李敏, 王国中, 赵海武. 基于 RoBERTa-BiLSTM-CRF 的 C 语言课程文本命名实体识别方法[J]. 智能计算机与应用,2025,15(10):16-23. DOI:10.20169/j.issn.2095-2163.251003

基于 RoBERTa-BiLSTM-CRF 的 C 语言课程文本命名实体识别方法

李 敏, 王国中, 赵海武
(上海工程技术大学 电子电气工程学院, 上海 201620)

摘 要: C 语言课程文本知识点的命名实体和关系识别任务存在其特有的挑战,特别是在捕获技术性文本中的精细语境和处理知识点的多义性上。为了更好地解决这些挑战,本文引入了基于 RoBERTa-BiLSTM-CRF 的模型进行命名实体识别。首先,使用 RoBERTa 对 C 语言课程文本进行预处理,从而生成了丰富的语义向量。这些向量随后被输入到 BiLSTM-CRF 模型中进行进一步的训练,有效地捕捉了知识点的上下文关系。在 C 语言课程数据集上的实验结果表现出色, *F*1 值达到了令人满意的水平。这进一步证明,利用 RoBERTa-BiLSTM-CRF 模型进行命名实体识别可以为 C 语言课程构建更为精确的知识图谱,这对于教育者提升教学效果、对于增强学习者的学习体验和效果都具有重要意义。

关键词: C 语言; 课程知识图谱; 实体命名识别; RoBERTa; 双向长短期记忆网络; 条件随机场
中图分类号: TP391.41 **文献标志码**: A **文章编号**: 2095-2163(2025)10-0016-08

Named Entity Recognition method for C language course text based on RoBERTa-BiLSTM-CRF

LI Min, WANG Guozhong, ZHAO Haiwu
(School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201620, China)

Abstract: The task of named entity and relationship recognition for knowledge points in C language course texts presents its unique challenges, especially in capturing fine-grained context in technical texts and dealing with the ambiguity of knowledge points. To better address these challenges, this study introduces a model based on RoBERTa-BiLSTM-CRF for Named Entity Recognition. Firstly, RoBERTa is used to preprocess the C language course texts, generating rich semantic vectors. These vectors are then fed into the BiLSTM-CRF model for further training, effectively capturing the contextual relationships of the knowledge points. The experimental results on the C language course dataset are outstanding, achieving a satisfactory *F*1 score. This further proves that utilizing the RoBERTa-BiLSTM-CRF model for named entity recognition can build a more accurate knowledge graph for the C language course, which is of great significance for educators to enhance teaching effects, and learners′ learning experience and effectiveness.

Key words: C language; course knowledge graph; Named Entity Recognition; RoBERTa; bidirectional Long Short-Term Memory network; Conditional Random Field

0 引 言

传统的教育模型在适应个性化学习逐渐强化的需求方面展现出明显的不足。面临人类知识体量的爆炸式扩张,个性化及高效学习的要求也相应提升。不断拓宽的知识领域意味着成为领域专家所需的学习时间也要延长,而从人们所接受的学历教育时长中就可反映出这一发展趋势。为了精简所需学习的知识体量,各学科和专业逐渐走向更为精细的划分,从而使个性化学习的需求更加凸显。尽管传统师徒教育模式能较好地实现个性化教学,其规模及效率的局限性却使其难以适应现代大规模教育的要求。

基金项目: 国家重点研发计划(2019YFB1802700); 上海市教育科学规划项目(C2023100); 上海工程技术大学项目(0232A1060123160431, 0232A1060123160432)。

作者简介: 李 敏(1997—),女,硕士研究生,主要研究方向:知识图谱,个性化推荐; 赵海武(1976—),男,博士,高级工程师,硕士生导师,主要研究方向:图像处理,视频处理。
通信作者: 王国中(1962—),男,博士,教授,博士生导师,主要研究方向:图像处理,视频处理。Email:wanggz@sues.edu.cn。

收稿日期: 2024-01-14

另一方面，传统的班级教学模式在兼顾学生间多样性的需求上常表现不足，这导致在实施过程中优秀学生的潜能可能未得到充分发挥，而学习进度相对较慢的学生也将面临跟不上进度的可能。

随着大数据的兴起，构建课程知识图谱并推动个性化学习已成为信息化教育时代的核心潮流。知识图谱能够将特定领域的知识以网络结构呈现出来，以直观的方式展现知识间的关联，构建了一种可视化的知识框架，提供了卓越的解释性。通过整合知识图谱与智能算法，个性化学习的实现成为可能。智能算法能够依据学习者的具体学习状况，推荐一系列后续的学习节点，直至确保学习者掌握所有关键的知识点。如同遍历一张网，可能的路线有很多，只要把每个节点都走到即可。

在推进知识图谱与教育领域的融合过程中，海量数据的整合成为推动人工智能深度整合教育的核心[1]。首步便是构建一个稳固的课程知识图谱。目前，建立知识图谱有 2 种主要策略：纯人工与半自动化。纯人工方法依赖于教育者或专家逐个构造“实体-关系-实体”或“实体-属性-属性值”的三元组。此方法得到的知识图谱精确且可信，适合作为初期应用的基石。然而，这种方法劳时劳力，难以应用于大范围。因此，许多学者转向研究自动构建知识图谱的方法。尽管目前的自动化方法并不完美，但通过后续的人工修正，可以进一步完善，形成半自动化的策略。

自动构建课程知识图谱的第一步是命名实体识别（Named Entity Recognition，NER）[2-3]，即从相关文本中提取出特定类型的实体信息，诸如人名、地名、组织机构名等，并为这些实体赋予相应的类别标签。C 语言是一门典型的计算机编程类课程，很多的研究工作都是以 C 语言课程为实例。教学者可通过 C 语言课程文本知识点实体命名识别构建课程知识图谱，可系统化地组织教材，设计针对性的教学策略，提高教学效率。学习者则可清晰掌握知识点的结构和关系，深化理解，并规划有效的学习路径。在处理 C 语言课程教学文本时，命名实体识别面临着一系列挑战，包括实体种类的多样性和不断变化，以及实体结构的复杂性，实体名可能存在嵌套、别名以及缩写等问题。

1　相关工作

早期的命名实体识别方法有基于规则和字典的方法、基于统计与机器学习的方法。基于规则和字典的方法需要人工编写实体规则和构建本体论。基于统计与机器学习的方法不需要人工标记规则，而是依赖大量的标注数据来学习模式。当提取的规则能比较精确地反映语言现象时，基于规则的方法性能要优于基于统计的方法。基于规则的方法依赖于领域专家通过对大量领域文本的总结归纳，结合语法知识和个人经验，制定识别规则，从而提高模型的可解释性和可信度。例如 NetOwl[4]、Facile[5] 和 SAR[5] 等系统。在教育领域的知识点提取方面，文献[6]分析了学科知识图谱的构建与创新应用。郭宏伟[7]分析了智能教育与知识图谱的关系，采用基于规则、即人工抽取的方法对《中国医学史》进行知识点抽取，构建了相应知识图谱。郎亚坤等学者[8]构建 C++课程知识图谱，对课本中的知识点采取人工抽取和专家审核的方法，高效抽取课程知识点。但是这些规则往往依赖于具体语言、领域和文本风格，编制过程耗时且难以涵盖所有的语言现象，特别容易产生错误，系统可移植性不好，对于不同的系统需要语言学专家重新书写规则。因此，这种方法所定义的规则和模板在泛化能力方面较为有限。

随着科技不断进步，基于统计和机器学习的方法[9-11]逐渐成为了替代基于规则方法的主流选择。采用机器学习技术，从标记文本中学习规则，能够利用大规模数据集进行模式识别和模型训练，从而更有效地处理复杂的自然语言处理任务。在知识点抽取方面，文献[12]采用 TF-IDF（Term Frequency-Inverse Document Drequency）和 K-Means 方法对关键词进行提取，建立了基于 MOOC 的高等教育知识图谱。研究中首先利用 TF-IDF 得到分词后的词向量，然后输入到 K-Means 聚类算法中得到不同的类，进一步得到课题组关键词，对知识点的提取则对文档基于规则进行过滤，再利用 TF-IDF 进行计算得到知识点关键词。文献[13]提出了教育领域知识图谱的构建方法。其中对于实体关系的抽取，改进了基于 Bootstrapping 的抽取方法，加入了语义约束的关系模型，减小了抽取关系实例的偏移问题。对于属性的抽取，以众包的形式完成对教育特点属性的确定。在此基础上，对于多源数据实体抽取中的实体对齐问题，提出了基于属性和类别的对齐方式，减少了加入知识库中的知识冗余度。然而，基于统计和机器学习的方法仍然要求人工创建大量特征集，这一过程需要耗费相当的人力和时间。

随着神经网络模型的不断发展和进步，深度学习等人工智能技术在自然语言处理（NLP）领域取得

了显著的进展。相较于传统的机器学习方法,基于深度学习的方法主要利用神经网络自动挖掘文本隐藏的特征,从而降低了人为主观因素对特征选择的影响,有效提升了实体识别的效果。早期 Huang 等学者[14]采用双向 LSTM 和 CRF 结合的神经网络模型,能良好解决序列标注的问题, $F1$ 值在 CoNLL-2003 语料上达到 88.83%。武国亮等学者[15]采取 FB-Latiice-BiLSTM-CRF 模型对中文突发事件进行事件抽取,获取语句语义特征,进行联合多任务学习, $F1$ 值达到 78.8%。许力等学者[16]提出了一个基于 BERT-BiLSTM-CRF 的模型。研究中还结合词性分析和组块分析特征来进一步提高精度。在 BC4CHEMD (BioCreative IV CHEMicalDisease relation)等数据集上,该研发模型的平均 $F1$ 值高达 89.45%,显著提高了生物医学命名实体识别的效果。张芳丛等学者[17]采用 RoBERTa-WWM-BiLSTM-CRF 模型更好利用各编码层的不同信息,使用动态掩码,提升下游实体识别任务效果,有效解决了中文电子病历命名实体识别中存在的一词多义和词识别不全的问题,使得实体识别效果 $F1$ 值达到 89.08%。张智源等学者[18]提出了 BERT 预训练模型和多窗口门控 CNN,通过生成句子的字向量序列并动态微调,增强语义表达,其 macro-$F1$ 值达到了 90.16%。尽管,该方法提升了识别的准确率,但是依旧存在一词多义和词识别不全的问题。

综上所述,文本实体识别和关系识别已有不错的发展,为教育领域 C 语言课程知识图谱构建奠定了基础。本文以教学课程 C 语言数据作为研究对象,提出基于 RoBERTa-BiLSTM-CRF 模型的 C 语言课程文本知识点命名实体和关系识别的方法,主要创新有以下几点:

(1)基于 BERT 模型进行优化和改进,采用 RoBERTa 预训练模型,对序列进行全词掩码,用动态遮蔽替换静态遮蔽,并且减去在一层的 NLP 预测任务,加强模型的深度语义理解能力,获取准确的词向量。

(2)采用 BiLSTM-CRF 作为 RoBERTa 的下游模型能捕获时间依赖性和提供连续文本信息,使得模型对文本序列中的长期依赖关系更为敏感。而 CRF 层为序列标注提供优化解码,并引入激活函数来进一步优化 CRF 的输出,确保整体标签的连贯性和合理性,有效提升命名实体识别任务的准确率和鲁棒性。

2 模型

2.1 RoBERTa-BiLSTM-CRF 模型介绍

近年来,语言预处理是实体识别领域的研究热点。本文提出基于 RoBERTa-BiLSTM-CRF 改进算法的 C 语言课程文本知识点实体抽取模型,其整体结构主要包括 3 个关键模块。首先优势在于 RoBERTa 预训练模型对 BERT 进行了优化和改进,获得准确的词向量,作为 BiLSTM 的输入进行词向量双向编码,提取更深层次的语义信息,最终将语义信息输入 CRF 模型,通过 Viterbi 算法进行解码,解码的结果通过归一化处理计算出全局最优结果,以提高 C 语言课本文本知识点命名实体和关系识别的准确性。详细模型结构框架图如图 1 所示。

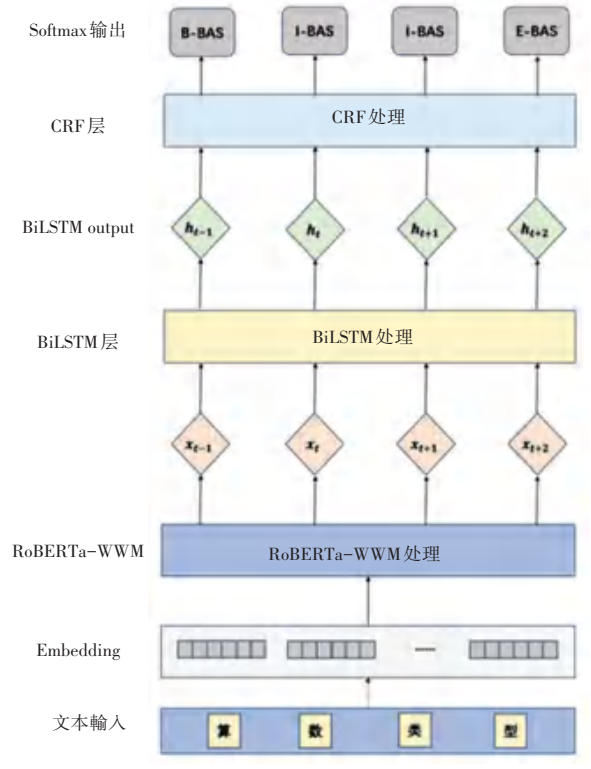

图 1 RoBERTa-BiLSTM-CRF 模型结构框架图

Fig. 1 Structure framework diagram of RoBERTa-BiLSTM-CRF model

2.2 RoBERTa 模型

语言模型的发展经历了一系列重要的里程碑,从最早的 one-hot 编码、Word2Vec,再到后来的 ELMO、GPT 和 BERT 及 BERT 优化模型。在这些模型中,前几种都存在一些局限性。例如,Word2Vec 难以处理一词多义的情况,而 GPT 受限于单向上下文信息的利用。BERT 模型则汲取了 ELMO 和 GPT 的优点,采用了双向 Transformer 模型和掩码语言模

型(MLM)的训练方式,从而更好地捕捉了词语的语义信息。而 RoBERTa 模型提升了模型的泛化能力和语言知识,使用动态掩码进行模型训练,并去掉了下一句 NSP 预测任务,极大提升了 BERT 模型的参数效率和模型性能。即如 Liu 等学者[19]通过对 BERT 模型的训练过程进行改进得到 RoBERTa,其中包括更长的训练时间、更大的批次、更多的数据以及优化的掩码模式,从而在多个任务上取得了最先进的性能表现。与 BERT 模型相比,在本文研究中借鉴了 RoBERTa 在词向量方面的先进性能来提高实体识别的准确率,采用全词掩码改进单字符掩码,用动态遮蔽替换静态遮蔽,并且减去在一层的 NLP 预测任务。RoBERTa 模型全词序列掩码见表 1。

表 1　RoBERTa 模型全词序列掩码

Table 1　Full-word sequence masking of RoBERTa model

对原始课程文本	算数类型和指针类型统称为纯量类型
分词课程文本	算数类型和指针类型 统称为纯量类型
BERT 掩码 MasK 输入	算数[Mask] 和 指针[Mask] 统称为 纯量[Mask]
RoBERTa 掩码 MasK 输入	[Mask] [Mask] 和[Mask] [Mask] 统称为[Mask] [Mask]

首先,对任意的课程文本序列执行分词处理。然后,随机地为部分分词进行 RoBERTa 全词 Mask 掩码,在序列开头添加[CLS]标记、句子间添加[SEP]进行分割,随着 tokenized 文本序列的不断输入,随之输出每个 token 的上下文感知的词向量表示,让模型在学习不同的语义特征时,适应不同的全词掩码策略,提升 C 语言课程文本知识点命名实体和关系识别效果。同时 RoBERTa 模型继承了 BERT 模型的优点,其主要结构由 12 层深度双向 Transformer 编码器组成。在这些编码器中,关键组成部分是注意力机制。此外还引入了卷积神经网络的残差连接,使得每个 Transformer 编码器层的输出都集中反映了句子中的语法、语义以及现实知识点的多样抽象表征。具体的 RoBERTa 模型结构如图 2 所示。计算过程可以表示为:

$$\boldsymbol{H} = \mathrm{RoBERTa}(\boldsymbol{X}) \tag{1}$$

其中,$\boldsymbol{X}$ 表示输入文本序列,$\boldsymbol{H}$ 表示词向量矩阵,每一行对应于输入文本序列中的一个 token。

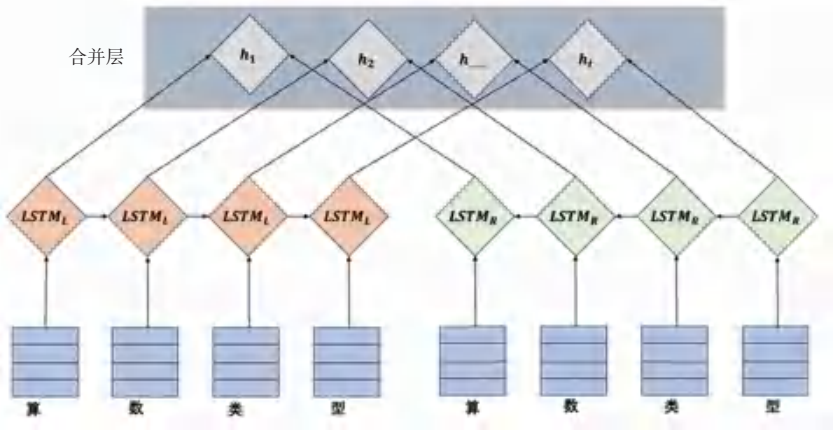

图 2　RoBERTa 模型结构图

Fig. 2　Structure diagram of RoBERTa model

2.3　BiLSTM 模型

传统的循环神经网络(RNN)虽然可用于命名实体识别任务,但却存在梯度消失和梯度爆炸等问题,由此导致模型难以有效捕捉到长期依赖关系。为了解决这一问题,长短时记忆网络(LSTM)应运而生,可通过引入门控机制,从而更有效地捕获和传递序列中的信息。而在长短时记忆网络(LSTM)这种 RNN 变体中,每个重复的神经元都包含 3 个门,这些门用于保护和控制信息状态。具体而言,LSTM 由 4 个关键组件构成,包括存储单元、输入门、遗忘门和输出门。这些组件协同工作,使网络在处理序列数据时能够更好地捕捉和利用上下文信息,从而显著提高了对长距离依赖性的建模能力。这种创新结构为深度学习在自然语言处理等领域的应用提供了坚实的支持。LSTM 结构公式表达如下:

$$\boldsymbol{i}_t = \sigma(\boldsymbol{W}_{ix}\boldsymbol{x}_t + \boldsymbol{W}_{ih}\boldsymbol{h}_{t-1} + \boldsymbol{b}_i) \tag{2}$$

$$\boldsymbol{f}_t = \sigma(\boldsymbol{W}_{fx}\boldsymbol{x}_t + \boldsymbol{W}_{fh}\boldsymbol{h}_{t-1} + \boldsymbol{b}_f) \quad (3)$$

$$\boldsymbol{m}_t = \tanh(\boldsymbol{W}_{cx}\boldsymbol{x}_t + \boldsymbol{W}_{ch}\boldsymbol{h}_{t-1} + \boldsymbol{b}_c) \quad (4)$$

$$\boldsymbol{c}_t = \boldsymbol{f}_t\boldsymbol{c}_{t-1} + \boldsymbol{i}_t\boldsymbol{m}_t \quad (5)$$

$$\boldsymbol{o}_t = \sigma(\boldsymbol{W}_{ox}\boldsymbol{x}_t + \boldsymbol{W}_{oh}\boldsymbol{h}_{t-1} + \boldsymbol{b}_o) \quad (6)$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t \odot \tanh(\boldsymbol{c}_t) \quad (7)$$

其中，σ 表示 sigmoid 激活函数；$\boldsymbol{W}$ 表示权重矩阵；$\boldsymbol{b}$ 表示偏移向量；$\boldsymbol{x}_t$ 表示 t 时刻的输入向量；$\boldsymbol{m}_t$ 表示待增加的内容；tanh 表示双曲正切函数；$\boldsymbol{c}_{t-1}$ 表示 t 时刻更新的记忆状态；$\boldsymbol{h}_{t-1}$ 表示 $t-1$ 时刻的输出；$\boldsymbol{i}_t$、$\boldsymbol{f}_t$、$\boldsymbol{o}_t$ 分别表示 t 时刻输入门、遗忘门及输出门的输出结果；$\boldsymbol{h}_t$ 表示整个 LSTM 单元 t 时刻的输出。

然而，随着研究的深入，单向 LSTM 模型无法同时处理上下文信息，为此 Graves 等学者[20]提出双向长短时记忆网络(BiLSTM)，将信息的双向流动纳入考虑，允许模型同时探索过去和未来的上下文信息。BiLSTM 模型由 2 个独立的 LSTM 层组成，一个用于正向传递信息，另一个用于反向传递信息，然后将其输出进行连接，即能更全面地捕捉序列中的语境。这种双向性质使得 BiLSTM 在自然语言处理任务中表现出色，尤其在命名实体识别等需要全局上下文信息的任务中，取得了显著的性能提升。因此，BiLSTM 模型代表着深度学习在自然语言处理领域的又一次重要进步，为解决复杂的语言理解问题提供了强大的工具。BiLSTM 单元结构如图 3 所示。

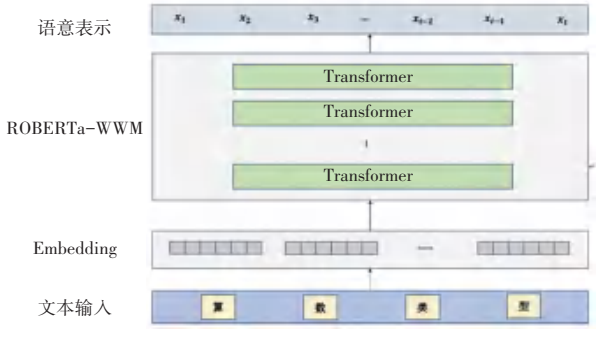

图 3 BiLSTM 单元结构图

Fig. 3 Structure diagram of BiLSTM unit

2.4 CRF 模型

在命名实体识别任务中，还面临一个序列标注问题，其中标签之间存在固定的顺序关系，如“I-BAS”标签必须跟随“B-BAS”标签，对于 BiLSTM 层输出的序列 $h = (h_1, h_2, \cdots, h_t)$，虽然能够有效地捕获课程文本知识点中的长距离信息并为每个标签提供上下文相关的隐藏特征，但在处理相邻标签的依赖关系方面存在不足。这却是 CRF 的强项，可专门处理标签之间的这种依赖，确保预测的序列是最佳的。因此，结合 BiLSTM 和 CRF 则能够更加全面地处理命名实体识别中的各种挑战，进一步将隐藏状态序列 $h = (h_1, h_2, \cdots, h_t)$ 转化为最佳标记序列 $Y = (y_1, y_2, \cdots, y_t)$。

CRF 模型的计算流程如下：对于 BiLSTM 层输出的序列为 $X = (x_1, x_2, \cdots, x_t)$，假设 $\boldsymbol{P}$ 是 BiLSTM 的输出的分数矩阵，$\boldsymbol{P}$ 的大小为 $n \times k$，其中 n 为词的个数，k 为标签个数，P_{ij} 表示第 i 个词的第 j 个标签的分数。对预测序列 $Y = (y_1, y_2, \cdots, y_t)$ 而言，得到的分数函数为：

$$s(X, Y) = \sum_{i=0}^{n} A_{yi,yi+1} + \sum_{i=1}^{n} P_{i,yi} \quad (8)$$

其中，$\boldsymbol{A}$ 表示转移分数矩阵；A_{ij} 表示标签 i 转移为标签 j 的分数；$\boldsymbol{A}$ 的大小为 $k+2$。预测序列 Y 产生的概率为：

$$P(X \mid Y) = \frac{e^{s(X,Y)}}{\sum_{\tilde{Y} \in Y_X} s(X, \tilde{Y})} \quad (9)$$

其中，$\tilde{Y}$ 表示真正的标签序列，Y_X 表示所有可能的标注序列。为了更好地计算 loss 值，利用式(9)得到预测序列的似然函数，研究中进一步推得：

$$\ln(P(Y \mid X)) = s(X, Y) - \ln\Big(\sum_{\tilde{Y} \in Y_X} s(X, \tilde{Y})\Big) \quad (10)$$

最后，使用 Viterbi 算法预测最佳标签序列，得到最大分数的输出序列，并引入激活函数来进一步优化 CRF 的输出，以增强其捕捉序列依赖性的能力。由此推得：

$$Y^* = \underset{\tilde{Y}}{\operatorname{argmax}} \max(0, s(X, \tilde{Y})) \quad (11)$$

其中，$\max(0, \cdot)$ 表示 ReLU 函数的标准形式。

3 实验结果

3.1 数据预处理和语料库构建

本文以谭浩强编写的《C 语言程序设计(第三版)》教材为主、百度百科、菜鸟教程和各类博客等网站教程为辅来对 C 语言知识点进行补充，共同构建 C 语言知识图谱。经过与专家的讨论，将 C 语言分为 5 个模块，包括基础知识模块、高级语法模块、开发环境模块、标准库模块、高级主题模块。这一细致划分的结构旨在帮助教育者与学习者系统性地管理和学习 C 语言，从基础知识到高级主题，逐步提升教育者与学习者 C 语言的知识掌握程度和编程能力。研究中用到了“B-N”、“I-N”、“O”和“E-N”

的实体标签，其中“B-N”表示每个知识点实体的起始位置，“I-N”表示中间位置，“O”表示不属于任何类型，“E-N”表示末尾位置，“N”表示知识点实体类型。C 语言课程文本知识点实体类型标注示例见表 2。

表 2　C 语言课程文本知识点实体类型标注示例

Table 2　Example of entity type annotation for knowledge points in a course text of C language

实体类型	实体	标签	示例标注
基础知识模块	数据类型和变量、运算符和表达式、输入和输出	BASIC	B-BAS、I-BAS、E-BAS
高级语法模块	递归、函数指针、数组、结构体和联合体	GRAMMER	B-GRA、I-GRA、E-GRA
开发环境模块	集成开发环境、编译和链接、调试技巧、版本控制工具	ENVIROMENT	B-ENV、I-ENV、E-ENV
标准库模块	字符串处理函数、数学库函数、输入输出库函数、内存管理函数	STOCK	B-STO、I-STO、E-STO
高级主题模块	多线程、数据结构、算法	THEME	B-THE、I-THE、E-THE

3.2　实验参数设置

在本文中，构建的实验模型利用深度学习框架 PyTorch2.13.0 和 Transformer 4.32.0 环境进行开发。经过初期实验，确定了每种模型的最佳参数设置，这些超参数的取值见表 3。本文的模型在训练集和测试集上表现良好。研究中采用了 Adam 优化器，对 RoBERTa 模型使用了 12 层 Transformer。为了防止过拟合问题，又在 BiLSTM 的输入和输出中引入了 Dropout，值为 0.5。

表 3　参数设置

Table 3　Masking parameter settings

相关参数	值
Transformer 层数	12
Dropout	0.5
隐藏层维度	768
优化器	Adam
Batch-size	32
Epoch	30
Max_seq_length	512
learning_rate	5e-5

3.3　模型评价指标

在本文中，使用了命名实体识别领域最常用的评估指标，包括精确率（Precision）、召回率（Recall）和 $F1$ 值。这些指标用于评估模型的性能。精确率（P）表示在模型预测为正的样本中、实际为正的样本的比例，衡量了模型预测为正的准确性。召回率（R）表示在实际为正的样本中，被模型正确预测为正的比例。该值衡量了模型对正样本的识别能力。$F1$ 值是精确率和召回率的调和平均值，用于综合衡量模型的性能。这些评估指标能够全面评估模型的能力，既考虑了模型的准确性，又考虑了识别能力，并将其综合在一起得出 $F1$ 值，以便更好地理解模型在命名实体识别任务中的表现。评价指标的计算方法如下：

$$P = \frac{T_P}{T_P + F_P} \tag{12}$$

$$R = \frac{T_P}{T_P + F_N} \tag{13}$$

$$F1 = \frac{2 \cdot P \cdot R}{P + R} \tag{14}$$

其中，T_P 表示能正确识别 C 语言课程文本知识点命名实体和关系标签的个数；F_P 表示能识别出 C 语言课程文本知识点命名实体和关系标签类别判定出现错误的个数；F_N 表示应该识别但没被识别的 C 语言课程文本知识点命名实体和关系的个数；$T_P + F_P$ 表示所有被预测的样本；$T_P + F_N$ 表示实际应该被预测的样本。

3.4　模型实验结果分析

为了更有效地评估本文提出的基于 RoBERTa-BiLSTM-CRF 的 C 语言课程文本命名实体识别方法，本文采用 BERT-CRF、BERT-BiLSTM、BiLSTM-CRF 和 BERT-BiLSTM-CRF 作为对比模型，用于验证本文提出的基于 RoBERTa-BiLSTM-CRF 的 C 语言课程文本命名实体识别方法性能更加优越。实验对比数据见表 4。

表 4　实验对比结果

Table 4　Comparison of experimental results　%

对比模型	P	R	$F1$
BERT-CRF 模型	85.11	87.26	86.17
BERT-BiLSTM 模型	86.28	86.96	86.62
BiLSTM-CRF 模型	85.46	86.35	85.91
BERT-BiLSTM-CRF 模型	88.67	89.54	89.15
RoBERTa-BiLSTM-CRF 模型	**90.83**	**91.24**	**91.04**

首先，比较 BERT 、BiLSTM、CRF 三种模型的组合，可以看出 BERT-BiLSTM-CRF 模型的效果是最好的，对比 BiLSTM-CRF 和 BERT-BiLSTM-CRF 这

2 个模型的实验结果,后者的 P、R、$F1$ 值分别高出 3.21%、3.19%、3.24%,加入的 BERT 模型能够提取从字符到句间关系的多级特征,使得预训练的词向量更好地捕捉句法和语义信息。其次,对比 BERT-BiLSTM-CRF 和 RoBERTa-BiLSTM-CRF 两个模型,后者的 P、R、$F1$ 值分别高出 2.16%、1.70%、1.89%。RoBERTa-WWM 预处理模型在提取语义特征方面表现出色,能够深入整合上下文信息,而 BiLSTM-CRF 模型则能够成功地解决实体的歧义问题,说明本文提出的方法提取的特征能力更强、效果更准确。

通过对比不同标注对实体标注的效果,验证本文提出的基于 RoBERTa-BiLSTM-CRF 的 C 语言课程文本命名实体识别方法识别效果更加精确。常见的标注体系有 BIO、BIOES 等。其中,BIO 标注方法简单明了、适用于多种类型的实体,标注的结果易于解释,但不适用于额外上下文信息处理;而 BIOES 标注在 BIO 基础上增加了实体尾部(E)和表示单个词的实体(S),允许更精确地处理单词级别和不规则实体,有效提高命名实体和关系识别的准确性。针对 2 种不同的标注方法,探索本文实体识别方法对 C 语言课程文本知识点不同实体类型识别结果的影响。结果见表 5。

表 5 C 语言课程文本知识点实体类型标注示例

Table 5 Example of entity type annotation for knowledge points in a course text of C language %

项目	RoBERTa-BiLSTM-CRF(BIO 标注)			RoBERTa-BiLSTM-CRF(BIOES 标注)		
	P	R	$F1$	P	R	$F1$
基础知识	88.97	89.52	89.24	90.61	91.38	**90.99**
高级语法	88.18	88.73	88.45	89.93	90.72	**90.32**
开发环境	89.41	90.23	89.82	91.34	91.72	**91.53**
标准库	88.26	89.14	88.70	89.79	90.13	**89.96**
高级主题	87.63	87.25	87.44	91.66	92.18	**91.92**

可以看到,本文实体识别方法在 C 语言课程文本知识点不同实体类型上表现尤为卓越。各个实体类型的识别效果都有了显著的提升。这进一步证明了,本文的方法在不同实体类型数据上的适应性和迁移性都达到了很高水平。

4 结束语

在本文中,研究采纳了 RoBERTa-BiLSTM-CRF 模型来专门针对 C 语言课程文本的命名实体及其关系进行识别。目标是通过自动识别这些实体和关系,进一步加深对 C 语言课程的洞察并实现更高效的知识管理。经由深入的分析,文中从教育者与学习者的视角均强调了实体与关系识别在教育文本中的关键作用。需要提及的是,不需依赖于复杂的特征工程,本文所提及的模型在 C 语言知识点的实体及关系识别上展现了优异性能,达到了高精准度和高召回率。与以往工作有别,研究引入了 BRERT 改进模型 RoBERTa 对数据进行预处理,并利用 BiLSTM-CRF 架构进行后续的数据处理,从而显著优化了 C 语言课程实体和关系的识别。

尽管本文中的 RoBERTa-BiLSTM-CRF 模型在 C 语言课程文本知识点的命名实体识别上表现优异,但在复杂描述中同时提取命名实体和其关系的任务上,模型仍有提升的潜力。首先,本文尚未深入研究特定语法和关键词对实体及关系识别的影响;其次,尽管已对命名实体标注进行了探索,但只涉及了 2 种,并仅对 BIOES 体系进行了检验。下一阶段将选取更多的标注规则进行对比,构建更加完善的 C 语言课程文本知识点语料库,为构建 C 语言课程知识图谱奠定基础。

参考文献

[1] 钟卓,唐烨伟,钟绍春,等. 人工智能支持下教育知识图谱模型构建研究[J]. 电化教育研究,2020,41(4):62-70.

[2] 刘浏,王东波. 命名实体识别研究综述[J]. 情报学报,2018,37(3):329-340.

[3] LI Jing, SUN Aixin, HAN Jianglei, et al. A survey on deep learning for Named Entity Recognition[J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 34(1): 50-70.

[4] KRUPKA G, ISOQUEST K. Description of the NetOwl extractor system as used for MUC-7[C]//Proceedings of the 7th Message Understanding Conference. ACL, 2005: 21-28.

[5] BLACK W J, RINALDI F, MOWATT D. FACILE: Description of the NE system used for MUC-7[C]//Proceedings of the 7th Message Understanding Conference (MUC-7). ACL, 1998: 1-10.

[6] 李艳燕,张香玲,李新,等. 面向智慧教育的学科知识图谱构建与创新应用[J]. 电化教育研究,2019,40(8):60-69.

[7] 郭宏伟. 基于智能教育的高校在线课程知识图谱构建研究—以中国医学史为例[J]. 中国电化教育,2021(2):123-130.

[8] 郎亚坤,苏超,王国中,等. 基于 Neo4j 的 C++课程知识图谱的构建和推理[J]. 智能计算机与应用,2021,11(7):144-150.
[9] LI Yinghao, SHETTY P, LIU L, et al. BERTifying the hidden Markov model for multi-source weakly supervised named entity recognition[C]// Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). ACL,2021: 6178-6190.
[10]余本功,范招娣. 面向自然语言处理的条件随机场模型研究综述[J]. 信息资源管理学报,2020,10(5):96-111.
[11]周强. 基于语料库和面向统计学的自然语言处理技术[J]. 计算机科学,1995,22(4):36-40.
[12]侯俊萌. 基于 MOOC 的高等教育知识图谱的构建[D]. 北京:北京邮电大学,2017.
[13]唐伟. 教育知识图谱的构建方法研究[D]. 哈尔滨:哈尔滨工程大学,2018.
[14]HUANG Zhiheng, XU Wei, YU Kai. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv preprint arXiv, 1508.01991, 2015.
[15]武国亮,徐继宁. 基于命名实体识别任务反馈增强的中文突发事件抽取方法[J]. 计算机应用,2021,41(7):1891-1896.
[16]许力,李建华. 基于 BERT 和 BiLSTM-CRF 的生物医学命名实体识别[J]. 计算机工程与科学,2021,43(10):1873-1879.
[17]张芳丛,秦秋莉,姜勇,等. 基于 RoBERTa-WWM-BiLSTM-CRF 的中文电子病历命名实体识别研究[J]. 数据分析与知识发现,2022,6(Z1):251-262.
[18]张智源,孙水华,徐诗傲,等. 基于 BERT 和多窗口门控 CNN 的电机领域命名实体识别[J]. 计算机应用研究, 2023, 40(1):107-114.
[19]LIU Yinhan, OTT M, GOYAL N, et al. Roberta: A robustly optimized Bert pretraining approach[J]. arXiv preprint arXiv, 1907.11692, 2019.
[20] GRAVES A, SCHMIDHUBER J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures[J]. Neural Network, 2005, 18(5/6): 602-610.