DOI:10.13878/j.cnki.jnuist.2019.06.011

何彬[1]　李心宇[1]　陈蓓蕾[2]　夏盟[1]　曾致中[1]

基于属性关系深度挖掘的试题知识点标注模型

摘要

在各类在线学习系统中,为了给学生提供优质的学习资源,一个基础性的任务是对大量未标注的试题进行知识点标注.已有标注方法通常基于人工专家标注或者采用传统机器学习方法.在实际应用中,这些方法普遍存在成本过高、标注精准度不足等局限.为此,本文提出了一种基于属性关系深度挖掘的试题知识点标注模型.首先,利用句法语义模型和结构语义模型分别从试题文本和试题图形中抽取试题的显性属性关系.然后,利用蒙特卡罗树搜索构建问题求解框架,挖掘试题的隐含属性关系.最后,结合学科知识图谱,将属性关系映射到知识图谱空间,生成试题知识点.实验结果表明,所提出的方法能够有效地进行试题知识点标注,将对学生认知诊断、个性化试题推荐等具有一定的实际应用价值.

关键词

知识点标注;属性关系挖掘;句法语义模型;结构语义模型;蒙特卡罗树搜索

中图分类号 TP391

文献标志码 A

收稿日期 2019-10-15

资助项目 国家自然科学基金(61877026);中央高校基本科研业务费资助项目(CCNU19QN036,CCNU19QN031)

作者简介

何彬,男,博士,讲师,主要从事智能教育相关理论、技术和系统研究.hebin@ mail.ccnu.edu.cn

1 华中师范大学 国家数字化学习工程技术研究中心,武汉,430079

2 湖北大学 教育学院,武汉,430062

0 引言

随着智能教育和在线教育的发展,学生能够非常便捷地从各类开放学习平台获取个性化学习所需的学习资源.试题作为一种重要的学习资源,在学生认知诊断和个性化学习推荐中被广泛使用.然而,随着在线试题资源的爆发式增长,如何自动准确和高效地进行在线试题知识标注,已成为智能教育背景下精细化适应性学习服务的研究热点.

试题的知识点是描述试题理解以及试题求解用到的知识的集合[1].准确地描述一道试题的知识点,对于根据学生的答题记录诊断其各个知识点的掌握程度,准确定位薄弱环节,进而开展个性化资源推荐和学习服务有极大帮助.本文将研究如何对包含文字和图形的试题进行精准的自动化知识点标注.目前主要存在两种试题知识点标注方法:一是完全由专家对试题进行分析并对试题知识属性进行标定[2].由于知识属性标定的复杂性,当知识属性较多或题量较大时,完全由专家来标定存在工作量大、主观性强、知识粒度太粗等问题.二是在部分试题知识属性标定结果的基础上,采用机器学习的方法对其他试题的知识属性进行估计[3-4].这类方法普遍存在未结合教研经验,知识标注的丰富度不足,尤其对标注语料少的知识点的预测效果极差,要想达到高质量的知识标注效果,语料库建设成本极高.

本文基于人工智能技术自动挖掘试题中的知识属性和知识关联关系,建立适用于基础教育工程学科试题的知识点自动标注框架.该框架使用题目理解技术实现试题知识的自动标注,构建句法语义模型来获取试题文本中显性和隐含的知识信息,利用图形关系抽取技术,提取图形中的知识信息,最后通过与标签库的映射完成知识点标签的自动生成.将本文方法实验结果与人工标注的结果分别与专家标准数据进行对比,验证了该方法的准确性、丰富性和学习性.

本文的主要贡献如下:

1) 提出了一种基于关系挖掘的试题知识点标注方法.该方法通过深度挖据题面直接陈述的以及求解题目过程中隐含的知识关系,全面精准标注试题知识点.

2) 设计了一种文字和图形通用的属性关系挖掘方法,并在初中物理电学试题集上开展了有效性验证.

3) 设计了一种试题隐含属性关系挖掘算法,利用蒙特卡罗树搜

索框架,实现初中物理电学问题逻辑关系推理,进而获取问题求解所需的试题隐含属性关系.

1 相关研究

试题知识点标注一般指的是为学习资源添加可以概括其知识内容的信息,包括对学习资源的含义进行理解、抽取学习资源中的核心文本或词汇、对学习资源做知识点的概括等.传统的标注方法以手工标注为主,人为对资源内容进行分析,确定其知识点标签.例如用户使用 DOMEO 工具[2]通过记笔记和标记等方式添加标签,各科习题集上题目的知识分类也是由人工编辑完成等.随着信息技术的发展,知识点标注技术逐渐转向自动标注,各种自动标注方法被提出,目前常见的方法可以分为以下几类:

1)基于词汇匹配的标注方法,例如词频统计、关键词匹配等.周菊明等[5]提出了学习资源智能标注系统,通过 TF-IDF 算法获取在当前学习资源中的出现的频率较高,同时在其他文本中相对较少的关键词,将该词作为分类的标签.Vanderwende 等[6]通过从文档集中选择句子来生成摘要,统计摘要中词汇出现的概率来完成主题聚焦,最后根据主题进行句子简化,完成简化摘要的自动生成.这类方法可以有效提取出文本资源中的关键词,但是对于文本表达中的隐含信息,比如需要用到的公式定理,却很难在缺少对应关键词的情况下成功挖掘.

2)基于词、本体、知识库、语义网的标注方法,通过文本资源中的特定词汇,构建元数据或知识库,并在此基础上实现标注.如戚欣等[7]提出了一种基于本体知识库的自动语义标注方法,首先根据语义词典的逻辑结构识别到文本中的命名实体,通过语义消歧,完成了校园新闻文本的语义标注.闫喜亮[8]利用表述逻辑的情感词汇本体,实现了网络教育资源情感属性的自动标注.蒋婷[9]通过抽取学科领域的本体术语,然后进行去重整合,结合模板实现本体概念的形成,并构建概念等级与非等级抽取模型,利用元数据和本体概念间的关联实现学术论文的语义标注.何中山[10]针对初中数学领域的学习资源,基于知识本体识别出其中的学科知识内容,然后通过本体距离计算出文档内容聚合度,获取到最相关的文档,根据词频统计提取元数据,实现语义标注.这些方法在各自的部分领域都取得了不错的效果,但在结构化领域知识的自动标注方面其性能还有待提升.

3)基于机器学习的标注方法.知识点标注过程可作为多标签文本分类问题,首先建立文本特征的描述模型,如基于词袋模型(bag-of-words)的向量空间模型(Vector-Space-Model,VSM),将文本表示成词表维度的向量,用来训练 SVM、朴素贝叶斯、决策树等分类模型[11-13].针对 VSM 特征稀疏问题,研究者们进一步提出了 LSA、LDA[14-15]等一系列隐语义分类方法,取得了不错效果.然而,试题具有文本短、领域属性强、样本分布不均、标签层次化等特点,无论是向量空间模型还是隐语义分析都仅利用了文本中的浅层语义信息,缺乏对试题短文本词序结构、实体关联等知识语义的有效表达和感知.同时,这些方法也缺乏教育专家知识和层次化标签结构的考虑,导致标注结果的可用性和可解释性不足.

对工程学科而言,图形是试题知识点的另一个主要载体.与传统图形理解不同,图形知识点挖掘需要识别图形符号并解析其表达的知识语义.传统图形理解多集中在图形符号的识别上,如几何图形中的基本几何元素(如点、线和圆)的检测和识别[16]、物理电路图符号识别[17-18]等.这类方法将传统图像处理算法与基于领域知识的结构分析相结合,易于实现.但是由于需要为不同学科设计单独分析算法和结构分析模型,不易扩展和维护.近年来,基于深度学习的 end-to-end 图形理解方法正逐步发展.通过神经网络模型实现图像元素的分割识别[19],特别是引入注意力机制后,图像元素之间的空间位置信息也能被正确检测,由此形成端到端的图形知识理解方法.该方法在公式理解、图像语义摘要等应用中取得了很好的效果,但是对于语义线索单一、语义模型复杂的领域知识挖掘,仍然面临挑战.

2 试题知识标注框架

试题的属性关系是题面显性和隐含表达的构建问题求解框架所需的知识语义.前期研究发现[20-22],利用文本句法语义模型和图形结构语义模型等题目理解技术挖掘试题属性关系,是获取问题求解所需知识信息的重要和有效技术手段.本文基于上述研究成果,将试题的知识点特征挖掘视为试题的属性关系挖掘问题,进而构建试题知识点标注模型.如图 1 所示,本文提出的试题知识点标注框架可分为题面关系抽取、隐含关系抽取和知识点标签生成 3 个主要阶段.

文本属性关系挖掘是指从题目文本中发现知识点语义特征.本文采用优化的句法语义混合模型

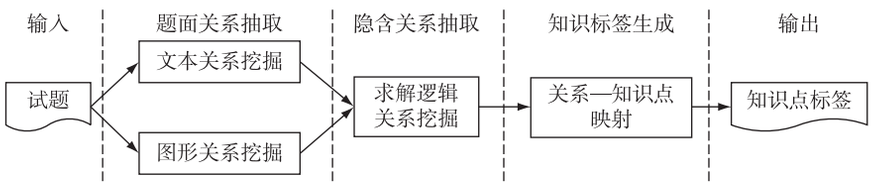

图 1　试题知识点标注框架

Fig. 1　Framework of knowledge points annotation

(Textural Syntax-Semantics, T-S^2),通过对学科知识表达中的名词、动词、数词、量词等统计学习,建立知识表达词类、句法模型,发现问题中的知识语义特征.句法分析主要是通过对句子或短语结构的分析,来确定句子中各个词和短语之间的关系以及在句子中的作用,并实现这些关系的层次化表达和句法结构的规范化.语义分析是指将分析得到的句法成分与应用领域中的目标表示相关联.如果依次独立分析,会使语法和语义分离,无法准确获取句法的结构.一个典型的 T-S^2 模型具有如下结构:

$$\boldsymbol{M}_{\mathrm{t}}=(\boldsymbol{K}_{\mathrm{t}},\boldsymbol{P}_{\mathrm{t}},\boldsymbol{R}_{\mathrm{t}}), \tag{1}$$

其中 $\boldsymbol{K}_{\mathrm{t}}$ 代表关键词元素,$\boldsymbol{P}_{\mathrm{t}}$ 是 POS 词性模式,$\boldsymbol{R}_{\mathrm{t}}$ 为相关实体之间的关系,$\boldsymbol{\Sigma}=\{\boldsymbol{M}_{\mathrm{t}i}=(\boldsymbol{K}_i,\boldsymbol{P}_i,\boldsymbol{R}_i)\mid i=1,2,\cdots,m\}$ 称为 T-S^2 模型池.

在图形属性关系挖掘上,基于学科知识指导图形要素的空间信息和局部连接关系(如电路图中的元件和元件之间的串并联关系等),可以"翻译"成一组确定的数学表达式序列[23],通过迭代的方式检索图形中的局部连接关系,实现题目图形中的知识关系抽取.本文采用优化的结构语义混合模型(Diagram Structure-Semantics, D-S^2),建立图形基元与知识语义之间的关联,通过检测图形中的图形基元,实现图形属性关系挖掘.一个典型的 D-S^2 模型具有如下结构:

$$\boldsymbol{M}_{\mathrm{d}}=(\boldsymbol{K}_{\mathrm{d}},\boldsymbol{P}_{\mathrm{d}},\boldsymbol{R}_{\mathrm{d}}), \tag{2}$$

其中,$\boldsymbol{K}_{\mathrm{d}}$ 代表图形基元,$\boldsymbol{P}_{\mathrm{d}}$ 为图形实体间连接结构基元,$\boldsymbol{R}_{\mathrm{d}}$ 为 $\boldsymbol{P}_{\mathrm{d}}$ 对应的相关实体之间的关系,$\boldsymbol{\Sigma}=\{\boldsymbol{M}_{\mathrm{d}i}=(\boldsymbol{K}_i,\boldsymbol{P}_i,\boldsymbol{R}_i)\mid i=1,2,\cdots,m\}$ 称为 D-S^2 模型池.

试题的隐含关系通常是指需要基于试题题面信息,借助知识库、推理引擎等辅助手段才能获取的知识信息,该类知识信息往往体现了试题的真正考察意图.因此挖掘试题的隐含关系对于试题知识点标注的精准度和全面度具有重要意义.本文在试题题面属性关系挖掘的基础上,通过建立学科定理-关系知识库,利用蒙特卡罗树搜索算法(MCTS)构建已知属性关系到待求解属性关系的试题解答关系序列,并将该试题解答关系序列翻译为试题知识点.

2.1　文本属性关系挖掘

用 X 表示一个题目文本,R 表示从题目文本 X 中抽取的知识关系集合,记为 $(r_1,r_2,\cdots,r_n)$,X 是一组自然语言描述的句子序列,记为 $(x_1,x_2,\cdots,x_m)$.则题目文本对应的知识关系 $R(T)$ 可如下表示为

$$R(T)=\sum_{i=1}^{m}r(x_i)+\sum_{i=1,j=i+k}^{i=m-k}r(x_i,\cdots,x_j),$$
$$\boldsymbol{T}=\{x_1,x_2,\cdots,x_m\mid x_i\in X\}, \tag{3}$$

其中,$r(x_i)$,$r(x_i,x_j)$ 分别表示 x_1 和 (x_i,x_j) 对应的知识关系,k 为自然语句层级步长因子.本文使用句法语义模型(T-S^2 模型)实现上述从文本到知识关系的映射.为此,本文设计了 T-S^2 模型池 $M=\{m_1,m_2,\cdots,m_j\}$,建立试题文本语义到属性关系关联.因此,试题文本的属性关系挖掘过程转化为建立从 X 中检测到自然语句 x_i 对应的定位模板 m_j 的过程,即:

$$p(m_j\mid x_i;n_t)=\frac{\exp(n_t\cdot f(x_i,m_j))}{\sum_{t_j'}\exp(n_t\cdot f(x,m_j'))}, \tag{4}$$

其中,n_t 是模型参数,$f(x_i,m_j)$ 为 m_j 对应的文本语义向量.

为此,我们借助 BERT 模型[20]来实现题目文本到模型 T-S^2 的结构化预测.在训练阶段,使用 BERT 中文预训练模型实现题目文本序列嵌入,将词向量、位置向量和句子切分向量之和作为模型输入.为了得到最终结构 T-S$_{\mathrm{o}}^2$,对 T-S$_{\mathrm{c}}^2$ 集合中的每个结构,利用学科词类结构相似性计算与基于 BERT 模型的语义计算在特征上的互补,进行词类结构相似性信度排名.

为了计算学科词类结构相似性信度,记 $A=\{a_1,a_2,\cdots,a_m\}$ 是自然语句 x_i 的 T-S$_{\mathrm{c}}^2$ 集和,则其最大优选概率函数:

$$p(a_k)=\frac{\exp(v_a\cdot f(x_i,a_k))}{\sum_{a_k'}\exp(v_a'\cdot f(x_i,a_k'))}, \tag{5}$$

其中,n_a 是模型参数.通过训练 SVM 分类器进行对齐排名,来获得每个 T-S_c^2 的优选概率.优选概率目标函数可如下定义:

$$\frac{1}{2}\| n_a \|^2 + C\sum_i l(n_a^M f(x_i,a_k)^+ - n_a^M f(x_i,a_l)^-), \quad (6)$$

其中,优选特征向量 $f(x_i,a_k)$ 与 T-S^2 中的关键词 K 的词类、句法语义模式 P 相关.+、-分别表示正确实例或错误实例.

2.2 图形属性关系挖掘

试题中的图形一般以图像的形式表示,但与图像语义特征不同,图形中的语义具有高度稀疏性.因此,本文先采用图像识别技术 Faster R-CNN [19] 构建对应图形(Graph),再采用 Graph2Seq[24] 将图形向量化并实现图形知识的结构化预测.

令 D 表示一个图形,记为 $(d_1,d_2,\cdots,d_m)$,包含多个图形元件、节点和连线等图形元素,R 表示从图形 D 中抽取的知识关系集合,记为 $(r_1,r_2,\cdots,r_n)''$.

基于上述 D-S^2 模型池,电路图 D 对应的知识关系可以表示为

$$R(D) = \sum_{i=1}^{m} r(d_i) + \sum_{i=1,j=i+k}^{m-k} r(d_i,\cdots,d_j),$$
$$D = \{d_1,d_2,\cdots,d_m\}, \quad (7)$$

其中,$r(d_i)$,$r(d_i,\cdots,d_j)$ 分别代表图形结构 d_i 和 $d_i,\cdots,d_j$ 对应的知识关系,k 为图形结构单元层级步长因子.

据此,建立基于 D-S^2 模型池的图形属性关系挖掘框架,即在传统图形理解的基础上,根据图形对象类别、空间坐标、连接关系等结构特征,建立图形拓扑结构基元的语义描述 D-S^2 模型,实现图形拓扑-结构基元-知识序列的描述模式.构建基于结构收缩的复杂结构图形的结构基元检测方法,实现复杂图形结构预测及图形属性关系抽取.

2.3 隐含关系挖掘

通过关系挖掘所得属性关系集是试题题面知识语义的形式化描述.为了全面描述试题的知识语义,需要进一步挖掘试题求解逻辑中的隐含知识语义.在 MCTS 框架下,关系序列的挖掘过程被视为关系树搜索过程,搜索从根节点(已知关系)出发,到目标节点(待求解关系)结束,最后将搜索路径转化为关系序列.在选择阶段,需要从当前解题状态 S 出发向下选择一个最需被拓展的节点 N,并选择值最大的子节点反复迭代;对于有未被拓展过的可行动作的节点,该点即是目标节点 N;对于已结束的节点,直接进行反向传播.每个被检查节点的被访问次数在该阶段都会增加.在反复迭代后,在底端将找到一个节点,来继续之后步骤.在解题状态 S 下,对于所有可行动作都已经被拓展过的节点,我们使用 SP-MCTS 下改进的 UCB 公式[25]:

$$UCB1 = \bar{X} + C\cdot\sqrt{\frac{\ln t(N)}{t(N_i)}} + \sqrt{\frac{\sum x^2 - t(N_i)\cdot\bar{X}^2 + D}{t(N_i)}}, \quad (8)$$

式(8)右边的前两部分为标准的 UCT 公式,$t(N)$ 代表节点 N 被访问的次数,$t(N_i)$ 代表子节点 N_i 对价值均值 $\bar{X}$ 的上界置信度贡献.式(8)的最后部分为对标准 UCT 的修正,代表到达子节点 N_i 的累积选择偏差,其中 $\sum x^2$ 为到达子节点 N_i 价值的评分和,$t(N_i)\cdot\bar{X}^2$ 为到达子节点 N_i 的期望值,D 为控制常量.

本文采用深度 Q 网络(DQN)来学习模拟过程,记 DQN 模型输出行动价值向量 $Q(s,\cdot,\theta)$,其中 θ 是网络参数.为了训练网络参数 θ,建立经验记忆内存 D 存储状态转换向量 (s,a,s',r),每次随机选择一组状态转换向量更新网络参数 θ,并使损失函数最小,

$$L_t(\theta_t) = E_{s,a}[(y_t - Q(s,a;\theta_t))^2], \quad (9)$$

其中,$y_t = r + \gamma\max_{a'}Q(s',a';\theta_{t-1})$ 为最优目标 Q 值矩阵,r 为当前反馈.为了确定网络参数在学习过程中的稳定,采用 Bellman 方程损失函数梯度下降逼近最优目标 Q 值:

$$\nabla_{\theta_t}L_t(\theta_t) = E[(y_t - Q(s,a;\theta_t))\nabla_{\theta_t}Q(s,a;\theta_t)]. \quad (10)$$

在 N_i 的模拟结束之后,从初始解题状态出发,到 N 的路径上的所有节点都会根据本次模拟的结果来增加自己的累计评分.在触发求解终止条件或达到迭代次数后结束,选择初始解题状态下的最优子节点作为属性关系连接序列.

2.4 知识点标签生成

由于试题属性关系的抽象性,需要进一步将属性关系映射到知识点空间,生成知识点标签.本文知识点标签生成过程如图 2 所示,其中的关键步骤是通过知识信息中的关键词和词性模式建立与权威知识库的映射,以此获取知识点标签.

标签模型由关键词、词性模式、数量关系和知识点标签组成,可将其看作一个四元组 $L = (K,P,R,W)$,K 是关键词,P 为词性结构模式,R 是数量关系中要求匹配的特定字符串,W 是当出现预设的关键词、词性模式和数量关系时对应的知识标签词集,模型

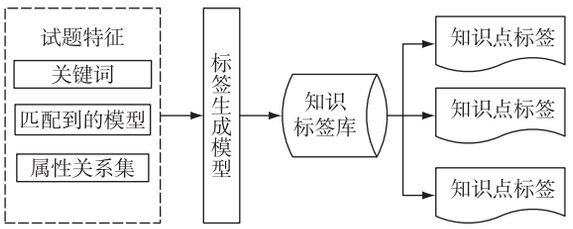

图 2 试题知识点标签生成

Fig. 2 Label generation of knowledge points

池定义为如下结构:

$$l=\sum_{i=1}^{m}L_i=(K_i,P_i,R_i,W_i). \tag{11}$$

标签生成的具体过程步骤为,依次将各个分句中提取出来的知识信息放入标签模型池中进行匹配,若词性模式 P、关键词 K,与数量关系 R 中定义的特定字符串均匹配成功,则与提取出该模型中对应的知识点标签;若三者中出现不同,则使用下一模型继续匹配,直至所有模型均匹配完为止.循环输入每个分句,直至每个分句均得到处理.通过与知识标签库的映射,提取出知识标签.

具体算法流程如下:

输入:知识信息集合 $\sum_{j=1}^{n}\Delta_j=(K_j,P_j,R_j,X_{\text{text}j})$

输出:各个分句对应的知识标签集 Γ.

1) 初始化标签模型池 $\sum_{i=1}^{m}L_i=(K_i,P_i,R_i,W_i)$.

2)loop:对知识信息集合 $\sum_{j=1}^{n}\Delta_j$,将每个知识信息单元 Δ_j 与标签模型池 $\sum_{i=1}^{m}L_i$ 中模型逐个匹配;如果匹配成功,则:使用知识信息单元 Δ_j 实例化知识标签 W_j;构建二元组 $\langle W_j,X_{\text{text}j}\rangle$ 并加入知识标签集 Γ.

endloop

3) 返回知识标签集 Γ.

3 实验结果与分析

3.1 实验数据集描述

为了保证测试数据的代表性和多样性、提高实验结果的可靠性,我们建立了一个名为 Circuit1K 的初中电学试题库来评估各算法的标注性能.Circuit1K 中的试题来自广泛使用的教科书、辅导书和入学考试试卷,通过这些渠道获得的试题并没有经过特意地筛选.这些试题的来源之一是目前七到九年级学生使用的 3 种主要物理教材,分别由人民教育出版社(人教版)、北京师范大学出版社(北师大版)和上海教育出版社(上教版)出版.另一个来源是人民教育出版社出版的 2 本广泛使用的辅导书(人教辅导书).最后一个来源是北京、湖北、上海等地 2014—2018 年的中考试卷.

Circuit1K 中共包含 1 012 道试题,分为 2 组.一组为纯文本试题("T"试题),这组试题只有文字描述.另一组为文本和电路图混合试题("T+S"试题),这组试题既有电路图又有文字描述.Circuit1K 中的试题组成详细信息如表 1 所示.

表 1 Circuit1K 中选自不同书籍和试卷的试题数量

Table 1 Numbers of questions in Circuit1K collected from textbooks and test papers

试题来源	数量			占比/%
	T	T+S	总计	
人教版	57	164	221	21.8
北师大版	51	131	182	18.0
上教版	35	132	167	16.5
人教辅导书	47	141	188	18.6
中考试卷	86	168	254	25.1
‖ 总计	276	736	1 012	100

根据初中物理电路题目中知识点特征以及物理学科特点,我们构建了初中电学知识层次体系,表 2 为部分初中物理电学知识点层次体系示例.为了保证各算法结果的一致性,本文暂不进行多层知识点标注的结果分析,而是统一转换到二级知识点上进行结果对比分析.

表 2 初中物理电学知识点层次体系示例

Table 2 An example of the hierarchy structure of knowledge points

一级	二级	三级	四级
电路和电流	电现象	摩擦起电	
		电荷种类	正电荷、负电荷
	电路	电路组成	电源、用电器、开关、导线
		电路符号	
		电路状态	通路、断路、短路
		电路种类	串联电路、并联电路
	电流	电流方向	
		电流产生条件	
		电流单位	
		电流测量	测量仪器、测量量程

3.2 对比试验方法

为了验证基于属性关系深度挖掘的知识点标注

效果,本文将与如下实验方法进行对比:

1)传统机器学习方法.此处选择了朴素贝叶斯(Native Bayes,NB)和支持向量机(Supported Vector Machine,SVM).对试题的题面抽取特征后,对每个知识点进行二分类,通过多个二分类进行多标签预测.

2)卷积神经网络(CNN).首先使用 Word2Vector 将试题的题面文本转换成词向量,然后通过 CNN 进行试题深层语义理解,进行多标签预测.考虑到本次测试数据集样本数量少,在词向量生成上,我们同时使用了 BERT 预训练模型增加模型的稳定性.

3.3 实验结果及分析

为了验证本文方法的效果,利用准确率 P(Precise)、召回率 R(Recall)以及 $F1$ 值(F1-Score)等数据来评估所提出的初中物理电路知识标注的有效性.其对应的公式分别为

$$P = \frac{N_{TP}}{N_{TP} + N_{FP}} \times 100\%, \tag{12}$$

$$R = \frac{N_{TP}}{N_{TP} + N_{FN}} \times 100\%, \tag{13}$$

$$F_1 = \frac{2 \times P \times R}{P + R} \times 100\%, \tag{14}$$

其中 N_{TP} 表示被正确标注的知识点标签数量,N_{FP} 为标注错误的标签数量,N_{FN} 指的是在标准数据集上正确但没有被标注上的标签数量.不同算法在 Circuit1K 上的标注结果比较如图 3 所示.

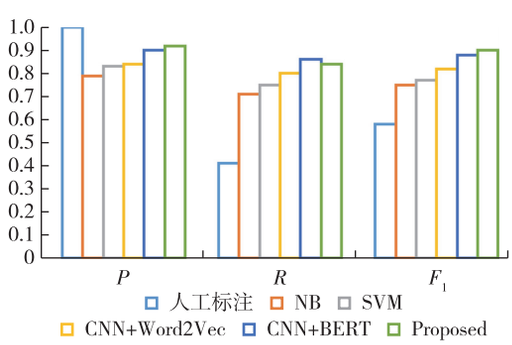

图 3 不同算法在 Circuit1K 上的标注结果

Fig. 3 Annotation results of different algorithms on Circuit1K

本论文提出的标注方法在初中物理电路知识点标注中,知识点的准确率达到了 92.26%,F_1 值也达到了 90.62%,均高于基于纯语义挖掘的 CNN 模型.虽然人工标注标签的准确率为 100%,但是与人工标注的数据相比,我们方法的知识点召回率达到了 84.08%,而人工方法只有 40.79%.这些数据证明了我们提出的学习资源自动标注方法在知识自动标注的准确性和丰富性上具有一定优势.

形成这些差距的原因是,人工方法标注的知识点主要集中在题目待求解的物理元素,以及主要涉及的定理知识中,忽略了题目直陈述信息中的知识点和隐含在定理转换间的知识信息.所以标注结果的准确率会达到 100%,但是由于缺少对部分信息的知识点标注,资源标注的全面性不够,造成召回率以及 F_1 值偏低.反之,本文提出的标注方法从题目中的直陈述与隐含信息两个层次进行知识点标注,图中的数据进一步证明了本文提出的方法在初中物理电路题目的资源标注中能获得相对人工标注方法而言更全面丰富的标签.另一方面,无论是传统机器学习方法和深度神经网络方法,由于其仅利用了题面语义、句法等浅层特征,对于试题所蕴含的知识逻辑以及领域知识的理解和挖掘能力有限.

为了进一步检验本文模型在不同级别试题上的知识点标注性能,分别对 736 个“T+S”型试题和 276 个“T”型试题进行了独立实验.表 3 为部分实验结果.

表 3 本文方法在不同类型试题集上的知识点标注结果

Table 3 Annotation accuracy of the proposed algorithm on different sets of questions %

	图形知识 P	文本知识 P	隐含知识 P
人教版	97.2	93.0	84.6
北师大版	96.7	94.1	85.2
上教版	96.1	93.6	83.2
人教辅导书	92.4	87.2	77.5
中考试卷	84.3	78.4	70.6

如表 3 中所示,本文模型在人民教育出版社出版的教科书试题的知识点标注正确率高达 97.2%,从北京师范大学出版社(96.7%)和上海教育出版社有限公司(96.1%)出版的教科书中收集的试题集也得到了类似的结果.此外,本文模型在辅导教材(92.4%)和中考试卷(84.3%)上都取得了令人满意的结果,证明了该算法对复杂问题的理解能力.

4 结束语

在线学习资源的试题知识点自动标注是智能教育领域中的重要问题.本文针对人工标注、传统机器学习在知识点标注任务上的不足,提出了基于属性关系深度挖掘的试题知识点标注方法,该方法分为 3 个步骤:第 1 步为题面关系抽取;第 2 步为隐含关系抽取.此 2 步分别从题面信息抽取和问题求解框架

构建的角度挖掘试题浅层语义和深层逻辑属性关系.第3步结合学科知识图谱,将属性关系映射到知识图谱空间,生成试题知识点.通过与多种传统方法对比实验,证明了本文所提出的基于属性关系挖掘的试题知识点标注方法的合理性和有效性.

本文所提出的试题显性属性关系挖掘方法,使用了专家经验设计的文本句法语义模型和图形结构语义模型,后续可考虑从试题信息和试题已标注知识点信息中自动抽取显性属性关系.本文对试题的知识点标注尚未引入知识点层次化信息,后续可考虑基于学科知识图谱对试题的知识点层次进行结构化描述和图谱可视化.

参考文献

References

[1] 魏伟,郭崇慧,邢小宇.基于语义关联规则的试题知识点标注及试题推荐[J/OL].数据分析与知识发现:1-14[2019-10-14]. http://kns.cnki.net/kcms/detail/10.1478.G2.20190909.1423.012.html
WEI Wei,GUO Chonghui,XING Xiaoyu.Knowledge point annotation based on semantic association rules and question recommendation[J/OL]. Data Analysis and Knowledge Discovery:1-14[2019-10-14]. http://kns.cnki.net/kcms/detail/10.1478.G2.20190909.1423.012.html

[2] Ciccarese P,Ocana M,Clark T.Open semantic annotation of scientific publications using DOMEO[J].Journal of Biomedical Semantics,2012,3(supl 1):S1

[3] 刘淇,陈恩红,朱天宇,等.面向在线智慧学习的教育数据挖掘技术研究[J].模式识别与人工智能,2018,31(1):77-90
LIU Qi,CHEN Enhong,ZHU Tianyu,et al.Research on educational data mining for online intelligent learning[J]. Pattern Recognition and Artificial Intelligence,2018,31(1):77-90

[4] 赵乐,张兴旺.面向LDA主题模型的文本分类研究进展与趋势[J].计算机系统应用,2018,27(8):10-18
ZHAO Le,ZHANG Xingwang. Research progress and trend of text classification for LDA topic model[J].Computer Systems & Applications,2018,27(8):10-18

[5] 周菊明,张良龙.学习资源智能标注系统的设计与实现[J].中国教育信息化,2018(7):41-44
ZHOU Juming,ZHANG Lianglong.Learning resources intelligent labeling system design and implementation[J]. China Education Info,2018(7):41-44

[6] Vanderwende L,Suzuki H,Brockett C,et al.Beyond SumBasic:task-focused summarization with sentence simplification and lexical expansion[J].Information Processing & Management,2007,43(6):1606-1618

[7] 戚欣,肖敏,孙建鹏.基于本体知识库的自动语义标注[J].计算机应用研究,2011,28(5):1742-1744,1747
QI Xin,XIAO Min,SUN Jianpeng. Automatic semantic annotation based on ontology and knowledge base[J]. Application Research of Computers,2011,28(5):1742-1744,1747

[8] 闫喜亮.基于情感本体的网络教育资源标注模型的设计与实现[D].南京:南京理工大学,2011
YAN Xiliang.Design and implementation of network education resource annotation model based on emotional ontology[D]. Nanjing: Nanjing University of Science and Technology,2011

[9] 蒋婷.学科领域本体学习及学术资源语义标注研究[D].南京:南京大学,2017
JIANG Ting. Discipline ontology learning and semantic annotation for scientific resources[D]. Nanjing: Nanjing University,2017

[10] 何中山.基于语义网的初中数学的自动语义标注方法研究与实现[D].成都:电子科技大学,2014
HE Zhongshan. Semantic web-based automatic semantic annotation of junior high school mathematics research and implementation[D]. Chengdu: University of Electronic Science and Technology of China,2014

[11] 朱远平,戴汝为.基于SVM决策树的文本分类器[J].模式识别与人工智能,2005,18(4):412-416
ZHU Yuanping,DAI Ruwei.Text classifier based on SVM decision tree[J].Pattern Recognition and Artificial Intelligence,2005,18(4):412-416

[12] 毛伟,徐蔚然,郭军.基于n-gram语言模型和链状朴素贝叶斯分类器的中文文本分类系统[J].中文信息学报,2006,20(3):29-35
MAO Wei, XU Weiran, GUO Jun. A Chinese text classifier based on n-gram language model and chain augmented naive Bayesian classifier[J].Journal of Chinese Information Processing,2006,20(3):29-35

[13] 胡于进,周小玲,凌玲,等.基于向量空间模型的贝叶斯文本分类方法[J].计算机与数字工程,2004,32(6):28-30,77
HU Yujin,ZHOU Xiaoling,LING Ling,et al.A Bayes text classification method based on vector space model[J]. Computer & Digital Engineering,2004,32(6):28-30,77

[14] 张玉峰,何超.基于潜在语义分析和HS-SVM的文本分类模型研究[J].情报理论与实践,2010,33(7):104-107
ZHANG Yufeng, HE Chao. Research on text categorization model based on latent semantic analysis and HS-SVM[J].Information Studies (Theory & Application),2010,33(7):104-107

[15] 杨萌萌,黄浩,程露红,等.基于LDA主题模型的短文本分类[J].计算机工程与设计,2016,37(12):3371-3377
YANG Mengmeng,HUANG Hao,CHENG Luhong,et al. Short text classification based on LDA topic model[J]. Computer Engineering and Design, 2016, 37(12):3371-3377

[16] Chen X Y,Song D,Wang D M.Automated generation of geometric theorems from images of diagrams[J].Annals of Mathematics and Artificial Intelligence,2015,74(3/4):333-358

[17] De P,Mandal S,Bhowmick P.Hierarchical vectorization of electrical drawings in document images by connectivity analysis of symbols and super-components[J]. Pattern

Recognition and Image Analysis,2017,27(2):309-325
[18] De Sekhar Mandal P,Bhowmick P,Chanda B.Topological simplification of electrical circuits by super-component analysis[C]// 2015 13th International Conference on Document Analysis and Recognition (ICDAR),2015:211-215
[19] Ren S Q,He K M,Girshick R,et al. Faster R-CNN:towards real-time object detection with region proposal networks[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2017,39(6):1137-1149
[20] Yu X G,Wang M S,Gan W B,et al.A framework for solving explicit arithmetic word problems and proving plane geometry theorems[J]. International Journal of Pattern Recognition and Artificial Intelligence, 2019, 33(7):1940005
[21] Jian P P,Sun C,Yu X G,et al.An end-to-end algorithm for solving circuit problems[J]. International Journal of Pattern Recognition and Artificial Intelligence,2019,33(7):1940004
[22] He B, Jian P P, Xia M, et al. Extracting algebraic relations from circuit images using topology breaking down and shrinking[C]// Pacific-Rim Symposium on Image and Video Technology,2018:116-130
[23] Devlin J,Chang M W,Lee K,et al.BERT:pre-training of deep bidirectional transformers for language understanding [J]. arXiv Preprint, 2018, arXiv:1810.04805
[24] Xu K,Wu L F,Wang Z G,et al.Graph2Seq:graph to sequence learning with attention-based neural networks[J]. arXiv Preprint,2018,arXiv:1804.00823
[25] Schadd M P D,Winands M H M,Herik H J V D,et al. Single-player Monte-Carlo tree search[J]. International Conference on Computers and Games,2008,34(5):3-11

Knowledge points annotation based on attribute relation mining

HE Bin[1] LI Xinyu[1] CHEN Beilei[2] XIA Meng[1] ZENG Zhizhong[1]
1 National Engineering Research Center for E-Learning,Central China Normal University,Wuhan 430079
2 Institute of Education,Hubei University,Wuhan 430062

Abstract Online learning systems need to perform the fundamental task of annotating a large number of raw questions to be able to provide students with learning materials of high quality.The existing methods used for this task rely either on labeling by human experts or traditional ways of machine learning.In practical applications,the existing methods are limited by being either labor intensive or inaccurate.In this paper,we propose a method based on the mining of attribute relations to annotate the knowledge points of questions.We first define and extract the explicit attribute relations from the text and diagram of a given question.We then extract the implicit attribute relations of the question using Monte Carlo Tree Search (MCTS) algorithm.Next,we map the attribute relations to the knowledge graph space using a transform model,to generate the knowledge points of the question.The experimental results confirm the effectiveness of the proposed method,which demonstrates practicality for the cognitive diagnosis of students and personalized questions recommendation.

Key words knowledge points annotation;attribute relation mining;syntax-semantics model;structure-semantics model;Monte Carlo Tree Search (MCTS)