

基于 BERT 模型的教育技术学领域实体抽取

胡慧婷, 李建平, 董振荣, 白欣宇

(东北石油大学 计算机与信息技术学院, 黑龙江 大庆 163318)

摘要:网络环境下资源丰富导致教育技术学信息量大,使得学习者认知效率低、注意力无法集中,最终偏离学习的目标并且无法完成特定的学习任务。为了解决学习者在网络学习中遇到的这些问题,该文提出一种结合 BERT-BiLSTM-CRF 的教育技术学主干课程命名实体识别方法,以提高学习者学习效率为目的。首先构建教育技术学主干课程命名实体识别数据集,将文本转换成计算机可识别的形式,使用 BERT 语言模型进行文本特征提取获取字粒度向量矩阵;然后使用双向长短期记忆网络(Bi-directional Long Short-Term Memory, BiLSTM)提取输入语句与上下文之间字与字的关系;最后使用条件随机场(Conditional Random Field, CRF)模型,根据标签之间的依赖关系提取全局最优的输出标签序列;最终得到教育技术学主干课程命名实体。实验结果表明,该模型的识别效果优于 CRF、BiLSTM-CRF,该模型的精确率、召回率和 F1 值均有提升,整体识别性能较高。

关键词:教育技术学;命名实体识别;BERT;双向长短期记忆网络;条件随机场

中图分类号:TP391.1

文献标识码:A

文章编号:1673-629X(2022)10-0164-05

doi:10.3969/j.issn.1673-629X.2022.10.027

Named Entity Recognition Method in Educational Technology Field Based on BERT

HU Hui-ting, LI Jian-ping, DONG Zhen-rong, BAI Xin-yu

(School of Computer and Information Technology, Northeast Petroleum University, Daqing 163318, China)

Abstract: Abundant resources in the network environment lead to a large amount of information in educational technology, which makes learners' cognitive efficiency low, unable to concentrate, and eventually deviates from the learning goal and fails to complete specific learning tasks. In order to solve these problems encountered by learners in online learning, we propose a named entity recognition method for the main curriculum of educational technology combined with BERT-BiLSTM-CRF, with the purpose of improving the learning efficiency of learners. First, a named entity recognition data set is constructed for the main course of teaching technology to convert the text into a computer-recognizable form, and the BERT language model is used for text feature extraction to obtain the word granularity vector matrix. Then BiLSTM is applied to extract the words and words between the input sentence and the context. Finally, the CRF model is used to extract the global optimal output tag sequence according to the dependency relationship between tags, and the named entity of the main course of education technology is obtained. Experimental results show that the recognition effect of such model is better than that of CRF and BiLSTM-CRF. The accuracy, recall and F1 value of such model are improved, and the overall recognition performance is higher.

Key words: education technology; named entity recognition; BERT; BiLSTM; CRF

0 引言

随着信息技术化被广泛应用于教育行业,计算机辅助教学是教育领域的主要方向以及大趋势^[1]。在互联网发展的大环境下,信息超限表现为信息迷航、信息爆炸、信息焦虑、信息污染等^[2]。尽管网络资源能够辅助学习者学习,但网络中海量的数据使教育技术学专

业学习者陷入信息超限,条理不清晰,学习者很难快速找到需要的信息,且不能帮助学生认清自身的薄弱之处。因此对教育技术学文本进行自动化的细致化知识点显得十分重要。使用结合命名实体识别(NER)教育技术学,提取出教育技术学中重要的术语,能有效提高学习者的学习效率。

收稿日期:2021-09-18

修回日期:2022-01-20

基金项目:黑龙江省高等教育教学改革项目(SJGY20190098)

作者简介:胡慧婷(1996-),女,硕士,研究方向为智能教育、教育技术管理理论及应用;李建平(1976-),男,博士,教授,硕导,研究方向为网络安全、物联网和智能计算;通讯作者:董振荣(1968-),男,硕士,研究员,硕导,研究方向为教育技术管理理论及应用。

教育技术学专业术语知识图谱可以从多源平台收集整理海量信息和知识,并能将知识及其关系可视化,为提高学习者学习效率提供了极大的帮助。教育技术学专业术语知识图谱主要包括实体抽取、关系抽取以及属性抽取等,实体抽取又称为命名实体识别(NER),是构建知识图谱的首要工作^[3]。

1 相关研究

NER 是自然语言处理任务中的基本步骤之一,主要是从非结构化文本中识别出句子中的人名、地名、机构名等实体^[4]。早期基于规则和词典的模式匹配方法,翟菊叶等人^[5]使用 CRF 与规则相结合的方法对中文电子病历进行命名实体识别,但该方法的缺点是需要领域专家制定大量的规则,领域词典需要定期维护,通用性不高,所以学者们使用机器学习方法来解决这一问题。传统机器学习的命名实体识别方法主要有隐马尔可夫模型、最大熵模型、支持向量机模型和条件随机场模型,王红斌等人^[6]将隐马尔可夫模型和条件随机场模型应用于泰语领域,尽管机器学习的方法避免使用手工构造规则模板,但是繁琐的特征工程依然需要大量人工参与。随着深度学习近几年的发展,由于其具有较强的泛化能力,使得命名实体识别领域逐渐使用该方法,取得了很好的效果^[7];石春丹等人^[8]提出一种基于双向门控循环单元的实体抽取模型,该模型结合门控循环单元结构简单、参数更少的特点,以 GRU 并发进行多尺度的处理加速,从而更加快捷地完成序列数据的计算;秦娅等人^[9]将 CNN-BiLSTM-CRF 应用于网络安全领域,大大提高了识别精度;Yu 等人^[10]采用 BERT 模型,提出了一种融合句子内容和上下文信息的隐式句子模型,对输入进行重构,有效提高了分类模型的性能;黄炜等人^[11]提出了一种基于 BiLSTM-CRF 的涉恐信息,获得了更高的分类准确率,但在文本数据中很多字词会根据文本语境的不同有不同的含义,该模型难以学习到字词的不同特征;李明扬等人^[12]在 BiLSTM-CRF 模型中加入了自注意力机制,在 Weibo NER 语料库上,能够捕捉上下文信息,提升模型的识别精度;刘鹏等人^[13]在提出矿山灾害模型时,提出 HIDCNN 模型,采用迭代堆叠 DCNN,避免了简单堆叠多个 DCNN 导致的模型参数量大进而使得模型训练困难的问题,提高了模型训练效率和检测的准确性。

因教育技术学专业术语识别是一种特定领域的命名实体识别,关于其研究相对较少,所以缺乏大量的专业语料库。针对以上问题,该文采用自制数据集,通过人工标注构建实体语料;再利用 BERT 模型在预训练数据集中获取词向量表示,然后将词向量输入到

BiLSTM 中提取特征,最后使用 CRF 进行实体标注修正后输出。以 BERT-BiLSTM-CRF 的命名实体识别方法,抽取教育技术学专业术语,具有较高的准确性。

2 模型设计

2.1 教育技术学文本语料特征分析

由于教育技术学领域没有开放的数据集,该文手动构建了一个语料集用于研究。因《教育技术学研究方法》是教育技术学科必修课程,对学生掌握该专业的技能具有承上启下的作用,该文以教育技术学专业教材《教育技术学研究方法》来构建命名实体识别数据集。

根据教学大纲以及目录,将实体分为 3 类:“研究概述类”、“研究方法类”与“数据分析类”。

教育技术学语料集共 10 350 句 320 140 个字,所用汉字 2 150 个,具体频率如表 1 所示。

表 1 教育技术学主干课程实体出现频率

实体类型	实体开始	实体内部及结尾	实体数量	出现次数
研究概述类	B-SUM	I-SUM	156	254
研究方法类	B-MET	I-MET	684	1 520
数据分析类	B-DAT	I-DAT	68	263

教育技术学语料通过 BIO 实现对序列数据的联合标注,其中,“B-”表示命名实体中的第一个字,“I-”表示命名实体中间字和结尾字,“O”表示非实体字符,教育技术学实体标注示例如图 1 所示。

调 查 研 究 法 是 对
B-MET I-MET I-MET I-MET I-MET O O
已 形 成 的 事 实 的 考 察 和 研 究
O O O O O O O O O O O O
方 差 分 析 又 称
B-DAT I-DAT I-DAT I-DAT O O
变 异 数 分 析 或 F 检 验
O O O O O O O O O O

图 1 实体标注方法及实体数量

2.2 整体框架

BERT-BiLSTM-CRF 教育技术学领域术语抽取模型整体结构如图 2 所示。

因为教育技术学主干课程实体的构建中,文字中的内容隐含于在上下文间、体现在字与字中的前后关系上。因此,首先使用 2.1 节生成的教育技术学命名实体识别数据库,作为训练特征输入到 BERT 预训练语言模型层中,在本模型层中被标注的字符集语料经过该层将每个字符转化为低维词向量。其次经过 BiLSTM 模块进行全局特征提取,将上一层输出的词向量序列输入到这一层进行语义编码,自动提取句子

特征。最后是 CRF 层,利用这一层解码输出概率最大的预测标签序列,实现教育技术学研究方法术语的抽取。

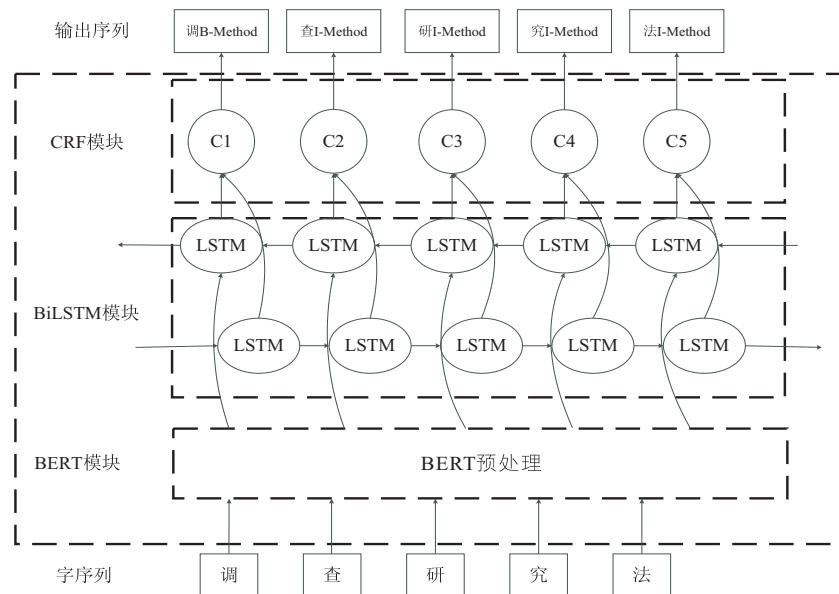


图 2 BERT-BiLSTM-CRF 模型

2.2.1 BERT

因为教育技术学主干课程的知识点分布跨度大,一个知识点涉及多个知识点的概念,主要知识点层级由多个分级的知识点构成。而 BERT 是一种自然语言处理预训练语言表征,能够捕捉到文本语料的上下文信息,学习连续文本片段之间的关系并能够计算词语之间的相互关系。以 BERT 进行教育技术学主干课程知识点特征提取,不仅包含词上下文的语境或语义,而且携带上下文语境信息的静态词向量。

BERT^[14] 预训练模型主要由双向 Transformer 编码结构组成,其中 Transformer 由自注意机制和前馈神经网络组成,其与 LSTM 相比能捕捉更远距离的序列特征。

首先教育技术学语料库向量经过三个不同的全连接层,在 Encoder 部分得到 Q (语料库中当前词的表示)、 K (Encoder 中语料库其他词的表示)、 V (Encoder 中其他词的表述) 三个向量;在 Decoder 部分,得到解码的 Q (Decoder 中当前词的表达)、 K (Encoder 结束后所有输入词的表达)、 V (Encoder 结束后所有输入词) 三个向量;然后 Q 和 K^T 进行矩阵相乘得到单词和其他单词相关程度的向量 QK^T ,最后将标准化的 K^T 放入到 Softmax 激活函数中,得到词与词之间的关联度向量,再乘以 V 得到最终向量。如公式所示:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

再通过多头结构拼接向量结果:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}^1, \dots,$$

$$\text{head}^h)W \quad (2)$$

$$\text{head}_1 = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \quad (3)$$

2.2.2 BiLSTM

在教育技术学主干课程实体的构建中,文字中的内容隐含于上下文间、体现在字与字中的前后关系上。而 BiLSTM 不仅可以保存短期的输入,对双向的语义关系也能够更好地捕捉。因此该模型以 BiLSTM 模型作为字处理器,提取单个字的信息以及输入语句内字与字之间的关系。

BiLSTM 由前向 LSTM 和后向 LSTM 组成用以提取全局的上下文特征^[15]。LSTM 是一种特殊的循环神经网络,相比于传统的 RNN, LSTM 神经元结构创新地采用了三个门控制单元,分别为输入门、输出门和遗忘门^[16]。

遗忘门决定遗忘神经元中的哪些信息:对前一时刻的隐层状态 h_{t-1} 与当前时刻的输入词 X_t , 选择要遗忘的信息,计算方式如公式(4)所示:

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \quad (4)$$

其中, σ 为激活函数, W_{xf} 为输入项 X_t ; W_{hf} 为输入项 h_{t-1} ; W_{xf} 和 W_{hf} 组成遗忘门的权重矩阵 W_f , b_f 为偏置项。

输入门控制当前信息:通过前一时刻的隐层状态 h_{t-1} 与当前时刻的输入词 X_t , 选择要记忆的信息,输出记忆门的值 i_t 与临时细胞状态 C_t , 计算公式如公式(5):

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \quad (5)$$

其中, W_i 为权重矩阵, b_i 为偏置项。当前时刻单元状态 c_t , 由上一轮的输出和当前的输入确定, 如公式

(6):

$$c_t = f_t c_{t-1} + i_t \tanh(W_{xc} x_t + W_{hc} h_{t-1} + b_c) \quad (6)$$

其中, c_{t-1} 为前一个的单元状态, f_t 为遗忘门。

输出门:决定的输出信息,计算如公式(7):

$$o_t = \sigma(W_{xo} x_t + W_{ho} h_{t-1} + W_{co} c_t + b_o) \quad (7)$$

输入门和单元状态确定了长短时记忆神经网络的输出,如公式(8):

$$h_t = o_t \tanh(c_t) \quad (8)$$

其中, h_t 表示 t 时刻的隐藏状态, \tanh 是正切激活函数。通过三个门的控制,使得 LSTM 具有长序列特征的内存功能,同时解决了 RNN 训练过程中出现的梯度消失和梯度爆炸问题。因此 BiLSTM 构建模型,并根据文本中词的分分布式自动提取特征,生成上下文预测的标签。

2.2.3 CRF

因为 BiLSTM 的分类方式忽略字符对应得分,会导致预测出非合法实体类型情况,而 CRF 的作用是对识别结果进行进一步的修正,即提取标签之间的依赖关系,使得识别的实体满足标注规则^[16]。其主要的实现方法是给定一个输入序列 $X = (x_1, x_2, \dots, x_n)$, 其对应的预测序列为 $Y = (y_1, y_2, \dots, y_n)$, 通过计算 Y 的评分函数,得到预测序列 Y 产生的概率,最后计算当预测序列产生概率的似然函数为最大时的预测标注序列作为输出^[17]。其中预测序列 Y 的评分函数的计算方法如公式(9)所示:

$$s(X, Y) = \sum_{i=1}^n X_{y_{i-1}, y_i} + \sum_{i=1}^n P_{i, y_i} \quad (9)$$

其中, X 表示转移分数矩阵, X_{y_{i-1}, y_i} 表示标签 y_{i-1} 到标签 y_i 的分数, P_{i, y_i} 表示第 i 个词映射到标签 y_i 的非归一化概率。该文以 Softmax 函数来计算教育技术学语料预测序列概率 $p(Y|X)$:

$$p(Y|X) = \frac{e^{s(X, Y)}}{\sum_{\tilde{y} \in Y_i} e^{s(X, \tilde{Y})}} \quad (10)$$

两头取对数得到预测序列的似然函数:

$$\ln(p(Y|X)) = s(X, Y) - \ln\left(\sum_{\tilde{y} \in Y_i} s(X, \tilde{Y})\right) \quad (11)$$

解码时, \tilde{Y} 表示真实的标注序列, Y_x 表示所有可能的标注序列 Y , 通过动态规划算法得到最大分数的输出预测标签序列,即为 CRF 预测的最终标注序列:

$$Y^* = \arg \max_{\tilde{Y} \in Y_x} s(X, \tilde{Y}) \quad (12)$$

3 实验分析

3.1 实验环境

实验模型的运行环境为 64 位 Ubuntu18.04 操作系统,具有实验的训练环境如表 2 所示。

表 2 实验环境

操作系统	CPU	GPU	Python	Tensorflow	内存
Linux	Interl(R)	NVIDIA RTX 3080 (11 GB)	3.8	1.14.0	64 GB
	Xeon(R)				
	Gold				
	5118				
	CPU @ 2.30 GHz (12-Core)				

3.2 数据集与评价指标

实验所用的数据集以教育技术学专业课本为例,对文本进行标注,根据教学大纲以及目录,将实体类别分为 3 种,分别为研究概论、研究方法以及数据分析。

该文采用准确率 P 、召回率 R 和 F1 值 3 个指标作为评价标准,计算公式如公式(13)~公式(15):

$$P = \frac{\text{正确的个数}}{\text{正确的个数} + \text{错误的个数}} \times 100\% \quad (13)$$

$$R = \frac{\text{正确标注的实体个数}}{\text{语料种实体的总个数}} \times 100\% \quad (14)$$

$$F1 = \frac{2PR}{P + R} \times 100\% \quad (15)$$

3.3 实验结果与分析

从表 3 可以看到,文中方法 P 为 81.72%,这是因为教育技术学领域中命名实体词组合比较灵活,相较于 CNN-BiLSTM-CRF 神经网络模型自动学习实体特征,采用自适应的特征模板从窗口提取的特征往往更有效。 R 为 75.73%,F1 值为 78.61%,因为文中方法相较于 BiFlaG 更侧重于字符级表示向量与词嵌入向量连接,同时 CRF 损失函数中转移概率矩阵可学习到很多约束的规则,使预测结果更加准确。文中方法相较于 IDCNN^[18],能够学习到字级词级的特征,充分考虑到字词在文本不同语境的不同含义,不存在深度神经网络带来的模型有效信息衰减问题。文中方法相较于 HIDCNN 模型,解决了长距离依赖的问题,不仅保存了模型前后时刻的状态信息,也保存了 label 之间的相互关系,因此在 R 值与 F1 值上高于 HIDCNN 模型。

表 3 命名实体识别实验结果

模型	$P/\%$	$R/\%$	$F1/\%$
BiLSTM-CRF	66.62	69.67	68.11
CNN-BiLSTM-CRF	65.56	67.05	66.30
BiFlaG	78.73	75.81	77.24
IDCNN	80.41	75.78	78.03
HIDCNN	81.80	75.35	78.44
Ours	81.72	75.73	78.61

所采用的基于 BERT-BiLSTM-CRF 的教育技术学专业术语抽取模型在 P 、 R 和 F1 值 3 个方面都优于

其他模型。

如表 4 所示,仅使用 BERT 模型时分类精度较低,因为只通过迁移学习了通用领域的词语信息,在加入了 BiLSTM 训练本文的教育技术学命名实体识别数据集后, P 、 R 、 $F1$ 值均有提高。原因有二,第一是因为文中教育技术学命名实体识别的有效性,第二是 BiLSTM-CRF 通过获取词语前后的信息融入词语的上下文信息,可以清楚地区分语料库中的多义词。经过 CRF 再次修正后,通过大规模语料的预训练,可以有效提高教育技术学领域命名实体的识别精度。证明了所采用的基于 BERT-BiLSTM-CRF 的教育技术学专业术语抽取模型的有效性。

表 4 实验结果有效性验证

模型	$P/\%$	$R/\%$	$F1/\%$
BERT	63.85	62.00	62.91
BiLSTM-CRF	66.62	69.67	68.11
Ours	81.72	75.73	78.61

4 结束语

对教育技术学领域命名实体识别进行了研究,设计了一种基于 BERT 的教育技术学文本命名实体识别方法。首先根据网络资料以及教育技术学主干课程的教材《教育技术学研究方法》完成了教育技术数据准备工作,提出了基于“研究概述”、“研究方法”以及“数据分析”三个大类的教育技术学命名实体识别数据集。然后,根据数据集,知识点跨度大,字与字之间联系紧密等特点,设计适用于文中的 BERT-BiLSTM-CRF 模型,完成对文本数据字级别的抽取,充分学习上下文的特征并且能提取出全局最优标注序列,最终得到教育技术学主干课程实体。在实验中进行了验证,为教育技术学主干课程知识图谱的构建提供了技术支撑。

参考文献:

- [1] 张剑平,陈仕品. 计算机辅助教学的智能化历程及其启示[J]. 教育研究,2008(1):76-83.
- [2] 肖丽平,姜策群. 互联网发展环境下“信息超限”问题研究[J]. 图书馆学研究,2018(10):16-21.
- [3] 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.
- [4] 焦凯楠,李 欣,朱容辰. 中文领域命名实体识别综述[J]. 计算机工程与应用,2021,57(16):1-15.
- [5] 翟菊叶,陈春燕,张 钰,等. 基于 CRF 与规则相结合的中文电子病历命名实体识别研究[J]. 包头医学院学报,2017,33(11):124-125.
- [6] 王红斌,邵洪奎,沈 强,等. 泰语人名、地名、机构名实体识别研究[J]. 系统仿真学报,2019,31(5):1010-1018.
- [7] 陈曙东,欧阳小叶. 命名实体识别技术综述[J]. 无线电通信技术,2020,46(3):251-260.
- [8] 石春丹,秦 岭. 基于 BGRU-CRF 的中文命名实体识别方法[J]. 计算机科学,2019,46(9):237-242.
- [9] 秦 娅,申国伟,赵文波,等. 基于深度神经网络的网络安全实体识别方法[J]. 南京大学学报:自然科学版,2019,55(1):29-40.
- [10] YU Gaihong, ZHANG Zhixiong, LIU Huan, et al. Masked sentence model based on BERT for move recognition in medical scientific abstracts[J]. Journal of Data and Information Science,2019,4(4):42-55.
- [11] 黄 炜,黄建桥,李岳峰. 基于 BiLSTM-CRF 的涉恐信息实体识别模型研究[J]. 情报杂志,2019,38(12):149-156.
- [12] 李明扬,孔 芳. 融入自注意力机制的社交媒体命名实体识别[J]. 清华大学学报:自然科学版,2019,59(6):461-467.
- [13] 刘 鹏,魏卉子,鹿晓龙,等. 基于新型卷积神经网络构建矿山灾害事件检测模型[J]. 中文信息学报,2020,34(10):59-68.
- [14] DEVLIN J, CHANG M W, LEE K, et al. Bert: pre-training of deep bidirectional transformers for language understanding [C]//Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies. Minneapolis: Association for Computational Linguistics,2019:4171-4186.
- [15] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation,1997,9(8):1735-1780.
- [16] GRAVES A, SCHMIDHUBER J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures[J]. Neural Networks,2005,18(5-6):602-610.
- [17] LAFFERTY J, MCCALLUM A, PEREIRA F. Conditional random fields: probabilistic models for segmenting and labeling sequence data[C]//Eighteenth international conference on machine learning. San Francisco: Morgan Kaufmann Publishers Inc,2001:282-289.
- [18] 刘玉成,王传生,杨 晶. “雨课堂”教学模式的“IDCNN+”结构化分析与实证研究[J]. 远程教育杂志,2019,37(1):94-103.