

文章编号: 1003-0077(2023)08-0075-11

基于依存结构学习的中学数学术语鲁棒抽取

罗凯威¹, 罗文兵^{1,2}, 黄琪^{1,2}, 王明文^{1,2}

(1. 江西师范大学 数字产业学院, 江西 上饶 334000;

2. 江西师范大学 计算机信息工程学院, 江西 南昌 330022)

摘要: 中学数学术语抽取方法未考虑句子的依存结构信息, 导致对句子的语义理解能力有限。此外, 由于依赖依存结构信息的术语抽取方法存在分词或依存结构错误, 导致术语抽取准确性和完整性不佳。为解决上述问题, 该文提出一种基于依存结构学习的中学数学术语鲁棒抽取模型。模型利用预训练模型得到文本向量语义表示, 并借助带有去噪注意力机制层的图神经网络和双向循环神经网络分别捕获文本的依存结构信息和上下文信息, 进一步采用注意力机制融合文本结构信息和上下文信息以实现在融入依存结构信息的同时缓解错误分词或依存结构的影响。模型在自建的中学数学术语数据集上抽取精度 P 和 F_1 值分别达到了 83.82%、82.51%, 相较于基准模型分别提升了 2.21%、1.22%, 表明该文所提方法能够鲁棒融合依存结构信息, 从而提升中学数学术语抽取的精确性和完整性。

关键词: 术语抽取; 依存结构; 图神经网络

中图分类号: TP391

文献标识码: A

Dependency Structure Learning for Robust Extraction of Middle School Math Terms

LUO Kaiwei¹, LUO Wenbing^{1,2}, HUANG Qi^{1,2}, WANG Mingwen^{1,2}

(1. School of Digital Industry, Jiangxi Normal University, Shangrao, Jiangxi 334000, China;

2. School of Computer and Information Engineering, Jiangxi Normal University, Nanchang, Jiangxi 330022, China)

Abstract: Term extraction methods in middle school mathematics do not consider the dependency structure information of the sentence, which leads to a limited semantic understanding of the sentence. In addition, the term extraction method of dependency structure information exists some errors in dependency structures, resulting in poor accuracy and completeness. This paper proposes a robust extraction model for middle school mathematical terms based on dependency structure learning. The dependency structure information and context information of text is captured by the graph neural network model with a denoising attention mechanism layer and bidirectional recurrent neural network, respectively. The attention mechanism is further used to integrate text structure information and context information to integrate the dependency structure information while alleviate the influence of incorrect word segmentation or dependency structure. The proposed model achieves 83.82% precision and 82.51% and F_1 value in the experiment, indicating 2.21% and 1.22% improvements compared with the baseline model, respectively.

Keywords: term extraction; dependency structure; graph neural network

收稿日期: 2022-11-20 定稿日期: 2023-02-07

基金项目: 国家自然科学基金(62266023); 江西省教育厅研究生创新基金项目(YC2022-s348); 江西省教育厅科学技术研究项目(GJJ210325, GJJ2200354)

0 引言

术语是专业领域内知识概念的语言指称和主要载体^[1],也是知识多模态表述中最重要的模态。术语抽取作为文本挖掘和信息抽取的首要步骤,也是机器翻译^[2]、信息检索^[3]、本体构建^[4]、知识图谱^[5]等领域的关键前提。自 20 世纪 30 年代初开始有研究者对术语相关领域展开广泛的研究^[6],其中包括从耗时耗力的人工术语抽取方法到自动术语抽取(Automatic Term Extraction, ATE)模型的探索。近些年,在线智慧教育的兴起推进了基础教育领域中术语抽取方法的研究,且已在多个学科领域中取得不错成果。中学数学术语及其术语之间的关联作为中学数学学科知识表示的核心^[7],其抽取的准确与否会直接决定所构建中学数学知识图谱的优劣。此外,学科知识库的不断丰富、中学数学教材的与时俱进,导致数学网络学习资源量迅猛增长。因此,如何从海量的非结构化中学数学知识中准确地自动抽取术语成为一大研究热点。

早期的术语自动抽取方法大多基于语言规则和统计,如术语抽取系统^[8]依靠专业领域内总结的规则模板对术语进行筛查、Bolshakova 等^[9]通过计算词的 TF-IDF(Term Frequency-Inverse Document Frequency)值评估某个词对某份文档的重要程度以实现术语抽取。两者在特殊学科领域内都取得了不错的抽取效果,但规则的不可穷尽性和统计值对语料库的依赖性分别导致这两类抽取模型的泛化能力较弱,迁移到中学数学术语抽取工作上的效果较差,且中学数学学科的知识表示大多又富有强逻辑性、结构性和多样性。这给中学数学术语抽取带来以下难点:①中学数学存在的“点”“面”“高”等单字多义术语难以精准抽取,如“计算高楼的楼高”中只有“楼高”一词中的“高”是术语;②中学数学存在的多重嵌套术语难以被完整抽取,如“无限不循环小数”中“无限”和“不循环”都是“小数”的修饰词,三者共同组成的“无限不循环小数”才是数学专业术语;③复杂语境下术语抽取难度较大,如“点动成线”这一简明的定理同时包含两个连续的动词“动”和“成”,这种隐含的双谓语结构不易被挖掘,且“成线”这个动作事件易被误解为术语。此外,还有一些如“上”“中”等的方位词也易因在句中所作成分不同而加大相关术语抽取的难度。

为了解决上述难点造成低频词难以抽取的问

题,李思良等^[10]结合了基础教育资源的学科特性,提出了以挖掘术语定义与术语关系为主的综合构词规则和边界检测的术语抽取方法 DRTE,并在中学数学数据集上取得了不错的抽取效果,但该方法较为耗时耗力,在复杂语境下术语抽取效果不佳。近年来,基于深度学习的术语抽取模型聚焦于丰富词嵌入表示或引入额外信息^[11]以提升术语抽取性能,如 Zhang 等^[12]提出了一种综合编码词级特征和字符特征的 Lattice-LSTM 模型,实现了句子语义表征的增强,且由此提高了抽取多义术语词的准确率。随着 Transformer^[13]模型的广泛应用,一些研究者将其与双向长短时记忆网络(Bidirectional Long Short-Term Memory, BiLSTM)、条件随机场^[14](Conditional Random Field, CRF)结合并应用于多个领域^[15-18]的专业术语抽取任务,其中中华鑫等^[18]提出了可增强标签预测的标签注意力层并结合了 Lattice-LSTM 模型,较好地解决了面向中学数学领域的部分多义术语词和嵌套术语词的抽取问题。但上述基于深度学习的自动术语抽取方法没有对句子本身存在的结构信息进行更深层的语义融合,导致模型对句子语义的理解有限,且没能很好地解决中学数学术语抽取存在的难点。

针对上述难点,有效捕获依存结构信息有助于提升术语抽取准确率。当前利用图神经网络获取依存结构信息表示的优异工作^[19-21]已有许多,如 Wu 等^[19]构建了一个语法融合编码器,通过整合依存边、依存标签以及词性标签等多种信息,增强了术语抽取模型的表现。但在复杂语境下由于依存结构分析器可能得到错误的分词结果或依存结构,导致图神经网络捕获的结构编码含有噪声。因此,为有效捕获句子依存结构信息并剔除其不利影响,本文提出了一种基于依存结构学习的中学数学术语鲁棒抽取模型 DSL-ATRE(Dependency Structure Learning on Automatic Term Robust Extraction)。该方法利用图神经网络模型捕获句子的依存结构信息,并通过注意力机制融合结构信息和上下文信息以实现融入句子依存结构信息的同时缓解错误信息造成的影响。本文的贡献主要有:

(1) 本文是术语抽取领域第一个同时考虑句子依存结构信息与其可能存在错误分词或依存结构而带来负面影响的工作;

(2) 本文所提的 DSL-ATRE 模型利用注意力机制融合由图神经网络编码表示的结构信息和上下文信息以实现融入依存结构信息的同时缓解错误分

词或依存结构信息的影响,从而提升模型在复杂语境下的术语抽取能力;

(3) 中学数学术语数据集上的实验结果表明本文所提的 DSL-ATRE: ①在评价指标 P 和 F_1 上比基准模型分别提升了 2.21%、1.22%; ②在复杂语境下比基准模型具有更好的准确性和完整性。

1 问题描述

中学数学术语抽取任务旨在从文本序列集中自动抽取数学领域相关知识概念的词或短语表示,本质上可看作是一个序列标注任务。任务的形式化定义如下: 给定包含 n 个句子的数据集 $D = \{s_1, s_2, \dots, s_n\}$, 每个句子 $s = \{w_1, w_2, \dots, w_m\}$ 由 m 个单字元素构成,再利用包含 k 个候选标签的集合 $\text{Label} = \{\text{Label}_1, \text{Label}_2, \dots, \text{Label}_k\}$ 对每个字进行标注。该任务在本文中的形式化定义如式(1)所示。

$$Y = f(s) = \{y_1, y_2, \dots, y_m \mid y_i \in \text{Label}\} \quad (1)$$

其中, Y 对应句子 s 的预测标签序列, m 为句子 s 所

包含字的个数。本文将候选标签集规定为 $\text{Label} = \{\text{B-T}, \text{M-T}, \text{E-T}, \text{S-T}, \text{O}\}$, 这五种标签可将序列中的词或短语标注为三类: 多字术语、单字术语和非术语。对于多字术语, B-T 表示术语的开始, M-T 表示术语的中间字, E-T 表示术语的结束; 对于单字术语, 标记为“S-T”; 对于构成非术语的其他汉字, 标记为“O”。由标签和术语词间的映射关系 g 可得句子 s 所含术语词的集合表示 T , 如式(2)所示。

$$T = g(Y) = \{\text{term}_1, \text{term}_2, \dots, \text{term}_r\} \quad (2)$$

其中, term_i 表示句子 s 中的第 i 个术语词, r 表示句子 s 所包含术语词的个数。

2 模型

基于依存结构学习的中学数学术语鲁棒抽取模型 DSL-ATRE 如图 1 所示, 该模型主要由六个模块组成: 语义表示、上下文语义表示、依存结构图构建、结构表示、语义融合和标签预测。下面详细介绍各个模块。

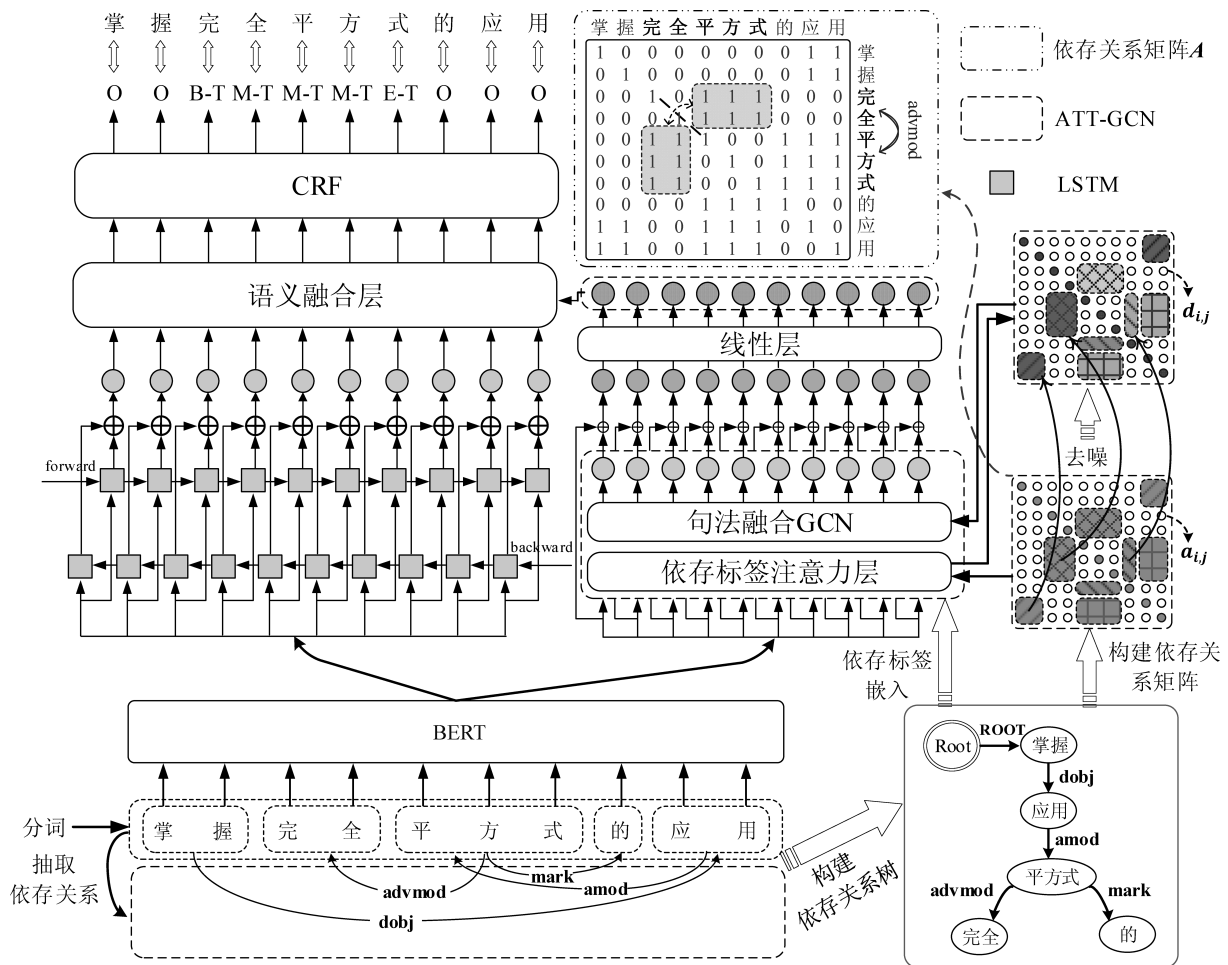


图 1 DSL-ATRE 总体框架图

2.1 语义表示

将中文预训练模型 BERT^[22] 作为语义表示层, 该模型沿用 Transformer 架构中的编码器模块, 所构成的多层双向编码模型能够很好地捕捉语句中的双向关系, 进而增强字的语义表示。因此我们将句子中的词例 w_i 输入 BERT 模型中获取对应的字嵌入向量 e_i^w , 计算如式(3)所示。

$$e_i^w = \text{BERT}(w_i) \quad (3)$$

2.2 上下文语义表示

利用 BiLSTM 捕获文本的上下文语义信息, 其中每个 LSTM 单元由遗忘门、输入门、输出门和记忆单元组成, 其输入包括前一时刻单元隐层输出 h_{t-1} 、前一时刻单元状态输出 C_{t-1} 以及当前时刻的输入 x_t 。接着经遗忘门 f_t 有选择地遗忘输入门信息 i_t 对单元状态的更新和输出门信息 o_t 对下一个隐层状态的确定, 由此得到最终隐藏层的输出 h_t 。LSTM 单元计算如式(4)所示。

$$\left. \begin{aligned} f_t &= \text{Sigmoid}(W_f[h_{t-1}, x_t] + b_f) \\ i_t &= \text{Sigmoid}(W_i[h_{t-1}, x_t] + b_i) \\ o_t &= \text{Sigmoid}(W_o[h_{t-1}, x_t] + b_o) \\ \tilde{C}_t &= \tanh(W_c[h_{t-1}, x_t] + b_c) \\ C_t &= f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \\ h_t &= o_t \odot \tanh(C_t) \end{aligned} \right\} \quad (4)$$

其中, W, b 分别表示门控单元的权重矩阵和偏置参数, $\odot$ 表示矩阵元素乘法。上下文语义表示模块将 BERT 词嵌入向量 e_i^w 作为 LSTM 单元当前的输入, 最终输出为双向隐藏层输出的拼接, 如式(5)所示。

$$h_i^{\text{LSTM}} = \text{LSTM}_f(e_i^w) \oplus \text{LSTM}_b(e_i^w) \quad (5)$$

其中, $\text{LSTM}_f, \text{LSTM}_b$ 分别表示前后向 LSTM 的输出向量, $\oplus$ 表示级联操作。

2.3 依存结构图构建

借助 StanfordCoreNLP 工具对中文文本进行分词处理和依存结构分析, 并将依存结构树转化为图结构以便模型后续的处理。

由于中文文本依存结构树中每个词节点可能由多个字构成, 因此在构建依存结构图的边集合时, 需将词与词间的连边转换成多字与多字的连边。为保证信息的双向流动和其语义表征的丰富性, 将存在依存关系的两对词各自所包含的字相互构建无向连

边, 可得依存关系矩阵为 $A = (a_{i,j})_{m \times m}$, 其结构如图 1 中上部分所示。其中 A 为对称矩阵, m 为图节点的个数, 即句子包含词例的个数, 且当 $i=j$ 或包含字节点的词间存在依存关系时有 $a_{i,j}=1$, 否则 $a_{i,j}=0$ 。

2.4 结构表示

受 Tian^[20] 的 A-GCN (Graph Convolutional Network, GCN) 模型的启发, 我们提出一种 ATT-GCN 模型(图 2)。该模型由依存标签注意力层和句法融合 GCN 层组成, 能够有效捕获图结构信息和缓解结构带来的噪声信息。下面详细介绍两层的结构。

依存标签注意力层 图中连边简单的二元表示会因存在一些无效的依存关系而产生噪声信息并影响图中节点信息的更新。因此本文借鉴 A-GCN^[20] 的思想, 利用依存标签信息和注意力机制优化邻接节点间的边权重, 从而重新计算图边的邻接矩阵以达到强化重要依存关系的同时缓解噪声信息影响的效果。依存标签注意力层的去噪过程如图 1 右侧中间部分所示, 每个小模块颜色的深浅对应表示每个节点在全局中的重要程度。我们首先从依存结构树中提取节点间的依存标签, 然后依照依存关系矩阵的构造方法构建依存标签矩阵, 再对矩阵中第 i 个节点与第 j 个节点间的依存标签进行嵌入表示得到 $e_{i,j}^{\text{dep}}$ 。接着如图 2(a) 中层的推理过程所示, 将 $e_{i,j}^{\text{dep}}$ 与 ATT-GCN 模型第 $l-1$ 隐层的第 i 个节点对应的输出 $h_i^{\text{GCN}(l-1)}$ 级联得到 $n_i^{(l)}$, 类似地得到第 j 个节点对应的级联结果 $n_j^{(l)}$, 形式化过程如式(6)、式(7)所示。最后由式(8)得到第 l 层的依存权重矩阵 $d_{i,j}^{(l)}$ 。

$$n_i^{(l)} = h_i^{\text{GCN}(l-1)} \oplus e_{i,j}^{\text{dep}} \quad (6)$$

$$n_j^{(l)} = h_j^{\text{GCN}(l-1)} \oplus e_{i,j}^{\text{dep}} \quad (7)$$

$$d_{i,j}^{(l)} = \frac{a_{i,j} \cdot \exp(n_i^{(l)} \cdot n_j^{(l)})}{\sum_{j=1}^m a_{i,j} \cdot \exp(n_i^{(l)} \cdot n_j^{(l)})} \quad (8)$$

其中, $a_{i,j} \in A$, $\cdot$ 表示内积操作。由于 ATT-GCN 模型有 L 层, 它的第 $l-1$ 隐层输出 $h_i^{\text{GCN}(l-1)}$ 会不断迭代以更新依存权重矩阵中的值, 使模型能更好地区分不同重要程度的依存关系。特别地, 当 $l=1$ 时, 我们将 BERT 词嵌入向量 e_i^w 作为 ATT-GCN 模型第一层的初始化表示, 即 $h_i^{\text{GCN}(0)} = e_i^w$ 。

句法融合 GCN 层 与标准的 GCN 不同的是, 本文引入的句法融合 GCN 层^[20] 加入了对依存标签嵌入向量 $e_{i,j}^{\text{dep}}$ 的计算。并且将依存标签注意力层输

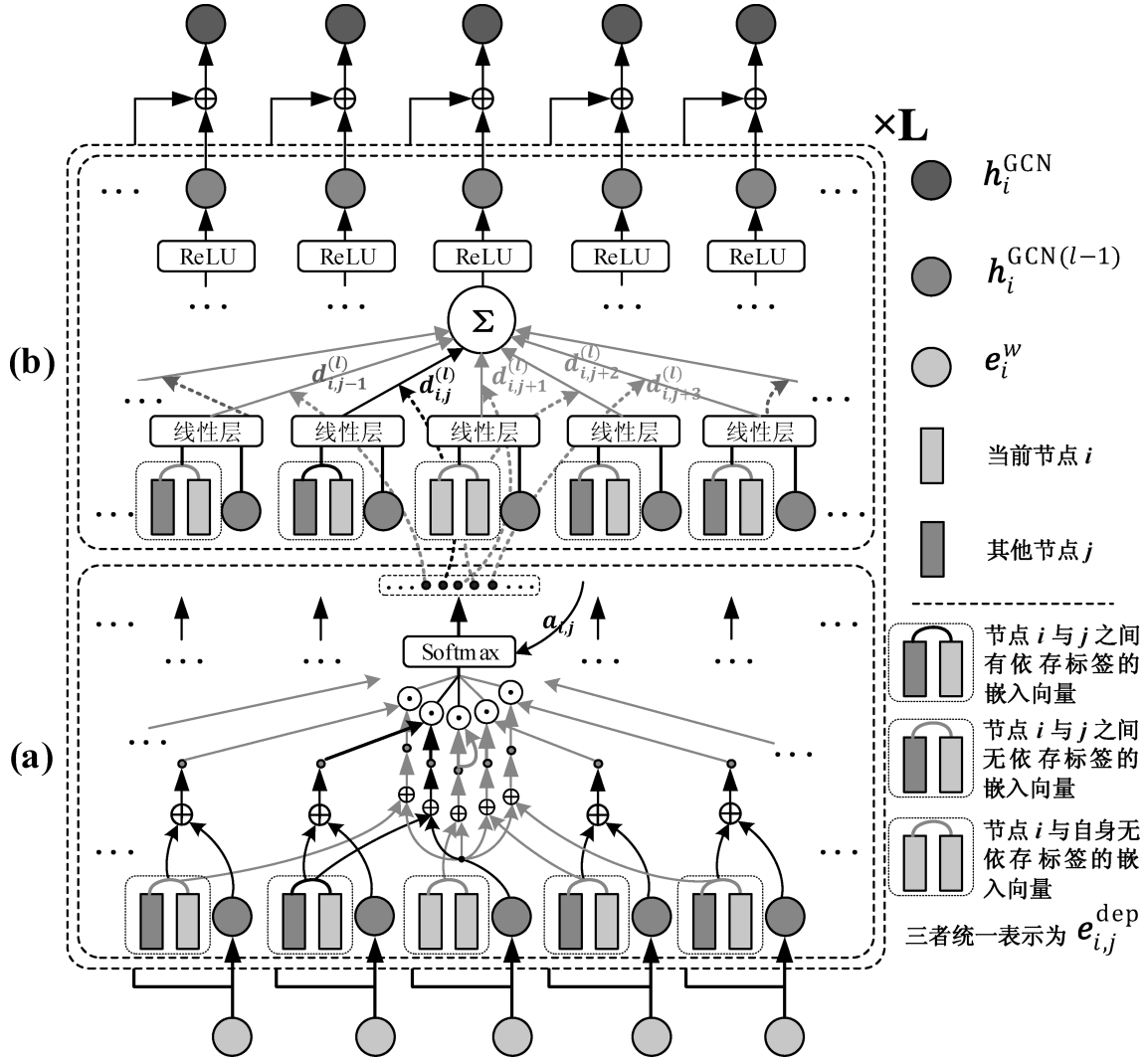


图2 ATT-GCN 模型结构图

(a) 依存标签注意力层 (b) 句法融合 GCN 层, 模型展示了当前节点 i 得到第 i 个隐层输出 $h_i^{GCN(l)}$ 的编码过程, 其中, $\odot$ 表示当前节点 i 的 $n_i^{(l)}$ 分别与其他所有节点 j 的 $n_j^{(l)}$ (包括与自身) 的内积, $\oplus$ 表示级联操作。

出的依存权重矩阵 $d_{i,j}$ 替代 GCN 输入的初始邻接矩阵 $a_{i,j}$, 以此有效捕获句子中的依存结构信息得到字节点的向量表示。改进后的 GCN 层的推理过程如图 2(b) 中层所示, 形式化过程如式(9)所示。

$$h_i^{GCN(l)} = \sigma \left(\sum_{j=1}^m d_{i,j}^{(l)} (W_a^{(l)} (h_j^{GCN(l-1)} + e_{i,j}^{dep}) + b^{(l)}) \right) \quad (9)$$

其中, σ 表示 ReLU 激活函数, $W_a^{(l)}$ 为 ATT-GCN 模型第 $l-1$ 隐层的第 j 个节点对应的隐层输出和第 i 个节点与第 j 个节点间的依存标签嵌入求和后共同的权重矩阵, $b^{(l)}$ 为它们共同的偏置项。结构表示模块最终的输出 h_i^{GCN} 为经过 L 层迭代的输出 $h_i^{GCN(L)}$ 与 e_i^w 的级联结果, 如式(10)所示。

$$h_i^{GCN} = h_i^{GCN(L)} \oplus e_i^w \quad (10)$$

2.5 语义融合

由于结构表示模块捕获了依存结构分析器可能存在的错误分词或依存结构信息, 导致所得结构编码含有噪声。为有效缓解上述噪声的影响, 本文采用注意力机制将句子上上下文信息与依存结构信息进行融合得到蕴含两者信息的语义向量表示。

将 BiLSTM 隐层输出 h_i^{LSTM} 与结构表示层输出 h_i^{GCN} 共同输入到语义融合层中进行注意力计算, 形式化如式(11)所示。

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax} \left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_K}} \right) \mathbf{V} \quad (11)$$

其中, $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ 分别表示查询向量、键向量和值向量, d_K 为 $\mathbf{Q}$ 和 $\mathbf{K}$ 的第二维度大小。为统一语义融合模块输入的维度, 本文令 $\mathbf{Q} = \mathbf{h}_i^{LSTM}, \mathbf{K} = \mathbf{V} = \mathbf{W}_g \mathbf{h}_i^{GCN}$,

其中, $\mathbf{W}_g$ 为线性层学习的参数矩阵。

为获得上下文和结构信息的稳定表示,采用多头注意力机制对上下文信息和结构信息表示进行融合,具体计算过程如式(12)、式(13)所示。

$$\text{head}_i = \text{Attention}(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V) \quad (12)$$

$$\text{Mhead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = (\text{head}_1 \oplus \cdots \oplus \text{head}_h) \quad (13)$$

其中, h 表示注意力头数, $\mathbf{W}_i^Q$ 、 $\mathbf{W}_i^K$ 、 $\mathbf{W}_i^V$ 分别表示第 i 次自注意力机制时的查询向量、键向量和值向量的投影参数矩阵,且 $\mathbf{W}_i^Q \in \mathbb{R}^{d_K \times \frac{d_K}{h}}$, $\mathbf{W}_i^K \in \mathbb{R}^{d_K \times \frac{d_K}{h}}$, $\mathbf{W}_i^V \in \mathbb{R}^{d_K \times \frac{d_K}{h}}$, $\oplus$ 表示结果级联操作。

因此由式(14)得到语义融合层的最终输出 $\mathbf{M}_i$ 。

$$\mathbf{M}_i = \text{Mhead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) \quad (14)$$

2.6 标签预测

深度神经网络可以捕获长距离的文本信息,但其无法对预测的标签序列添加约束条件。为有效避免上述情况的出现,本文采用 CRF 对标签序列进行约束和预测。

如图 3 所示,将融合上下文和结构信息的输出 $\mathbf{M}_i$ 作为状态特征即字节节点的状态分数表示,然后直接输入 CRF,且经过训练能够学习一个转移分数矩阵,矩阵中的元素为 $\{\text{B-T}, \text{M-T}, \text{E-T}, \text{S-T}, \text{O}\}$ 集合中任意两种标签组合的分数,表示各种组合的可能性。随着模型的迭代训练,CRF 可以通过转移分数矩阵逐渐地判断出标签组合是否符合语料中的标注规则,并在预测标签序列时进行规则约束,如标签序列中不可能出现连续的“B-T”或“E-T”。

综上预测过程可得句子 s 预测标签序列 $Y = \{y_1, y_2, \dots, y_m\}$ 的概率如式(15)所示。

$$P(Y | s) = \frac{\exp\left(\sum_i (\mathbf{W}_c^{y_i} \mathbf{M}_i + \mathbf{b}_c^{(y_{i-1}, y_i)})\right)}{\sum_{\tilde{Y}} \exp\left(\sum_i (\mathbf{W}_c^{\tilde{y}_i} \mathbf{M}_i + \mathbf{b}_c^{(\tilde{y}_{i-1}, \tilde{y}_i)})\right)} \quad (15)$$

其中, $\tilde{Y}$ 是所有有效标签序列集合且 $\tilde{y}_i \in \tilde{Y}$, $\mathbf{W}_c^{y_i}$ 和 $\mathbf{b}_c^{(y_{i-1}, y_i)}$ 分别表示 CRF 层的权重矩阵和偏置项,在预测最优标签时,使用 Viterbi 算法搜索标签组合概率最大的标签序列。

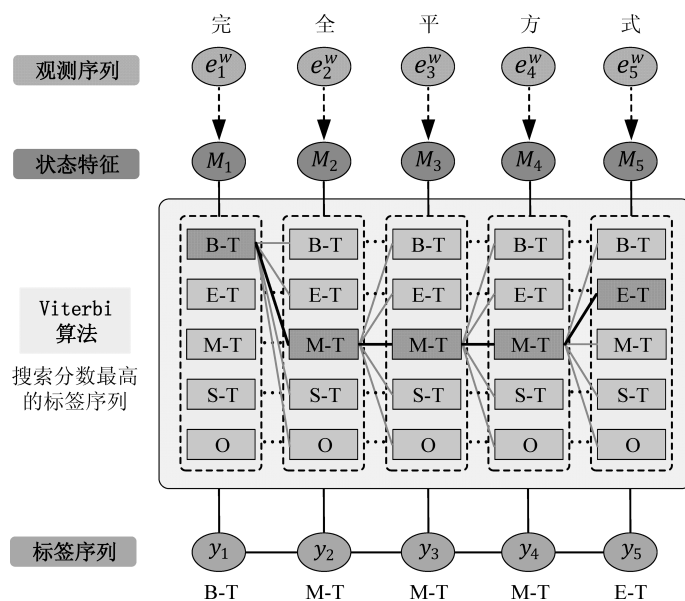


图3 CRF模型预测过程图

3 实验结果及分析

3.1 实验数据集及评价指标

实验在我们已构建的中学数学术语数据集^[18]上进行,该数据集内容来源于新人教版中学教材、试卷考纲、教案等数学相关文本,总共收集了 10 934 条句子,且考虑到语料中术语词分布的稠密性,绝大

部分句子都包含术语词。再通过编写数据预处理程序对这 10 934 条句子依次进行句子去重、数据集划分、自动标注等操作,最后对标注结果进行人工检查矫正得到中学数学术语数据集。

在不改变原数据集规模的情况下,我们综合借鉴了最新的百度百科、数学教材以及术语词典^[23]对原数据集的术语标注进行改进和优化,由此得到最终的中学数学术语数据集,详细信息如表 1 所示。

另外,我们构建了一个具有复杂语境的实例集,用于验证所提模型的实用性。

表 1 中学数学术语数据集详细信息

统计量	训练集	验证集	测试集	实例集
句子数	3.5 k	0.22 k	0.25 k	54
字数	12.5 k	8.9 k	9.5 k	890
术语数	13 909	231	400	188

标签一致性检验 为验证优化前后数据标注的客观性,我们借助第三方标注员的标注结果并采用三种指标来衡量不同标注结果之间的标签一致性。三个指标分别为一致率(C-rate)、肯德尔相关性系数(Kendall)和 Fleiss Kappa 系数(κ)。其中 C-rate 表示两种标注结果中相同字符标签占总字符的百分

比;Kendall 表示两种标注结果的统计依赖性,是一个从 0 到 1 范围内的非参数统计量,且 Kendall 值越接近 1 表示两个标注结果的相关性越强;Fleiss Kappa 是针对多个标注者进行一致性检验的方法,且当 $\kappa > 0.8$ 时表示标注员的标注有几乎完美的一致性。

我们将优化前后的数据标注结果分别与第三方标注员的标注结果进行 C-rate、Kendall、 κ 值的计算,且为避免标签不平衡的干扰,在计算 κ 值时不考虑所有标注者共同标为“O”的字符。图 4 中的对比结果表明我们优化后的标注结果与第三方标注员的标注结果更为贴切,其中 C-rate、Kendall、 κ 值较优化前分别高出 2.48%、11.44%、25.72%。特别地,对于优化后 $\kappa = 0.9960 > 0.8$,表示我们的标注结果与第三方标注员的结果有着几乎完美的一致性。

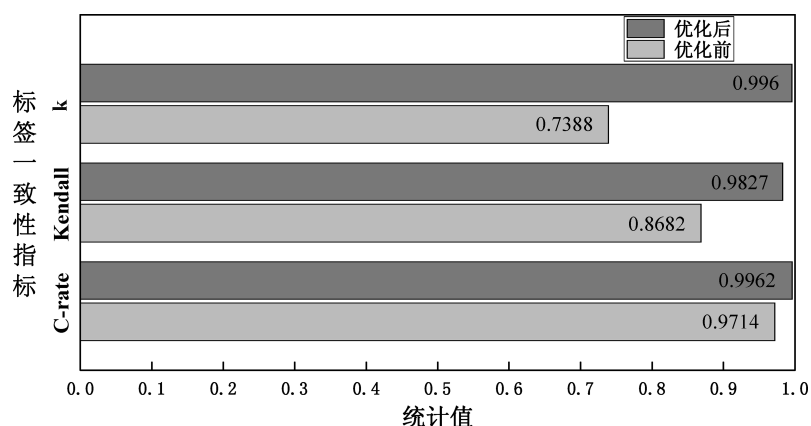


图 4 数据优化前后的标签一致性结果对比图

评价指标选取 实验采用准确率 (Accuracy, Acc.)、精确率 (Precision, P)、召回率 (Recall, R) 以及 F_1 -score (F_1) 值作为模型算法的评价指标。为保证实验的可靠性和减少噪声的干扰,我们对所提模型和其他基准模型都进行了 3 次重复实验,并取四种指标的平均值作为最后的验证结果。

3.2 实验参数设置

实验使用的深度学习模型框架为 PyTorch,使用的中文预训练模型中词向量嵌入维度为 768, BiLSTM 的神经元数设为 256,依存标签嵌入向量的维度设为 768 维,经过线性层处理后的 ATT-GCN 隐层输出为 200 维。ATT-GCN 的迭代层数 L 为 1,多头注意力机制的头数 h 为 5。为防止在训练过程中模型过拟合,设置 Dropout 的值为 0.4,初始学习率设为 0.001。

3.3 对比实验及结果分析

3.3.1 基准对比实验

为验证所提模型的有效性,本文选择以下模型作为基准方法:

BERT-BiLSTM-CRF 吴俊等^[15] 事先利用 BERT 预训练模型获取具有丰富语义的中文字词向量,并结合 BiLSTM-CRF 实现了中文专业术语抽取。

Lattice-LSTM Zhang 等^[12] 提出的一种综合编码词级特征和字符特征的 LSTM 模型,该模型基于外部词典匹配,对文本序列中所有潜在单词进行编码,并将得到的词级特征编码融入处理对应字符特征的 LSTM 编码单元。

BERT-BiGRU-CRF GRU (Gate Recurrent Unit) 与 LSTM 性能相近但参数更少,Lample 等^[24] 通过构建 BiGRU-CRF 模型解决了英文命名实体识

别任务。

LLA-CRF 华鑫等^[18]引用了 Lattice-LSTM 并提出了增强标签预测的标签注意力层,解决了中学数学部分多义术语词和嵌套术语词的抽取问题。

LLA-CRF 在 Lattice-LSTM 的基础上增加了标签注意力层,更好地学习了序列与标签的相关性。由表 2 可见,在 Acc.、 P 、 F_1 值上超过了 Lattice-LSTM。而 BERT-LLA-CRF 较 LLA-CRF 性能又有着大幅度的提升,这说明 BERT 能够大大提高传统深度学习模型的性能。进一步来看,在基于 BERT 的术语抽取模型中,BERT-BiGRU-CRF 和 BERT-BiLSTM-CRF 都是基于字实现的序列标注模型,虽然两者可以避免序列分词错误对模型性能的影响,但其向量表示欠缺更深层语义的表征,如依存结构信息。BERT-LLA-CRF 在融入 Lattice-LSTM 的词级语义表示基础上增设了计算序列与标签直接相关性的注意力机制,使模型更关注术语词的抽取,但 BERT 的引入反而干扰了词级信息在 LSTM 中的表示,并导致注意力层学习了序列与标签之间错误的相关性。从表 2 可以看出,Acc.、 P 、 F_1 值较 BERT-BiLSTM-CRF 有所下降。而本文所提的 DSL-ATRE 模型利用图神经网络捕获了句子的依存结构信息,并设计语义融合层来减少融合依存结构信息时可能存在错误分词和依存结构分析的干扰,实验结果也表明 DSL-ATRE 在四个指标上均优于基准模型 BERT-BiLSTM-CRF,其中 P 、 F_1 值分别提升了 2.21%、1.22%。

表 2 各模型术语抽取性能对比

(单位: %)

模型		Acc.	P	R	F_1
传统深度学习模型	Lattice-LSTM	94.02	62.30	69.67	65.78
	LLA-CRF	94.81	65.52	68.38	66.92
基于 BERT	BERT-BiGRU-CRF	96.62	80.68	79.43	80.05
	BERT-BiLSTM-CRF	96.97	81.61	80.98	81.29
	BERT-LLA-CRF	96.39	78.33	81.75	80.00
	DSL-ATRE*	97.03	83.60	79.95	81.73
	DSL-ATRE	97.06	83.82	81.23	82.51

此外,表 2 中 DSL-ATRE* 表示未引入依存标签注意力机制对输入的依存关系矩阵进行去噪,虽然模型性能相比 DSL-ATRE 有所下降,但 P 、 F_1 值还是比基准模型 BERT-BiLSTM-CRF 分别高出 1.99%、0.44%,这进一步表明所提框架是有效的。

3.3.2 消融实验

为进一步验证所提方法的有效性,我们在优化后的中学数学术语数据集上构建了两个消融实验,并将 BERT-BiLSTM-CRF 作为基准进行对比:①移除了 ATT-GCN 模块,直接将 BERT 字嵌入输入语义融合模块中与 BiLSTM 隐层输出进行多头注意力机制计算;②移除了语义融合模块,直接将 BiLSTM 与 ATT-GCN 的输出进行级联。

消融实验对比结果如表 3 所示,实验结果表明,在移除 ATT-GCN 模块后,模型没有融入依存结构信息,只是以 Attention 机制的形式对 BERT 字嵌入向量进行残差计算,这样得到的字向量语义依旧不足,导致此时模型的整体指标较基准略低。

表 3 消融实验结果

(单位: %)

模型	Acc.	P	R	F_1
基准	96.97	81.61	80.98	81.29
DSL-ATRE	97.03	83.82	81.23	82.51
-(ATT-GCN)	96.86	82.74	77.63	80.11
-(Attention)	96.97	83.24	79.52	81.33

对于去除语义融合模块后的模型在 P 值上较基准提升了 1.63%,这表明 ATT-GCN 输出与 BiLSTM 简单的级联确实也可以丰富字向量的语义表示信息,但 F_1 值只提升了 0.04%,这说明在去除语义融合模块后,模型在最后的标签预测阶段受到了错误信息的干扰。

而本文将语义融合模块与 ATT-GCN 共同考量,实现了一个既可为字向量丰富句法语义也可缓解错误分词或依存结构的中学数学术语抽取模型,实验结果也表明所提的 DSL-ATRE 模型在 F_1 值上较两个消融实验方法分别提升了 2.40%、1.22%。对于中学数学表达中一些极简句型,在分词任务中会出现错误的分词结果,从而产生无效的依存结构信息,这些噪声信息会被本文引入的去噪机制事先削减,而残余噪声会在语义融合模块计算上下文信息与依存结构信息的相关性权重时被减弱。如图 5 所示,语义融合模块依据上下文语境增强“成”与“线”的谓宾关系,赋予蕴含了“动成线”或“成线”等错误成词的结构信息低权重,从而达到在不同分词情况下成功抽取“点”和“线”两个单字术语的目的。这也进一步表明了模型在含有噪声的依存结构学习中有较好的鲁棒性。

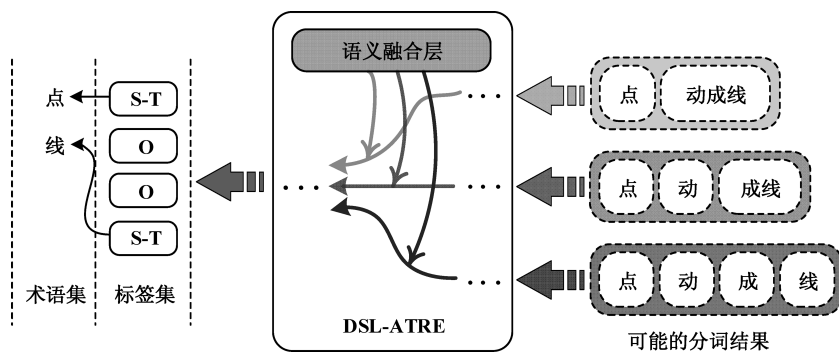


图5 不同分词环境下的 DSL-ATRE 术语抽取过程

3.3.3 实例抽取对比实验

为验证所提方法的实用性，我们在测试集上分别对基准模型 BERT-BiLSTM-CRF 和 DSL-ATRE 进行实例术语抽取，如表4展示了部分实例的嵌套术语抽取结果。分析发现 DSL-ATRE 模型对多重嵌套术语的抽取能力优于基准模型，比如表4中的“无限不循环小数叫无理数”，基准模型忽略了“无限”与“小数”之间的依存关系导致错

误抽取。而 DSL-ATRE 模型由于采用了图神经网络捕获句子的依存结构信息，有效捕获了“无限”“不循环”“小数”三者间的关联信息，从而实现准确抽取。但由于数学语言的精简性，一些实例在基准模型和 DSL-ATRE 上的抽取结果会表现为“过度嵌套”，如模型从“对立事件概率之间的关系”中抽取出了“对立事件概率”的错误结果。

表4 部分实例的嵌套术语抽取结果

实例片段	基准	DSL-ATRE	实际标注术语
除数概念是四则运算之一……	除数、运算	除数、四则运算	除数、四则运算
……则称 A 是完全平方式	平方式	完全平方式	完全平方式
……互为有理化因式	有理化、因式	有理化因式	有理化因式
对立事件概率之间的关系……	对立事件概率	对立事件概率	对立事件、概率
……与几个二次根式互为……	根式	二次根式	二次根式
……所以就有了阳性素数定理	素数定理	阳性素数定理	阳性素数定理
……若干个简单变元的……	简单变元	简单变元	变元
无限不循环小数叫无理数	不循环小数、无理数	无限不循环小数、无理数	无限不循环小数、无理数

此外，我们在由复杂语境例句构成的实例集上进行了术语抽取实验，结果如表5所示。实验结果表明在一些复杂语境下 DSL-ATRE 模型抽取术语的完整性和准确性均优于基准模型。其中复杂语境常常由中文存在的省略词造成，如表5中第三个实例，“等腰梯形”与“上底”之间以及“下底”与“中点”之间都省略了结构助词“的”“上底”之后也省略了“中点”，这种省略形式易导致基准模型对句子语义解析混乱，即将“下”错误地理解为方位词，将“底中点”错误解析成术语。同时，这类例句由于分词结果不佳导致在依存结构分析阶段产生较多的无效依存关系，即第三个实例在真实的依存结构分析结果中包含“梯形”和“下底”之间错误的并列关系，“下底”

作为形容词修饰“点”的关系，等，这易导致一些基于依存结构分析的模型忽略“上底”或无法完整抽取“中点”的情况，这是一种缺乏对上下文语义信息深层考量的表现。而 DSL-ATRE 模型综合考量了依存结构信息和上下文信息的重要程度，即模型融入“梯形”“和”“下底”作为分词结果中独立属性以及依存关系信息的同时，通过借助上下文信息，不仅增强了位于“梯形”与“和”之间的“上”“底”共同构成并列实体的可能性，还强化了“下底”对“中点”在语义上的修饰关系，从而正确预测出了“上底”“下底”“中点”才是实例所包含的真正术语。这进一步表明 ATT-GCN 与语义融合模块的组合可提升模型术语抽取的性能和鲁棒性。

表 5 基于复杂语境的部分实例术语抽取结果

实例	模型	抽取结果	实际标注术语
点动成线,线动成面,面动成体	基准 DSL-ATRE	点动成线,线动成面,面动成体 点动成线,线动成面,面动成体	点、线、面、体
底边上的中点和底边上的高三线合一	基准 DSL-ATRE	底边上的中点和底边上的高三线合一 底边上的中点和底边上的高三线合一	底边、中点、高、线
等腰梯形上底和下底中点的连线既是高也是平分线	基准 DSL-ATRE	等腰梯形 上底和下底中点的连线 既是高也是平分线 等腰梯形 上底和下底 中点的连线 既是高也是平分线	等腰梯形、上底、下底、中点、连线、高、平方线

补充: 第三个实例的分词结果为[等腰 梯形 上 底 和 下底 中 点 的 连 线 既 是 高 也 是 平 分 线]。

注: 加粗的黑色字体表示模型预测出的术语。

4 总结与展望

针对中学数学术语抽取的难点,本文提出了一种基于依存结构学习的中学数学术语鲁棒抽取模型,该方法利用图神经网络模型捕获依存结构信息,并通过注意力机制融合结构信息和上下文信息以实现融入依存结构信息的同时缓解错误分词或依存结构的影响。在中学数学术语语料集上的实验结果表明,本文所提模型性能优于 BERT-BiLSTM-CRF。进一步,我们打算提出针对中文依存关系矩阵更优的构造方式,同时考虑以多模态的形式融入与文本相关联的符号、公式和图片所蕴含的信息。

参考文献

- [1] 周长青.术语工作原则与方法(续四)(ISO DIS704)[J]. 中国科技术语, 1999(4): 24-32.
- [2] YUAN Y, GAO Y, ZHANG Y, et al. Cross-lingual terminology extraction for translation quality estimation[C]//Proceedings of the 11th International Conference on Language Resources and Evaluation, 2018: 3774-3780.
- [3] LIS, LI J, SONG T, et al. A novel topic model for automatic term extraction [C]//Proceedings of the 36th international ACM SIGIR Conference on Research and Development in Information Retrieval, 2013: 885-888.
- [4] RANI M, DHAR A, VYAS O. Semi-automatic terminology ontology learning based on topic modeling [J].Engineering Applications of Artificial Intelligence, 2017,63: 108-125.
- [5] PAULHEIM H. Knowledge graph refinement: A survey of approaches and evaluation methods[J]. Semantic Web,2017,8(3): 489-508.
- [6] 张雪,孙宏宇,辛东兴,等. 自动术语抽取研究综述[J]. 软件学报,2020,31(7): 2062-2094.
- [7] 周璐. 基础教育数学教材语言研究[D]. 厦门: 厦门大学博士学位论文,2019.
- [8] NPTOOL A. A detector of english noun phrases[C]// Proceedings of the Workshop on Very Large Corpora, 1995:48-57.
- [9] BOLSHAKOVA E, LOUKACHEVITCH N, NO-KEL M. Topic models can improve domain term extraction[C]//Proceedings of the European Conference on Information Retrieval, 2013: 684-687.
- [10] 李思良,许斌,杨玉基. DRTE:面向基础教育的术语抽取方法[J]. 中文信息学报,2018,32(3): 101-109.
- [11] 刘宇瀚,刘常健,徐睿峰,等. 结合字形特征与迭代学习的金融领域命名实体识别[J]. 中文信息学报, 2020,34(11): 74-83.
- [12] ZHANG Y, YANG J. Chinese ner using lattice lstm [C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018: 1554-1564.
- [13] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st Conference on Neural Information Processing Systems, Long Beach,CA,USA, 2017:1-11.
- [14] ZHENG D, ZHAO T, YANG J. Research on domain term extraction based on conditional random fields [C]//Proceedings of International Conference on Computer Processing of Oriental Languages. Springer, 2009: 290-296.
- [15] 吴俊,程垚,郝瀚,等. 基于 BERT 嵌入 BiLSTM-CRF 模型的中文专业术语抽取研究[J]. 情报学报,2020,

- 39(4): 10.
- [16] 翟羽佳, 田静文, 赵玥. 基于 BERT-BiLSTM-CRF 模型的算术语抽取与创新演化路径构建研究[J]. 情报科学, 2022, 40(4): 8-16.
- [17] 李韧, 李童, 杨建喜, 等. 基于 Transformer-BiLSTM-CRF 的桥梁检测领域命名实体识别[J]. 中文信息学报, 2021, 35(4): 83-91.
- [18] 华鑫. 基于序列标注模型的中学数学术语抽取方法研究[D]. 南昌: 江西师范大学硕士学位论文, 2021.
- [19] WU S, FEI H, REN Y, et al. Learn from syntax: Improving pair-wise aspect and opinion terms extraction with rich syntactic knowledge[C]//Proceedings of International Joint Conference on Artificial Intelligence, 2021: 3957-3963.
- [20] TIAN Y, CHEN G, SONG Y, et al. Dependency-driven relation extraction with attentive graph convolutional networks[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the Eleventh International Joint Conference on Natural Language Processing, 2021: 4458-4471.
- [21] WU S, FEI H, REN Y, et al. High-order pair-wise aspect and opinion terms extraction with edge-enhanced syntactic graph convolution[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 2396-2406.
- [22] DEVLIN, JACOB, CHANG, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [C]//Proceedings of Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2019: 4171-4186.
- [23] 王元. 数学大辞典[M]. 北京: 科学出版社, 2010: 100-316.
- [24] LAMPLE G, BALLESTEROS M, SUBRAMANIAN S, et al. Neural architectures for named entity recognition[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016: 260-270.



罗凯威(1999—), 硕士研究生, 主要研究领域为自然语言处理、术语抽取。

Email: luokaiwei@jxnu.edu.cn



黄琪(1993—), 通信作者, 博士, 讲师, 硕士生导师, 主要研究领域为社交网络分析、谣言检测、图神经网络、自然语言处理、信息抽取。

Email: huangqi@jxnu.edu.cn



罗文兵(1984—), 硕士, 高级实验师, 主要研究领域为自然语言处理、知识图谱。

Email: lwb@jxnu.edu.cn