文章编号: 1671-251X(2024)01-0088-08 DOI: 10.13272/j.issn.1671-251x.2023060014

基于煤矿井下不安全行为知识图谱构建方法

付燕， 刘致豪， 叶鸥

（西安科技大学 计算机科学与技术学院，陕西 西安 710054）

摘要: 虽然知识图谱已广泛应用于各个领域，但在煤矿安全方面，尤其在煤矿井下不安全行为方面的研究较少。构建了一种自底向上的煤矿井下不安全行为知识图谱。首先，采用传统机器学习和深度学习算法相结合的方法进行命名实体识别，采用 RoBERTa 进行词语向量化，采用双向长短时记忆网络(BiLSTM)对向量进行标注，提高网络模型对上下文特征的捕捉能力，通过多层感知机(MLP)解决煤矿井下不安全行为数据集数据量不足的问题，采用条件随机场(CRF)模型解决前面存在的单词关系不识别问题，并捕获全文信息和预测结果。其次，根据语句的结构特点，设计了基于知识“实体-关系-实体”三元组的依存句法树结构，对井下不安全行为领域的知识资源进行知识抽取与表示。最后，构建面向井下不安全行为的知识图谱。实验结果表明：① RoBERTa-BiLSTM-MLP-CRF 模型对于导致结果、违反性行为、错误性行为及粗心性行为 4 类实体类别具有较好的识别效果，其准确率分别为 86.7%，80.3%，80.7%，77.4%。② 在相同的数据集下，RoBERTa-BiLSTM-MLP-CRF 模型训练的准确率、召回率、F_1 值较 RoBERTa-BiLSTM-CRF 模型分别提高了 1.6%，1.5%，1.6%。

关键词: 井下不安全行为；知识图谱；依存句法；命名实体识别；知识三元组；知识融合；知识存储；词语向量化

中图分类号: TD79 文献标志码: A

A method for constructing a knowledge graph of unsafe behaviors in coal mines

FU Yan, LIU Zhihao, YE Ou

(College of Computer Science and Technology, Xi'an University of Science and Technology, Xi'an 710054, China)

Abstract: Although knowledge graphs have been widely applied in various fields, there is relatively little research on coal mine safety, especially in the area of unsafe behavior underground. A bottom-up knowledge graph of unsafe behaviors in coal mines has been constructed. Firstly, a combination of traditional machine learning and deep learning algorithms is used for named entity recognition. RoBERTa is used for word vectorization. The bidirectional long short term memory network (BiLSTM) is used to annotate the vectors, improving the network model's capability to capture contextual features. To solve the problem of insufficient data volume in the dataset of unsafe behaviors in coal mines, a multi-layer perceptron (MLP) is used. The conditional random field (CRF) model is adopted to solve the problem of unrecognized word relationships and capture full-text information and prediction results. Secondly, based on the structural characteristics of the statements, a dependency syntax tree structure based on the knowledge "entity - relationship - entity" triplet is designed to extract and represent knowledge resources in the field of unsafe behavior underground. Finally, a knowledge graph of unsafe behaviors underground is constructed. The experimental results show that the RoBERTa-

收稿日期: 2023-06-06; **修回日期**: 2024-01-08; **责任编辑**: 王晖，郑海霞。

基金项目: 中国博士后科学基金项目(2020M673446)。

作者简介: 付燕(1972—)，女，河南鹤壁人，教授，博士，主要研究方向为计算机图形图像处理技术、科学计算及其可视化技术等，E-mail: 942542352@qq.com。通信作者：刘致豪(1997—)，男，河南商丘人，硕士研究生，主要研究方向为知识图谱，E-mail: 2267318289@qq.com。

引用格式: 付燕，刘致豪，叶鸥. 基于煤矿井下不安全行为知识图谱构建方法[J]. 工矿自动化，2024，50(1)：88-95.

FU Yan, LIU Zhihao, YE Ou. A method for constructing a knowledge graph of unsafe behaviors in coal mines[J]. Journal of Mine Automation, 2024, 50(1): 88-95.

扫码移动阅读

BiLSTM-MLP-CRF model has good recognition performance for four types of entity categories: results, violating behavior, erroneous behavior, and careless behavior, with accuracy rates of 86.7%, 80.3%, 80.7%, and 77.4%, respectively. ② Under the same dataset, the accuracy, recall, and F_1 value of the RoBERTa-BiLSTM-MLP-CRF model training are improved by 1.6%, 1.5%, and 1.6%, respectively, compared to the RoBERTa-BiLSTM-CRF model.

Key words: unsafe underground behavior; knowledge graph; dependency syntax; named entity recognition; knowledge triplet; knowledge fusion; knowledge storage; word vectorization

0　引言

近年来虽然煤矿井下事故发生率逐年降低，但每年仍有较多的煤矿井下安全生产事故发生。据相关统计，由于工作人员的不安全行为导致的安全生产事故在中国煤矿井下安全生产事故中占比高达97.67%[1]。因此，研究井下工作人员的不安全行为对降低事故发生率、实现煤矿井下安全生产具有重要意义。

由于煤矿数据的复杂性，利用大数据安全管理系统难以实现结构化不安全行为知识的语义关联及知识推理。知识图谱拥有较好的知识结构性及较强的表达性，能更加直观地描述各类概念之间的关系，从而实现井下不安全行为数据挖掘。知识图谱按照构造方式的不同可分为基于规则的知识图谱构建方法、基于统计的知识图谱构建方法和基于深度学习的知识图谱构建方法 3 类。① 基于规则的知识图谱构建方法。N. Guarino 等[2]提出基于本体学的知识表示和推理方法 OntoClean，其通过定义本体的基本概念、属性和关系等方式来表示和推理知识，OntoClean 已广泛应用于语义 Web 和知识图谱的构建。但 OntoClean 只能处理简单、单一的知识，难以应用于丰富、复杂的知识领域中。Horrocks 等[3]提出 SWRL（A Semantic Web rule language combining OWL and RuleML），该方法可与 OWL（Web Ontology Language）等本体语言结合使用，以表示更加丰富和复杂的知识，可处理多层次和不对称的语义关系。但 SWRL 和 OWL 这 2 种基于规则的方法需领域专家对知识进行抽象和分类，且需手动构建规则和逻辑表达式，知识图谱的构建过程较耗时和复杂，且缺乏自适应性。② 基于统计的知识图谱构建方法。A. Bordes 等[4]提出了一种基于超平面转换的知识图谱嵌入方法，称为 TransE，该方法使用向量空间中的超平面来表示实体和关系之间的转换，以便在低维空间中对知识图谱进行建模。但该方法只能处理单一类型的实体。 Wang Zhen 等[5]对 TransE 进行了扩展，提出了一种适用于含有异质实体的知识图谱嵌入方法，称为 TransH，该方法将实体投影到不同的超平面上，以处理不同类型的实体。但基于统计的知识图谱构建方法只能对语言表面的信息进行提取，难以理解语言中的隐含信息和语义，难以准确捕捉实体之间的关系。③ 基于深度学习的知识图谱构建方法。刘文聪等[6]采用双向长短时记忆（Bidirectional Long Short-Term Memory, BiLSTM）模型与条件随机场（Conditional Random Field, CRF）模型相结合的方式抽取中文地质时间信息，在一定程度上解决了传统方法特征提取不足的问题。吴闯等[7]利用 BERT（Bidirectional Encoder Representations from Transformers）−BiLSTM−CRF 模型对航空发动机设备润滑系统进行命名实体识别，先利用 BERT 模型进行词向量化，再进行实体识别，在一定程度上改善了实体识别的效果。然而，传统的 BERT 模型在进行词语向量化时易造成大量实体和语义丢失。

虽然知识图谱已广泛应用于各个领域，但在煤矿安全方面，尤其在煤矿井下不安全行为方面的研究较少。因此，本文提出了一种基于煤矿井下不安全行为知识图谱构建方法。首先，针对煤矿井下不安全行为的命名实体识别问题，结合现有的知识，用传统机器学习和深度学习算法相结合的方法进行命名实体识别，采用 RoBERTa（Robustly Optimized BERT pretraining Approach）进行词语向量化后，通过 BiLSTM 对向量进行标注，提高网络模型对上下文特征的捕捉能力。其次，根据语句的结构特点，设计了基于知识三元组的依存句法树结构，并根据该数据结构对井下不安全行为领域的知识资源进行知识抽取与表示。最后，利用图数据库 Neo4j 存储煤矿井下不安全行为知识，形成井下不安全行为知识图谱。

1　相关理论方法

知识图谱的主要任务是使用符号的方式去描述本体的概念及其相互之间的关系，其本身是具有属性的实体通过关系链接而成的网状知识库。其基本组成单位是“实体−关系−实体”及“实体−属性−属性值”三元组[8-10]。当前，知识图谱主要分为自顶向下

及自底向上 2 种构建方式。

1.1　自顶向下的知识图谱构建方法

自顶向下的知识图谱构建方法是从较高质量的结构化数据源中获取数据资源，并根据结构化数据源中预先定义的实体关系来构建完整的知识图谱[11-12]。自顶向下的知识图谱构建分为以下 3 个步骤：① 通过大量结构化数据源完成本体知识库的构建，包括本体学习和相应规则制定。② 进行实体学习，主要包括实体链接和实体填充 2 项任务。③ 构建图谱。

1.2　自底向上的知识图谱构建方法

自底向上的知识图谱构建方法是从大量知识密度小且没有固定关系的半结构化[13-14]、非结构化数据源中获取知识资源，从而完成知识图谱的构建。自底向上的知识图谱构建主要包含知识抽取、知识融会及图谱构建 3 个步骤。其中知识抽取包含实体识别、关系抽取及属性抽取 3 个任务，知识融会的主要任务是进行实体消歧。

2　知识图谱构建方法

由于本文采用的是开放数据源，其中包含大量半结构化、非结构化数据，故而采用自底向上的知识图谱构建方法。

2.1　数据的获取、预处理

本文采用的数据源主要为开放的文献知识资源及《煤矿安全规程》中的相关规定。其中文献知识资源是从知网中主题或关键词为“不安全行为 煤矿”检索得到的文献。经筛选，保留其中 210 篇作为实验数据。本文采用 BIO（Beginning−Inside−Outside）标准标注策略对不安全行为实体进行标注。通过参考中国国家标准化管理委员会发布的煤矿科技术语汇总表，对文献[1]、文献[15]中关于不安全行为的研究内容进行分析，将井下不安全行为实体分为遗忘性行为、粗心性行为、错误性行为、违反性行为、关联因素影响行为及导致后果 6 种，见表 1。将属于一个命名实体开始的 token 标记为 B−label，对于属于命名实体类型但不是第 1 个字的 token 标记为 I−label，其他不属于命名实体范围的统一用 O 进行标记。

2.2　实体识别

针对井下不安全行为实体识别中实体数量庞大、交替频繁、语义复杂等问题，需选择合适的命名实体识别方法。基于监督的统计学习方法在实体识别过程中依赖大型标注语料库进行模型训练，不适合没有专业大型语料库的井下不安全行为，容易出现实体识别不准确的情况。因此，本文采用改进神经网络模型实现井下不安全行为实体识别。在 BiLSTM−CRF 基础上引入 RoBERTa 及多层感知机（Multilayer Perceptron，MLP）作为井下不安全行为命名实体识别模型（RoBERTa−BiLSTM−MLP−CRF）。将预处理后的数据分为训练集和测试集，训练集通过 RoBERTa 模型将输入的文本序列转换为具有丰富上下文语义的词向量，RoBERTa 模型的输出向量作为 BiLSTM 模型的输入，以提取上下文的特征值。由于所获得的煤矿井下不安全行为语料数据量少，为了获得更好的模型训练效果，在 BiLSTM 层与 CRF 层中间加入 MLP，并将开源数据集的输出维度与煤矿数据集输出维度进行统一，达到迁移学习的目的。CRF 模型用于标注输入注释序列的实体。具体实体识别流程如图 1 所示。

表 1　实体待预测标签

Table 1　Entity to be predicted labels

实体类型	开始标签	中间或结尾标签
遗忘性行为	B−forget	I−forget
粗心性行为	B−careless	I−careless
错误性行为	B−error	I−error
违反性行为	B−violate	I−violate
关联因素影响性行为	B−factor	I− factor
导致后果	B−cause	I−cause

2.2.1　RoBERTa 模型

RoBERTa 模型是一种基于 Transformer 神经网络的预训练模型。当前，基于神经网络的预训练技术主要分为静态词向量与动态词向量 2 大类。① 静态词向量。Word2Vec[16]词向量模型能从大规模语料库中得到高精度的词向量。Glove[17]模型结合了 Word2Vec 及矩阵分解模型（Singular Value Decomposition，SVD）的优点，训练速度显著提高。静态词向量模型在一定程度上可得到较为精准的词向量，但无法解决一词多义的问题。② 动态词向量。ELMo 模型[18]采用长短时记忆（Long Short-Term Memory，LSTM）模型，在一定程度上解决了一词多义的问题。但 ELMo 模型采用的双向拼接特征融合方式比一体化的融合方式要弱。BERT 模型[19]采用双向语言模型、掩码语言模型（Masked Language Model，MLM）和 NSP（Next Sentence Prediction）3 种技术，在现阶段自然语言领域中被广泛应用，但 BERT 庞大的参数量使得实际应用面临困难。RoBERTa 模型对 BERT 模型的超参数进行改进，与 BERT 模型相比，RoBERTa 模型拥有更优越的模型性能。RoBERTa采用动态掩码的方式学习不同的特征，解决了传统 BERT 训练时大量短语和实体丢失

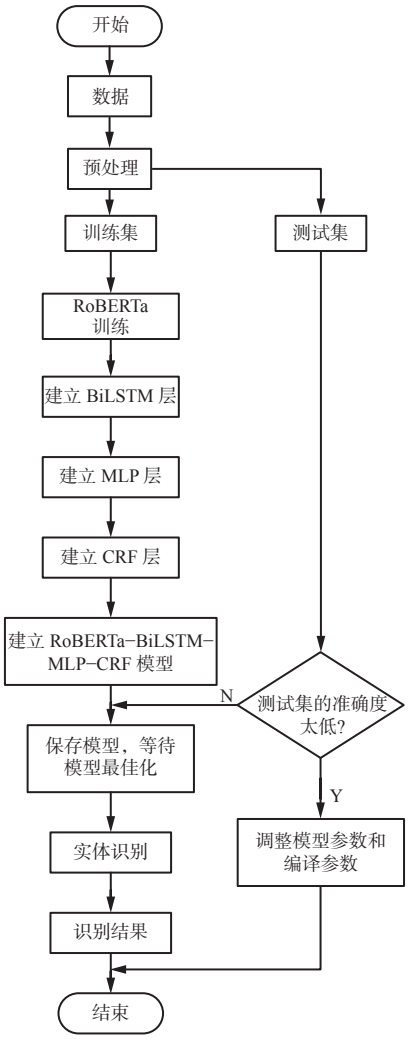

图 1　基于 RoBERTa−BiLSTM−MLP−CRF 实体识别过程

Fig. 1　RoBERTa-BiLSTM-MLP-CRF based entity recognition

的问题。由于煤矿井下不安全行为文本数据比较复杂，存在大量一词多义的现象，导致实体识别效果较差，因此，本文选择 RoBERTa 作为词向量抽取模型，其模型如图 2所示，其中 X_1—X_4 为词的向量化特征，E_1—E_4 为输入文本序列。

2.2.2　BiLSTM 模型

LSTM 模型在进行文本特征提取时，利用其复杂的网络结构可较好地捕获长距离依赖关系，但对于输入信息无法进行反方向解码，不能捕获双向语义依赖关系。煤矿井下不安全行为文本数据具有冗余特性，其数据文本语句通常较长且关系复杂。因此，提出 BiLSTM 模型，如图 3 所示，X_t 为当前时刻 t 的词向量化特征，h_t 为当前时刻 t 的隐藏状态，表示 BiLSTM 模型的输出结果。BiLSTM 模型在命名实体识别模型中的作用是捕获文本序列的上下文特征，对双向语义依赖关系进行捕捉。

2.2.3　MLP 模型

由于煤炭领域数据的复杂性，能够收集到的煤

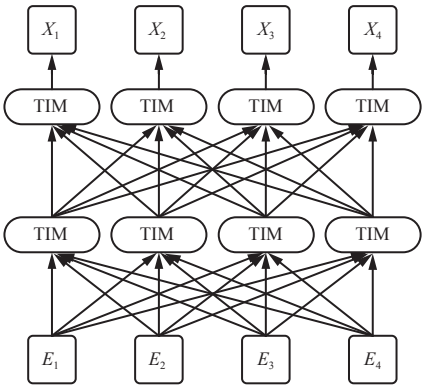

图 2　RoBERTa 模型

Fig. 2　RoBERTa model

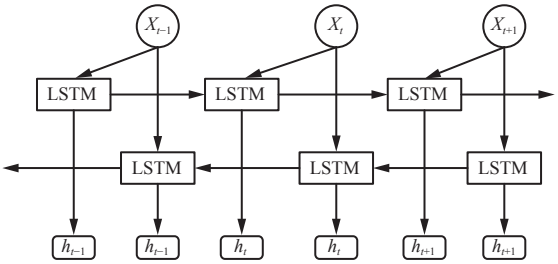

图 3　BiLSTM 模型

Fig. 3　BiLSTM model

矿井下不安全行为数据量较小，模型训练结果相对较差。为解决该问题，本文在 BiLSTM 层与 CRF 层中间加入 MLP[20]，将开源数据集输出维度与煤矿数据输出维度进行统一，利用知识迁移的方式弥补数据量不足的问题。首先，通过 RoBERTa、BiLSTM 与清华大学的开源数据集 THUCNews 进行训练，得到 1 个初始模型，该模型已获得 THUCNews 数据集中包含的一些特征参数，将其作为煤矿数据集训练初始模型参数；其次，通过 MLP 将开源数据集 THUCNews 输出维度与煤矿数据集输出维度进行统一。MLP 模型结构如图 4 所示。

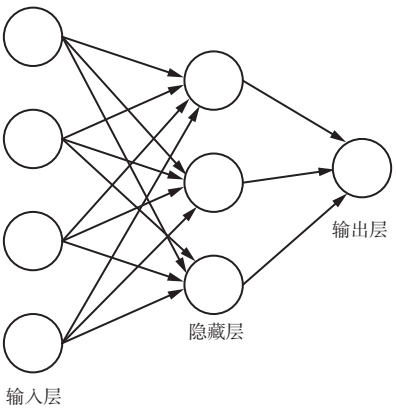

图 4　MLP 模型

Fig. 4　MLP model

2.2.4　CRF 模型

虽然经过 BiLSTM 及 MLP 模型之后输出的信息

是选择输出概率最高的标签，但没有考虑到不同单词之间的关系，输出的标签可能会混淆且缺乏逻辑。因此，引入 CRF 模型来解决单词关系不识别问题，并捕获全文信息和预测结果。该模型可表示为 $P(x|y)$，其中，x 为输入变量，表示输入的观测序列；y 为输出序列，表示对应 x 的标签序列。假设给定一个输入序列 $x=(x_1,x_2,\cdots,x_n)$和相应的标注序列 $y=(y_1,y_2,\cdots,y_n)$，且每个(x_i, y_i)对是线性链中最大团，若同时满足式(1)，则称$P(x|y)$为线性链的条件随机场。

$$P(y_i|x,y_1,y_2,\cdots,y_n)=P(y_i|x,y_{i-1},y_{i+1}) \tag{1}$$

式中：i 为当前字符所在位置；n 为输入句子长度；y_i 和y_{i-1}分别为当前单词的标签及前一个单词的标签。

给定预设的观测序列 x，CRF 模型求解隐态序列 y 的公式为

$$P(y|x)=\frac{1}{z(x)}\left(\sum_{i,j}\lambda_j t_j(y_{i-1},y_i,x,i)+\sum_{i,l}\mu_l s_l(y_i,x,i)\right) \tag{2}$$

$$z(x)=\exp\left(\sum_{i,j}\lambda_j t_j(y_{i-1},y_i,x,i)+\sum_{i,l}\mu_l s_l(y_i,x,i)\right) \tag{3}$$

式中：t_j为 i 处的传递特征；λ_j为t_j对应的权重；s_l为 i 处的状态特征；μ_l为s_l对应的权重；j 和 l 为特征函数的数量；$z(x)$为归化因子。

线性链 CRF 模型(图 5)对标签之间的约束关系进行预测，以此提高命名实体识别的准确性。

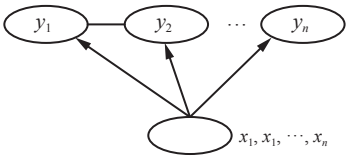

图 5　线性链 CRF 模型

Fig. 5　Linear chain CRF model

对每个单词进行评分，条件概率模型$P(x|y)$通过最大似然估计来计算。在实际预测过程中，对于给定的观测序列，计算其最大标签序列。评分公式为

$$s(y|x)=\sum_{j=1}^{m}\sum_{i=1}^{n}u_i f_j(x,i,y_i,y_{i-1}) \tag{4}$$

式中：u_i 为 i 处词向量的特征；f_j 为 u_i 对应的权重；m 为特征函数的总数量。

2.2.5　RoBERTa−BiLSTM−MLP−CRF 模型

RoBERTa−BiLSTM−MLP−CRF 模型如图 6 所示，其中 x_t 为当前时刻的输入特征。模型从下往上依次是字向量层 RoBERTa、融合层、Farward LSTM−Backward LSTM、输出层、MLP 和 CRF 层。该模型输入的是序列化文本，如图中输入层输入的文本“井下打架”。在 CRF 层输出相应的注释序列，输出序列采用 BIO 标注方式进行标注。

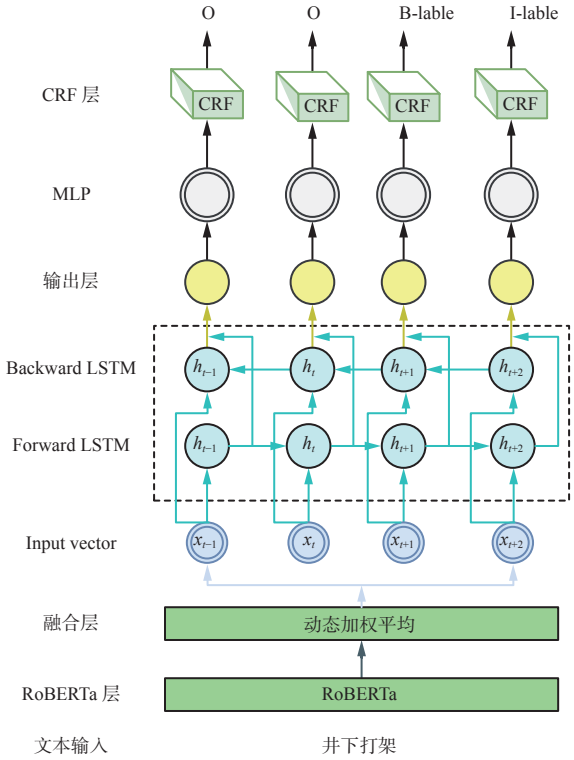

图 6　RoBERTa−BiLSTM−MLP−CRF 模型

Fig. 6　RoBERTa-BiLSTM-MLP-CRF model

2.2.6　模型评估标准

采用精确率 P、召回率 R 和 F_1 值 3 个标准来评价 RoBERTa−BiLSTM−MLP−CRF 模型对井下不安全行为实体识别的效果。

$$P=\frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}}+N_{\mathrm{FP}}} \tag{5}$$

$$R=\frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}}+N_{\mathrm{FN}}} \tag{6}$$

$$F_1=\frac{2PR}{P+R} \tag{7}$$

式中：N_{TP}为被预测为正样本的正样本数量；N_{FP}为被预测为正样本的负样本数量；N_{FN}为被预测为负样本的正样本数量。

2.3　关系抽取

本文数据来源于开放的相关文献及《煤矿安全规程》，其中《煤矿安全规程》中的文本数据为一条条规章制度，满足依存句法的单句中只能存在一个核心成分、每一个词语仅有一个依存对象、核心词不可与其两边的词产生依存关系等条件，且开放的文献文本知识一般高度凝练，故采用依存句法进行关系抽取。王志广等[21]在进行地址领域实体关系抽取

时提出联合抽取模型，该方法在一定程度上解决了并列句三元组抽取丰富的问题，但依然比较容易出现模式不匹配的现象，会造成大量知识不能被抽取。针对该问题，本文将句子的依存关系转换为语法树，分析比对三元组知识的枝条结构，利用树的遍历去搜索整个句子的语法树结构；并将每个并列句视为单独存在的句子，分步对其进行三元组抽取，更深度地抽取语句知识。

2.4　知识融会

知识融会的主要任务是对知识信息进行有效融合统一，将上述流程中得到的一些缺乏层次性与逻辑性的冗余信息及错误概念剔除，从而提高知识图谱数据库的知识质量[22]。知识融会主要包含实体消歧[23]和共指消解 2 个任务。实体消歧的任务是解决相同表述指代不同实体的问题。例如，“煤炭运输”在本文中指的是“井下劳作中的煤炭运输”，有的描述则是指“运货火车的煤炭运输”，因此，要联系上下文的语义，明确命名实体的确切含义。共指消解的任务是处理多种描述指代同一实体的问题，例如，“个体因素”“个体原因”“单人因素”均对应的是“个体因素”这一单元实体，在人工撰写的安全报告、事故报告中，用语不规范现象普遍存在。为解决此问题，本文采用余弦距离和 Jaccard 相关系数相结合的方式计算井下不安全行为实体之间的相似度。通过相似度确定对齐实体是否匹配，从而实现知识融会，得到统一规范的井下不安全行为实体名称。

$$S_{\text{consine}}(A,Q)=\frac{AQ}{\|A\|\|Q\|}=\frac{\sum_{i=1}^{n}A_iQ_i}{\sqrt{\sum_{i=1}^{n}A_i^2\sum_{i=1}^{n}Q_i^2}} \tag{8}$$

$$S_{\text{Jarccard}}(A,Q)=\frac{|A\cap Q|}{|A\cup Q|} \tag{9}$$

式中：S_{consine} 为余弦相似度；S_{Jarccard} 为 Jarccard 相似度；A 和 Q 为 2 个实体的属性字符串。

任意 2 个实体之间的语义相似度大小与余弦相似度和 Jarccard 相似度的大小成正比。井下不安全行为文本知识实体表述见表 2，可看出对于“不安全动作”和“不安全行为”2 个不同表述的实体，其 Jarccard 相似度 S_{Jarccard} 为 0.43，余弦相似度 S_{consine} 达到 0.60，进而得到“不安全动作”和“不安全行为”2 个实体实际上为同一概念，应该融合为同一实体。

2.5　知识存储

井下不安全行为文本数据经过上述流程处理后，从多元异构状态转换为结构化状态。知识存储的任务就是将各类知识存储为“实体−关系−实体”或“实体−关系−属性”的三元组形式。

表 2　实体相似度计算实例

Table 2　Example of entity similarity calculation

实体1	实体2	S_{consine}	S_{Jarccard}
粉尘瓦斯爆炸	粉尘瓦斯事故	0.67	0.50
违章指挥	违章命令	0.67	0.60
不安全动作	不安全行为	0.60	0.43
安全培训	安全训练	0.67	0.60

本文采用图数据库 Neo4j 来实现井下不安全行为知识的存储。考虑 Neo4j 只需插入节点与边就可实现数据的高效存储和查询[24]，利用带属性的图模型将实体存储为节点，实体属性存储为节点属性，边和边的属性表示关系与关系属性，标签表示描述知识的概念。基于 Neo4j 的知识存储方案见表 3。

表 3　基于 Neo4j 的知识存储方案

Table 3　Neo4j-based knowledge storage solutions

类型	作用	对象范围
节点	描述知识实体	井下扒车、穿化纤衣入井等
标签	描述知识概念类	违章指挥、违规操作等
边	描述实体关系	包含关系、关联关系等

3　实验结果与分析

3.1　模型参数设置

本次实验采用 TensorFlow1.15.5 框架进行模型的搭建，实验中批尺寸为 32，学习率为 0.001，迭代次数为 50。

3.2　实体识别结果

实验采用预处理的井下不安全行为文本语料库进行训练。基于该文本数据集，本文预定义了遗忘性行为、粗心性行为、错误性行为、违反性行为、关联因素影响行为、导致后果 6 种实体类型，识别效果见表 4。

由表 4 可看出，本文模型对于导致结果、违反性行为、错误性行为及粗心性行为 4 类实体具有较好的识别效果，其准确率分别为 86.7%，80.3%，80.7%，77.4%，对于遗忘性行为及关联因素影响性行为识别效果较差，其准确率分别为 63.5%，73.0%。这是因为导致后果、违反性行为、错误性行为及粗心性行为包含的实体表达形式较为固定，而遗忘性行为及关联因素影响性行为包含的实体语义复杂且较长，从而导致识别效果较差。

为了验证本文模型的有效性，将本文模型与 BiLSTM−CRF，BERT−BiLSTM−CRF，RoBERTa−

表 4　实体类型识别效果

Table 4　Entity type identification effect　　%

实体类别	P	R	F_1
遗忘性行为	63.5	67.4	65.4
粗心性行为	77.4	84.1	80.6
错误性行为	80.7	83.1	81.9
违反性行为	80.3	83.7	82.0
关联因素影响性行为	73.0	76.0	74.5
导致后果	86.7	90.0	88.3

BiLSTM−CRF 模型进行对比，结果见表 5。

表 5　模型对比结果

Table 5　Model contrast results　　%

模型	P	R	F_1
BiLSTM−CRF	71.2	74.8	73.0
BERT−BiLSTM−CRF	74.9	79.1	77.0
RoBERTa−BiLSTM−CRF	75.6	79.1	77.3
RoBERTa−BiLSTM−MLP−CRF	77.2	80.6	78.9

由表 5 可看出，BERT−BiLSTM−CRF 模型的准确率比 BiLSTM−CRF 模型提高了 3.7%，这表明进行实体识别之前进行词向量化是必要的；RoBERTa−BiLSTM−CRF 模型的准确率较 BERT−BiLSTM−CRF 模型提高了 0.7%，这表明 RoBERTa 模型比 BERT 模型更适合本次任务；RoBERTa−BiLSTM−MLP−CRF 模型的准确率、召回率、F_1 较 RoBERTa−BiLSTM−CRF 模型分别提高了 1.6%，1.5%，1.6%，这表明添加 MLP 后能够学习更多公共数据集的特征，用此模型对公共数据集进行训练，对于本次实验有正确的导向作用。

3.3　知识图谱构建结果

以井下不安全行为文本中的实体为节点，以实体之间的关系为边，将其存储在 Neo4j 图数据库中，从而构成煤矿井下不安全行为知识图谱。部分煤矿该图谱井下不安全行为知识图谱如图 7 所示。可看出该图谱通过“包含”“关联”等关系将不安全行为与影响因素及行为类别连接起来，通过“违规作业”等关系将行为实体与发生部门连接起来，构建了井下不安全行为不同实体间的相关关系，为煤矿井下进行员工管理提供了强有力的支持，进而提高了井下安全管理效率 。

4　结论

1）提出将句子的依存关系转化为语法树，分析比对三元组知识的枝条结构，利用树的遍历去搜索整个句子的语法树结构，实现煤矿井下知识三元组抽取。

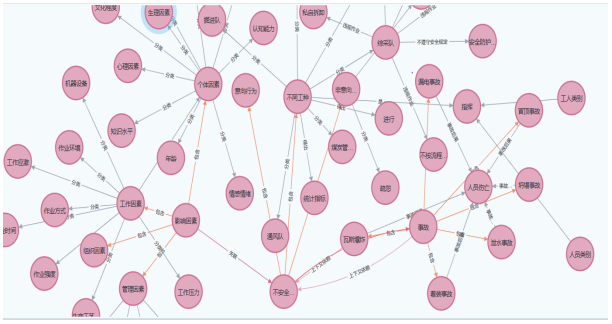

图 7　部分煤矿井下不安全行为知识图谱

Fig. 7　Knowledge graph of underground unsafe behavior in some underground coal mines

2）构建了煤矿井下不安全行为知识图谱，为煤矿井下进行安全管理提供了强有力的支持，进而提高了煤矿井下安全管理效率。

3）在构造煤矿井下不安全行为命名实体识别与知识三元组抽取时，由于收集文本数据集只包含部分煤矿井下不安全行为，使得命名实体识别与知识三元组抽取具有局限性且不可避免地会出现缺失和错误。因此，下一步将逐步补充和完善煤矿井下不安全行为知识体系。

参考文献（References）：

[1] 黄辉，张雪. 煤矿员工不安全行为研究综述[J]. 煤炭工程，2018，50(6)：123-127.
HUANG Hui，ZHANG Xue. Review of research on unsafe behavior of miners[J]. Coal Engineering，2018，50(6)：123-127.

[2] GUARINO N，WELTY C. Evaluating ontological decisions with OntoClean[J]. Communications of the ACM，2002，45(2)：61-65.

[3] HORROCKS，IAN，PATEL-SCHNEIDER，et al. SWRL：a semantic web rule language combining OWL and RuleML[J]. W3C Member Submission，2004，21(79)：1-31.

[4] BORDES A，USUNIER N，GARCIA-DURAN A，et al. Translating embeddings for modeling multi-relational data[C]. Neural Information Processing Systems，South Lake Tahoe，2013：1-9.

[5] WANG Zhen，ZHANG Jianwen，FENG Jianlin，et al. Knowledge graph embedding by translating on hyperplanes[C]. The 28th AAAI Conference on Artificial Intelligence，2014.

[6] 刘文聪，张春菊，汪陈，等. 基于 BiLSTM−CRF 的中文地质时间信息抽取[J]. 地球科学进展，2021，36(2)：211-220.
LIU Wencong，ZHANG Chunju，WANG Chen，et al. Geological time information extraction from Chinese

text based on BiLSTM-CRF[J]. Advances in Earth Science, 2021, 36(2): 211-220.

[7] 吴闯, 张亮, 唐希浪, 等. 航空发动机润滑系统故障知识图谱构建及应用[J/OL]. 北京航空航天大学学报: 1-14[2023-05-22].https://doi.org/10.13700/j.bh.1001-5965.2022.0434.
WU Chuang, ZHANG Liang, TANG Xilang, et al. Construction and application of fault knowledge graph for aero-engine lubrication system[J/OL]. Journal of Beijing University of Aeronautics and Astronautics: 1-14[2023-05-22]. https://doi.org/10.13700/j.bh.1001-5965.2022.0434.

[8] SHAO Zhou, YUAN Sha, WANG Yongli, et al. ELAD: an entity linking based affiliation disambiguation framework[J]. IEEE Access, 2020, 8: 70519-70526.

[9] FANG Yuan, CHANG Mingwei. Entity linking on microblogs with spatial and temporal signals[J]. Transactions of the Association for Computational Linguistics, 2014, 2: 259-272.

[10] SIMONE F, ANSALDI S, AAGNELLO P, et al. Industrial safety management in the digital era: constructing a knowledge graph from near misses[J]. Computers in Industry, 2023, 146. DOI: 10.1016/j.compind.2022.103849.

[11] 尉桢楷, 程梦, 周夏冰, 等. 基于类卷积交互式注意力机制的属性抽取研究[J]. 计算机研究与发展, 2020, 57(11): 2456-2466.
WEI Zhenkai, CHENG Meng, ZHOU Xiabing, et al. Convolutional interactive attention mechanism for aspect extraction[J]. Journal of Computer Research and Development, 2020, 57(11): 2456-2466.

[12] 刘峤, 李杨, 段宏, 等. 知识图谱构建技术综述[J]. 计算机研究与发展, 2016, 53(3): 582-600.
LIU Qiao, LI Yang, DUAN Hong, et al. Knowledge graph construction techniques[J]. Journal of Computer Research and Development, 2016, 53(3): 582-600.

[13] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[14] SEKI K. On cross-lingual text similarity using neural translation models[J]. Journal of Information Science, 2020, 27: 315-321.

[15] 李红霞, 樊欣怡. 人因视角下国内煤矿安全领域研究现状与发展趋势[J]. 煤炭工程, 2022, 54(1): 181-186.
LI Hongxia, FAN Xinyi. Status and development trend of coal mine safety research from the perspective of human factors[J]. Coal Engineering, 2022, 54(1): 181-186.

[16] BENGIO Y, DUCHARME RVINCENT P. A neural probabilistic language model[J]. Journal of Machine Learning Research, 2003, 3: 1137-1155.

[17] PENNINGTON J, SOCHER R, MANNING C. Glove: global vectors for word representation[C]. Conference on Empirical Methods in Natural Language Processing, Doha, 2014: 1532-1543.

[18] PETERS M E, NEUMANN M, LYYER M, et al. Deep contextualized word representations[C]. Conference of the North American Chapter of the Association for Computational Linguistics, New Orleans, 2018: 2227-2237.

[19] DEVLIN J, CHANG Mingwei, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]. Conference of the North American Chapter of the Association for Computational Lingristics, Jill Burstein, 2019: 4171-4186.

[20] XU Wencong, HU Yue, LI Jianxun. A data-driven Dir-MUSIC method based on the MLP model[J]. IET Science, Measurement & Technology, 2022(6): 367-376.

[21] 王智广, 文红英, 鲁强, 等. 地质领域开放式实体关系联合抽取[J]. 计算机工程与设计, 2021, 42(4): 996-1005.
WANG Zhiguang, WEN Hongying, LU Qiang, et al. Joint extraction of open entity relation in geological field[J]. Computer Engineering and Design, 2021, 42(4): 996-1005.

[22] 赵晓娟, 贾焰, 李爱平, 等. 多源知识融合技术研究综述[J]. 云南大学学报(自然科学版), 2020, 42(3): 459-473.
ZHAO Xiaojuan, JIA Yan, LI Aiping, et al. A survey of the research on multi-source knowledge fusion technology[J]. Journal of Yunnan University(Natural Sciences Edition), 2020, 42(3): 459-473.

[23] 乔骥, 王新迎, 闵睿, 等. 面向电网调度故障处理的知识图谱框架与关键技术初探[J]. 中国电机工程学报, 2020, 40(18): 5837-5849.
QIAO Ji, WANG Xinying, MIN Rui, et al. Framework and key technologies of knowledge-graph-based fault handling system in power grid[J]. Proceedings of the CSEE, 2020, 40(18): 5837-5849.

[24] 曹现刚, 张梦园, 雷卓, 等. 煤矿装备维护知识图谱构建及应用[J]. 工矿自动化, 2021, 47(3): 41-45.
CAO Xiangang, ZHANG Mengyuan, LEI Zhuo, et al. Construction and application of knowledge graph for coal mine equipment maintenance[J]. Industry and Mine Automation, 2021, 47(3): 41-45.