黄斌，刘一琪，耿欣，等. 长江大保护工程经验知识抽取及策略智能生成[J]. 水利水电技术(中英文)，2025，56(S2)：27-31.
HUANG Bin，LIU Yiqi，GENG Xin，et al. Knowledge extraction and strategic intelligence generation in the Yangtze River Protection Project[J]. Water Resources and Hydropower Engineering，2025，56(S2)：27-31.

长江大保护工程经验知识抽取及策略智能生成

黄　斌[1]，刘一琪[2]，耿　欣[1]，徐　正[1]，常倩然[2]
(1. 长江三峡技术经济发展有限公司，北京　101100；2. 天津大学 水利工程智能建设与运维全国重点实验室，天津　300372)

摘　要：长江大保护工程采用EPC承包模式，属于涉及多利益相关者和复杂环境的大型工程。该项目面临复杂的协同关系和难度较大的协同管控挑战。因此，如何利用积累的经验知识实现高效、高质量辅助协同管控是各方面关注的焦点。为了有效利用文本知识，构建了BERT-BiLSTM文本分类模型和RoBERTa-BiLSTM-CRF实体识别模型。依托宜昌两网二期项目设计了知识抽取试验，进一步对长江大保护EPC项目文本资料进行知识抽取，经过实体对齐和结构化存储，最终形成了可复制、可推广的结构化协同策略库。此外，利用实体识别模型获得了问句关键实体信息，并通过匹配检索获得待定策略集。最后利用策略句融合方法，实现基于句子语义关系对待定策略集的语义融合，提出协同策略，为后续项目协同管控提供重要参考。

关键词：长江大保护；知识抽取；协同管控；策略智能生成
DOI：10.13928/j.cnki.wrahe.2025.S2.007
中图分类号：TV213.4　　**文献标志码**：A　　**文章编号**：1000-0860(2025)S2-0027-05

Knowledge extraction and strategic intelligence generation in the Yangtze River Protection Project

HUANG Bin[1], LIU Yiqi[2], GENG Xin[1], XU Zheng[1], CHANG Qianran[2]
(1. The Three Gorges Economic and Technical Development Co., Ltd., Beijing　101100, China; 2. State Key Laboratory of Hydraulic Engineering Intelligent Construction and Operation, Tianjin University, Tianjin　300372, China)

Abstract: The Yangtze River Protection Project adopts the EPC contracting mode, which is a large-scale project involving multi-stakeholders and a complex environment. The project faces complex synergistic relationships and difficult synergistic control challenges. Therefore, how to utilize the accumulated empirical knowledge to realize efficient and high-quality assisted collaborative control is the focus of attention of all parties. In order to effectively utilize the text knowledge, BERT-BiLSTM text classification model and RoBERTa-BiLSTM-CRF entity recognition model were proposed. Knowledge extraction experiments were designed relying on the Yichang two-network phase II project, and further knowledge extraction was carried out on the textual information of the Yangtze River Great Protection EPC project, and after entity alignment and structured storage, a replicable and scalable structured collaborative strategy library was finally formed. In addition, the entity recognition model is utilized to obtain the key entity information of the interrogative sentence, and the pending strategy set is obtained through matching retrieval. Finally, the strategy sentence fusion method is utilized to realize the semantic fusion of

收稿日期：2024-10-16
基金项目：中国长江三峡集团有限公司科研项目(202103551)
作者简介：黄　斌(1973—)，男，高级工程师，硕士，研究方向为工程建设管理。E-mail：huang_bin@ctg.com.cn
通信作者：刘一琪(2000—)，女，硕士研究生，研究方向为工程建设协同管控。E-mail：liuyiqi@tju.edu.cn

the pending strategy set based on sentence semantic relations, and the synergistic strategy is proposed to provide an important reference for the synergistic control of subsequent projects.

Keywords: Yangtze River Protection; knowledge extraction; collaborative management; intelligent strategy generation

0　引　言

长江大保护工程作为长江经济带的重要组成，采用 EPC 模式建设时面临时空干扰强、利益关系复杂等挑战，亟需利用经验知识实现智能化协同管控。国内外已开展诸多相关研究：耿飙等[1]以文本数据为对象，构建了 BERT-BiLSTM-CRF 模型，将非结构化的医学文本转换成结构化的数据。何江等[2]则对跨境电商与跨境物流的协同机理进行了总结，并基于现状与困境提出了协同策略。但现有策略多缺乏对历史经验的借鉴，实用性不足。

为此，本研究提出融合文本分类与实体识别的知识抽取法，以 BERT-BiLSTM 模型抽取协同策略句，RoBERTa-BiLSTM-CRF 模型识别关键实体，提出句子融合优化法生成协同策略。依托宜昌两网二期项目实验，抽取知识形成策略库，经实体匹配与语义融合优化，为长江大保护 EPC 项目协同管控提供技术支持。

1　文本分类和实体识别知识抽取模型

1.1　BERT-BiLSTM 文本分类模型

构建 BERT-BiLSTM 模型用于策略句分类。该模型先利用预训练的 BERT 获取文本向量，再将其输入 BiLSTM 网络捕捉上下文信息，最后经全连接层和 *Softmax* 函数完成文本分类，如图 1 所示。

1.2　RoBERTa-BiLSTM-CRF 实体识别模型

构建 RoBERTa-BiLSTM-CRF 模型进行命名实体识别。此模型由三模块构成，先通过 RoBERTa 预训练模型处理标注语料获取文本向量、捕捉语义，将词向量输入 BiLSTM 捕捉上下文，最后经线性分类器结合 CRF 模块对 BiLSTM 输出解码，生成预测标注序列，如图 2 所示。

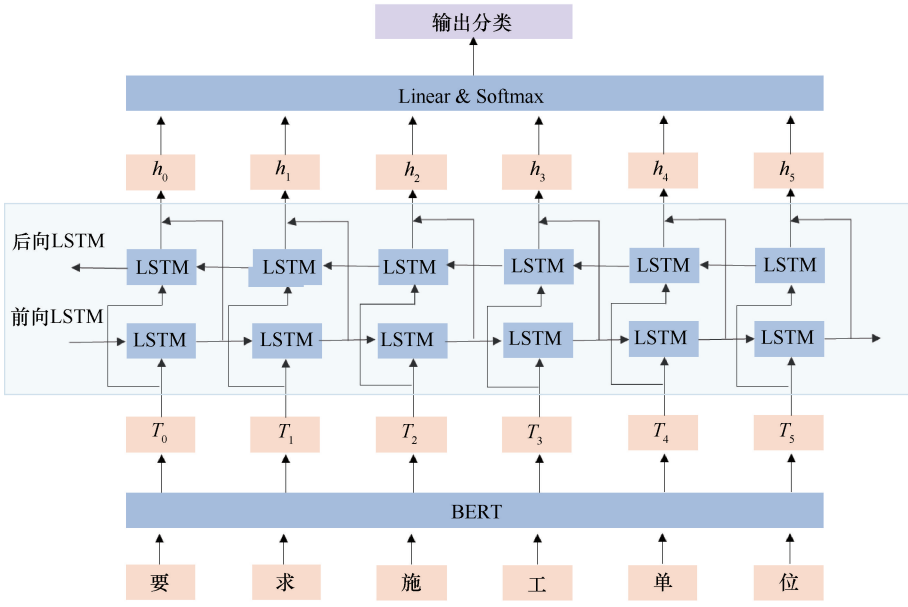


图 1　BERT-BiLSTM 文本分类模型结构

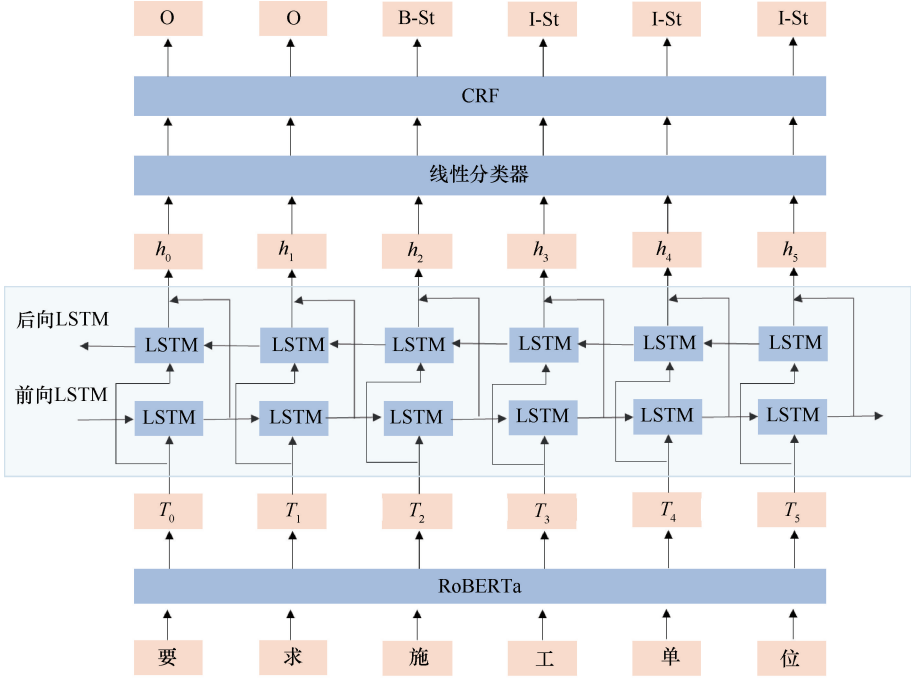


图 2　RoBERTa-BiLSTM-CRF 命名实体识别模型结构

2　知识抽取试验及分析

2.1　文本分类试验过程及效果分析

本研究详细梳理了长江大保护宜昌二期项目在2020年至2024年间的会议与周报资料，经过预处理，利用Label Studio软件将文本标注为策略句和其他句，划分训练集、验证集测试集。经过文本分类实验得到BERT-BiLSTM模型的精度、召回率及$F1$值均为99.03%。该模型在BERT的基础上引入了BiLSTM，BiLSTM结构能够同时捕获文本的前向和后向信息，从而更全面地理解文本语义。

2.2　命名实体识别试验过程及效果分析

基于文本分类后的策略句，构建命名实体识别的数据集。将长江大保护宜昌二期项目关键因素的实体类别划分为利益相关者、EPC项目、子项及位置、工序、自然环境、社会环境、技术管理、安全文明施工、质量管理、进度及目标管理、合同商务财务管理、沟通管理、资源管理、项目管理共14个大类，采用Label Studio标注数据集中的实体。

经过命名实体识别试验，RoBERTa-BiLSTM-CRF命名识别模型在14个实体类别中表现如表1所列，其中后续作为重点信息的利益相关者类其$F1$值达到90.38%。

2.3　经验知识实体对齐及结构化存储

通过RoBERTa-BiLSTM-CRF模型完成命名实体识别，获得结构化知识后利用Word2Vec将实体映射成融合上下文语义的词向量，计算实体的特征向量表征，通过计算实体向量间的余弦相似度，获得实体语义相似度。设定相似度阈值，指导实体融合的判断。

2.4　问句关键实体识别

上述构建的实体识别模型同样适用于问句文本。利用训练好的RoBERTa-BiLSTM-CRF模型，将输入数据传入模型进行预测得到标签对应的ID，解析预测结果为BIO标签数据，最后得到问句Q中的实体Q_E。

3　考虑实体语义相似度的策略句匹配

3.1　模型假设与参数设置

实体类别集合为$C=\{c_1, c_2, \cdots, c_{14}\}$，分别对应14个实体类别。假设问句$Q$排除利益相关者类后的实体集为$Q_E=\{q_1, q_2, \cdots, q_m\}$。知识抽取格式化储存的协同管控经验知识库策略句共有N条，$S=\{S_1, S_2, \cdots, S_n\}$。$x_i$为二元变量，表示第$i$条协同策略$S_i$是否放入待定策略集$Su$。

3.2　正则表达式匹配

正则表达式依据预定义规则模式，在策略句匹配时，用于快速检索协同管控经验知识库中与问句Q实体集匹配的策略句。

3.3　实体语义相似度匹配

计算问句Q与第i条协同策略S_i中实体相似度时，基于语义相似度，利用Roberta获取实体单字向量，叠加得到实体特征向量表征，再通过计算实体向量间余弦相似度，得出实体语义相似度。

3.4　策略句匹配检索

检索知识抽取格式化储存的协同管控经验知识库协同策略S_i的实体集，与问句Q的实体进行正则匹配。如果存在$\forall q_j \in Q_E$正则匹配成功，那么$x_i=1$第i条协同策略S_i被放入待定策略集Su；如果不存在$\forall q_j \in Q_E$与S_i的实体集中的实体正则匹配成功，对每个问句实体q_j找到其所属的实体类别c_h，将策略句S_i中实体类别为c_h的每个实体与问句实体q_j进行语义相似度计算，如果相似度大于阈值vt，那么问句实体q_j与该条协同策略S_i语义相似度匹配成功。如果存在$\forall q_j \in Q_E$与该条协同策略S_i语义相似度匹配成功，那么$x_i=1$第i条协同策略S_i被放入待定策略集Su；如果正则匹配和相似度匹配都不成功，$x_i=0$第i条协同策略S_i不被放入待定策略集Su。

4　基于实体规则匹配的策略句融合方法

4.1　基于实体信息的策略句融合

(1)符号定义。S_a、S_b分别为句子a、b分割后的短句集；E_a、E_b分别为句子a、b中所有实体的列表；E_{ab}为句子a中非重合实体的列表；FE_a、FE_b分别为句子a、b中非重合实体的列表；$Len(s)$为短句s的长度(字符)；$Ent(s)$为短句s中包含的实体数量；$Eet(E)$为实体列表E中包含的实体数量；c为融合基础句，初始化为空。

(2)融合规则。如果$Ent(FE_b) \geqslant Ent(FE_a)$，则令$c=S_b$(即句子$b$作为融合基础句)。否则，令$c=S_b$(即句子$a$作为融合基础句)。对于每个重合实体

表1　RoBERTa-BiLSTM-CRF命名实体识别模型在不同实体类别中$F1$值的实验结果

实体类别	利益相关者	EPC项目	子项位置	工　序	自然环境	社会环境	技术管理	质量安全	资源管理	商财合同	内部沟通	外部沟通	项目管理	设计进度
$F1$/%	90.38	62.50	73.33	45.76	75	58.82	45.1	65.57	43.90	62.86	51.06	47.06	62.07	59.79

$e \in E_{ab}$，在 S_a 和 S_b 中找到包含实体 e 的短句 S_{ae} 和 S_{be}，计算它们的优先级 $P(S_{ae})$ 和 $P(S_{be})$，选择优先级更高的短句替换融合基础句 c 中对应位置的短句。假设 b 为融合基础句，对于每个非重合实体 $e \in FE_a$，在 S_a 中找到包含实体 e 的短句 S_{ae}，如果 S_{ae} 在重合实体融合阶段未被使用(即不在 c 中)，则将其添加到融合基础句 c 的末尾。

4.2　基于实体的句子语义关系定义

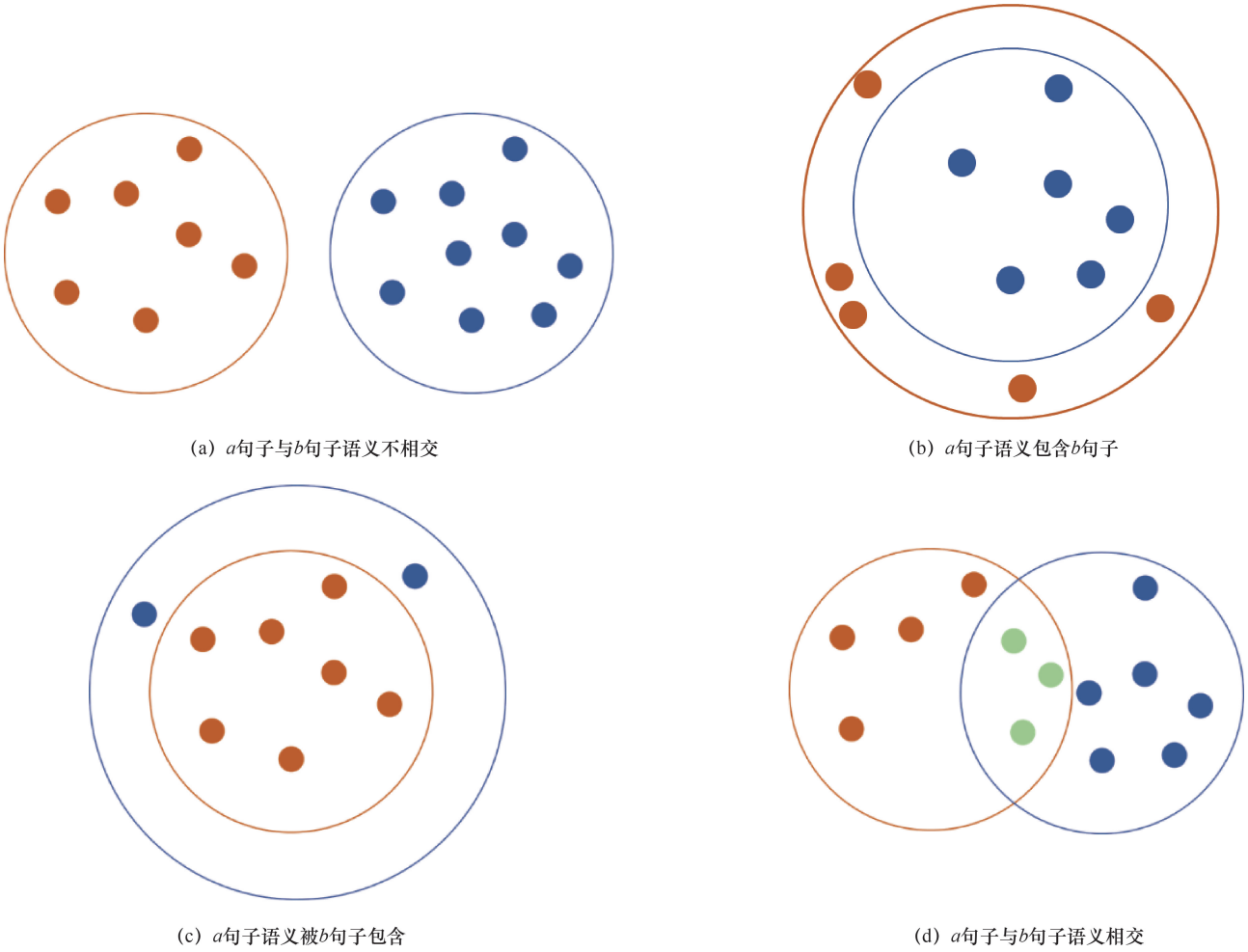
(1)符号定义。A 和 B 分别表示策略句 a 和策略句 b 中排除问句 Q 中实体的实体列表；$allEnRE(a, b)$为遍历 A 中每个元素与 B 中元素进行正则匹配。$allEnRE(a, b) = NE$ 表示 $allEnRE(a, b)$的结果为 A 中没有实体匹配成功；$allEnRE(a, b) = PE$ 表示 $allEnRE(a, b)$的结果为存在部分 A 中的实体匹配成功但不是 A 中所有实体都匹配成功；$allEnRE(a, b) = AE$ 表示 $allEnRE(a, b)$ 的结果为 A 中所有实体匹配成功。

(2)基于实体的句子语义关系。如图 3 所示，可以看出对于任意两个待定协同策略集中的策略句，其各个包含了一定量和类别的实体，会产生四种基于实体的句子语义关系。

4.3　利益相关者协同策略融合

假设已经匹配的待定协同策略集 Su 共有 m 条策略，$Su = \{Su_1, Su_2, \cdots, Su_m\}$。

取出第 i 条策略句 Su_i，从第 $i+1$ 条到第 m 条 Su_m，依次与 Su_i 关系判断。假设与 Su_i 进行语义判断的是第 r 条策略句 Su_r，$i+1 \leqslant r \leqslant m$，如果两个句子的利益相关者主体不同，则 Su_i 继续与下一条策略句 Su_{r+1} 进行判断；如果相同，计算 $allEnRE(Su_i, Su_r)$ 和 $allEnRE(Su_r, Su_i)$：若 $allEnRE(Su_i, Su_r) = NE$，则 Su_i 继续与 Su_{r+1} 语义判断；若 $allEnRE(Su_i, Su_r) = PE$，$allEnRE(Su_r, Su_i) = AE$，剔除 Su_r，Su_i 继续判断；若 $allEnRE(Su_i, Su_r) = AE$，$allEnRE(Su_r, Su_i) = PE$，剔除 Su_r，判断停止；若 $allEnRE(Su_i, Su_r) = PE$，$allEnRE(Su_r, Su_i) = PE$，进行语义融合 $Su_i = SFusion(Su_i, Su_r)$，剔除 Su_r，用



(a) a句子与b句子语义不相交　(b) a句子语义包含b句子

(c) a句子语义被b句子包含　(d) a句子与b句子语义相交

图 3　基于实体的句子语义关系

融合后的 Su_i 继续判断。重复上述步骤，完成所有策略句的语义关系判断，更新待定协同策略集。

5 结　论

本文提出融合文本分类与实体识别的知识抽取方法，采用 BERT-BiLSTM 模型进行文本分类、RoBERTa-BiLSTM-CRF 模型进行实体识别，在实验中均展现优秀性能，实现对长江大保护 EPC 项目协同管控策略句及其关键实体的高效抽取。通过知识抽取与实体对齐，完成施工协同经验知识的结构化存储。

在协同策略生成方面，构建句子融合优化的智能生成逻辑：基于实体识别结果与问句实体匹配检索策略句；针对工程文本信息冗余问题，依据实体定义四种语义关系，设计语义相交句融合方法，消除重复信息，提升策略质量。该方法为项目协同管控提供了可复用的技术方案与实践参考。

参考文献：

[1] 耿飙，梁成全，魏炜，等. 基于深度学习的非结构化医学文本知识抽取[J]. 计算机工程与设计，2024，45(1)：177-186.

[2] 何江，钱慧敏. 跨境电商与跨境物流协同策略研究[J]. 物流科技，2017，40(7)：1-6.

（责任编辑　王　璐）