文章编号: 1000-5641(2019)05-0053-13

面向初等数学的知识点关系提取研究

杨东明, 杨大为, 顾 航, 洪道诚, 高 明, 王 晔

(华东师范大学 数据科学与工程学院, 上海 200062)

摘要: 随着互联网技术的发展, 在线教育已经改变了学生的学习方式. 但由于缺乏完整的知识体系, 在线教育存在着智能化程度低和“信息迷航”的问题. 因此, 构建知识体系成为在线教育平台的核心技术. 知识点间的关系提取是知识体系构建的主要任务之一, 目前比较高效的关系提取算法主要是监督式的. 但是这类方法受限于文本质量低、语料稀缺、标签数据难获取、特征工程效率低、难以提取有向关系等挑战. 为此, 基于百科语料和远程监督思想, 研究了知识点间的关系提取算法. 提出了基于关系表示的注意力机制, 该方法能够提取知识点间的有向关系信息. 结合了 GCN 和 LSTM 的优势, 提出了 GCLSTM, 该模型更好地提取了句子中的多点信息. 基于 Transformer 架构和关系表示的注意力机制, 提出了适用于有向关系提取的 BTRE 模型, 降低了模型的复杂度. 设计并实现了知识点关系提取系统. 通过设计 3 组对比实验, 验证了模型的性能和效率.

关键词: 知识体系构建; 关系提取; 注意力机制; 远程监督; Transformer

中图分类号: TP301.6 **文献标志码**: A **DOI:** 10.3969/j.issn.1000-5641.2019.05.004

Research on knowledge point relationship extraction for elementary mathematics

YANG Dong-ming, YANG Da-wei, GU Hang, HONG Dao-cheng, GAO Ming, WANG Ye

(*School of Data Science and Engineering, East China Normal University, Shanghai* 200062, *China*)

Abstract: With the development of Internet technology, online education has changed the learning style of students. However, given the lack of a complete knowledge system, online education has a low degree of intelligence and a “knowledge trek” problem. The relation-extraction concept is one of the key elements of knowledge system construction. Therefore, building knowledge systems has become the core technology of online education

收稿日期: 2019-07-29

基金项目: 国家重点研发计划(2016YFB1000905); 国家自然科学基金(U1811264, 61672234, 61502236, 61877018, 61977025); 上海市科技兴农推广项目 (T20170303)

第一作者: 杨东明, 男, 硕士研究生, 研究方向为面向新硬件的大数据系统.
E-mail: y1271752959m2@yahoo.com.

通信作者: 王 晔, 男, 研究员, 研究方向为 Web 数据管理、海量数据挖掘、分布式系统等.
E-mail: ywang@dase.ecnu.edu.cn.

platforms. At present, the more efficient relationship extraction algorithms are usually supervised. However, such methods suffer from low text quality, scarcity of corpus, difficulty in labeling data, low efficiency of feature engineering, and difficulty in extracting directional relationships. Therefore, this paper studies the relation-extraction algorithm between concepts based on an encyclopedic corpus and distant supervision methods. An attention mechanism based on relational representation is proposed, which can extract the forward relationship information between knowledge points. Combining the advantages of GCN and LSTM, GCLSTM is proposed, which better extracts multipoint information in sentences. Based on the attention mechanism of Transform architecture and relational representation, a BTRE model suitable for the extraction of directional relationships is proposed, which reduces the complexity of the model. Hence, a knowledge point relationship extraction system is designed and implemented. The performance and efficiency of the model are verified by designing three sets of comparative experiments.

Keywords: knowledge system construction, relation-extraction, attention mechanism, distant supervisor, Transformer

0 引 言

近年来, 随着互联网技术的快速发展, 越来越多的在线教育平台出现在大众的视线中, 如网易云课堂、腾讯课堂、好未来、阿凡题等. 这些平台依托互联网将教育资源整合起来, 并向广大学生提供课程视频、音频、课辅材料等教学资源, 促进了教育资源的快速传播. 这种新型的教育形式突破了传统教育在地域、时间、场景上的限制, 缓解了传统教育中教育资源不平衡的问题. 换言之, 这种教育形式利用先进的技术手段不仅为学生提供了更多受教育的机会, 而且为学生提供了更为优质的学习资源. 智研资讯网对过去3年中国在线教育发展情况的统计数据显示, 2016年我国在线教育用户在9 000万人左右, 同比增长21.5%, 且自2016年以来, 每年都保持20%以上的增长速度, 预计2019年将达到1.6亿名用户. 更值得一提的是, 中小学在线教育占比从2015年的25.3%, 增长到2018年的36.8%. 由此可见, 随着在线教育模式的不断成熟, 教学方式的不断优化, 这种教育模式也得到了越来越多人的青睐.

在线教育平台虽然为学生学习提供了便利、缓解了教育资源不平衡的问题, 但也带来了诸多问题, 主要包括以下两方面:

- 造成“信息迷航”. 由于各大在线教育平台之间数据不是互通的, 很多内容详尽的教学资料在不同的教学平台中被归为不同类别, 但一些内容相差甚远的学习资料反而被归结在相似的类别中. 每个平台都有一套自己的分类标准, 会造成“信息迷航”的问题, 人们花费了大量的精力却难以找到想要的资源, 极大地浪费了学习者的时间与精力, 打击了他们的学习热情.

- 智能化程度低. 千百年来, 因材施教是教育一直未能实现的目标, 随着互联网技术的发展, 实现因材施教成为可能. 例如, 根据学生的在线学习行为可以发现学生的薄弱知识点以及能力缺陷, 进一步为学生设计更为合理的学习路径, 从而提升学生的学习效率, 提升学生的学习效果. 可惜的是, 目前各大平台大多是一个资源库, 用户可以按照平台给定的分类体系找到自己所需要的资源, 但无法享受智能化服务, 最终导致用户的个性化需求得不到满足. 具体表现为: ① 各类教育资源缺乏有效整合; ② 学生还在使用题海战术, 致使学习效率

低下.

上述问题产生的原因是没有一套层次分明、关系明确的知识体系. 有了完备的知识体系, 平台就可以将原本离散的资源有效整合起来, 通过知识点之间的关系来发现课程、题目、教案等教育资源之间的关系, 也能够从用户的浏览记录、做题情况中挖掘出用户的真实需求, 真正赋能教育.

但是, 现在主流的关系提取方法, 如基于特征[1]、基于核[1-2]的机器学习方法和深度学习构造端对端的关系提取模型[3]的方法, 都有难以提取长依赖的局部依赖信息和有向关系的问题. 因此, 本文以初中数学为切入点, 聚焦关系提取的难点问题, 研究数学知识点之间的关系提取问题.

1　研究设计

1.1　研究对象

关系提取是自然语言处理中实体识别基础上的一个任务, 其核心是抽取一个句子中包含实体对之间的关系.

本文的目的是构建初中数学知识体系, 初中数学知识体系由知识点与关系组成, 知识点可以从教学大纲中直接获得, 因此关系提取成为了初中知识体系构建的关键. 在初中知识体系中, 本文定义了 8 种知识点之间的有向关系, 它们是依赖、被依赖、属于、包含、拥有、被拥有、同义、反义, 再加上一些不属于这八种的其他关系, 初中数学知识体系就是由这 9 种关系构成的有向图.

表 1 展示了每种关系的样例. 第一列为共现句 s, s 中包含了两个目标知识点. 第二列 e_1 与第三列 e_2 为两个目标知识点, 它们必须出现在共现句中. 最后一列为关系, 它表示 e_1 到 e_2 的有向关系.

表 1　关系样例

Tab. 1　Relationship example

共现句 s	知识点 e_1	知识点 e_2	关系 r
两个整数的最大公因子可用于计算两数的最小公倍数	最小公倍数	整数	依赖
求几个整数的最大公因数, 只要把它们所有共有的质因数连乘	质因数	最大公因数	被依赖
代数式根据它所包含的运算可以分为有理式和无理式,	术语		
而有理式又可以分为整式和分式	整式	有理式	
正比例函数为特殊的一次函数	一次函数	正比例函数	包含
实数根也经常被称为实根	实数根	实根	同义
分数分为两类: 真分数和假分数	真分数	假分数	反义
直角三角形的外心在三角形斜边中点	直角三角形	外心	拥有
二元二次方程是含有两个未知数且未知数的最高次数为 2 的整式方程	未知数	二元二次方程	被拥有

1.2　关系数据集介绍

为了完成本文的研究任务, 我们构造了初中数学知识点关系数据集 MKR, 该数据集包含 336 个知识点, 覆盖数与运算、方程与代数、函数与分析、数据整理与概率统计、图形与集合 5 大方面, 包含了 9 种关系. 这 336 个知识点是从上海市初中数学大纲与教材中人工整理得到的, 然后根据这些知识点从百度百科、维基百科、搜狗百科等质量较高的百科网站上爬取相关的页面. 在爬取页面的过程中, 部分知识点是多义词, 存在多个百科页面, 如直线、函数、抛物线等知识点具有除了数学之外的其他含义, 因此需要根据百科标签对页面进行筛选. 其中部分知

识点没有直接对应的百科页面, 由多个相关百科词条构成, 因此 336 个知识点共涉及 362 个百科页面. 大部分百科页面会包含发展历史、相关内容、背景等模块, 因此将会存在部分噪声语料. 另外, 文本中的数学公式会以图片或特殊字符的形式插入, 因此爬取下来的文本会存在缺失或不同符号, 这些问题会影响语料的质量. 去除噪声语料后, 总共爬取 21 445 句有效句子, 构成了 6 258 个未标记实例.

通过上述步骤, 我们得到了没有关系标签的实例. 为了进一步工作, 本文采用了基于远程监督思想[4]的数据标记方法, 并通过在线标注模块, 来为实例打上标签.

由于数据存在"有偏性"以及数学符号不统一的问题, 通过调换实体对中知识点的位置来构造新的训练样本, 采用上采样的方法来缓解有偏性, 并通过直接替换和形式化替换这两种标准化方法来解决数学符号不统一的问题.

2 研究成果

2.1 基于关系表示的注意力机制

关系提取任务有一个很大的特点, 对相同的输入, 但目标实体对不一样时, 模型需要能够提取到不同的信息, 并且一句话中不同单词提供的信息的权重也是不同的. 为了能够让模型根据任务目标进行选择性学习, 有人提出了注意力机制 (AttentionMechanism)[5-6].

本文运用了 Luong 提出的三种 NLP (Natural Language Processing) 中通用的注意力得分[7]计算方法:

$$score(q,k_i)=\begin{cases} q^{\mathrm{T}}k_i, & dot;\\ q^{\mathrm{T}}Wk_i, & genneral;\\ v^{\mathrm{T}}\tanh(W[q;k_i]), & concat. \end{cases}$$

score 表示向量 k_i 对于 q 的重要程度, 其中 q 为目标向量, 在关系提取中一般由两个目标实体构成; 而 k_i 为信息向量, 如单词向量 x_i 或者 RNN 网络输出的向量 h_i. 第一种方式称为 *dot*, 即直接计算点积, 使用这种方式的前提是 q 与 k_i 的维度相同. 第二种方式称为 *general*, 这种方式不要求 q 与 k_i 的维度相同, 因为它通过一个线性变换将 q 的维度与 k_i 对齐, 然后进行点积. 第三种方法称为 *concat*, 将 q 与 k_i 进行拼接, 用双曲正切函数进行非线性变换, 最后与 $v^{\mathrm{T}}\in\mathbf{R}^{d^p+d^k}$ 相乘得到分数信息.

得到单词向量 x_i 或者 RNN (Recurrent Neural Network) 网络输出向量 h_i 的分数之后使用 Softmax 函数计算他们的权重, 则最终每个权重为

$$w_i=\frac{\exp(score(q,k_j))}{\sum_{j=1}^{d^k}\exp(score(q,k_j))}\quad (i=1,2,\cdots,d^k).$$

最后将权重向量 w 与单词向量 x_i 或者 RNN 网络输出的向量 h_i 相乘得到带权的向量.

2.2 基于 GCN 与 LSTM 的 GCLSTM 模型

CNN (Convolutional Neural Network) 模型善于提取句子的局部特征模型, 但难以提取多点信息[8]. RNN 模型能够学习到单词之间的顺序信息[9-10], 但当句子长度较长时, 前面的信息会因为一层一层传递而丢失, 难以捕捉单词之间的长距离依赖信息[11]. 因此, 本文采用 GCN (Graph Convolutional Neural) 网络来提取局部多点信息, 并使用注意力机制得到加权的局部信息, 最后使用双向 LSTM (Long Short-Term Memory)[12-14]将加权局部信息整合成关

系信息, 把这称为 GCLSTM 模型 (见图 1).

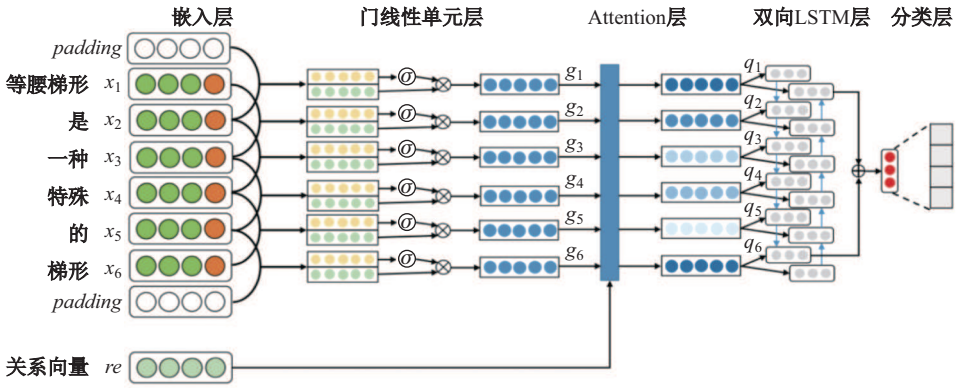

图 1　GCLSTM 模型架构图

Fig. 1　GCLSTM Model Architecture

嵌入层

该层的作用与传统 CNN 和 RNN 一样, 将自然语言的单词或短语映射到连续的高维空间, 得到单词或短语的向量表示[15], 该层主要涉及三种向量: 词向量、位置向量、关系向量.

假如句子的长度为 l, 则将其表示为 $S=(x_1,x_2,\cdots,x_l)$, 每个单词被转化为长度为 d^x 的向量 x_i, x_i 由单词向量 we 与位置向量 pe 构成. 对于单词向量而言, 首先在词表中查找单词的位置, 然后根据位置去向量矩阵 $W^{wrd}\in\mathbf{R}^{d^{we}\cdot|v|}$ 中查找与这个单词对应的单词向量, 其中 $|v|$ 为整个语料中的单词数量, 包括符号、字母等.

在关系提取任务中, 训练任务的目标是寻找两个目标实体之间的关系, 因此考虑每个单词与两个目标实体的相对位置具有十分重要的意义. 但在语料中, 一句话中的两个目标实体可能会重复出现, 因此本文采用了最近相对位置作为特征.

本文提出的关系是具有方向性的, 这一点与通用领域的关系提取略有不同. 关系向量通过减法操作来构造两个实体之间的有向性, 再通过双曲正切函数进行非线性变换, 从而让网络更容易捕获到关系的有向性信息.

门线性单元层

GCLSTM 模型采用门线性 (GLU)[16]来提取多点局部信息. GLU 的结构相比于传统 CNN 网络的池化层, 能够更加高效地通过反向传播学习参数, 并且在前馈传播中能够保留所有的信息.

门线性单元层放弃了池化操作, 池化操作在反向传播时不能让所有参数都有效地学习, 而且它只能提取单点信息 (最大池化只取卷积后向量中的最大值), 不能很好地满足关系提取任务的需求. GLU 的做法是受到 LSTM 模型中门机制的启发而被提出的, 通过 σ 激活函数控制信息在网络中的传播, 在前向传播中每个参数都能被传递到下一层, 在反向传播时每个参数都能有效地得到学习, 因此比传统 CNN 具有更高的信息传播及参数学习效率.

Attention 层

Attention 层的作用是根据关系向量为门线性单元层提取的局部信息进行加权, 使得与预测目标关系有用的信息以较大权重往后传递, 而没用的信息以较小权重往后传递, 这样能够让模型更加准确地预测关系.

双向 LSTM 层

经过 Attention 层处理之后就得到了带权的局部特征信息 Q, 局部特征信息只考虑相邻单

词之间的语义信息, 还是缺少整体的顺序性信息[14,17]. 另外 Q 的维度受到了句子长度的影响, 但是分类层需要有固定维度的输入, 因此双向 LSTM 层还能够将不同句长的语义信息以维度相同的向量 h 输出.

当局部特征信息 Q 经过双向 LSTM 层以后, 就能够得到一个固定长度为 d^h 的向量 h_l, 并且 h_l 包含了整句句子的双向信息.

分类层

最后是分类层, 该层的输入为双向信息 h_l, 首先经过一个全连接得到所有关系的评分向量 $s \in \mathbf{R}^{|Y|}$, 它可以按照如下的公式计算: $s = W_o \cdot h_l + b_o$, 其中 $W_o \in \mathbf{R}^{|Y| \cdot q}$, $b_o \in \mathbf{R}^{|Y|}$ 为参数, 其中 $|Y|$ 表示候选关系的数量.

得到所有关系的评分向量之后, 通过 Softmax 函数对所有得分进行归一化, 得到所有候选关系的条件概率分布 $p \in \mathbf{R}^{|Y|}$.

GCLSTM 模型结合了 GCN 与双向 LSTM 的优点, 通过嵌入层将自然语言转为高维向量, 然后通过门线性单元层提取相邻单词之间的局部信息, Attention 层根据关系向量为局部信息赋以权重, 双向 LSTM 层将带权的局部信息整合为最终的关系信息, 最后通过分类层实现关系预测.

2.3 基于 Transformer 的 BTRE 模型

受到 Transformer 的启发, 提出了基于 Transformer 与向量关系 re 的关系提取模型 (简称 BTRE).

Transformer 模型完全采用注意力机制, 主要由编码器 (Encoder) 与解码器 (Decoder) 组成, 如图 2 所示. 编码器通过自注意力机制 (Self-Attention) 从句子中提取单词之间的长依赖多点信息[18], 解码器则根据关系向量, 从编码器输出的长依赖多点信息中提取最终的关系信息. 而编码器与解码器主要有多头注意力机制[19]、残差标准化层[17]以及位置前馈网络组成.

为了充分考虑单词之间的局部依赖信息, 并且加快模型计算速度, BTRE 模型也采用了与 Transformer 模型类似的编码器与解码器架构, 编码器用来提取单词之间长依赖的局部信息, 解码器则从编码器提取的局部信息中抽取出与目标关系最相关的信息.

嵌入层

BTRE 模型的嵌入层与 GCLSTM 的嵌入层类似, 将自然语言的句子映射到高维空间, 得到句子表示 $S = (x_1, x_2, \cdots, x_l)$, 其中 x_i 为每个单词的向量. x_i 由词向量 $we \in d^{we}$ 与位置向量 $pe \in d^{we}$ 组成, 词向量 we 也是使用 word2vec 方法预训练得到, 但是位置向量采用基于正弦与余弦函数的位置嵌入方式得到.

由于 BTRE 没有类似循环神经网络中的串行依赖结构, 也没有卷积神经网络的卷积结构, 因此在 BTRE 模型中, 位置向量是让模型学习到顺序信息的唯一来源. BTRE 模型采用基于正弦与余弦函数的位置嵌入方法, 其定义如下:

$$PE(pos, 2i) = \sin\left(\frac{pos}{10000^{2i/d^{pe}}}\right),$$
$$PE(pos, 2i+1) = \cos\left(\frac{pos}{10000^{2i/d^{pe}}}\right).$$

其中, d^{pe} 为位置向量的维度, pos 表示当前单词的位置, i 表示向量的维度位置. 通过上述公式可知, 位置向量与该单词的相对位置和位置向量的维度位置有关, 如果位置向量的维度位置是偶数

则使用 cos 函数, 如果是奇数则使用 sin 函数.

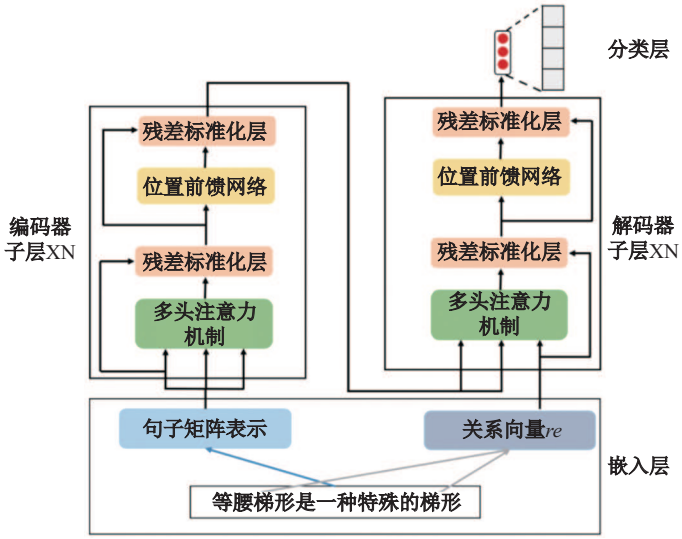

图 2　BTRE 模型

Fig. 2　BTRE model

这种位置向量表示方式同时考虑了单词的绝对位置和相对位置信息, 根据如下正弦与余弦的和角公式:

$$\sin(\alpha+\beta)=\sin(\alpha)\cos(\beta)+\cos(\alpha)\sin(\beta),$$
$$\cos(\alpha+\beta)=\cos(\alpha)\cos(\beta)-\sin(\alpha)\sin(\beta).$$

容易看出, 某个单词的位置向量可以表示为 $PE(abs_{pos}+rel_{pos})$, 其中 abs_{pos} 表示绝对位置, rel_{pos} 表示相对位置. 由上述公式可知, $PE(abs_{pos}+rel_{pos})$ 可以有 $PE(abs_{pos})$ 与 $PE(rel_{pos})$ 表示, 因此这种位置嵌入方式能够同时考虑某个单词在句子中的绝对位置与相对位置. Vaswani 通过实验证明, 基于正弦与余弦函数的位置嵌入方式与基于其表示学习的位置嵌入方式 (即 GCLSTM 使用的位置嵌入) 能够取得几乎相同的效果, 但是基于正弦与余弦函数的位置嵌入能够提前生成可能用到的向量值, 并固定下来, 而不像表示学习过程中随时发生改变, 从而减少网络参数, 进一步提高训练速度[20].

关系向量 re 的生成方式与 GCLSTM 相同, 由两个目标实体 e_a、e_b 根据先后顺序相减再加上 tanh 非线性变换得到.

编码器

在 BTRE 模型中, 编码器用来提取单词之间的局部依赖信息, 这个结构类似于 GCLSTM 中的门线性单元. 编码器由 N 个子层构成, 每个子层都由多头注意力机制、残差标准化层与位置前馈网络组成, 子层之间的参数是独立的, 使得模型可以学习到更加丰富的信息.

编码器采用自注意力机制, 即参与 DotAttention 操作的 Queries、Keys、Values 都来自同一个矩阵 H. 编码器的第一个子层运算结束之后, 其输出为第二个子层的输入, 并且矩阵 H 的每个向量都会与其他所有向量进行 Attention 操作, 这样就能充分考虑每个向量与其他所有向量之间的关系. 由于编码器中前一个子层的输出是后一个子层的输入, 因此子层输出与输入的维度需要一样, 都是 $d^x\cdot l$.

编码器使用自注意力机制来取代循环神经网络中的循环结构与卷积神经网络模型中的卷

积操作[15], 因此获得了更高的计算效率. 为了更好地对比自注意力机制、循环结构、卷积操作之间的计算复杂度, 表 2 给出了每层的计算复杂度、串行化操作数与最大路径长度对比, 其中 l 表示句子长度, d^x 表示单词向量的维度, k 表示卷积操作中卷积核的大小.

表 2 自注意力机制、循环结构、卷积操作复杂度对比

Tab. 2 Comparison of self-attention mechanism, circulation mechanism, and convolution operation complexity

运算方式	计算复杂度	串行化操作数	最大路径长度
自注意力机制	$O(l \cdot d^x)$	$O(1)$	$O(1)$
循环结构	$O(l \cdot (d^x)^2)$	$O(l)$	$O(l)$
卷积操作	$O(k \cdot l \cdot (d^x)^2)$	$O(1)$	$O(\log_k(l))$

在一般情况下, 单词向量 d^x 是大于句子长度 n的, 因此从每层计算复杂度来看, 自注意力机制最优, 卷积操作最耗时. 串行化操作数用于衡量每种运算方式的并行能力, 串行化操作数越少, 其并行化能力越强. 由于循环结构的后一步操作依赖前一步的输出, 因此其并行能力最差. 最大路径长度是指考虑句子中所有单词之间的依赖关系需要的操作次数, 最大路径长度越小, 则越容易学习到长距离依赖的信息. 自注意力机制每次都会考虑句子中所有单词之间的两两关系, 而循环结构需要的次数与句子长度相等, 卷积操作则与卷积核大小以及句子长度有关. 自注意力机制只要一次操作就能获取句子中任意两个单词之间的信息, 因此最容易学习到长依赖信息.

解码器

解码器用于从编码器输出的局部依赖信息中, 提取与目标关系相关的信息. 解码器的输入由编码器的输出 $O_{encoder}$ 和关系向量 re 组成. 在解码器的多头注意力机制中, 关系向量 re 组成 Query, 而 $O_{encoder}$ 组成 Keys 与 Values, 其含义为从 $O_{encoder}$ 中找到与关系向量 re 最相关的信息.

解码器也由多个子层组成, 与编码器一样, 每个子层都有属于自己的参数, 目的是让解码器能够充分提取到关系信息. 在解码器中, 前一个子层的输出是后一个子层的 Keys 和 Values, 因此解码器的输出维度与编码器的输出维度相同, 为 $d^x \cdot l$.

分类层

BTRE 模型的最后一层是分类层, 用来预测最终的关系, 定义如下:

$$s = W_o \cdot O^l_{decoder} + b_o,$$
$$p(y = i|s) = \frac{\exp(s_i)}{\sum_{j=1}^{t} \exp(s_j)}.$$

其中, $O^l_{decoder}$ 为解码器输出矩阵的最后一列向量, 包含了关系信息. $W_o \in \mathbf{R}^{|Y| \cdot n}$ 与 $b_o \in \mathbf{R}^{|Y|}$ 是全连接的参数.

BTRE 模型采用编码器-解码器架构. 编码器通过自注意力机制来提取单词之间的依赖信息, 解码器根据关系向量 re, 从编码器输出的依赖信息中进一步提取与目标关系的最相关信息, 从而让分类层能够更加准确地进行关系分类.

2.4 关系提取系统

为了更好地进行关系提取实验与结果展示, 本文构建了关系提取系统. 该系统由预处理、模型训练与结果展示三部分构成, 实现了数据爬取、预处理、候选关系对生成、在线标记、模

型训练与可视化、知识体系可视化等功能, 图 3 展示了关系提取系统的整体业务流程.

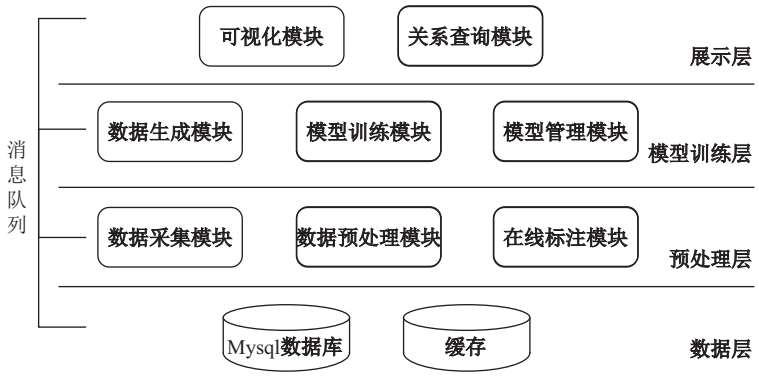

图 3　关系提取系统架构

Fig. 3　Relationship Extraction System Architecture

架构图中最底层是数据层. 该层的作用是存储系统中的所有数据, 并支持模块间进行数据交换.

架构图中的第二层是预处理层. 该层主要是为模型训练层做数据准备的, 包括数据采集模块、数据预处理模块与在线标注模块. 数据采集模块由 Python 实现, 可以与系统分开部署. 该模块从各大百科页面爬取语料, 先存储到本地, 然后进行分句后发送给数据预处理模块. 数据预处理模块会对句子进行分词, 然后统一符号与公式表达, 最后生成未标记实例保存到数据库中. 在线标注模块从数据库中查询出未标记实例交给用户标记, 该模块支持多人在线同时标记.

架构图中的第三层是模型训练层. 该层由数据生成模块、模型训练模块与模型管理模块组成. 数据生成模块将标注好的实例转换为模型输入需要的格式. 模型训练模块由 Python 实现, 部署在带有 GPU 的环境中, 通过消息队列与数据生成模块和模型管理模块进行数据交互, 该模块用于模型训练与验证, 用户可以通过选择模型来选择 GCLSTM、BTRE、PCNN (Partition Convolutional Neural Network)[19]等不同模型, 来对数据集进行训练.

架构图中的第四层是展示层. 该层分关系查询模块与可视化模块. 其中, 关系查询模块用于查询知识点之间的关系, 如相关知识点或某两个知识点之间的路径. 可视化模块则将查询到的结果以图的方式展示给用户.

为了让各个模块之间能够进行跨语言实时通信, 各模块之间还采用消息队列来作为数据交互通道, 还能够有效降低各模块间的耦合程度, 让系统非常方便地部署在分布式环境.

3　结果与讨论

验证 BTRE 性能的对比实验

本实验从百度百科、搜狗百科等较有权威性的网站中爬取了上海市初中数学知识点有关的语料, 并通过远程监督的思想标记数据, 生成带标签的数据集.

实验所用设备为一台配有英伟达 GTX1070 显卡的台式机, 系统为 ubuntu 16.04, 软件 python 版本为 3.6.3, anaconda 版本为 4.4.1, tensorflow 版本为 1.8.0, pytorch 版本为 0.3.0.

为了更好地验证 BTRE 模型的有效性以及计算复杂度, 本文设计了以下 3 组对比实验:

- **与基准算法的比较.** 将 BTRE 模型与 GCLSTM、PCNN 模型进行对比, 其中 GCLSTM、PCNN 模型为基准算法, 从而验证 BTRE 模型的性能.
- **BTRE 算法效率分析.** 对比 BTRE、GCLSTM、PCNN、Attn_BiLSTM (去掉了门线性单元并增加了自注意力机制的双向 LSTM 网络) 这 4 个模型在训练时迭代10次所消耗的时间,

从而分析这几个模型的计算复杂度差异. 在这个对比实验中, 所有模型采用相同的优化函数, 从而避免因为优化函数不同造成的差异.

- **子层数量对 BTRE 性能的影响.** 比较 BTRE 模型中编码器与解码器的子层数量对模型表现的影响. 在实验设置中, 除了子层数量以外, 保证了其他超参数是相同的, 从而排除其他因素的干扰.

上述对比实验将按照对应的指标分别进行比较.

验证 BTRE 模型的性能. 该组实验将 BTRE 模型与 GCLSTM 模型、PCNN 模型进行对比, 结果如表 3 所示, 从最终结果中我们可以发现, 相比于 GCLSTM 模型, BTRE 模型在准确率上提升了 0.62 个百分点, 在召回率上提升了 0.74 个百分点, F1 值提升了 0.68 个百分点, 在这三个指标上提升不大. 但是在 AUC 指标上提升了 0.054, 突破了 0.6.

其中, F1 值时准确率与召回率的调和平均值, 它能够兼顾准确率与召回率. AUC 是 Precision-Recall 曲线与横坐标轴行程的面积, AUC 越大, 模型表现越好.

表 3 BTRE 模型与 GCLSTM 模型、PCNN 模型实验结果对比

Tab. 3 Comparison of BTRE model with GCLSTM model and PCNN model

模型名称	准确率	召回率	F1	AUC
PCNN	74.15%	65.83%	69.74%	0.548
GCLSTM	74.73%	69.33%	71.93%	0.569
BTRE	75.35%	70.07%	72.61%	0.623

这 3 个模型的准确率-召回率曲线如图 4 所示, 从图 4 中可以看到, 相比于 GCLSTM 与 PCNN 这两个模型, BTRE 模型在低召回率处拥有非常好的表现, 而召回率是根据模型最终对关系预测的评分逆序排序的, 低召回率处的曲线表示模型对正确关系拥有很高的预测评分, 因此 BTRE 模型在高评分处的预测结果的准确率非常高. 从整体曲线来看, BTRE 模型几乎完全超越 GCLSTM 与 PCNN 模型.

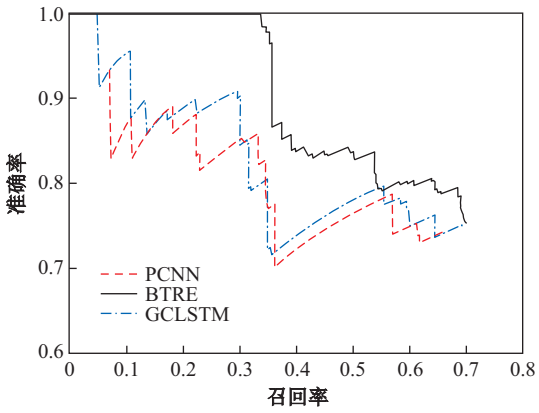

图 4 BTRE、GCLSTM、PCNN 的准确率-召回率曲线

Fig. 4 BTRE, GCLSTM, PCNN accuracy-recall rate curve

BTRE 模型之所以能取得如此好的表现, 多头注意力机制发挥了重要作用. 编码器中的多头注意力机制让 BTRE 模型能够更容易地提取到单词之间的长距离局部依赖信息, 而CNN 或者 GCN 都只能提取相邻单词之间的局部信息. LSTM 能够提取长距离依赖信息, 但是没法很好地考虑单词间的局部信息, 而 BTRE 模型的编码器弥补了它们的缺陷. 解码器中的多头注意力

机制考虑了关系向量与局部依赖信息之间的相关度, 从而提取最终的关系信息.

分析 BTRE 算法效率. 该组实验对比了BTRE、GCLSTM、PCNN、Attn_BiLSTM这四个模型的训练速度.

图5展示了上述四个模型在训练中迭代10次的时间消耗, 从图5中可以看到, PCNN模型耗时最短, GCLSTM模型耗时最长.

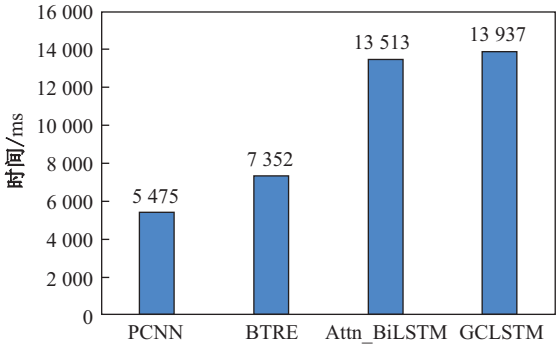

图5　BTRE、GCLSTM、PCNN、Attn_BiLSTM时间消耗对比图

Fig. 5　Comparison of time consumption for BTRE, GCLSTM, PCNN, and Attn_BiLSTM

BTRE模型比PCNN模型慢是因为BTRE的编码器-解码器架构是串行计算的, 而PCNN模型采用一层卷积结构, 通过分段之后有3个卷积操作是并行的. GCLSTM模型本身就包含注意力机制与双向LSTM结构, 并且还增加了门线性单元层, 因此时间消耗是最大的. 但是从数值上来看, GCLSTM模型也并没有比Attn_BiLSTM模型慢很多, 这也从另一方面说明门线性单元层也是一个时间复杂度较低的结构.

子层数量对 BTRE 性能的影响. 该组实验结果如表4所示, 当子层数量N=1时, 其准确率、召回率、F1值、AUC都是最低的, 两层子层相比于一层子层而言, 准确率提升了6.57个百分点, 召回率提升了5.23个百分点, F1值与AUC也提升了很多. 但是当子层数量继续增加, 模型的表现反而开始下降, 但训练集上的F1值仍在上升.

表 4　BTRE 模型中不同子层数量结果对比

Tab. 4　Comparison of the number of different sub-layers in the BTRE model

子层数量 N	准确率	召回率	F1	AUC	训练集 F1
1	68.78%	64.84%	66.75%	0.543	91.35%
2	75.35%	70.07%	72.61%	0.624	93.26%
3	71.05%	67.33%	69.14%	0.602	95.33%
4	69.90%	68.33%	69.10%	0.573	95.78%

当子层只有1层时, 只进行一次多头注意力机制, 这样只能计算出句子中任意两个单词之间的依赖信息, 没能充分挖掘多点相互依赖信息. 当子层数量达到2层时, 准确率、召回率、F1值、AUC以及训练集F1相比于1层有了很大的提高, 因为第二层子层会接着前一层的结果再进行一次多头注意力机制, 这样就能够考虑任意三个单词之间的关系, 从而提取到更加丰富的依赖信息, 这也从另一方面说明1层子层所提取的局部依赖信息是不够的. 但是当N>2时, 随着子层数量的增加, 模型在训练集上的F1值在上升, 验证集上的性能反而开始下降, 这是因为数据集规模较小, 模型发生了过拟合, 导致验证集准确率下降. 因此需要根据数据规模来设置最优的子层数量.

4 结 论

本文的主要工作如下:

(1) 构建初中数学知识体系语料 MKR. 本文从百科网站上爬取了上海市初中数学知识点有关的语料, 通过直接替换和形式化替换来统一语料中的符号与公式表述, 并采用基于远程监督的关系标记方法来生成带标签数据集.

(2) 提出基于关系向量的注意力机制. 知识体系中的关系是有向的, 为了更好地表示两个实体之间的有向性, 本文提出了关系向量, 并使用关系向量作为注意力机制的目标, 从而让模型能够捕获句子中体现知识点间关系的信息.

(3) 提出基于GCN与LSTM的GCLSTM模型. 该模型受到PCNN模型、GCN模型与LSTM模型的启发, 利用门线性控制单元提取单词之间的多点局部信息, 然后利用注意力机制寻找与目标关系最有关的局部信息, 最后通过双向LSTM层整合最终的关系信息.

(4) 为了提取单词间的长依赖局部信息和提升关系提取的效率, 本文提出基于编码器-解码器架构的BTRE模型. 该模型受到Transformer模型的启发, 编码器使用自注意力机制来提取单词之间的长距离局部依赖信息, 解码器根据关系向量从长距离局部依赖信息中提取关系信息. 编码器与解码器只使用了注意力机制, 这种结构使得模型具有更低的时间复杂度.

(5) 设计并实现了关系提取系统. 该系统分为4层: 数据层用于数据存储与数据交换; 预处理层用于数据采集、预处理以及标注数据集; 模型训练层用于生成指定格式的数据, 交给模型训练, 并记录中间结果进行可视化, 方便用户调参与模型对比; 展示层用于在构建好的知识体系上进行关系查询等操作; 通过队列来实现各模块之间的信息交换.

关于知识体系构建问题的研究, 有以下3个方面值得进一步完善:

(1) 考虑图片语义. 对于数学这个学科而言, 由于图片格式的公式和图形包含了丰富的信息, 因此图片也是十分重要的信息来源. 对于知识体系构建而言, 并不需要将图片转换成文字, 将图片语义分类后整合到模型中, 从而引入更多有意义的数据来提升关系提取算法的性能.

(2) 通过题目与解答语料来构建知识体系. 题目解答过程包含非常丰富的知识点关系, 但前提是需要对题目进行知识点分类, 题目解答过程中知识点间的相互作用关系将会对知识体系构建有非常大的帮助.

(3) 其他学科的知识体系构建. 本文仅以初等数学为切入点, 研究了知识点间的关系提取问题, 然而不同学科的知识点体现方式不同, 如在物理中公式是知识点的重要载体、在化学中反应方程式是知识点的重要载体、在英语中语法结构和固定搭配等为主要载体. 数学学科知识点间的关系提取不能简单地迁移到其他学科.

[参 考 文 献]

[1] LIU H, MA W, YANG Y, et al. Learning concept graphs from online educational data [J]. Journal of Artificial Intelligence Research, 2016, 55: 1059-1090.

[2] NOVAK J D, BOB GOWIN D, JOHANSEN G T. The use of concept mapping and knowledge vee mapping with junior high school science students [J]. Science education, 1983, 67(5): 625-645.

[3] MIWA M, BANSAL M. End-to-End Relation Extraction using LSTMs on Sequencesand Tree Structures [C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2016: 1105-1116.

[4] MINTZ M, BILIS S, SNOW R, et al. Distant supervision for relation extraction without labeled data [C]// Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2, Association for Computational Linguistics, 2009: 1003-1011.

[5] ZHOU P, SHI W, TIAN J, et al. Attention-based bidirectional long short-term memory networks for relation classification [C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2016: 207-212.

[6] LIN Y, SHEN S, LIU Z, et al. Neural relation extraction with selective attention over instances [C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2016: 2124-2133.

[7] LUONG M T, PHAM H, MANNING C D. Effective approaches to attention-based neural machine translation [C]// Empirical Methods in Natural Language Processing , 2015: 1412-1421.

[8] NGUYEN T H, GRISHMAN R. Relation extraction: Perspective from convolutional neural networks [C]// Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing, 2015: 39-48.

[9] ZHANG D, WANG D. Relation classification via recurrent neural network [J]. CoRR abs/1508.01006, 2015.

[10] HASHIMOTO K, MIWA M, TSURUOKA Y, et al. Simple customization of recursive neural networks for semantic relation classification [C]// Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013: 1372-1376.

[11] EBRAHIMI J, DOU D. Chain based RNN for relation classification [C]// Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2015: 1244-1249.

[12] SUNDERMEYER M, SCHLÜTER R, NEY H. LSTM neural networks for language modeling [C]// Thirteenth annual conference of the international speech communication association, 2012.

[13] XU Y, MOU L, LI G, et al. Classifying relations via long short term memory networks along shortest dependency paths [C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015: 1785-1794.

[14] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. Neural computation, 1997, 9(8): 1735-1780.

[15] ZENG D, LIU K, LAI S, et al. Relation classification via convolutional deep neural network [C]// 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014: 2335-2344.

[16] DAUPHIN Y N, FAN A, AULI M, et al. Language modeling with gated convolutional networks [C]// Proceedings of the 34th International Conference on Machine Learning-Volume 70, 2017: 933-941.

[17] KINCHIN I M, HAY D B, ADAMS A. How a qualitative approach to concept map analysis can be used to aid learning by illustrating patterns of conceptual development [J]. Educational research, 2000, 42(1): 43-57.

[18] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate [C]// International Conference on Learning Representations, 2015.

[19] ZENG D, LIU K, CHEN Y, et al. Distant supervision for relation extraction via piecewise convolutional neural networks [C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015: 1753-1762.

[20] MNIH V, HEESS N, GRAVES A. Recurrent models of visual attention [C]// Advances in neural information processing systems, 2014: 2204-2212.

(责任编辑: 李万会)