Education and Learning, 教育与学习(3)2021, 10
ISSN:2705-0408(P); 2705-0416(O)

基于知识图谱的诗词学习研究

刘亚娟
（哈尔滨师范大学　黑龙江哈尔滨　150025）

【摘　要】目的：基于公开的古诗相关资源，运用自上而下和自下而上相结合的知识图谱构建技术实现自动化构建古诗知识图谱，为传统的古诗词研究提供新的视角与工具。应用背景：自动抽取古诗译文中的诗歌、诗人和意象并构建三者实体之间的关系。应用一种可视化呈现的方式将碎片化的古诗词资源有序地整合到一起，为古诗词资源的推理和分析提供了理论基础。方法：通过python爬虫技术获取古诗资源数据后，使用BERT-BiLSTM-CRF模型进行实体识别，通过定义规则与融合外部领域知识构建实体间关系，运用图数据库Neo4j实现实体及实体间关系的存储与图谱可视化呈现。结果：构建古诗知识图谱诗歌—诗人—意象的本体模型，将古诗词的内容通过语义之间的内在关联，以一种结构化的形式联系在一起，能够支持多角度分析与挖掘。结论：本研究构建的知识图谱发掘了古诗文本间的内在关联，为基于诗词资源数据的研究应用提供了一个新颖易用的方案，但暂缺乏精细化的实体对齐处理及机构实体之间所属关系的构建。

【关键词】古诗词；知识组织；知识图谱

“熟读唐诗三百首，不会作诗也会吟”，古诗词不仅被认为是继承了我国优秀的传统文化的特殊艺术和教育形式，也被认为是当今中小学在现代语文课程教学中的重要资源。将知识图谱与古诗词的研究相结合已成为一种普遍趋势。现如今，中高考学生在学习古诗词时贯穿应试教育的思想，在学习和掌握古诗词时采用机械化强行记忆的方式，不仅耗时耗力，效果也很不理想。出现这种现象的原因大致有如下两点：第一，古诗词的发展和产生都是基于其当下的历史文化背景，如果不依托原有的历史背景去理解古诗词，对于初学者会存在一定的难度。第二，古诗词的相关内容存在严重的碎片化现象。分散于古诗本身、注释、翻译、鉴赏、互联网等多个出处。而知识图谱以一种智能化的方式整合诗词资源，理解数据其实是创建数据映射到知识库中对应的实体、概念和关系的过程，解释现象其实是运用知识库中对应的实体、概念和关系解释回归现象本质的过程[1]。而概念、属性、关系是人类认知事物的基础[2]。所以，将古诗词的内容通过语义之间的内在关联，以一种结构化的形式联系在一起，是一种较合理的掌握古诗词的方式。

为了解决上述的问题，本文构建基于古诗资源的知识图谱。通过从古诗的译文、赏析等内容中提取每一段的主体内容，识别主体内容中的实体以及实体之间的关系。将古诗中提取出的主体内容用一种结构化的方式组织起来。

1　相关工作

目前现有的古诗资源通常以结构化和半结构化的形式存在，但目前的古诗资源存在组织架构较为松散、知识体系错综复杂和资源异构等特点[3]，并且古诗词在其用词、用句方面的要求本就高于自然语言，且其行文习惯较为特殊。因古诗文特有的属性造成当下现有的自然语言处理技术无法直接处理，这给大数据环境下的古诗文的命名实体识别研究带来了巨大的考验与挑战[4]。但就当下的教育行业对古诗领域的智能化服务提出更高的要求[5]，目前已有的对于古诗词的研究虽在可视化展示方面取得了一定的成就，但都没有基于诗句本身进行分析和研究，未实现对于古诗词资源进行系统化的讨论。本研究采用 RDF 三元组作为描述古诗资源之间关联关系的基本表示框架，对其他多个数据源中的古诗知识也加以组织和利用。目前针对古诗资源的智能化服务主要集中在古诗词知识图谱、创作诗词和诗词知识智能问答三个方面。

目前，自然语言处理技术被广泛应用于识别出古诗文中的人名、地名和机构名信息的重要领域中，通过分析古诗文本中潜在的语义关系，进而提升对古诗文本的解析能力。同时对先秦典籍中的命名实体识别的相关研究积累了相关的技术经验。其结果也经常应用于文本分类、关系挖掘和自动文摘等，构建知识图谱正在逐渐成为一个新的探索方向。

在古诗词知识图谱领域，“唐诗别苑”[6]在古诗智能化服务虽在语义检索和可视化方面做出一定贡献，但也仅涉及诗人的社交、迁徙和作品三个特殊角度来构建诗词知识图谱，忽略了三者之间的内在联系。“宋代学术传承知识图谱”[7]，从中国历代人物传记资料库中抽取出宋代学术人物之间的传承关系和学术人物内部之间存在的亲属关系，但其知识图谱在内容结构完整性和条理清晰性有所欠缺。周莉娜[8]等人一起设计的 KnowPoetry（基于唐诗知识图谱的智能知识服务平台），该平台就唐诗知识的个性化搜索、诗人的时空轨迹和匹配语义查询三种智能化服务为用户带来了不一样的体验效果。唐宋文学编年地图也为用户提供了人的时空轨迹可视化，但其主要依赖人工整理和统计，并没有深入到古诗中实体与实体之间语义关系的

判定与处理。“宋词缱绻，何处画人间”，主要是针对宋词中的生平有礼路线和所处年代的可视化展示，但也只是简单的词频分析，不涉及语义关联的研究。王兆鹏等人共同搭建了“搜韵”平台，不仅为用户专门提供了关键词检索入口的个性化服务，还选用一种可视化方式呈现诗人的时空轨迹分布信息。

在诗词创作领域，百度新推出的“微软绝句”“微软小冰”以及清华大学孙茂松推出的“九歌”“猎户星在线写诗”。在智能问答领域，微软公司研发推出的小娜 Cortanta，MIT 的 Start 等。小娜 Cortanta 是一款主打助理软件的产品，其通过自然语言处理技术识别用户意图，不仅能实现帮助用户打电话发短信等基本操作，对于用户提出的问题还能做出简单的回应。Start 是首个基于 Web 端的智能问答，其涉及的领域范围广，大多数的英语问题都能通过 Start 找到答案。

古诗词领域知识图谱构建的研究近况（如图 1）。

分类	知识图谱	全名	URL	机构/作者
诗词知识图谱	唐诗别苑	-	http://poem.studentsystem.org.	北京师范大学
	Know-Poetry	基于唐诗知识图谱的智能知识服务平台	-	武汉大学--周莉娜
	搜韵	搜韵: 诗词门户网站	https://sou-yun.cn/	-
	宋代学术传承知识图谱	-	-	北京大学信息管理系KVision 实验室
	唐宋文学编年地图	-	https://sou-yun.cn/poetlifemap.aspx	王兆鹏
	“宋词缱绻，何处画人间”	“宋词缱绻，何处画人间”平台	-	新华网&浙江大学
诗词创作	Desktop Poet	桌面诗人	https: //www.file-extensions.org/desktop-poet-file-extensions.	-
	微软绝句	微软亚洲研究院律诗绝句	http://duilian.msra.cn/jueju/.	百度
	微软小冰	微软六代小冰	http://www.xiaoice.com/	百度
	九歌	清华 AI 九歌计算机诗词创作系统	https://jiuge.thunlp.cn//.	清华大学孙茂松
	猎户星在线写诗	-	http: //www.dopoem.com.	-
智能问答	Cortanta	小娜	-	微软公司
	Start	-	-	MIT
	阿里小蜜	-	-	阿里巴巴
	小度机器人	小度人工智能音箱	-	百度
	BabyQ	QQ群聊天机器人	-	腾讯

图 1　古诗词领域知识图谱研究现状

2　研究思路与系统方案

2.1　系统架构

网页中的古诗文本数据常以一种以数据库驱动半结构化的方式来组织，因此结合正则表达式及实体识别算法，依据 Web 页面结构的模式匹配实现古诗资源的信息采集与抽取[9]。基于以上思路，由于古诗资源分布的特殊性，本研究提出了一种将自顶向下和自底向上相结合的古诗词知识图谱构建框架方式，主要包括数据源获取、知识抽取、知识融合、存储和可视化四层，如图 2 所示。

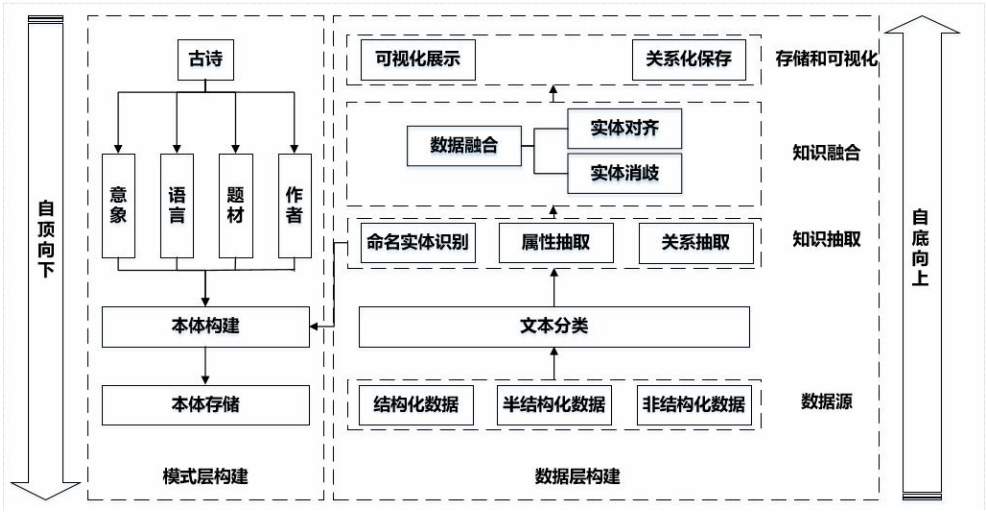

图 2　古诗知识图谱架构

在数据获取层，本研究使用网络爬虫采集古诗文网站公开的古诗资源数据，通过古诗名、URL 及人工检查的方式去重建并保存为 JSON 格式文件。在信息抽取层，一方面借助正则表达式抽取出古诗中的人物姓名、籍贯、朝代、字、号等信息；另一方面，通过 BERT-BiLSTM-CRF 深度学习算法抽取每一个事件中的地点以及人物关系。在关系抽取层，对抽取出的信息，确立人、机构、地点三类实体及其对应属性，针对特殊表述情况，对抽取的实体执行共指消歧操作[10]。引入古诗领域的知识，在实体间建立表达关系的语义三元组，并存储为实体关联文档。最终再存储为可视化层，将古诗词语义三元组批量导入到图数据库 Neo4j 中，实现古诗知识图谱的关联化存储及可视化展

现，支持用户通过点击或 Cypher 查询语言的方式浏览与检索图谱。

2.2 基于BERT-BiLSTM-CRF模型的命名实体识别

对于古诗译文中的意象、情感和主题的表述方式多样、长短不一，并无统一规律可循，故基于正则表达式匹配的抽取方法在此并不适用，需要借助算法进行实体识别。本研究采用 BERT-BiLSTM-CRF 模型识别诗歌实体、诗人实体和意象实体。将 BiLSTM（Bi-directionalLong Short Term networks，BiLSTM）和 CRF（Conditional Random Fields，CRF）两者的结合的 BiLSTM-CRF 模型最先由百度提出并应用在序列标注任务上取得良好的结果。BERT 模型则是谷歌在 2018 年发布的自然语言处理模型，其强大的性能优势刷新了近 11 项自然语言处理的子任务的相关记录。将 BERT 与 BiLSTM-CRF 相结合的模型以提升命名实体识别的性能，成为近期研究的热点。

BERT-BiLSTM-CRF 模型包括 BERT 层、BiLSTM 层和 CRF 层三个子部分，总体架构逻辑如图 3 所示，其大概思路为，首先将已有的语料库输入到 BERT 预训练语言模型中，在获得诗句中每个字的字向量后再将字向量输入到双向的 BILSTM 中提取特征，最后，将得到的 BILSTM 特征向量由 CRF 模块解码并输出频率较高的标记序列。

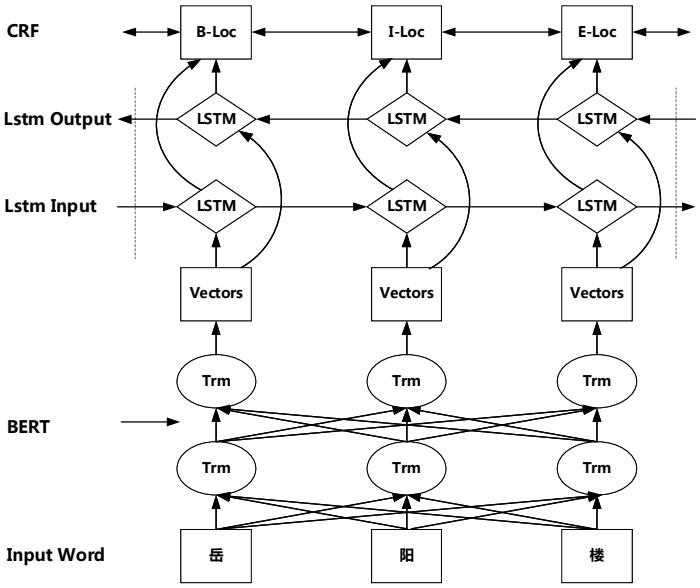

图 3 BERT-BiLSTM-CRF 模型示意图

BERT 其实是一种采用双向 Transformer 编码结构的预训练语言模型，可以得到每个字的上下文信息，在一定程度上不仅能够表征字的多义性还能体现句子的句法特征。Transformer 编码结构的核心单元是自注意力机制及前馈神经网络，单元可以连续堆叠。在自注意力机制中，每个输入的词都将转为 3 个不同的向量，分别是查询向量 $\boldsymbol{Q}$、键向量 $\boldsymbol{K}$、值向量 $\boldsymbol{V}$[11]。核心思想是计算每个输入序列中不同单词和其他单词之间的相关程度，并根据相关性来调整每个单词的重要性权重，计算方式如公式（1）所示：

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

BERT 的输入数据由字向量（TokenEmbedding）、位置向量（PositionEmbedding）、段向量（SegmentEmbedding）三部分构成，通过深层双向编码生成最终的字向量，输入到 BiLSTM 中进一步学习这个字向量上下文信息特征。

BiLSTM 作为模型重要的中间部分，能够快速准确地捕获到诗句的上下文特征获得更加全面的语义信息。LSTM 克服了传统 RNN 模型在处理长序列数据时可能会发生梯度消失或爆炸的现象，它由一个记忆单元和更新门、输出门、遗忘门组合而成，通过三个门的限制机制来达到控制记忆单元中信息的遗忘、更新和传递，最终达到使有用的信息即使经过较长的序列也能保存在记忆单元中不被遗忘的功能。t 时刻细胞的状态更新可由以下公式（2）（3）（4）（5）（6）（7）得到：

$$i_t=\sigma\left(W_i\left[h_{t-1};x_t\right]+b_i\right) \quad (2)$$

$$f_t=\sigma\left(W_f\left[h_{t-1};x_t\right]+b_f\right) \quad (3)$$

$$o_t=\sigma\left(W_o\left[h_{t-1};x_t\right]+b_o\right) \quad (4)$$

$$g_t=\tanh\left(W_c\left[h_{t-1};x_t\right]+b_c\right) \quad (5)$$

$$c_t=f_t*c_{t-1}+i_t*g_t \quad (6)$$

$$h_t=o_t*\tanh\left(c_t\right) \quad (7)$$

其中，（2）、（3）、（4）中含有的 σ 表示 sigmoid 激活函数；（5）、（7）中包含的 tanh 表示双曲正切激活函数，（2）中的 W_i 表示输入门权重，（3）中的 W_f 表示遗忘门权重，（4）中的 W_o 表示输出门权重，（2）、（3）、（4）中包含的 b_i，b_f，b_o 代表偏置项。在应用中，LSTM 只能获得前文的信息而无法有效利用后文信息，BiLSTM 由正反序两个 LSTM 构成，分别存储前文和后文的隐藏信息，

Education and Learning, 教育与学习(3)2021, 10
ISSN:2705-0408(P); 2705-0416(O)

然后计算两个隐藏单元的结果作为最终输出。

CRF 作为模型的最后一部分，其条件为输入给定的一组随机变量，输出另一组随机变量的条件概率分布模型，模型通过考虑了标签与标签之间的相互依赖性来获得全局最优的标签序列[12]。

本文使用的是线性链条件随机场，可以用条件概率 P（Y|X）表示。其中 X 是输入变量，Y 是预测序列，则有公式（8）：

$$P(Y|X) \propto \exp\left(\sum_{k=1}^{K} w_k f_k(y,x)\right) \quad (8)$$

其中：f 表示特征函数，w 表示特征函数对应的权重。在模型训练时，采用极大似然估计得到条件概率模型（Y|X）。

3 系统设计与实现

3.1 古诗译文数据的获取与简单抽取

本文使用 Python 中的 Beautifulsoup 库对互联网数据进行抽取。Beautifulsoup 能够将每一个古诗资源的 Html 文档转换成一个树形结构，将文档代表的树形结构细分拆解到 Python 对象可以处理的程度。在分析了互联网网页的结构后，本文提出了三种数据抽取方法[13]。

第一种，根据关系抽取。在分析出网页数据间存在的关系后，在指定的关系下抽取相关的古诗信息。在网页的类型栏下会将古诗分为多个类型，每个具体类型名下的古诗都属于该类型，这种数据抽取方式的优点是抽取过程简单，缺点是同一古诗可能会对应多种类型，在对每一种类型下的古诗进行抽取时会出现同一古诗多次抽取的情况，不便于后续的数据处理。

第二种，根据古诗抽取。在网页的古诗栏中，古诗分页存储，每一页有 20 首古诗，通过更改网页的 URL 信息能够抽取到每个页面中所有的古诗数据。这种数据抽取方式的优点是抽取过程简单，缺点是存在诗人和古诗的重名问题，会造成一定的数据混乱，给后续的数据处理带来不便。

第三种，根据诗人抽取。在网页的诗人栏下有诗人的基本信息以及该诗人的古诗链接，在网页的诗人部分首先获取到诗人的 ID 信息，随后根据诗人信息下方的古诗链接抽取相关的古诗数据。这种抽取方式的优点是诗人和古诗间的关系固定，结构清晰，易于抽取。

在对比了每种方法的特点后，本文选择使用第三种方法进行古诗数据抽取。

3.2 训练数据及模型参数设置

为获得质量较高的训练语料，本文选取了古诗文网中 63 首古诗数据，人工标注了 1571 个机构、职位实体，后使用百度 LAC（LexicalAnalysisofChinese）工具[14]、BIOES 的标注模型进行命名实体识别标注：B 代表一个实体的起始，I 代表一个实体的内部，E 代表一个实体的结束，S 代表单字构成的实体，O 代表文本中非所关注的实体部分；标签类型包括姓名（name）、地点（address）、机构（organization）、职位（position）四类。标注完毕后，共产生 5810 条以句为单位的数据，分为 4841 条训练数据、969 条测试数据。

将诗句“故人西辞黄鹤楼，烟花三月下扬州。”按字级粒度读入到 BiLSTM 模型中并转化为字向量，并通过 BiLSTM 中双层的 LSTM 网络学习诗句中前后文的语义信息特征，使用 BIESO 标注体系最终输出诗句中各个字的标签。“故”对应“B-People”“人”对应“E-People”“黄”对应“B-LOC”“鹤”对应“I-LOC”“楼”对应“E-LOC”“扬”对应“B-LOC”“州”对应“E-LOC”，其他标签为“O”，其中存在“故人”一个人物实体，此处的故人是指李白的老朋友孟浩然，BILSTM-CRF 模型在 LSTM 中添加了双向特征，并在 BILSTM 上添加了 CRF 模型层，依赖 CRF 模型计算转移特征从整体角度考虑句子的标签信息。

BERT 模型是一种基于 Transformer 的深度双向预训练模型，而 Transformer 其本质又是一种注意力机制[15]。BERT 利用编码方式学习诗句中词和词之间与之相关的上下文逻辑关系，当碰到一个词的时候，应该从整体角度考虑这个词的前后语义的信息。在模型训练中，BERT 模型分别从位置字向量（TokenEmbedding）、位置向量（PositionEmbedding）、段向量（SegmentEmbedding）三个方面将输入的诗句，如“故人西辞黄鹤楼，烟花三月下扬州。”按照字级粒度进行向量的转化并充分获取上下文的语义信息。最终将上面提到的三个向量求和并传送到双向 Transformer 层中，经过模型的反复训练最终输出每个字对应的标签信息，从而可以识别出诗句中包含的“故人”这个人物实体。

本研究选用了 BERT 中文预训练向量中的谷歌开源数据，参数如下：BERT-Chinese 共 12 层，隐层为 768 维，采用 12 头模式，共计 110M 个参数。训练时，最大序列长度采用 60，rain _ batch _ size 为 64，其他参数均采用默认值。BiLSTM 层设置其隐藏层维数为 64 层。

4 结果与应用

4.1 模型性能及识别结果

为检验 BERT-BiLSTM-CRF 模型的有效性，本文采用准确率 P、召回率 R 以及 F1 值三个指标对该模型针对不同类型实体识别的实验结果进行评价，评价指标的计算如公式（9）（10）（11）所示，可以看到对于各类型实体识别，该模型都具有较好的效果，BERT-BiLSTM-CRF 模型的准确率 P、召回率 R 和 F1 值表现最佳且显著高于其他两个模型。

$$P = \frac{\text{正确识别出的实体数目}}{\text{所有识别出的实体数目}} \times 100\% \quad (9)$$

$$R = \frac{\text{正确识别的实体数目}}{\text{所有标记的实体数目}} \times 100\% \quad (10)$$

$$F_1 = \frac{2 \times P \times R}{P + R} \quad (11)$$

将训练好的模型应用到实际的古诗文的文本数据中，共抽取出 4296 个不同的诗人实体与 12241 个地点信息。实体识别总体效果较好，基本能够准确地分出诗人及地点信息。相比于其他模型，BERT-BiLSTM-CRF 模型在实体的起止位置判断上较准确。

4.2 图谱应用

本次研究选用图数据库 Neo4j 作为古诗词知识图谱存储与可视化展示的工具，用户可以通过 Neo4j 交互式界面轻松地对图谱进行搜寻和探查，亦可以通过 Cypher 查询语言对图数据库进行检索。

5 结语

本研究基于公开的古诗资源数据，构建了一个古诗知识图谱，实现了古诗分析技术与知识图谱技术的结合。同时用图谱以可视、可交互的方式呈现古诗词资源的信息，不仅能够弥补了现有古诗信息分散、无法查询、无法揭示深层关联的不足，也可以为其他基于古诗数据的研究与决策提供了一个直观而有效的方案和工具。同时，也需要看到，本研究在如下方面尚有优化的空间：其一，现有意象实体之间相对独立，缺乏层级间关系描述，进而一定程度上制约了意象之间上下级等关系的揭示。其二，由于地名变更、写作不规范等因素，部分同一地点实体在不同诗文中写作不同的地点名称，现有图谱仍无法完全精确识别其中关联，导致将其视为不同的地点实体。针对上述问题，未来仍可聚焦领域知识引入关系补全、实体对齐等任务，进一步完善基于古诗数据的知识图谱构建工作。

作者简介：刘亚娟（1995.2—），女，新疆哈密人，研究方向：自然语言处理，知识图谱。

【参考文献】

[1] 徐增林，盛泳潘，贺丽荣，等.知识图谱技术综述[J].电子科技大学学报，2016，45（4）：589-606.

[2] 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.

[3] Zhao，M.，Wang，H.，Guo，J.，*et al*.Construction of an Industrial Knowledge Graph for Unstructured Chinese Text Learning[J].Applied Sciences，2019，9（13）.

[4] Hongwei Wang，Miao Zhao，Xing Xie，*et al*.Knowledge Graph Convolutional Networks for Recommender Systems[J].In Proceedings of the 2019 World Wide Web Conference，2019.

[5] 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.

[6] 李振，周东岱.教育知识图谱的概念模型与构建方法研究[J].电化教育研究，2019，40（8）：78-86+113.

[7] 曹欣怡.古诗知识搜索系统的设计和实现[D].杭州：浙江大学，2016.

[8] 周莉娜，洪亮，高子阳.唐诗知识图谱的构建及其智能知识服务设计[J].图书情报工作，2019，63（2）：24-33.

[9] 冯俐.基于 Neo4j 图数据库构建中学语文诗词知识图谱[D].西安：陕西师范大学，2019.

[10] 刘昱彤，吴斌，白婷.古诗词图谱的构建及分析研究[J].计算机研究与发展，2020，57（6）：1252-1268.

[11] 黄华升.基于知识图谱的个性化学习资源推荐研究[J].软件工程，2018，21（10）：24.

[12] 刘昱良.基于知识图谱的个性化学习资源推荐研究[D].新乡：河南师范大学，2018.

[13] Huang Z，Xu W，Yu K.Bidirectional LSTM-CRF models for sequence tagging[J].arXiv preprint arXiv，2015.

[14] Lample G，Ballesteros M，Subramanian S，*et al*.Neural Architectures for Named Entity Recognition[C].ProceedingsofNAACL-HLT，2016：260-270.

[15] Collobert R，Weston J，Bottou L，*et al*.Natural language processing （almost） from scratch[J].Journal of Machine Learning Research，2011，12（8）：2493-2537.