdots,v_n)$,同时字向量输出作为改进BiGRU-IDCNN编码模型的输入.

2）改进BiGRU网络:得到前向传播的模型输出序列$(\vec{h}_1,\vec{h}_2,\cdots,\vec{h}_n)$以及后向传播的模型输出序列$(\overleftarrow{h}_1,\overleftarrow{h}_2,\cdots,\overleftarrow{h}_n)$,再将其前向输出序列与后向输出序列进行信息拼接.由于本文提出了贡献因子,因此最后的输出是加入贡献因子后的输出$h_t^{\mathrm{B}}=\alpha\vec{h}_t+(1-\alpha)\overleftarrow{h}_t$,图1中的$\alpha'=1-\alpha$,贡献因子设置为0.8.经过训练最终得到改进BiGRU编码模型的输出$h^{\mathrm{B}}=[h_1^{\mathrm{B}},h_2^{\mathrm{B}},\cdots,h_n^{\mathrm{B}}]$.

3）IDCNN（iterated dilated convolutional neural network）网络:当BERT层数据进入IDCNN时,首先到达DCNN-1层.DCNN-1层的输出向量会分为2个部分:一部分与其他DCNN分层的输出拼接成向量直接输出;另一部分作为DCNN-2的输入.依此类推,最终得到IDCNN网络的输出$h^{\mathrm{I}}=[h_1^{\mathrm{I}},h_2^{\mathrm{I}},\cdots,h_n^{\mathrm{I}},]$.将此输出与改进BiGRU网络输出拼接在一起,得到该层的输出$h=[h^{\mathrm{B}},h^{\mathrm{I}}]$.此层的输出经过线性映射后得到各个标签的分数后输入到CRF（conditional random field）层中,线性映射公式如式(2)所示:

$$\boldsymbol{P}_i=W_{\mathrm{s}}h^{(t)}+b_{\mathrm{s}}. \tag{2}$$

其中:$\boldsymbol{P}_i$是标签得分向量;$h^{(t)}$是t时刻改进BiGRU和IDCNN网络层的最后输出;$W_{\mathrm{s}},b_{\mathrm{s}}$是线性映射的参数.

4）CRF:各序列的标签通过CRF模型中的状态转移矩阵$\boldsymbol{A}_{ij}$得到,$\boldsymbol{A}_{ij}$是表示第i个标签转移到第j标签的得分,因此对于输入文本为$\boldsymbol{M}=\{m_1,m_2,\cdots,m_n\}$的最终得到的标签序列为$N^{m_i}=(n_1,n_2,\cdots,n_{|m_i|})$,标签分数如下:

$$\mathrm{score}(x,y)=\sum_{i=1}^{|m_i|}\boldsymbol{P}_{i,y_i}+\sum_{i=1}^{|t_i|+1}\boldsymbol{A}_{y_{i-1},y_i}. \tag{3}$$

其中:x表示输入序列;y表示输出序列;$\sum_{i=1}^{|m_i|}\boldsymbol{P}_{i,y_i}$表示改进BiGRU和IDCNN网络的共同输出结果;$\sum_{i=1}^{|t_i|+1}\boldsymbol{A}_{y_{i-1},y_i}$表示CRF中状态转移矩阵$\boldsymbol{A}$的和.对标签分数采用softmax函数进行归一化处理,得到标签序列的概率P如式(4)所示:

$$P(y|x)=\frac{\exp(\mathrm{score}(x,y))}{\sum_{y'}\exp(\mathrm{score}(x,y'))}. \tag{4}$$

式中y'是所有可能的标签序列.

损失函数L如式(5)所示:

$$L=-\frac{1}{N}\sum_{n=1}^{N}\log P(y|x). \tag{5}$$

将本文提出的方法通过训练集对模型进行训练,然后利用测试集完成测试,最终得到能够识别航空事故文本的实体识别模型.

2 基于强化学习的聚类远程监督关系抽取模型

在第1章的基础上继续对实体对之间的关系进行研究.由于航空事故文本中含有深层语义的句子较多,且航空事故文本的专业性强,知识库与文本集的不对等导致关系标签中出现大量“NA(无关系)”样本,所以提出基于强化学习的聚类远程监督关系抽取模型.

2.1 文本向量化

为了使句子中的句法和语义信息得到更好的表达,利用Word2vec方法将句子中的字向量化,同时使用Position Embeddings将位置信息向量化嵌入,将两者拼接作为文本的向量表示,作为模型的输入.

2.2 关系抽取模型的设计

为了解决航空事故文本语料库不足的问题,本文采用远程监督来获取数据进行关系抽取.然而,远程监督获取的数据存在大量噪声,影响了关系抽取的精确率.通过分析航空事故文本,发现了大量被错误标注为“NA”关系的实体对,因此本文提出聚类远程监督方法来减少噪声.同时,引入强化学习去除数据集中的负实例噪声,以提高关系抽取模型的准确性.强化学习去噪部分与关系抽取模型相辅相成,具体过程如图2所示.

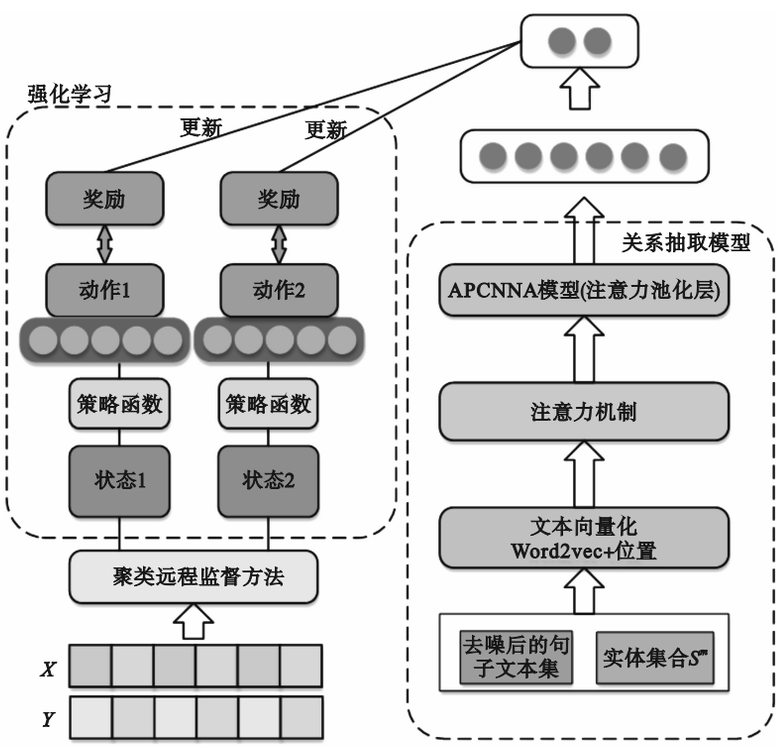

图2 基于强化学习的聚类远程监督关系抽取

Fig. 2 Reinforcement learning-based clustering distant supervision relation extraction

2.2.1 聚类远程监督标注语料

远程监督主要是通过启发式匹配完成的,将文本集中的实体对与知识库中的实体对进行匹配,并标注相应的关系.然而,这种方法存在强假设,会导致很多句子被错误标注,产生噪声.为了减少这些噪声,本文提出了聚类远程监督方法,将改进的K-means聚类算法和远程监督相结合,完成数据集标注,从而减少部分噪声的产生.

使用聚类远程监督方法进行数据标注的过程主要分为两个步骤:首先,将句子根据语义相似度聚类成多个簇;然后,通过将每个簇中的实体对与知识库对齐,找出出现次数最多的关系来标注每个簇的关系标签.最终目标是通过最小化簇对象与簇质心之间距离的平方和来优化目标函数Q,具体计算如式(6)所示:

$$Q=\sum_{i=1}^{K}\sum_{x\in C_i}\mathrm{dis}(C_i,\mu_i)^2. \tag{6}$$

式中:C_i表示每个簇对象;μ_i表示每个簇类中心,

即簇质心;$\mathrm{dis}(i)^2$表示簇对象到簇质心的距离平方;K表示聚类簇数.实现步骤见表1.

表1　聚类远程监督方法实现步骤

Table 1　Implementation steps of clustering distant supervision method

输入:句子样本集$\boldsymbol{M}=\{\boldsymbol{m}_1,\boldsymbol{m}_2,\boldsymbol{m}_3,\cdots\}$,$K$是聚类簇数,$N$是最大迭代次数

输出:聚类划分后的句子集$C=\{C_1,C_2,C_3,\cdots,C_K\}$

开始:

1. 从数据集D中随机选取K个句子样本作为初始聚类中心:$\{\mu_1,\mu_2,\mu_3,\cdots,\mu_K\}$
2. 初始簇划分C为$C_t=\varnothing,t=1,2,\cdots,K$
3. 对于样本集$\boldsymbol{M}$中的每一个样本$\boldsymbol{m}_n$执行
4. 计算句子向量$\boldsymbol{m}_n$到各聚类中心μ_i的距离:$\mathrm{dis}(\boldsymbol{m}_n,\mu_i)$
5. 将样本分配给具有最小距离$\mathrm{dis}(\boldsymbol{m}_n,\mu_i)$的中心点$\mu_\lambda$,并更新簇划分$C_\lambda=C_\lambda\cup\{\boldsymbol{m}_n\}$
6. 为每个簇重新计算质心μ
7. 若所有K个聚类中心的位置均未发生变化,或目标函数已收敛,或迭代次数已达到设定上限N,则跳转至步骤8
8. 输出簇划分$C=\{C_1,C_2,C_3,\cdots,C_K\}$

基于以上过程,数据集已经分为了K个簇,这时,每个簇中所包含句子的关系标签不再相同,所以会对同一个簇中需要重新标注的句子进行再次的标注,而对于每一个簇C需要重新标注关系r的可能性通过式(7)和(8)进行计算.

$$P(c\in r)=\begin{cases}0,\text{如果实体对 } c\in r,\\1,\text{如果实体对 } c\notin r,\end{cases}\tag{7}$$

$$P(C\in r)=\frac{\sum_{c\in C}P(c\in r)}{|C|}.\tag{8}$$

其中:$P(c\in r)$表示1个簇中的句子c属于关系r的概率;$P(C\in r)$表示整个簇是关系r的概率.整个过程将概率P最大的关系标签r作为重新标注的标签.因此该方法在一定程度上减少了“NA”关系的噪声以及漏标注的数据.

2.2.2　强化学习去噪

强化学习(reinforcement learning,RL)分为四部分,即状态(state)、动作(action)、策略函数(policy)以及奖励(reward).本文将聚类远程监督数据去噪过程看作是智能体(agent)和外部环境的交互过程,来完成对数据的降噪工作.通过将噪声数据输入到agent中,利用其内部的动作、状态、策略函数以及奖励来判断句子标注正确性,将错误标注的句子划分到负样本中,能够有效达到降噪效果.

本文在强化学习整个过程中涉及的状态、动作、策略函数以及奖励具体定义如下:

状态:$s_i=[m'_{\text{present}},m'_{\text{select}},(e_1,e_2)]$,$m'_{\text{present}}$是当前的句子,$m'_{\text{select}}$是选择出来的句子,$(e_1,e_2)$是目标的实体对.环境状态引导agent作出最佳的动作.

动作:判断句子的正确标注,并将动作a_i保留或重新分配到负样本中.本文主要控制这2个动作的选择来完成标签的去噪工作.

策略函数:用于确定句子中的词语是否与预定义的关系类型相关,并通过随机策略$\pi_\theta(s_i,a_i)$选择最佳动作,如式(9)所示:

$$\pi_\theta(s_i,a_i)=P(a_i|s_i;\theta).\tag{9}$$

其中:θ是需要学习的参数;$P(a_i|s_i;\theta)$是在给定状态s_i和参数θ的情况下,执行某个特定动作a_i的概率.

奖励:本文使用$F1$值(精确率和召回率的调和平均数)作为奖励,通过策略梯度对策略函数进行优化.在损失函数的计算公式中考虑了奖励值,以提高关系抽取模型的精确率和整体抽取效果.奖励值与$F1$值差值成正比,通过每5个时期的平均$F1$值来计算奖励,以减少随机性.奖励R_t如式(10)所示:

$$R_t=\beta(F1^i-F1^{i-1}).\tag{10}$$

其中β是奖励缩放系数,在预训练过程中使用策略梯度优化策略函数,并通过Reward进行系数加入以增强或减少动作的发生.损失函数的表达式如下:

$$L(\theta)=-\frac{1}{N}\sum_{\tau}R(\tau)\log\pi_\theta(\tau).\tag{11}$$

其中:$\tau=\{s_1,a_1,s_2,a_2,\cdots,s_t,a_t\}$是行为状态序列;$\pi_\theta(\tau)$是当前动作发生的概率;$R(\tau)$是获取到的奖励.

实验中,将带有标签的正样本和负样本分别分为训练集和测试集,并通过迭代过程中重新分配样本来优化模型.最终得到的$F1$值差值作为奖励,用于优化和调整策略.同时,提出2个集合来

计算损失函数,分别为Ω_i和Ω_{i-1},表示去掉公共识别部分后留下的各自识别的部分,以优化agent的策略.

$$\Omega_{i-1}=\psi_{i-1}-(\psi_i\cap\psi_{i-1}), \tag{12}$$

$$\Omega_i=\psi_i-(\psi_i\cap\psi_{i-1}), \tag{13}$$

$$L(\theta)=\sum^{\Omega_i}\log\pi(a|s;\theta)+\sum^{\Omega_{i-1}}\log\pi(a|s;\theta). \tag{14}$$

其中:ψ_i表示在第i轮次被移除的实例;$\psi_i\cap\psi_{i-1}$表示在当前的轮次和前一个轮次中所选取移除实例的交集.

2.2.3 APCNNA

获得的去噪后数据$\boldsymbol{M}'=\{m'_1,m'_2,m'_3,\cdots\}$输入到关系抽取模型中进行关系提取.PCNN在远程监督关系抽取方面展现了良好的性能,能够挖掘文本中高层语义信息.考虑到航空事故实体复杂且高层语义信息较多,本文提出APCNNA关系抽取模型如图3所示,其具体构造描述如下.

输入层:将2.2.2节获取到的去噪数据集和实体对集合作为关系抽取的输入,考虑到航空事故文本在去噪后可能存在一些无关的词语和符号,因此引入注意力机制来增加预定义实体在句子中的占比.预定义的实体类型作为目标词,通过注意力机制得到每个句子中的词语的权重,最终得到分段卷积网络的输入向量$\boldsymbol{W}=[w_1,w_2,w_3,\cdots,w_l]^{\mathrm{T}}$.

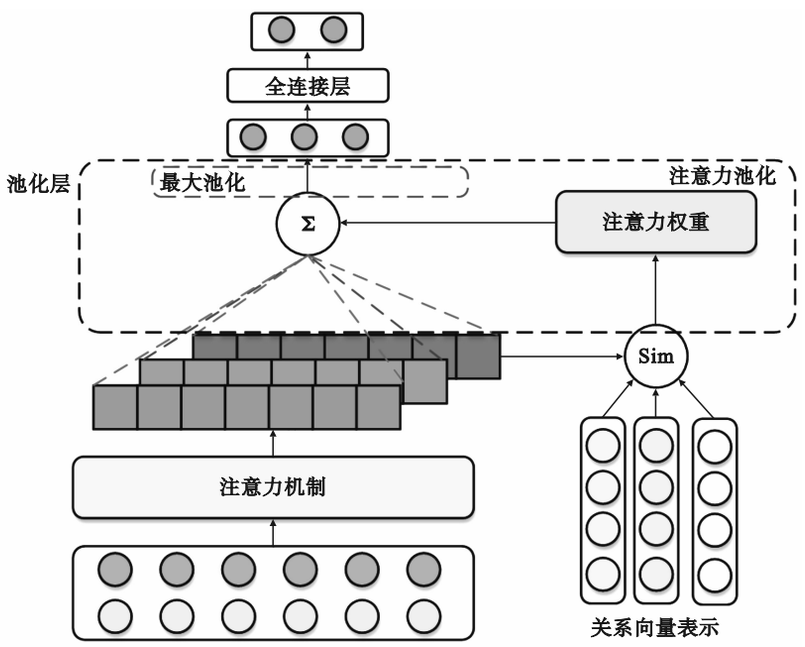

图3 APCNNA模型图

Fig. 3 APCNNA model diagram

APCNNA模型:它是由PCNN和两层注意力机制组成.一层注意力机制是在输入层,通过增加注意力机制能够减少无关词语的干扰,快速抓取重点实体,得到输入向量.另一层在PCNN的池化层,形成注意力池化层,用于获取不同语义特征和关系之间的相关性.这样能够增强相关特征的作用,抑制不相关特征,提升模型的抽取效果.

经过注意力机制得到的向量输入到卷积层中,则句子长度为l,卷积核为g,填充长度为$g-l$,卷积层输出特征向量为$\boldsymbol{c}$.在池化层中,由卷积层的特征向量$\boldsymbol{C}$与关系类型向量$\boldsymbol{R}$计算相似度,得到注意力权重向量$\boldsymbol{e}$,通过将$\boldsymbol{e}$和$\boldsymbol{W}$进行乘法运算得到向量$\boldsymbol{Y}$,进行池化操作得到长度为3的向量.最终将分段池化输出拼接并经过tanh激活函数得到最终输出$\boldsymbol{z}$.具体计算公式如下:

$$\boldsymbol{c}\in\mathbf{R}^{l+g-1},c_j=gq_{j-l+1},1\leqslant j\leqslant l+g-1, \tag{15}$$

$$\boldsymbol{e}_n=\mathrm{softmax}(C_k\boldsymbol{F}R_n), \tag{16}$$

$$\boldsymbol{Y}=\boldsymbol{e}_i^{\mathrm{T}}\cdot\boldsymbol{W}_i, \tag{17}$$

$$p_{ij}=\max(Y_{ij}),1\leqslant i\leqslant n,1\leqslant j\leqslant 3, \tag{18}$$

$$p_i=\{p_{i1},p_{i2},p_{i3}\}, \tag{19}$$

$$\boldsymbol{z}=\tanh(p_{1:n}),\boldsymbol{z}\in\mathbf{R}^{3n}. \tag{20}$$

其中:k是使用卷积核的数量;n是定义的关系向量总数,本文中n=7;$p_{i:n}$为分段后的所有池化输出p_i进行拼接后的池化层总输出;$\boldsymbol{F}$为参数向量;$\boldsymbol{W}_i$为句子特征向量.

输出层:使用softmax函数,将结果转化为概率分数,输出记为o,具体表达式如下:

$$o=\mathrm{soft\,max}(w'\boldsymbol{z}+b). \tag{21}$$

其中:w'表示权重;b表示偏置.

通过输出层得到句子的关系标签,计算$F1$值作为评价指标,将结果反馈给强化学习部分进行优化,避免局部最优情况.同时,通过二者的交互,提高关系抽取模型的精确率.为防止过拟合,引入Dropout策略,随机隐藏神经元节点,提高模型的泛化能力.设置Dropout策略值为0.5,完成整个模型的训练过程.最终使用训练好的模型进行航空事故文本的关系抽取.

2.3　关系抽取结果和存储

在完成实体对之间的关系抽取后,要将获取的实体对和关系通过三元组形式来完成数据的整合和存储.使用Neo4j图基于键值对完成数据存储.将航空事故文本经实体关系抽取后得到的实体数据和关系数据转换成csv表格数据,分别命名为Air.csv和Relation.csv,将这2个csv格式的数据导入数据库,得到航空事故的知识图谱.

3　实验与分析

本文的所有实验都基于Ubuntu 18.04的操作系统,内存为512 GB,显卡为NVIDA RTX 3090,显存为24 GB.本实验所用的数据主要是通过查阅中国民用航空安全信息系统和航空安全自愿报告系统上的公开航空安全事故报告构建的.程序基于Python3.6及TensorFlow1.14.0进行仿真.

3.1　基于BERT嵌入的改进BiGRU-IDCNN-CRF模型仿真分析

3.1.1　评价指标

模型的评价指标有:精确率(precision,记为P)、召回率(recall,记为R)和$F1$值($F1$-measure),公式如下:

$$P=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \tag{22}$$

$$R=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \tag{23}$$

$$F1=\frac{2PR}{P+R}. \tag{24}$$

其中:TP表示实际为正类的样本被正确地分类为正类的数量;FP表示实际为负类的样本被错误地分类为正类的数量;FN表示实际为正类的样本被错误地分类为负类的数量.

3.1.2　参数设置与数据集说明

实体识别模型的参数设置如下:BERT_base层数为12,BERT_base隐层为768,最大序列长度为128,隐层大小为120,学习率为0.005,Dropout值为0.5,优化方法采用Adam算法.

DatasetQ数据集构建基于中国民用航空安全信息系统网站和航空安全自愿报告系统网站上的公开信息.首先,通过对原始文本进行清洗与标注,形成基础语料.随后,该基础语料经历了两阶段的数据增强流程:第一阶段采用EDA(exploratory data analysis)方法以增加数据的多样性;第二阶段在此基础上,进一步应用TF-IDF方法进行深度语义层面的扩充与生成,最终形成了总规模达43 224条数据的数据集,该数据集按4∶1的比例划分为训练集与测试集,用于后续的模型训练与性能验证.

3.1.3　不同实体模型的对比实验

为了能够更好地体现本文提出的识别模型性能,将其分别与CRF,GRU,IDCNN,BiGRU等9种模型进行对比,在其他条件都相同的情况下,使用DatasetQ数据集来验证本文提出的模型.实验中各项参数由上述实验获取到的数据设置,实验最终结果如表2所示.

表2　不同命名实体识别模型的对比结果

Table 2　Comparison results of different named entity recognition models　　%

模型	P	R	$F1$
CRF	83.68	84.76	84.22
GRU	85.48	86.62	86.05
IDCNN	83.23	79.56	81.35
BiGRU	86.56	87.23	86.89
BiGRU-CRF	88.80	87.16	87.97
IDCNN-CRF	87.39	81.64	84.41
BERT-BiGRU-CRF	92.20	91.00	91.59
BERT-IDCNN-CRF	91.80	88.10	89.91
BERT-BiGRU-IDCNN-CRF	93.15	90.39	91.74
BERT-改进BiGRU-IDCNN-CRF	**94.69**	**91.72**	**93.18**

注:黑体表示结果最好.下同.

实验结果表明,基于神经网络的实体识别方法比基于统计机器学习的方法效果更好,精确率和$F1$值有所提升,因为神经网络能够学习到更多信息并提取更精确的特征.此外,BiGRU网络的效果略好于GRU,因为它能够同时考虑句子的上下文信息,实现更准确的全局特征提取.CRF模型能够提升模型效果,因为CRF能够解决神经网络输出标签不合理的问题,提高精确率和$F1$值.引入BERT模型能够显著提高模型的精确率、召回率和$F1$值.将BiGRU网络与IDCNN网络结合并加入贡献因子改进后,最终BERT-改进BiGRU-IDCNN-CRF模型表现出很好的性能,精确率提高了1.54%,相比传统CRF模型提高了

11.01%,$F1$值同时也提高了8.96%.整体模型性能较好,实体识别的效果也有所提升.

3.1.4 稳定性对比实验

为了验证BERT-改进BiGRU-IDCNN-CRF模型的稳定性,随机选取6 000,8 000,10 000,12 000,14 000条训练数据进行实验,测试数据为统一的10 000条数据.在这个环境下将本文提出的模型与BERT-BiGRU-CRF模型以及BERT-IDCNN-CRF模型进行了对比,实验结果见图4.

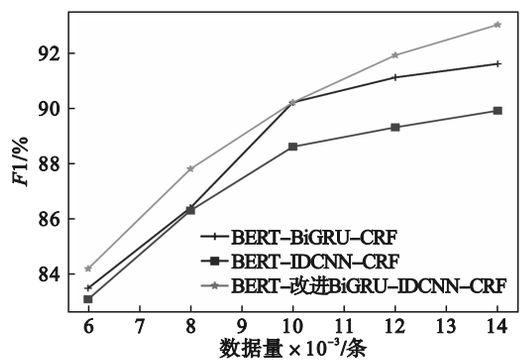

图4 稳定性实验结果对比

Fig. 4 Comparison of stability experimental results

通过图4可以看出,在5组训练数据中,BERT-改进BiGRU-IDCNN-CRF模型除了在数据集为10 000条时出现了和BERT-BiGRU-CRF模型相同的指标值,其余情况均优于对比模型.无论数据集为多大,其都能表现出很好的性能,且数据集数量越多,其性能越好.因此,验证了本文提出的方法具有一定的稳定性.

3.2 基于强化学习的聚类远程监督关系抽取仿真分析

本节采用的数据集与3.1节数据集相同,共有56 900条数据,包含预定义的3 000组正样本实体对和2 900组随机选取的负样本.将数据按照8∶2的比例划分为训练集和测试集.预定义的六类关系包括实体间包含关系、同义关系、原因关系、结果关系、属性关系和其他关系,数据量分别为8 265,9 256,12 653,11 636,10 089,9 578条.

3.2.1 评价指标

评价指标采用微平均值Micro_P,Micro_R和Micro_$F1$.微平均值是分析多实例多标签的重要指标,其在P,R,$F1$的基础上分别计算每一个TP,FP,TN,FN类的值,再将它们加和得到新的值,计算公式如下:

$$\text{Micro\_}P=\frac{\sum_j \text{TP}_j}{\sum_j \text{TP}_j+\text{FP}_j}, \tag{25}$$

$$\text{Micro\_}R=\frac{\sum_j \text{TP}_j}{\sum_j \text{TP}_j+\text{FN}_j}, \tag{26}$$

$$\text{Micro\_}F1=\frac{2\times\text{Micro\_}P\times\text{Micro\_}R}{\text{Micro\_}P+\text{Micro\_}R}. \tag{27}$$

式中,j代表第j个实际值或第j个预测值.

3.2.2 参数设置

关系抽取模型的学习率和Dropout值与实体识别模型一致,所使用的参数是根据APCNNA关系抽取模型所设定的,参数设置如下:句子最大长度为120,卷积窗口大小为3,字向量维度为100,位置向量维度为5,卷积核数量为100.

3.2.3 消融实验

为了验证双层注意力机制的有效性,本文进行了4种模型的消融实验:PCNN本身、PCNN加上字向量注意力机制(ATT_Z+PCNN)、注意力机制池化层的PCNN(PCNN+ATT_C)以及本文提出的模型APCNNA.实验结果如表3所示.

表3 有无注意力机制模型的不同表现对比

Table 3 Comparison of different performance of models with and without attention mechanism %

模型	Micro_P	Micro_R	Micro_$F1$
PCNN	75.27	70.18	72.48
ATT_Z+PCNN	76.27	81.03	78.50
PCNN+ATT_C	80.12	78.86	79.48
APCNNA	**81.59**	**83.26**	**82.42**

根据表3可知,加入字向量注意力机制和池化层注意力机制都能提高模型性能.在PCNN池化层加入注意力机制后,其性能提升较ATT_Z+PCNN略高,因为该机制能使整个池化层更加精准地输出航空事故实体关系;与其他模型相比,APCNNA模型在Micro_$F1$值上表现最优,因为它综合了前两者的优点并进一步提升了整体性能.

3.2.4 数据标注方法对比实验

本文提出的聚类远程监督方法主要目的是提供一个噪声较低的数据集,提高整体模型的有效性.因此暂不考虑后续的神经网络模型与强化学习部分,采用普通的远程监督方法Mintz[21](采用强假设标记的传统远程监督方法)、MultiR[22]和MIMLRE[23](多实例方法和多实例多标签方法)在改进前后的数据集上进行对比.实验结果如表4所示.

表4　聚类远程监督方法效果

Table 4　Effect of clustering distant supervision method

方法	传统远程监督				聚类远程监督			
	Micro_*P*/%	Micro_*R*/%	Micro_*F*1/%	*t*/h	Micro_*P*/%	Micro_*R*/%	Micro_*F*1/%	*t*/h
Mintz	39.80	35.20	37.36	2	40.15	38.23	39.17	2.2
MultiR	54.20	40.15	46.13	1	50.16	48.74	49.44	1.2
MIMLRE	42.60	36.90	39.54	4	45.20	40.10	42.50	4.2

由表4可以看出,本文提出的聚类远程监督方法对实验数据集具有一定的改善作用,*F*1值相比于传统远程监督在各个方法上都有3%左右的提升.说明在通过聚类之后,数据集中的一些错误标注有所减少,提高了抽取的精确率;同时改进K-means算法中聚类中心的选取考虑了航空文本句子全局和局部联系,使得召回率在一定程度上得到提高,但是标注时间会略长一些,然而相比于标注数据完成后使用的神经网络关系抽取模型训练时间来说,这部分标注时间的消耗对整体影响是十分微小的.因此,就整体效果而言,本文提出的聚类远程监督获取的标注数据具有一定的优势.

3.2.5　关系抽取方法对比实验

为了验证APCNNA+RL模型性能,在本文提出的聚类远程监督获取的标注数据集基础上,将提出的模型(APCNNA+RL)与主流的关系抽取模型进行对比.实验结果如表5所示.

表5　不同模型的对比结果

Table 5　Comparison results of different models

%

模型	Micro_*P*	Micro_*R*	Micro_*F*1
MultiR	50.16	48.74	49.44
MIMLRE	45.20	40.10	42.50
BGWA	69.46	71.24	70.34
PCNN+ONE	72.56	74.92	73.72
PCNN	75.27	70.18	72.48
APCNNA	81.59	83.26	82.42
APCNNA+RL	**84.16**	**83.41**	**83.96**

由表5可以看出,根据前2个基于特征方程模型和后面4个神经网络模型对比说明神经网络方法在解决有噪声标签的关系抽取问题时效果更好;对比PCNN+ONE和PCNN可知,利用多个句子降噪比仅使用概率最大的句子更有效;APCNNA与PCNN相比,APCNNA的特征提取效果更好;APCNNA和APCNNA+RL相比,结合强化学习的APCNNA网络模型降噪效果很好,精准率相比于单独的APCNNA模型提升了2.57%,这是因为强化学习与关系抽取模型的交互能够减少大量的噪声.

4　结　　语

构建航空事故的知识图谱对提升航空事故应急管理能力具有重要意义.航空事故知识抽取是构建航空事故知识图谱的关键.本文以航空事故知识图谱构建的知识抽取为切入点,提出了基于BERT的改进的BiGRU-IDCNN-CRF命名实体识别模型,并且整体模型的融合改进能够更准确地识别出航空事故文本实体,在解决航空事故文本存在问题的同时提高了识别精确率;提出了基于强化学习的去噪APCNNA关系抽取模型.实验结果表明,提出的命名实体识别模型实现高达94.69%的精确率,关系抽取模型实现84.16%的精确率,因此,本文的方法在航空事故文本中具有有效性和实用性.

参考文献:

[1] Pujara J, Miao H, Getoor L, et al. Knowledge graph identification[C]//The Semantic Web-ISWC 2013. Berlin, Heidelberg: Springer, 2013: 542-557.

[2] 张汝佳, 代璐, 王邦, 等. 基于深度学习的中文命名实体识别最新研究进展综述[J]. 中文信息学报, 2022, 36(6): 20-35.
(Zhang Ru-jia, Dai Lu, Wang Bang, et al. Recent advances of Chinese named entity recognition based on deep learning [J]. *Journal of Chinese Information Processing*, 2022, 36(6): 20-35.)

[3] Goller C, Kuchler A. Learning task-dependent distributed representations by backpropagation through structure[C]// Proceedings of International Conference on Neural Networks (ICNN'96). Washington DC, 1996: 347-352.

[4] Gers F A, Schmidhuber J, Cummins F. Learning to forget: continual prediction with LSTM[J]. *Neural Computation*, 2000, 12(10): 2451-2471.

[5] Qin Q L, Zhao S, Liu C M. A BERT-BiGRU-CRF model for entity recognition of Chinese electronic medical records [J]. *Complexity*, 2021, 2021: 6631837.

[6] Li Z R, Hu C M, Guo X H, et al. An unsupervised multiple-task and multiple-teacher model for cross-lingual named entity recognition [C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Dublin, 2022: 170-179.

[7] Boudjellal N, Zhang H P, Khan A, et al. ABioNER: a

BERT-based model for Arabic biomedical named-entity recognition[J]. *Complexity*, 2021, 2021: 6633213.
[8] Zhou R, Li X, Bing L D, et al. Improving self-training for cross-lingual named entity recognition with contrastive and prototype learning [C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Toronto, 2023: 4018-4031.
[9] Ma J, Ballesteros M, Doss S, et al. Label semantics for few shot named entity recognition [C]//Findings of the Association for Computational Linguistics: ACL 2022. Dublin, 2022: 1956-1971.
[10] 鄂海红, 张文静, 肖思琪, 等. 深度学习实体关系抽取研究综述[J]. 软件学报, 2019, 30(6): 1793-1818.
(E Hai-hong, Zhang Wen-jing, Xiao Si-qi, et al. Survey of entity relationship extraction based on deep learning [J]. *Journal of Software*, 2019, 30(6): 1793-1818.)
[11] He H, Ganjam K, Jain N, et al. An insight extraction system on BioMedical literature with deep neural networks [C]//Proceedings of the 2017 Conference on Empirical Methods in NaturalLanguage Processing. Copenhagen, 2017: 2691-2701.
[12] Zhang M S, Zhang Y, Fu G H. End-to-end neural relation extraction with global optimization [C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, 2017: 1730-1740.
[13] Yang S Z, Liu Y X, Zhang K W, et al. Overview of remote supervision relationship extraction [J]. *Journal of Computer Science*, 2021, 44(8): 1636-1660.
[14] 许鑫冉, 王腾宇, 鲁才. 图神经网络在知识图谱构建与应用中的研究进展[J]. 计算机科学与探索, 2023, 17(10): 2278-2299.
(Xu Xin-ran, Wang Teng-yu, Lu Cai. Research progress of graph neural network in knowledge graph construction and application [J]. *Journal of Frontiers of Computer Science and Technology*, 2023, 17(10): 2278-2299.)
[15] 郑志蕴, 徐亚媚, 李伦, 等. 融合位置特征注意力与关系增强机制的远程监督关系抽取[J]. 小型微型计算机系统, 2023, 44(12): 2678-2684.
(Zheng Zhi-yun, Xu Ya-mei, Li Lun, et al. Distantly supervised relation extraction with position feature attention and relation enhancement [J]. *Journal of Chinese Computer Systems*, 2023, 44(12): 2678-2684.)
[16] Wang G Y, Zhang W, Wang R X, et al. Label-free distant supervision for relation extraction via knowledge graph embedding [C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, 2018: 2246-2255.
[17] Ji G L, Liu K, He S Z, et al. Distant supervision for relation extraction with sentence-level attention and entity descriptions [C]// Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. San Francisco, 2017: 3060-3066.
[18] Chen T, Shi H Z, Tang S L, et al. CIL: contrastive instance learning framework for distantly supervised relation extraction [C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics. Online, 2021: 6191-6200.
[19] Luo B F, Feng Y S, Wang Z, et al. Learning with noise: enhance distantly supervised relation extraction with dynamic transition matrix [C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, 2017: 430-439.
[20] Zhou Y R, Pan L M, Bai C Y, et al. Self-selective attention using correlation between instances for distant supervision relation extraction [J]. *Neural Networks*, 2021, 142: 213-220.
[21] Mintz M, Bills S, Snow R, et al. Distant supervision for relation extraction without labeled data [C]//Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. Singapore, 2009: 1003-1011.
[22] Hoffmann R, Zhang C L, Ling X, et al. Knowledge-based weak supervision for information extraction of overlapping relations [C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland, 2011: 541-550.
[23] Surdeanu M, Tibshirani J, Nallapati R, et al. Multi-instance multi-label learning for relation extraction [C]// Conference on Empirical Methods in Natural Language Processing. Jeju Island, 2012: 455-465.

(上接第 88 页)

[17] 黄锦灏, 丁钰真, 肖亮, 等. 一种基于强化学习的嵌入式系统抗拒绝服务攻击的缓存调度方案[J]. 计算机科学, 2020, 47(7): 282-286.
(Huang Jin-hao, Ding Yu-zhen, Xiao Liang, et al. Reinforcement learning based cache scheduling against denial-of-service attacks in embedded systems [J]. *Computer Science*, 2020, 47(7): 282-286.)
[18] 欧东阳, 张开强, 陈圣蕾, 等. 基于PCPEC的数据中心功耗攻击防御策略[J]. 计算机科学, 2022, 49(12): 374-380.
(Ou Dong-yang, Zhang Kai-qiang, Chen Sheng-lei, et al. Data center power attack defense strategy based on PCPEC [J]. *Computer Science*, 2022, 49(12): 374-380.)
[19] 周知, 刘方明. 面向多租户数据中心资源回收利用的能效激励机制[J]. 中国科学: (信息科学), 2021, 51(5): 735-749.
(Zhou Zhi, Liu Fang-ming. An energy-efficient incentive mechanism for resource recycling in multi-tenant datacenters [J]. *Scientia Sinica* (*Informationis*), 2021, 51(5): 735-749.)
[20] 杜楠, 彭文成, 谭亚新. 基于SAAS的LVC装备试验资源实时调度模型研究[J]. 系统工程与电子技术, 2023, 45(12): 3924-3931.
(Du Nan, Peng Wen-cheng, Tan Ya-xin. Research on real-time scheduling model of LVC equipment test resources based on SAAS [J]. *Systems Engineering and Electronics*, 2023, 45(12): 3924-3931.)