本科生毕业论文（设计）

毕业论文（设计）题目：

基于自然语言处理的知识图谱半自动化构建研究

部　　院　系：　教育学部

专　　　业：　教育技术学

学　　　号：　201711010109

学 生 姓 名：　简靖琳

指 导 教 师：　郑勤华

指导教师职称：　教授

指导教师单位：　教育学部教育技术学院

2021 年 5 月 24 日

北京师范大学本科生毕业论文（设计）诚信承诺书

本人郑重声明： 所呈交的毕业论文（设计），是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

本人签名： 2021 年 5 月 24 日

北京师范大学本科生毕业论文（设计）使用授权书

本人完全了解北京师范大学有关收集、保留和使用毕业论文（设计）的规定，即：本科生毕业论文（设计）工作的知识产权单位属北京师范大学。学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，允许毕业论文（设计）被查阅和借阅；学校可以公布毕业论文（设计）的全部或部分内容，可以采用影印、缩印或扫描等复制手段保存、汇编毕业论文（设计）。保密的毕业论文（设计）在解密后遵守此规定。

本论文（是、否）保密论文。

保密论文在__/___年__/__月解密后适用本授权书。

本人签名： 2021 年 5 月 24 日

导师签字： 2021 年 5 月 24 日

基于自然语言处理的知识图谱半自动化构建研究

摘　要

在这项研究工作中，从文本数据中自动化地提取知识图谱的问题得到了一定程度的解决。以往从文本这类非结构化数据中抽取知识图谱主要依赖专家进行人工处理，这种方法十分费时费力。而且随着文本信息的快速增多，人工处理已经逐渐变得不可行。而在教育领域中存在大量文本数据，缺乏对文本信息的自然语言理解会阻碍人工智能技术在教育领域的应用。

本研究以教学视频转录的文本记录作为数据源，设计并实现了一套基于自然语言处理技术、采用了深度学习、统计学、规则方法的知识图谱半自动化构建方案。该方案注重了构建成本效益并检验了提取结果的有效性。不仅降低了从文本中构建知识图谱的成本，还能为更多教育技术应用提供结构化的数据基础。

关键词：知识图谱；深度学习；人机协同；BiLSTM-CRF

Research on Semi-automatic Construction of Knowledge Graph Based on Natural Language Processing

ABSTRACT

In this work, the problem of automatically extracting knowledge graph (KG) from text data is partially addressed. In the past, extracting knowledge graph from unstructured data such as texts mainly relied on experts by their manual processing, which was not only time-consuming and laborious, but also has gradually become infeasible with the rapid increase of text information. However, in the education field, there is a large amount of text information and the lack of natural language understanding of text data will hinder the educational application of artificial intelligence technology.

This research uses transcripts of instructional videos as the data source to design and implement a semi-automatic knowledge graph construction method based on natural language processing, using deep learning technology, statistical method and rule-based approach. In addition, this work focuses on the cost-effectiveness of the construction of KG and the validity of the extraction results. So it could not only reduce the cost of constructing KG for texts, but also provides a structured data foundation for more educational technology.

KEY WORDS： Knowledge graph; deep learning; human-machine collaboration； BiLSTM-CRF

目 录

1 引言

1.1 研究背景

1.1.1 自适应学习是智慧教育和智慧学习的根本

《2020 年地平线报告：教与学版》指出教育教学技术的三大趋势之一就是人工智能技术（陈新亚 & 李艳，2020）[1]，而其中自适应学习技术被报告介绍为对教育未来发展会有重要影响的六项新兴技术和实践之一（Feldstein & Hill，2016）[2]，自适应学习技术能为学生提供与其学习水平适应并且精确的学习支持。目前主要的应用场景是具有相对清晰的知识体系与结构的科学与工程领域。

然而目前的自适应学习技术无法应用于许多知识领域，尤其是没有课程标准的、概念交叉复杂的知识领域，我们可称其为泛知识领域。如社会科学领域可被认为是一个泛知识领域，社会科学中概念众多且概念之间的关系交叉复杂；又如高等教育领域，前沿领域知识更新速度快，高校课堂教学及讲座也不断地对知识体系进行新的改进和补充，课堂教学中的内容很可能远远超过了原定大纲所呈现的知识结构，而大部分讲座没有提供知识大纲。我们能够看到，目前由于难以对在泛知识领域的知识体系进行表征，因此在这样的知识学习领域中，自适应学习技术仍然无法介入。

1.1.2 自适应学习的核心是知识图谱

近年来，自适应学习技术是智慧学习的重要体现形式。自适应学习系统能够根据学生的水平为学习者智能化推送学习资源、测评手段、练习内容。根据朱的研究，自适应学习系统中的底层模型可以分为四种：内容模型，学生模型，界面模型和推荐引擎（朱艳茹等，2018）[3]。其中，内容模型刻画了学习内容对应的知识体系，是自适应学习系统的重要组成部分。可以说，内容模型的质量决定了自适应学习系统的质量。

---

[1] 陈新亚, & 李艳. (2020). 《2020 地平线报告: 教与学版》 的解读及思考--疫情之下高等教育面临的挑战与变革. 远程教育杂志, 38(2), 3-16.

[2] Feldstein, M., & Hill, P. (2016). Personalized learning: What it really is and why it really matters. Educause review, 51(2), 24-35.

[3] 朱艳茹, 范亚芹, & 赵洋. (2018). 基于知识图谱的自适应学习系统知识模型构建. 吉林大学学报: 信息科学版, 36(3), 345-350.

而知识图谱（Knowledge Graph，KG）作为一种重要的人工智能技术，在技术应用到教育的时候成为了知识在计算机中的重要表征手段。知识图谱作为一种特定结构的信息库，能够集成多种异构数据。同时，其能进行支持一定的自动化数据处理，具有大量应用方面的优势：推荐、评估、管理等，已经有大量例子证明了它在支持应用方面的优势。

在基础教育领域，现在已有部分平台对学习资源以及知识点抽取知识图谱，为学习者提供一定程度的知识导航、知识推荐或者个性化学习支持，取得了良好的效果。这些知识图谱通常通过经验丰富的专家或教师以手工的方式构建。

但专家人工提取知识图谱的方法很难迁移到泛知识领域的图谱构建中，一方面由于泛知识领域的复杂程度高，人工处理是容易出错的，而且随着数据量爆炸式增长正在逐渐变得不可行；另一方面抽取知识图谱这样的工作需要高水平的领域专家才能胜任，而他们的时间和精力是宝贵的。在这样的情况下，自动化或者半自动化构建知识图谱的方案应该被提出来，用于解决这类泛知识领域抽取知识图谱的问题。

1.1.3 新的技术为自动化构建知识图谱创造可能

从 2012 年 Google 公司提出知识图谱以后，知识图谱因其具有的巨大应用价值受到了广泛关注。知识图谱因其能够结构化地表示信息以及信息之间的语义关系，被认为是人工智能中认知智能的基础，也有学者认为结合知识图谱和机器学习的人工智能技术有潜力产生强人工智能。同时，在工业界，近年来在医疗、金融等领域，许多企业把构建知识图谱作为重要的人工智能技术发展战略规划，而特定领域的知识图谱也不断增强了相关服务和应用的智能化。

而构建知识图谱的技术和方法也在不断发展，早期的知识图谱主要由人工构建，后来众包构建成为一种有效的构建办法，但人工构建的巨大成本以及注定无法规模化的问题无法忽视。而近几年随着自然语言处理技术（Natural Language Processing, NLP）的发展，不断有新的技术被引入了知识图谱构建，例如图像识别、语音识别、监督学习、增强学习等，不断提高知识图谱构建的效率和准确率，也带来了许多振奋人心的结果。

1.2 论文的主要工作以及创新之处

1.2.1 主要工作

在这项研究工作中，从文本数据中自动化地提取知识图谱的问题得到了一定程度的解决。以往从文本这类非结构化数据中抽取知识图谱主要依赖专家进行人工处理，这种方法不仅费时费力，而且随着文本信息的快速增多，人工处理已经逐渐变得不可行。而在教育领域中文本数据大量存在，缺乏对文本信息的自然语言理解会阻碍人工智能技术在教育领域的应用。

本研究以教学视频转录的文本记录作为数据源，设计并实现了一套基于自然语言处理技术、采用了深度学习、统计学、规则方法的知识图谱半自动化构建方案。该方案注重了构建成本效益并检验了提取结果的有效性。不仅降低了从文本中构建知识图谱的成本，还能为更多教育技术应用提供结构化的数据基础。

1.2.2 创新之处

本研究设计并实现了一套半自动化抽取知识图谱的方案，是第一个为中文教学视频构建知识图谱的研究，此方案的创新性在于：

(1) 在知识图谱模式层的设计中，本研究针对知识实体抽取采用了自下而上的识别方法，能够尽可能大范围地抓取知识要素；而在实体关系连接中设计并采取了自上而下的识别方法，有利于抓取具有教育学意义的实体关系。

(2) 在知识实体识别部分，针对文本数据，本研究采取了人机协同和深度学习的半自动化抽取方法，在能够平衡良好的准确率和召回率的情况下尽可能地减少对专家人工的依赖。

(3) 在知识实体关系连接部分，本研究创新地提出先对文本数据进行小节分割，提取文本信息结构后再从结构中获得实体之间的关系，是解决非结构化文本中知识实体的先序、包含、相关**关系识别问题**的一次尝试。

1.3 论文组织结构以及研究框架

1.3.1 组织结构

论文各章节概述如下：

第一章：引言，统领全文。首先介绍了引入新技术自动化构建知识图谱的研究背景和研究意义，接着阐述了本研究的主要工作和创新点，最后介绍了本文的组织结构及研究框架。

第二章：文献综述，为后文技术设计以及研究设计提供一定的基础。首先回顾了知识图谱应用到教育领域的价值；接着阐述了教育领域中知识图谱的已有研究，特别聚焦于已有研究中知识图谱的构建方法、知识表征粒度、关系连接方法及教育学意义；最后阐述了和本研究最相关的基于视频转录的文本数据的知识图谱构建研究。

第三章：相关技术基础，为研究设计、实现和检验提供技术基础。一方面阐述了知识图谱构建技术的理论和方法，尤其是针对文本数据的知识图谱构建技术；另一方面阐述了本研究采用的其他算法，包括：基于互信息和左右熵的新词挖掘算法、基于 BiLSTM 和 CRF 的实体识别算法、实体识别效果检验指标和基于滑动区间和词向量的小节分割算法。

第四章：研究设计，设计每个环节的实现与检验方案。首先说明了这项研究的研究目

标以及构建思路；接着详细地阐述了每一部分的构建方案设计；最后说明了数据来源和构建技术框架。

第五章：知识图谱半自动化构建的实现与检验。根据第四章的研究设计方案，本文以《教育大数据》为例，进行了数据处理、命名实体识别及效果检验、基于统计方法和规则方法的实体关系连接以及知识图谱的储存及可视化呈现。

第六章：总结与展望。一方面对这项研究工作进行了总结和分析；另一方面阐述了知识图谱的半自动化构建研究未来的改进方向。

1.3.2 研究框架

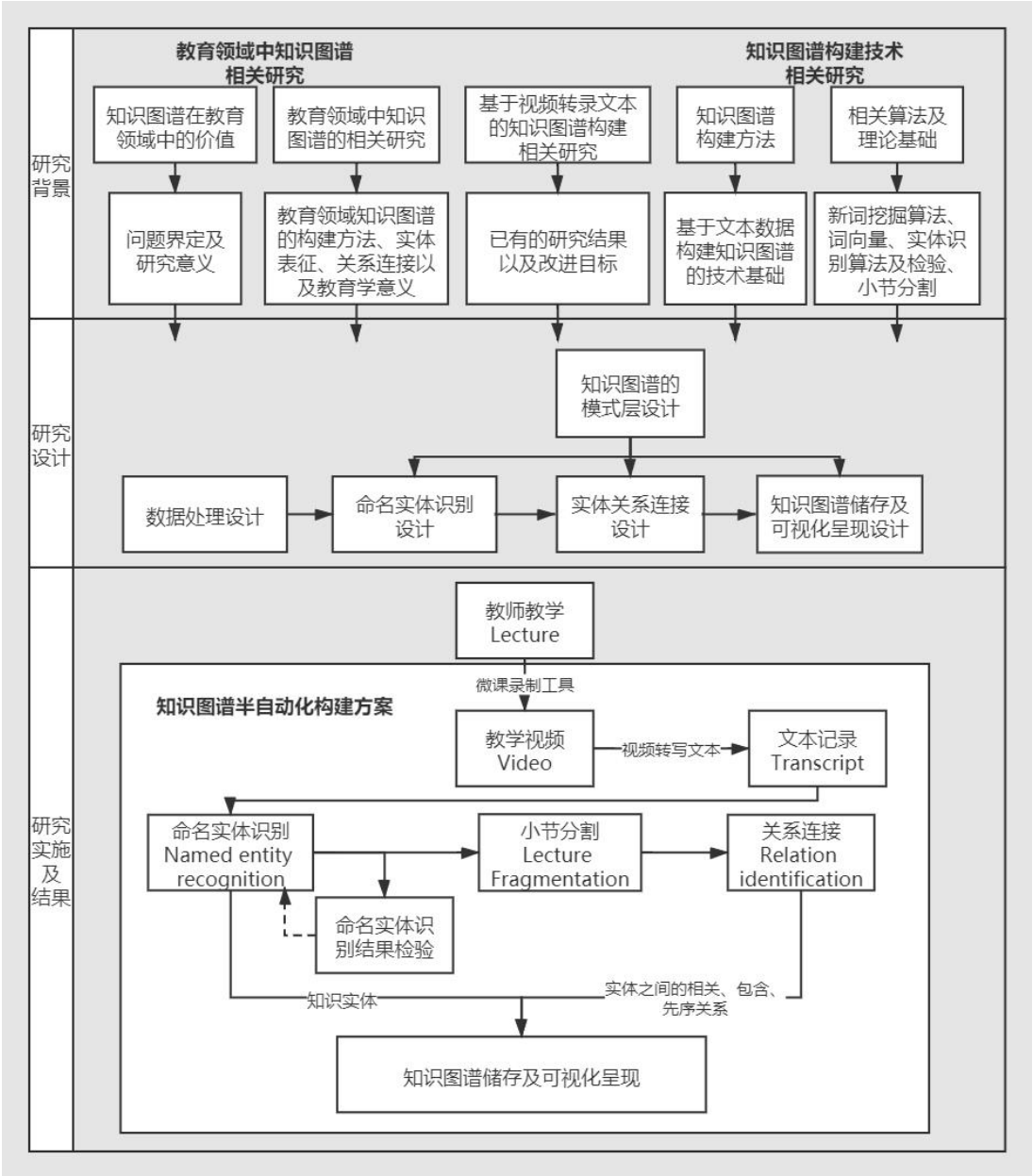


图 1 本文整体研究框架

2 文献综述

本章结构如下：首先回顾了为什么要研究知识图谱在教育领域的应用，即知识图谱教育应用的价值；接着阐述了教育领域中知识图谱的已有研究，特别聚焦于知识图谱的构建方法、实体表征、关系连接以及教育目的；最后阐述了和本研究最相关的基于视频转录的文本数据的知识图谱构建研究。

2.1 知识图谱在教育领域中的价值

在一方面，教育领域知识图谱为教育数据的集成和解构提供了可能。知识图谱作为一个集成信息库，能够将不同领域的异构数据连接在一起。而随着多媒体教学的普及，讲座以及课程中的教学大多从传统黑板板书转向了多媒体演示，在这过程中产生多模态的数据，如图片、文本、视频、音频等。例如，演讲者的演讲语音就是一种语音模态的教学数据，包含了对重要知识点的讲解，以及知识点的组织逻辑，然而这些数据往往由于无法处理而无法挖掘出其中的价值（祁晓慧，2020）[4]。另一方面，教育领域中的知识图谱具有大量应用方面的优势。比如它是能够辅助和促进智慧教育（Smart education）的有效工具（Shi et al., 2019）[5]，尤其是个性化教育与学习（Education individualization）（Rizun, 2019）[6]。

教育领域的知识图谱就如同教育领域的数据电网，有极其丰富的研究价值和使用价值，近年来逐渐引起教育学术界和工业界的关注。

2.2 教育领域中知识图谱的相关研究

本节将从四个不同的角度回顾已有的教育知识图谱相关研究：不同的自动化程度构建方法、不同的实体粒度、不同的关系连接以及特定目的的教育知识图谱相关研究，以深入了解知识图谱在教育领域的构建方法以及构建成果，为本研究打下基础。

---

[4] 祁晓慧.(2020).多模态课程知识图谱构建与应用研究(硕士学位论文,吉林大学).

[5] Shi, W., Liu, X., Gong, X., Niu, X., Wang, X., Jing, S., ... & Luo, J. (2019, November). Review on Development of Smart Education. In 2019 IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI) (pp. 157-162). IEEE.

[6] Rizun, M. (2019). Knowledge graph application in education: a literature review. Acta Universitatis Lodziensis. Folia Oeconomica, 3(342), 7-19.

2.2.1 不同构建方法的知识图谱

从知识图谱的构建方法上来说，目前主要的教育知识图谱构建方法有四种：人工、众包构建、机器、人机协同的方法，而且随着人工智能技术不断更新，越来越多的研究采用了机器的方法尝试自动化地构建知识图谱。但由于机器目前的自然语言处理技术有限，在特定领域的应用中，为了提高知识图谱的可用性，人工需要在一定程度适度介入以提高图谱的精准度。下面分别介绍了四种构建方法的相关研究。

(1) 采用人工的方法具有比较深入的知识结构，但缺乏灵活性，这里不进行主要介绍；

(2) 采用众包构建的方法：有研究者设计并运用了众包构建的方法构建了具有学校-学生-教师-课程-知识点-练习类型节点构成的某高校教育技术学领域知识图谱，并能够根据应用场景分离出三种子图：教师-课程-类型的子图、学生-课程子图以及知识历程子图。在构建过程中，研究者运用了分组构建、多重验证、逆向生成验证码和松弛策略来消除不相关答案等方法进行任务检查。

(3) 采用机器方法进行知识图谱构建的研究：如朱福军采用 TF-IDF 算法和信息增益算法对学习元平台上的学习资源进行实体识别，用潜在语义分析计算实体之间的相关度从而构建知识图谱（朱福军，2018）[7]的研究等。

(4) 采用人机协同的方法构建知识图谱的研究有：朱艳茹先基于该领域内专家的教学经验完成初始知识图谱的搭建，再使用 FP-Growth 算法对学习者行为数据进行知识节点间关系的挖掘，最终完成知识图谱的动态更新（朱艳茹，2018）[8]。还有国外研究者采用用户自主导入高质量的教科书的方法，系统自动整合 web 上的学术文章以及 Wikipedia 中的信息集成到知识图谱的构建中，形成了半自动化的知识图谱构建工具（Wang, 2017）[9]。

2.2.2 不同实体粒度的知识图谱

大多数教育知识图谱研究在进行知识模型建立的时候会根据不同需求设计不同粒度的实体，主要有三种不同的层次，从宏观到微观分别是：学习资源粒度、知识点粒度、概念粒度，还有一些特殊的知识组织模型的工作。

(1) 学习资源粒度

2016 年有研究者集成了教育类的电子博客和视频资源构建了能够为学习者推荐相关学习资源的知识图谱。其中他们将电子博客的文本资料以及视频资源的隐藏字幕（Closed captions，也叫 CC 字幕，是通过手动创建而成的）映射到主题的公共语义空间并进行匹配，

---

[7] 朱福军. (2018). 基于学习元的领域知识图谱自动构建研究 (Master's thesis, 四川师范大学).

[8] 朱艳茹. (2018). 基于知识图谱的自适应学习系统设计与实现 (Master's thesis, 吉林大学).

[9] Wang, S. (2017). Knowledge Graph Creation from Structure Knowledge.

即运用 LDA 进行主题建模并根据信息论和概率论进行了两个文本主题相似性的度量（Basu et al., 2016）[10]。还有侯在基于 MOOC 的高等教育知识图谱的构建中，根据具体的 MOOC 网站的网页上的课程信息，定义该知识图谱的实体类型有以下几种：课程组、课程、授课老师、参考书籍、网站及学校（侯俊萌，2017）[11]。

(2) 知识点粒度

有研究者整合了以知识点多媒体资源为单元的知识图谱（Ilkou & Signer, 2020）[12]，他们利用在 2007 年开发的一个基于资源链接的超媒体元模型（Resource-selector-link, RSL）（Signer & Norrie, 2007）[13]并开发了 EduKnow 平台，但 RSL 模型并不涉及知识点下位概念的连接。

(3) 概念粒度

2016 年卡耐基梅隆大学的研究者尝试自动化地构建一个能够表征课程先决关系的有向图，他们将来自不同来源的在线课程所涵盖的知识概念映射到一个通用的概念空间，并利用概念之间的关系预测课程之间潜在的先决关系。

由 Chen 等人在 2018 年开发的 K12EduKG 系统，可以自动为 K-12 阶段的数学学科构建知识图谱，他们利用 K12 阶段国家数学课程标准提取了数学知识点并识别具有教育意义的关系，如“实数”“包含”“整数”等。接着使用了概率关联规则挖掘 p-Apriori 算法来识别数学概念之间的认知先决关系（Chen et al., 2018）[14]。

2020 年蒋琪构建并实现了半自动化提取概念粒度的 MOOC 知识地图的方案（蒋琪，2020）[15]，研究者从 MOOC 已有的章节结构以及先修文本中提取知识概念，如“二叉树”等，并基于规则提取了概念间的包含关系和先序关系，基于词向量表征提取了概念间的相关关系。

(4) 知识组织模型

通用的知识组织模型（Hodge, 2000; Bollacker et al., 2008; Wu et al., 2012; Dong et al.,

---

[10] Basu, S., Yu, Y., Singh, V. K., & Zimmermann, R. (2016, January). Videopedia: Lecture video recommendation for educational blogs using topic modeling. In International Conference on Multimedia Modeling (pp. 238-250). Springer, Cham.

[11] 侯俊萌. (2017). 基于 MOOC 的高等教育知识图谱的构建 (Doctoral dissertation, 北京: 北京邮电大学).

[12] Ilkou, E., & Signer, B. (2020). A Technology-enhanced Smart Learning Environment based on the Combination of Knowledge Graphs and Learning Paths. In CSEDU (2) (pp. 461-468).

[13] Signer, B., & Norrie, M. C. (2007, November). As we may link: a general metamodel for hypermedia systems. In International Conference on Conceptual Modeling (pp. 359-374). Springer, Berlin, Heidelberg.

[14] Chen, P., Lu, Y., Zheng, V. W., Chen, X., & Li, X. (2018, June). An automatic knowledge graph construction system for K-12 education. In Proceedings of the Fifth Annual ACM Conference on Learning at Scale (pp. 1-4).

[15] 蒋琪. (2020). 基于 MOOC 的交互式领域知识地图半自动化构建研究 (Master). Beijing Normal University.

2014）[16][17][18][19]能够在一定程度上表征知识单元，也有一部分教育领域知识工程的研究者构建了属于教育领域的知识组织模型，如知识森林（Knowledge Forest）的构建就是其中的一项研究。知识森林由方面树（Facet trees）和路径组成。知识森林使用了资源描述框架来表示和储存知识。此外，该项工作的研究者还提出了一种自动构建知识森林的流程，包括方面树构建、路径抽取、知识片段组装等（Zheng et al., 2019）[20]。

2.2.3 不同关系连接的知识图谱

在教育领域，许多带有教育目的的知识实体关系比较受关注，主要是这四种：包含关系（inclusion relation）、递进关系（causal relation）、因果关系（progressive relation）、先决关系（prerequisite relation）。其中先决关系和包含关系是大多数研究者比较关注的。

(1) 先决关系

在以上这些关系中，先决关系是最隐含的关系，因此相对难以识别。而且先决关系符合知识空间理论（knowledge space theory）（Doignon & Falmagne, 1985）[21]，该理论认为人们认知的过程中在概念之间是存在着自然依存关系的，从概念 A 到概念 B 的先决关系也就意味着学习者在掌握概念 B 之前需要掌握概念 A，也就是概念 B 先决于概念 A。

有研究者通过 MOOC 平台、学习测试平台中的学生测试数据（基于概念的前后测数据）并利用概率关联规则挖掘 p-Apriori 算法（Sun et al., 2010）[22]自动识别概念之间的先决关系

---

[16] Hodge, G. (2000). Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files. Digital Library Federation, Council on Library and Information Resources, 1755 Massachusetts Ave., NW, Suite 500, Washington, DC 20036.

[17] Bollacker, K., Evans, C., Paritosh, P., Sturge, T., & Taylor, J. (2008, June). Freebase: a collaboratively created graph database for structuring human knowledge. In Proceedings of the 2008 ACM SIGMOD international conference on Management of data (pp. 1247-1250).

[18] Wu, W., Li, H., Wang, H., & Zhu, K. Q. (2012, May). Probase: A probabilistic taxonomy for text understanding. In Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data (pp. 481-492).

[19] Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., ... & Zhang, W. (2014, August). Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 601-610).

[20] Zheng, Q., Liu, J., Zeng, H., Guo, Z., Wu, B., & Wei, B. (2019). Knowledge forest: A novel model to organize knowledge fragments. arXiv preprint arXiv:1912.06825.

[21] Doignon, J. P., & Falmagne, J. C. (1985). Spaces for the assessment of knowledge. International journal of man-machine studies, 23(2), 175-196.

[22] Sun, L., Cheng, R., Cheung, D. W., & Cheng, J. (2010, July). Mining uncertain data with probabilistic guarantees. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 273-282).

（Chen et al., 2018）[23]。

(2) 包含关系

包含关系是指一个概念从属于另一个概念，这通常被教育者用来构建知识的层次结构。与先决关系相比，它更容易识别，因为这些信息通常保存在教科书或者教程的原始 s 结构中。

有研究者已经做了相关的工作，例如梁等人从大学课程的依赖中提取先决关系（Liang et al., 2017）[24]，卡耐基梅隆大学的研究者利用观察到课程之间的关系创建了一个有向概念图（Liu et al., 2016）[25]，王等人从教科书中提取概念层次结构（Wang et al., 2015）[26]，Chaplot 和 Koedinger 尝试在一门课程中提取多个实体之间的结构（Chaplot et al., 2016）[27]，KnowEdu 采用了基于相似性的方法从教科书的目录中提取了层次结构以及概念之间的包含关系。

2.2.4 特定教育目的的知识图谱相关研究

有一些知识图谱是为特定教育目的而构建的，具有比较特殊的知识表征模型。这里简要介绍三个例子：《三国演义》图谱 CKGHV（Zhu et al.,2016）[28]，教育舆情图谱 EduVis（Sun et al., 2016）[29]，高中数学题解析系统 MathGraph（Zhao et al., 2019）[30]。

朱等人在 2016 年开发了《三国演义》的知识图谱 CKGHV。研究者利用传统的可视化方法呈现人物关系和事件，比如不同的颜色来代表不同的角色类型（敌人或者兄弟）；使

---

[23] Chen, P., Lu, Y., Zheng, V. W., Chen, X., & Yang, B. (2018). KnowEdu: a system to construct knowledge graph for education. Ieee Access, 6, 31553-31563.

[24] Liang, C., Ye, J., Wu, Z., Pursel, B., & Giles, C. L. (2017, February). Recovering Concept Prerequisite Relations from University Course Dependencies. In AAAI (pp. 4786-4791).

[25] Liu, H., Ma, W., Yang, Y., & Carbonell, J. (2016). Learning concept graphs from online educational data. Journal of Artificial Intelligence Research, 55, 1059-1090.

[26] Wang, S., Liang, C., Wu, Z., Williams, K., Pursel, B., Brautigam, B., ... & Giles, C. L. (2015, September). Concept hierarchy extraction from textbooks. In Proceedings of the 2015 ACM Symposium on Document Engineering (pp. 147-156).

[27] Chaplot, D. S., Yang, Y., Carbonell, J., & Koedinger, K. R. (2016). Data-Driven Automated Induction of Prerequisite Structure Graphs. International Educational Data Mining Society.

[28] Zhu, Y., Cao, X., Bian, Y., & Wu, J. (2014, September). CKGHV: a comprehensive knowledge graph for history visualization. In IEEE/ACM Joint Conference on Digital Libraries (pp. 437-438). IEEE.

[29] Sun, K., Liu, Y., Guo, Z., & Wang, C. (2016). Visualization for Knowledge Graph Based on Education Data. International Journal of Software and Informatics, 10(3).

[30] Zhao, T., Huang, Y., Yang, S., Luo, Y., Feng, J., Wang, Y., ... & Zhu, F. (2019, April). Mathgraph: A knowledge graph for automatically solving mathematical exercises. In International Conference on Database Systems for Advanced Applications (pp. 760-776). Springer, Cham.

用线条的粗细来表示关系的强度。通过故事人物之间的可视化技术，用户可以清晰、直接地分析历史事件的动态过程（Zhu et al., 2014）[31]。

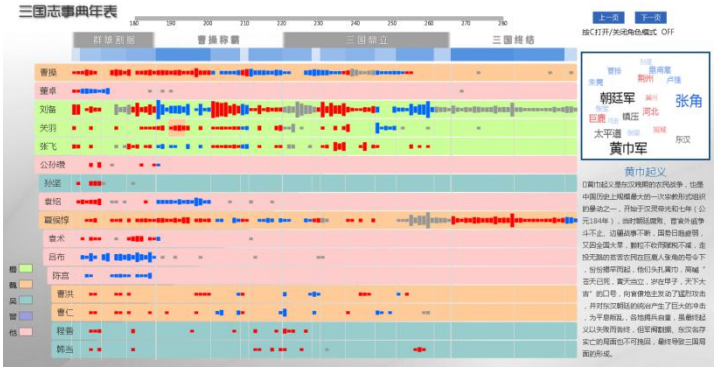


图 2 CKGHV 中《三国》故事概览（来源：Zhu, Y. et al., 2016）

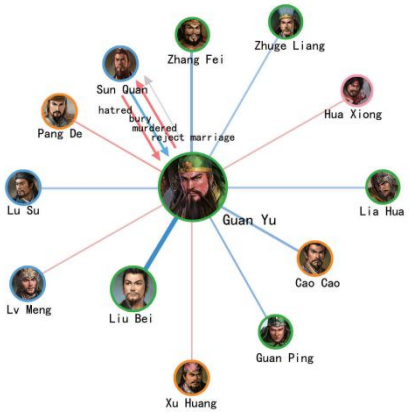


图 3 CKGHV 中《三国》人物关系（来源：Zhu et al., 2016）

华东师大的研究者在 2016 年开发了一个对教育舆情事件的可视化知识图谱 EduVis（Sun et al., 2016）[32]，基于网络上收集海量的教育舆情数据，自主设计并构建了一个可视化面板，包括一个词云树图、一个基于拓扑结构和时间线的教育事件关系网络和一个帮助用户回溯的跟踪路径，并证明能够较好地帮助教育管理者完舆情分析任务。

---

[31] Zhu, Y., Cao, X., Bian, Y., & Wu, J. (2014, September). CKGHV: a comprehensive knowledge graph for history visualization. In IEEE/ACM Joint Conference on Digital Libraries (pp. 437-438). IEEE.

[32] Sun, K., Liu, Y., Guo, Z., & Wang, C. (2016). Visualization for Knowledge Graph Based on Education Data. International Journal of Software and Informatics, 10(3).

图 4 EduVis 系统界面（来源：Sun, K. et al., 2016）

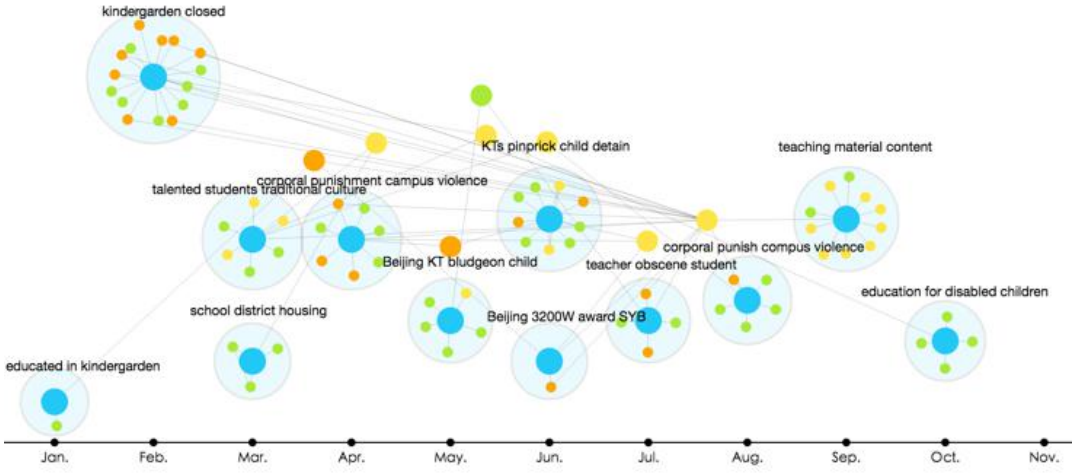


图 5 EduVis 中可根据时间线线性排列的教育事件视图（来源：Sun, K. et al., 2016）

清华大学的研究者构建了高中数学题的有向知识图谱 MathGraph，并开发了一个基于 MathGraph 的自动解析高中数学题的系统（Zhao et al., 2019）[33]。他们认为在特定领域构建教育知识图谱（比如数学领域）必须经过专门的设计，并与其他知识图谱区分开，才能满足深度应用的需求。他们为 MathGraph 设定了强约束的节点类型和边类。

[33] Zhao, T., Huang, Y., Yang, S., Luo, Y., Feng, J., Wang, Y., ... & Zhu, F. (2019, April). Mathgraph: A knowledge graph for automatically solving mathematical exercises. In International Conference on Database Systems for Advanced Applications (pp. 760-776). Springer, Cham.

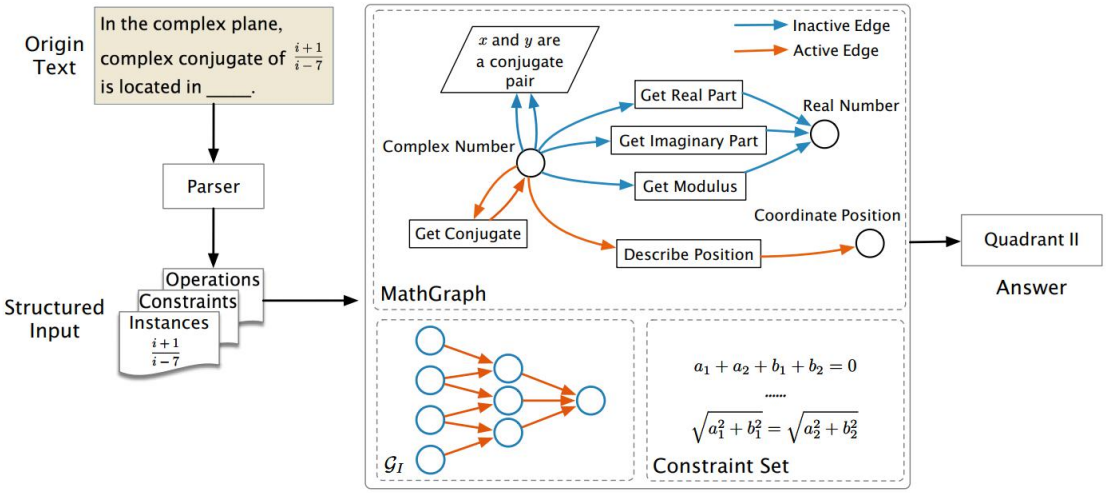


图 6 使用 MathGraph 系统解析一道高中数学题的流程概览（来源：Zhao, T. et al., 2019）

MathGraph 的原理大致描述如下：

（1）节点有三类：对象节点（比如复数）、操作节点（比如加法）、约束节点（包括四种约束条件：描述性约束（如复数 x 和 y 共轭）、等式约束（如 a+2=b）、不等式约束（如 a^2<10）、集合约束（如 a∈N））；

（2）边有两种：衍生边（the Derive edge）（表示从一般对象到特殊对象的关系，如三角形指向等腰三角形）和流边（the Flow edge）（表示操作或者约束关系，比如从“实数”指向“加法”一共会有两条流边，代表两个实数相加，而结果会从“加法”指向“实数”，代表结果返回一个实数）。

接着他们使用基于规则的语义解析器（Semantic parser），将习题文本映射到 MathGraph 相应的节点来解析每一个句子，解析内容包括实例、操作、约束。

2.3 基于教学视频转录文本的知识图谱相关研究

过往研究已经在非结构化文本数据的知识图谱构建方面做了一些工作，而其中基于教育领域的文本数据构建知识图谱最相关的工作是由印度计算机领域的研究者在 2020 年发表的（Shanmukhaa et al., 2020）[34]，他们基于英文的在线学习视频构建了对应视频内容的知识图谱。他们从视频中提取音频转录成文本，接着将文本输入到命名实体识别、共指消解、三元组提取的环节中，最后储存到 Neo4j 图数据库中。

印度研究者所做的研究是第一次对提取在线学习视频文本信息的知识图谱所做的努力，同时他们指出了知识图谱提取的准确性取决于从视频中提取的文本记录的准确性和质

[34] Shanmukhaa, G. S., Nandita, S. K., & Kiran, M. V. K. (2020, March). Construction of Knowledge Graphs for video lectures. In 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS) (pp. 127-131). IEEE.

量，为后续该方向的研究提供了较好的参考意义。

该项研究也有一定的局限性。一方面，该研究在进行音频转录为文字时，仅做了去除无关词和无意义语气词的数据处理，没有考虑由于语音识别技术水平有限带来的转录文本错误会较大地影响知识图谱质量的问题。另一方面，由于该项研究中采用的 NER 模型属于通用提取规则，如提取名词短语（Noun phrase），这样提取出来的实体不完全适用于教育领域，在研究报告的结果中提取出了“when we”“before you”等不具有实际学习意义的词汇（如下图）。

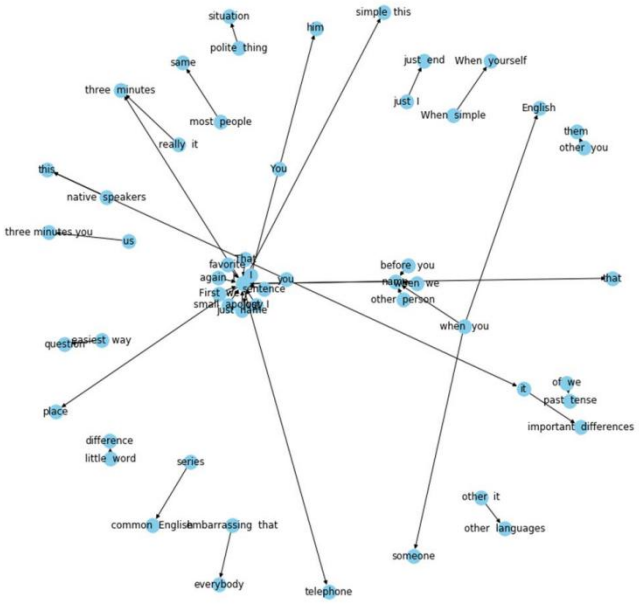


图 7 印度研究者构建的在线学术视频知识图谱

（来源：Construction of Knowledge Graphs for video lectures）

3 相关技术基础

本章主要介绍了半自动化构建知识图谱的相关技术基础，为后续研究设计、实现及检验提供技术基础。一方面阐述了知识图谱构建技术的理论和方法，尤其是针对文本数据的知识图谱构建技术；另一方面阐述了本研究采用的其他算法，包括：基于互信息和左右熵的新词挖掘算法、基于 BiLSTM 和 CRF 的实体识别算法、实体识别效果检验指标和基于滑动区间和词向量的小节分割算法。

3.1 知识图谱构建技术

3.1.1 知识图谱概述

3.1.1.1 知识图谱的定义

知识图谱的定义可以在各种综述文献中找到（Rizun, 2019; Ehrlinger & Wöß, 2019）[35][36]。整体来说，知识图谱是一个结构化的储存实体以及实体之间存在关系的网络（Paulheim, 2017）[37]，它能够结构化地描述现实世界中的信息及其关系。

根据 2020 年 4 月斯坦福大学知识图谱课程（Stanford, 2020）[38]介绍，知识图谱主要有三种具体定义方式：

(1) 数学形式：某一实体集合 E 和某一关系集合 R，知识图谱是一个包含了若干三元组的有向关系图。

(2) 本体和实例：针对不同的主题领域，定义一个本体，其外延涵盖的实例之间相互关联的图即为知识图谱。

(3) RDF 和 LPG 都是知识图谱具体的储存格式。RDF 和 LPG 是过往主流的三元组数据格式，而近年来 Neo4j 提出了为 LPG 创建图数据库文件格式并设定标准查询语言。所以目前 Neo4j 也是一种主流的储存知识图谱的图数据库文件，而且由于其可

---

[35] Rizun, M. (2019). Knowledge graph application in education: a literature review. Acta Universitatis Lodziensis. Folia Oeconomica, 3(342), 7-19.

[36] Ehrlinger, L., & Wöß, W. (2016). Towards a Definition of Knowledge Graphs. SEMANTiCS (Posters, Demos, SuCCESS), 48, 1-4.

[37] Paulheim, H. (2017). Knowledge graph refinement: A survey of approaches and evaluation methods. Semantic web, 8(3), 489-508.

[38] CS 520: Knowledge Graphs. (2020). Retrieved 10 December 2020, from https://web.stanford.edu/class/cs520/

以直接可视化呈现，得到了研究者的广泛使用。

3.1.1.2 知识图谱的分类

从涵盖内容方面来说，知识图谱可分为通用知识图谱（Generic knowledge graphs）和领域知识图谱（Domain knowledge Graph）（Qingjie et al., 2014）[39]，两者区别具体如表 1 所示。

表 1 通用知识图谱和垂直领域知识图谱的比较（肖仰华，2020）[40]

比较层面	比较维度	领域知识图谱	通用知识图谱
知识表示	广度	窄	宽
	深度	深	浅
	粒度	细	深
知识获取	质量要求	苛刻	高
	专家参与	程度高	程度低
	自动化程度	低	高
知识应用	推理链条	长	短
	应用复杂性	复杂	简单

通用知识图谱规模庞大，主要是工业界的互联网企业才有能力进行构建和储存。例如谷歌的 Knowledge Vault（Dong et al., 2014）[41]和微软的 Microsoft's Probase（Wu et al., 2012）[42]中涵盖的实体数量都达到了亿级，其中的实体关系也达到了百亿级。而且随着信息不断增多，知识图谱整体规模也不断在增大。构建这些图谱的方法有很多，主要是基于结构化

---

[39] Qingjie, L., Lingyu, X., Jie, Y., Lei, W., Yunlan, X., Suixiang, S., & Yang, L. (2014, September). Research on domain knowledge graph based on the large scale online knowledge fragment. In 2014 IEEE Workshop on Advanced Research and Technology in Industry Applications (WARTIA) (pp. 312-315). IEEE.

[40] 肖仰华. (2020). 领域知识图谱落地实践中的问题与对策_实体. Retrieved 10 December 2020, from https://www.sohu.com/a/280006592_100099320

[41] Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., ... & Zhang, W. (2014, August). Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 601-610).

[42] Wu, W., Li, H., Wang, H., & Zhu, K. Q. (2012, May). Probase: A probabilistic taxonomy for text understanding.

的外部词典如 Wikipedia 或者百科全书等，采用了人工、社区众包的方式构建的。而通用知识图谱能够提供的应用主要也是通用服务，例如 Google 的 Knowledge Vault 图谱支持了 Google Search 搜索引擎服务，Wolfram Alpha 图谱支持了 Apple Siri 语音助手服务。

但通用知识图谱通常不能很好地支持许多特定领域的应用，因为这些应用需要深入的领域信息和专业知识，因此领域知识图谱应需求而生。目前大型的领域知识图谱主要有医学领域知识图谱、地理领域知识图谱例如 Geonames（Frontini et al., 2013）[43]，而为了满足领域中的特定应用，知识图谱还进一步细化，如医学领域知识图谱目前可以细化为生物医学知识图谱、中医学科领域知识图谱、中文疾病知识图谱等（刘烨宸 & 李华昱，2020）[44]。本研究构建的知识图谱也属于教育领域中的知识图谱。

3.1.2 知识图谱的构建方法

依据数据的特点以及应用的需求，知识图谱的构建方式主要有自上向下（Top-Down）、自下向上（Bottom-Up）以及二者结合的方法（王昊奋，2019）[45]。

自上向下的构建方法适用于处理数据具有已知的知识体系或者可由该领域专家梳理之后定义出本体和模式信息，流程为：首先为知识图谱定义数据模式层，即构建知识本体，其次逐步细化本体中各个实体的定义，最后在命名实体识别的环节中将信息按照实体定义的实例存入知识图谱中。

在各企业和机构刚开始构建知识图谱初期，由于技术受限，主要都通过定义知识本体的方式构建知识图谱（刘峤等，2016）[46]，如许多知识图谱项目都使用了 wikipedia 中已经结构化的数据作为数据源进一步构建知识图谱。

自下而上的构建方法适用于缺少知识体系的数据或者是缺乏规则和范式的信息，主要是一个迭代更新的过程。这种构建方法通过自动抓取实体从而构建出模式，可以依赖开放数据集和百科类资源库。主要构建流程为信息抽取、外部词典引入、知识融合和知识模式抽取等

随着自动化的知识挖掘技术的发展，目前也有许多知识图谱采用自下而上的方式构建，包括 Knowledge Vault 和 Microsoft' s Satori 都通过自动化识别实体和实体关系，进而自动化更新并完善知识图谱（刘峤等，2016）[47]。而自下而上的构建方法主要依靠迭代来完成的，在程序自动化完成抽取和更新之后，大多数需要人工介入以判断知识抽取以及关系

---

[43] Frontini, F., Del Gratta, R., & Monachini, M. (2013, December). GeoDomainWordNet: Linking the geonames ontology to WordNet. In Language and Technology Conference (pp. 229-242). Springer, Cham.

[44] 刘烨宸, & 李华昱. (2020). 领域知识图谱研究综述. 计算机系统应用, 29(6), 1-12.

[45] 王昊奋. (2019). 知识图谱 方法、实践与应用. 北京: 电子工业出版社.

[46] 刘峤, 李杨, 段宏, 刘瑶, & 秦志光. (2016). 知识图谱构建技术综述. 计算机研究与发展, 53(3), 582.

[47] 刘峤, 李杨, 段宏, 刘瑶, & 秦志光. (2016). 知识图谱构建技术综述. 计算机研究与发展, 53(3), 582.

抽取的质量，进而迭代知识图谱。

3.1.3 知识图谱的构建环节

本节将介绍知识图谱构建中的基本环节。主要分为四个基本板块：知识建模、实体抽取（Entity extraction）、关系抽取（Relation extraction）、知识图谱存储和管理。

3.1.3.1 知识建模

知识建模并不是每个知识图谱所必须的。在采用自上而下的方法构建知识图谱的流程中，知识图谱的架构包括了知识的模式层（Knowledge Schema）与数据层。而采用自下而上的方法构建的知识图谱中不存在模式层。模式层即知识图谱的知识本体（Ontology），也指目标数据中存在的知识结构。而数据层是根据模式层的规则和定义抓取得到的实例数据。

3.1.3.2 实体抽取

实体抽取是指通过自动化、半自动化、人工等方式从结构化、半结构化和非结构化数据中抽取信息的环节。这个环节的关键问题是：如何从数据中获得知识实体？

对于不同类型的数据源有不同的知识抽取方法。如针对垂直网页类的结构化数据，可以采取基于模板的方法（Pattern-based method）抽取知识实体，针对百科全书知识库这样的半结构化数据源，可以采用网页处理（Web page processing）的方法。而针对文本这类非结构化数据则需要采用命名实体识别（Named Entity Recognition, NER）的方法（Yan et al., 2018；王昊奋，2019）[48][49]。不同数据源及对应的实体抽取方法如表 2 所示。

表 2 不同数据源及对应的实体抽取方法

数据源	垂直网站	百科全书知识库	文本
数据源类型	结构化数据	半结构化数据	非结构化数据
方法	基于模板的方法	网页处理	命名实体识别
抽取属性	可抽取	可抽取	不可抽取
人工注释	不需要	不需要	需要
实体覆盖范围	特殊领域	好	取决于文本

[48] Yan, J., Wang, C., Cheng, W., Gao, M., & Zhou, A. (2018). A retrospective of knowledge graphs. Frontiers of Computer Science, 12(1), 55-74.

[49] 王昊奋, 漆桂林, & 陈华钧. (2019). 知识图谱: 方法, 实践与应用. 北京: 电子工业出版社.

实体质量	高质量	高质量	中等质量

由于本研究基于文本进行实体抽取，因此本节重点介绍在文本中抓取实体的技术基础：**命名实体识别**。

命名实体识别在实体抽取中起着重要作用，它能够识别并分类出句子中的信息元素，如位置、人员、组织等这些对构建知识图谱有效的信息。这些信息也被称为命名实体（Named Entity, NE）。

命名实体识别的方法经过一定的发展，逐渐提高了识别的效率和准确率，主要形成了三种方法如下陈述。三种方法并非哪一种是最好的，而是需要研究者在不同的情况下，根据自身情况自行选择更适合的方法。

(1) 基于规则以及词典

基于规则以及词典的方法是最早被引入到 NER 中的尝试。

在人工智能发展初期，人们编写大量规则，通过将目标文本与规则匹配来识别实体。如在 1991 年 Rau 提出了启发式算法并结合了人工编写规则的方式能够自动化地抽取公司名称实体。但规则的方法具有明显的局限性，首先是规则的构建需要大量的人力、物力和时间，成本高；其次是规则的可扩展性差，同一份规则在不同数据上的效果会大打折扣，因此基于规则的实体抽取方法效果逐渐退出历史舞台。

而后来人们发现基于词典的实体抽取效果比编写规则要好一些，即自定义词典作为参考，程序将特定字符识别成实体，而不对它们进行切分。但是该方法也有较大的局限性，即只能识别词典中存在的实体，不利于挖掘文本中可能出现的新词。

(2) 基于机器学习技术

目前相对主流的实体抽取方法之一就是基于监督学习的机器学习技术。这种方法将实体抽取任务被视为序列标注任务。该算法需要研究者大规模地给语料库中每个文字标注标签（即该文字是否为实体），机器学习目标是在这些已有标注的文本数据上学习出文字标注模型，能够给剩下没有标注的文本按照标签进行分类。这种方法至今仍是主流的 NER 方法之一，并与后续的深度学习技术结合能够在许多 NER 任务中取得良好效果。

对文字进行标注的标签主要有 BIO、BIOES 两种主流的标注体系。其中 B 代表实体的开头字符，I 表示实体内部字符，O 表示非实体字符，E 表示实体字符的结尾，S 表示单个字符即为实体。而 BIO 标注体系由于简单有效被更广泛地采取，在 BIO 体系下，如“小靖在北京师范大学观看了中国女排的比赛”的标注结果可能为“B-PER, I-PER, O, B-ORG, I-ORG, I-ORG, I-ORG, I-ORG, I-ORG, O, O, O, B-ORG, I-ORG, I-ORG, I-ORG, O, O, O”。

尝试解决 NER 任务的机器学习模型有许多。如隐马尔可夫模型（HMM）（周和苏[18]）、条件随机场（Conditional random field, CRF）（芬克尔（Finkel）等人[19]）。其中 CRF 是十分成功的方法之一，被广泛地应用到特定文字识别（Leaman et al., 2015）[50]和中文实体

[50] Leaman, R., Wei, C. H., & Lu, Z. (2015). tmChem: a high performance approach for chemical named entity recognition and

识别任务（Han et al., 2013）[51]中。

然而，该方法面临了一个问题：每个领域建立语料库的构建者主要是大量的领域专家，而标注语料将会花费他们大量时间。因此针对每个领域和实体分别建立单独的语料库作为机器学习的训练集，成本也将是巨大的。那么如何从少量的语料库中自动地学习到能够有力区分实体与非实体的模型成为该领域研究的重要方向。如部分研究者提出了迭代扩展实体语料库的解决方法（Whitelaw et al., 2008）[52]，还有研究者引入了一些复杂的机器学习技术，如刘等人[20]将 K 近邻（K-nearest neighbors, KNN）分类器与 CRF 模型相结合，它能够通过一个半监督的框架解决语料库缺乏和训练数据不可用的问题。同时，迁移学习（transfer learning）也被用来减少不同领域中语料库的标注工作（Pan et al. [21]）。

(3) 基于神经网络的深度学习技术

目前，另外一种十分主流 NER 技术结合了深度学习模型（Lample et al., 2016）[53]。基于神经网络的深度学习模型仍把 NER 任务视作序列标注任务。而该模型的优点是可以通过神经网络（Neural Network）学习到语料库中对模型有用的更复杂的特征，可能有更好的表现。

结合传统机器学习的研究结果，基于深度学习技术的 NER 主要以神经网络为中间层，把 CRF 作为输出层，这样能够通过具有有效约束条件的 CRF 层来修正神经网络输出的结果。目前比较常用的模型结构为：双向长短期记忆网络-条件随机场（BiLSTM-CRF）（Huang et al., 2015）[54]、卷积神经网络-条件随机场（CNN-CRF）（Huang et al., 2015）[55]、循环神经网络-条件随机场（GRU-CRF）（苏丰龙等，2016）[56]等。也出现采用强化学习（Feng et al., 2018）[57]的解决方法。

---

normalization. Journal of cheminformatics, 7(S1), S3.

[51] Han, A. L. F., Wong, D. F., & Chao, L. S. (2013, June). Chinese named entity recognition with conditional random fields in the light of Chinese characteristics. In Intelligent Information Systems Symposium (pp. 57-68). Springer, Berlin, Heidelberg.

[52] Whitelaw, C., Kehlenbeck, A., Petrovic, N., & Ungar, L. (2008, October). Web-scale named entity recognition. In Proceedings of the 17th ACM conference on Information and knowledge management (pp. 123-132).

[53] Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural architectures for named entity recognition. arXiv preprint arXiv:1603.01360.

[54] Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:1508.01991.

[55] Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:1508.01991.

[56] 苏丰龙, 谢庆华, 邱继远, & 岳振军. (2016). 基于深度学习的领域实体属性词聚类抽取研究. 微型机与应用, 35(1), 53-55.

[57] Feng, J., Huang, M., Zhao, L., Yang, Y., & Zhu, X. (2018). Reinforcement learning for relation classification from noisy

3.1.3.3 关系抽取

关系抽取的目的是构建实体之间的语义信息，即能够描述实体之间的事实。而目前主要提取的是二元关系，即两个实体之间的关系，能够形成<实体，关系，实体>三元组。该环节的关键问题是：如何得到三元组？

关系抽取技术的发展也从基于模板的方法开始，逐渐引入了传统机器学习和深度学习的方法。此外，开放关系抽取（open relation extraction, ORE）可以在没有人工监督的情况下识别出三元组，也是一种较为特殊的关系抽取框架。目前主要的关系抽取技术如表 3 所示。

表 3 目前主要的关系抽取技术

方法	基于模板的方法	分类	深度学习	开放关系抽取（ORE）
学习方法	半监督的	有监督的	有监督的	无监督的
特征空间	文本模板	文本特征	高维空间特征	文本特征
人工注释	不需要	需要	需要	不需要
预先定义关系	需要	需要	需要	不需要
准确率	中等	高	高	低
召回率	中等	低	中	高

(1) 基于模板的方法

基于模板的方法是由人工预先构造模板跟语料库中的句子进行匹配抓取实体关系。模板主要有两种，一种是触发词，另一种是语义规则。这里具体说明一下语义规则的方法，需要人工通过分析句子的结构，根据词性预定义句子里关系提取规则。这种方法具有较高的准确率，但其缺点有两方面，一方面是需要人工定义许多精准的符合句子结构的规则，而这需要模板制定者同时具备语言学知识以及相应领域知识，通常为专家；另一方面基于模板的方法可迁移性低，预定义的模板难以处理丰富多样的自然语言和文本。

(2) 基于传统机器学习和深度学习的方法

基于机器学习的方法是将实体关系识别任务当作关系分类任务（Nadeau & Sekine, 2007）[58]进行处理。传统的方法是预定义好实体关系，进一步对语料库进行关系标注，通

---

data. arXiv preprint arXiv:1808.08013.

[58] Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. Lingvisticae Investigationes, 30(1), 3-26.

过分类器（如朴素贝叶斯、最大熵模型（Ratnaparkhi, 1996）[59]、虚拟感知器（Collins & Duffy, 2002）[60]、线性对数（Knoke et al., 1980）[61]等）学习语料库中的文本特征。根据输入数据的性质，有监督的关系抽取方法进一步分为了基于特征的关系抽取（feature based extraction）和基于核函数（kernel based extraction）（刘克彬等，2007）[62]。有监督的关系抽取方法提升了关系抽取的准确性。而有监督的方法也需要大量标注的语料库，成本也相对较高。因此半监督（Carlson et al.,2010; Liwei et al., 2013）[63][64]、无监督（Zhang & Zhou, 2000）[65]的方法逐渐被引入到关系抽取中。

(3) 基于自监督的开放关系抽取的方法

传统的关系抽取系统使用监督或者半监督的方法需要以预先定义的实体关系、种子数据以及训练集作为输入。然而获取这些信息对于实际应用中的大规模、非结构化的数据来说是困难的。因此研究者对于这样的数据正在尝试开放关系抽取的方法。最早由 Banko 提出了一个自监督（self-supervised）的系统 TextRunner，它能够在没有任何人工输入的情况下从语料库中提取大量的关系实例，然后使用一个自监督的学习和一个标记候选关系是否可信的分类器遍历关系实例，使用概率模型计算关系的正确概率。这种方法在开放知识域的关系抽取中取得了显著的效果。

3.1.3.4 知识图谱存储

目前由于图数据库的发展和广泛使用，大多数知识图谱主要储存在图数据库中。不同的图数据库有各自的特点，而且不一定具有统一的图模型。因此针对不同的需求，工业界和学术界使用的数据库稍有不同。本节在这里介绍两种图数据库。

(1) 基于关系型数据库的图数据库

---

[59] Ratnaparkhi, A. (1996). A maximum entropy model for part-of-speech tagging. In Conference on empirical methods in natural language processing.

[60] Collins, M., & Duffy, N. (2002, July). New ranking algorithms for parsing and tagging: Kernels over discrete structures, and the voted perceptron. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (pp. 263-270).

[61] Knoke, D., Burke, P. J., & Burke, P. J. (1980). Log-linear models (Vol. 20). Sage.

[62] 刘克彬, 李芳, 刘磊, & 韩颖. (2007). 基于核函数中文关系自动抽取系统的实现. 计算机研究与发展, 44(8), 1406-1411

[63] Carlson, A., Betteridge, J., Wang, R. C., Hruschka Jr, E. R., & Mitchell, T. M. (2010, February). Coupled semi-supervised learning for information extraction. In Proceedings of the third ACM international conference on Web search and data mining (pp. 101-110).

[64] Liwei, C., Yansong, F., & Dongyan, Z. (2013). Extracting relations from the web via weakly supervised learning. Journal of Computer Research and Development, 50(9), 1825.

[65] Zhang, Y., & Zhou, J. (2000, October). A trainable method for extracting Chinese entity names and their relations. In Second Chinese Language Processing Workshop (pp. 66-72).

传统的关系数据库管理系统（RDBMS）对于处理图形化的任务是非常强大的工具，它在查询、插入、删除和更新数据库信息的速度都非常快。许多图数据库是基于关系型数据库的基础上实现的。如 G-shore、Filament 等。

(2) 基于三元组的图数据库

在互联网中表示信息的统一框架为资源描述框架（resource description framework, RDF）。通过 RDF 储存知识图谱一般基于 W3C 的 RDF 标准以及 SPARQL 标准。在 RDF 描述中，每一个有向边关系将会被表示成<主语，谓词，宾语>。而一个 RDF 图可被看作是一个大型三元组的集合。有许多图数据库是以 RDF 三元组的形式储存数据的，如 AllegroGraph、Neo4j、DEX、Sones、HyperGraphDB 等。

3.2 本研究中采用的其他算法

3.2.1 基于互信息和左右熵的新词挖掘算法

3.2.1.1 为什么采用基于互信息和左右熵的新词挖掘算法？

本研究采取互信息和左右熵的新词挖掘算法来初步挖掘文本记录中的知识实体。由于文本记录中实体的表达方式具有口语化的特征，可能具有较大的不确定性，因此可以把想要挖掘的知识实体视为未登录词（unkown words），于是考虑采用一种基于信息论的新词挖掘算法对文本记录进行初步的命名实体识别。

3.2.1.2 新词挖掘算法的工作原理

基于互信息和左右熵的新词挖掘算法最早是由 Matrix67 提出的（Matrix67, 2012）[66]，这两个概念都出自信息论，互信息可以表征两个中文字之间的内部凝合度，左右熵可以表征词的外部自由度。下面对这两个概念进行具体阐释。

互信息（Mutual Information）表征了获得 Y 信息后对获得另一信息 X 的不确定性的减少，也可以说互信息可以表征两个变量之间的依赖程度，可定义为 X 的信息熵和在 Y 条件下的 X 的信息熵之间的差值。若互信息越大，则说明 X 与 Y 之间的凝固程度越大，X 与 Y 组成一个新词的可能性也越大（Stanford, 2009）[67]。互信息计算公式如式（3-1）所示。

$$H_{MI}(X,Y) = log_2 \frac{P(X,Y)}{P(X)\cdot P(Y)} \qquad \text{式（3-1）}$$

---

[66] 互联网时代的社会语言学：基于 SNS 的文本数据挖掘 | Matrix67: The Aha Moments. (2012). Retrieved 27 April 2021, from http://www.matrix67.com/blog/archives/5044

[67] Mutual information. Retrieved 27 April 2021, from https://nlp.stanford.edu/IR-book/html/htmledition/mutual-information-1.html

左右熵（Information Entropy）是指一个候选词左右两边的信息熵，能够反映候选词是否具有丰富的左右搭配（周奥特，2017）[68]。候选词的左右熵越大，则说明候选词具有越丰富的搭配，作为一个独立词语的概率就更大。左熵计算公式如下：

$$H_L(W) = -\sum_{\forall a \in A} P(aW|W)\, log_2 P(aW|W) \qquad \text{式（3-2）}$$

3.2.1.3 在本研究中如何实现基于互信息和左右熵的新词挖掘算法？

本研究中使用互信息和左右熵的新词挖掘算法包括以下五个步骤：

(1) 首先使用 jieba 包进行中文分词；

(2) 利用互信息和左右熵挖掘新词；

将相邻分词结果组合作为候选词储存片段。

(3) 使用字典搜索树（Trie 树）存储单词和统计词频；

选用 Trie 树储存一定长度的片段，用 n-gram 序列构建 Trie 树。Trie 树中每个节点还储存了从根节点到当前节点所有节点构成的词汇在文本中出现的频率。

(4) 使用外部数据源修正新词结果；

(5) 取概率最高的 N 个候选词作为新词；

在使用该算法挖掘知识实体时，由于人类语言中常存在常用无意义词，如“来说”“的话”，并不符合对知识实体的定义，因此可考虑采用人机协同的方式对挖掘结果进行多轮迭代从而得到更好的实体识别效果。

3.2.2 基于 BiLSTM-CRF 的命名实体识别算法

3.2.2.1 为什么使用 BiLSTM-CRF？

本研究中需要基于已标注的文本数据训练一个能够良好抓取教育领域知识实体的模型（model），后续即可应用该模型自动化地识别其他目标文本中的知识实体。

而对于命名实体识别（Named Entity Recognition, NER）任务，基于神经网络的方法非常流行和普遍。一种名为（Bi-directional Long Short-Term Memory，BiLSTM）叠条件随机场（Conditional Random Field，CRF）的 NER 模型具有良好的任务完成效果，其中 BiLSTM 和 CRF 是模型中不同的两层，以下简称 BiLSTM-CRF 实体识别算法。本节将具体介绍 BiLSTM-CRF 的工作原理及在本研究中如何应用。

3.2.2.2 BiLSTM-CRF 工作原理

在工作开始之前，需要准备一个数据集，其中一列为文本数据，一列为文本标签。其

---

[68] 反作弊基于左右信息熵和互信息的新词挖掘. (2017). Retrieved 27 April 2021, from https://zhuanlan.zhihu.com/p/25499358

中文本标签通常为 BIO 或 BIOES 的标注方式。本研究使用了 BIO 的标注方式，其中 B 标注了一个实体的开头字符，I 标注了一个实体的中间字符以及结尾字符，O 标注其他非实体的字符。如“从大数据到人工智能”对应的标注序列为 OBIIOBIII。

BiLSTM-CRF 的输入是词向量或者字符向量，输出是目标句子中每个字对应的标签序列。其原理如图 8：

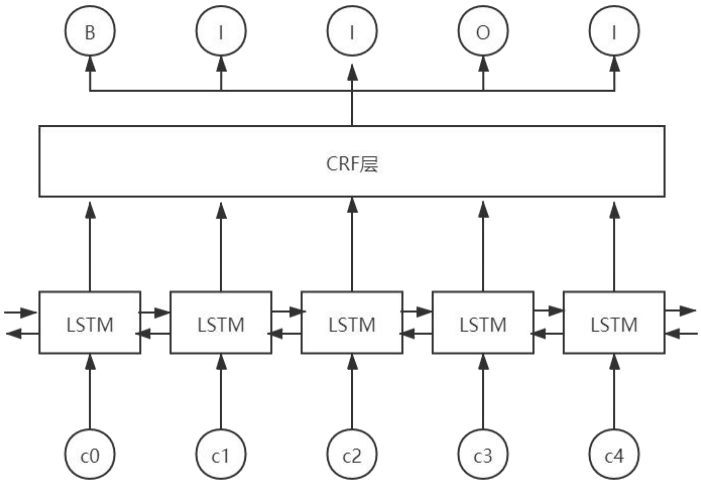


图 8 BiLSTM-CRF 模型示意图

下面将针对 LSTM 单元、LSTM 单元构成的 BiLSTM 层以及 CRF 层进行原理说明。

(1) LSTM 神经单元

LSTM（长短期记忆网络，Long-Short Term Memory）是一种特殊的 RNN 神经单元，将它们连起来能够组成一种特殊的递归神经网络。LSTM 单元解决了 RNN 的梯度问题和远程依赖性的问题，即 LSTM 单元能对有价值的远距离信息进行长期记忆。LSTM 神经单元在整个链上运行，只有很少的线性相互作用，因此保留了长距离的信息流。

具体来说，每个 LSTM 单元的工作原理参见下图，相比于 RNN，LSTM 单元巧妙地增加了一个记忆单元C_t和三个门来维护和调整单元状态，分别为输入门I_t（Input Gate）和输出门O_t（Output Gate），以及特别设置的遗忘门f_t（Forget Gate）；其中I_t和O_t的功能是舍去无效信息，将有效信息传递到下一 LSTM 单元和下一时间（谢腾等，2020）[69]，而 LSTM 单元的输出为h_t。

[69] 谢腾, 杨俊安, & 刘辉. (2020). 基于 BERT-BiLSTM-CRF 模型的中文实体识别. 计算机系统应用, 29(7), 48-55.

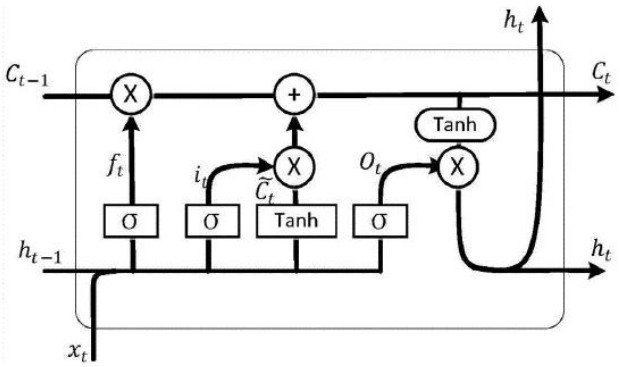


图 9 LSTM 单元工作示意图

(2) BiLSTM 层

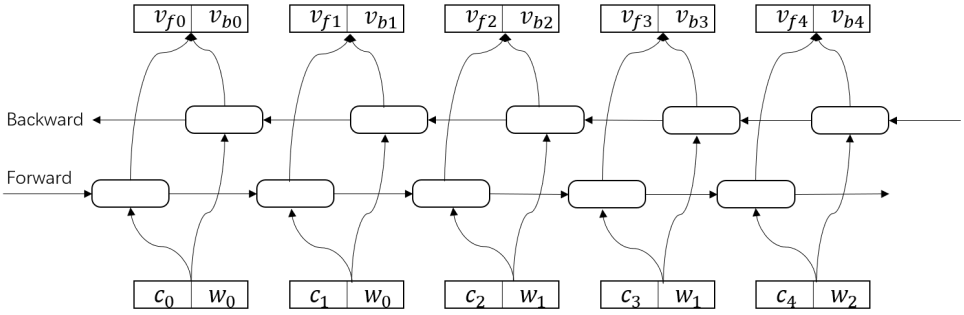


图 10 BiLSTM 层工作示意图

BiLSTM 层的输入为词向量或字符向量，输出是一串该词或字符对应各个标签的分数（Emission score）。如对某个词或字符，BiLSTM 节点的输出是 0.4 (B), 1.9 (I), 0.008 (O)。其实已经可以得到每个词或字符对应的标签，但往往这些分数是有失误的，因此这些分数将会输入到 CRF 层进行矫正。

(3) CRF 层

CRF 是 NER 目前主流的模型，它由 Lafferty[70]等人（2001）首次提出，被 McCallum（McCallum，2003）[71]在 2003 年应用在了 NER 任务中。CRF 能在给定一组输入序列 X 的条件下得到输出序列 Y 的条件概率分布，它的优势在于它可以为一个位置标记的过程中利用丰富的内部以及上下文特征信息，因此在自然语言处理的序列识别任务中得到了最广泛的应用。

在 BiLSTM-CRF 模型中，CRF 层将 BiLSTM 层的 Emission score 作为输入，并输出满

---

[70] Lafferty, J., McCallum, A., & Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

[71] McCallum, A., & Li, W. (2003). Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons.

足该词或字符位置约束的最大可能的预测标签序列。

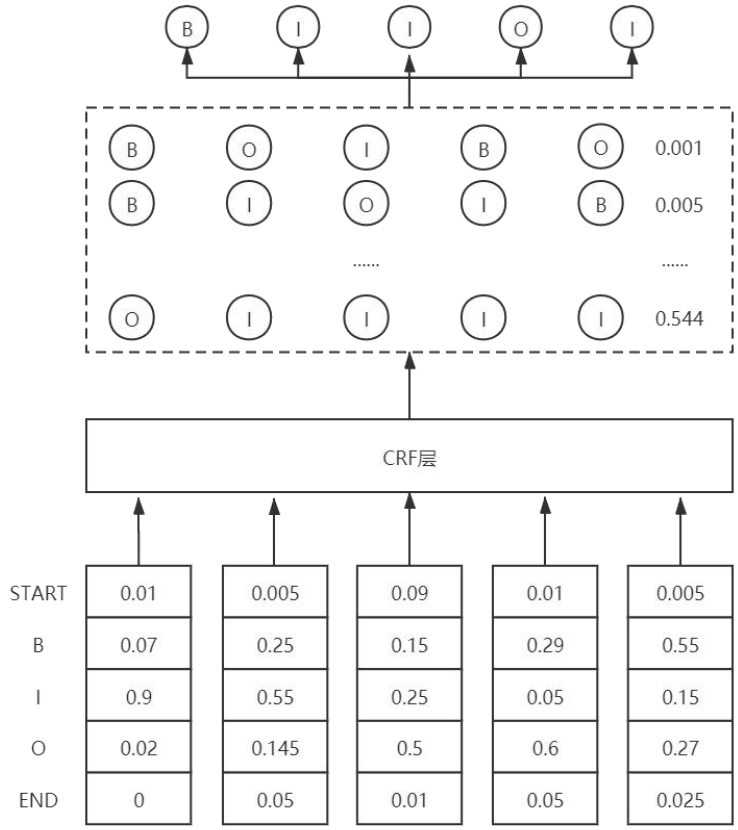


图 11 CRF 层工作示意图

CRF 层向预测标签中增加了一些约束来确保标签预测的有效性，主要是位置约束，如“句子中第一个字符的标签应该是 B 或者 O，而不是 I”，在这些有用的约束下，无效的标签预测将被剔除并得到更有效的标签预测。

3.2.2.3 在本研究中如何实现？

有许多 Github 项目已经实现了训练模型、输出模型以及模型调用的接口，本研究使用了 luopeixiang 贡献的 Github 项目（luopeixiang，2019）[72]进行 BiLSTM-CRF 实体识别模型的训练、保存及调用。

3.2.3 实体识别效果检验指标

这项研究使用三个指标来评估命名实体识别训练模型的效果，分别是准确率（Precision，P），即计算被程序判定为阳例（TP+FP）中有多大比例是真的阳例（TP）；

[72] luopeixiang/named_entity_recognition. (2019). Retrieved 1 May 2021, from https://github.com/luopeixiang/named_entity_recognition

召回率（Recall，R），即计算在测试集中所有阳例（TP+FN）中被预测对了（TP）的比例；F 值（F-score，F），即计算准确率和召回率的调和平均数，用以综合考虑准确率和召回率。其计算公式分别如下：

$$P = \frac{TP}{TP+FP} \qquad \text{式（3-3）}$$

$$R = \frac{TP}{TP+FN} \qquad \text{式（3-4）}$$

$$F = \frac{2*P*R}{P+R} \qquad \text{式（3-5）}$$

若三个指标均超过 0.75，即说明训练模型可接受；若三个指标均超过 0.8 则说明训练效果较好。

3.2.4 基于滑动区间和句向量的小节分割算法

3.2.4.1 为什么需要进行小节分割？

人类在对知识实体的关系判断过程中主要依赖的是信息结构，因此基于文本的组织结构成为自动化对文本中的实体进行关系连接的一种可能的方法。而连续的文本大多数不会具有明确的结构（如在一场演讲文本记录中，演讲者不一定在主题小节之间使用固定格式的连接词如“第一部分”“第二部分”等），因此本研究先采用一种自动化方法对文本记录进行小节分割，为知识实体的连接打下基础。

3.2.4.2 小节分割方法的工作原理

以往该领域的研究者对教学视频结构的自动化提取主要依靠对视频分段（Lecture Segmentation 或 Lecture Fragmentation）的方法。在过去，基于幻灯片（PPT）对提取学术视频片段曾经被证实是有效的，我们可以获得视频中子主题的线索，但在许多情况下，讲座并不一定提供了 PPT 或者没有使用 PPT，可见基于视觉信息的小节分割方法局限性较大。相比之下，演讲人的演讲文本记录（Lecture video transcripts）很容易从语音转文本技术（ASR）中生成，而且演讲记录中包含了整个演讲的关键信息，因此相比于 PPT、手势等，这是演讲者传递出的最详细的信息（Galanopoulos & Mezaris, 2019）[73]。随着技术进步，研究者开始结合演讲的视觉信息、听觉信息和文本特征进行视频分割。最先进的研究使用了监督学习方法，但这需要大量已标记的视频数据集，而且往往十分耗时耗力。

为解决这个问题，研究者 Damianos Galanopoulos 和 Vasileios Mezaris（2019）提供了一

[73] Galanopoulos, D., & Mezaris, V. (2019, January). Temporal lecture video fragmentation using word embeddings. In International Conference on Multimedia Modeling (pp. 254-265). Springer, Cham.

个更具有成本效益的方法（Galanopoulos & Mezaris, 2019）[74]，他们只以演讲文本记录为材料，借助预训练语言模型（如 Bert 一类的语言模型）分析文本中相邻句子相似度的方法，顺利解决了英文在线讲座视频的小节分割问题。这个方法效果取得了比较好的成效，而且已经被应用在了一些视频网站上，如著名的学术在线视频资料库 VideoLectures.NET[75]等。由于本研究问题所属领域同样缺乏大规模数据集，因此本研究参照这种算法来解决文本记录的小节分割问题。

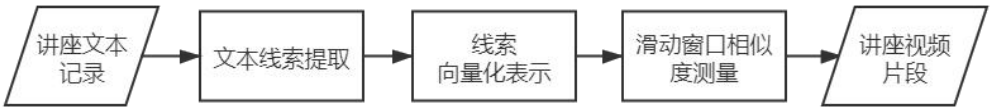


图 12 Damianos Galanopoulos 和 Vasileios Mezaris 的研究框架

（摘自《Temporal Lecture Video Fragmentation using Word Embeddings》）

他们对在线学术视频分割小节的流程如上图所示。

首先利用 ASR 技术从视频中转录出演讲文本。其次，这些文本信息被用来提取有意义的文本线索（Cue extraction），这些线索是原始文本包含的短语或术语。这些线索是原文的特征，它们非常简洁地抓住了文本的本质和意义。他们的研究提供了两种不同的线索提取方法。第一种方法是参考了当时最先进的研究工作，使用名词短语（Noun Phrases）作为线索进行视频分割；第二种是基于自组织视频搜索系统（an ad-hoc video search system）的文本分析和文本分解的组件进行的。第三步，这些线索被向量化，每个文本部分被表示为一个向量。同时检查两种不同的方法向量空间中提取的线索。最后，为了分割每个演讲视频，研究使用了基于滑动窗口的方法来检测可以被分割的时间边界，这些边界定义了最终的时间视频片段集。

3.2.4.3 在本研究中如何实现这个方法？

前人的研究工作对本研究有较大启发和参考价值，本研究主要参考了研究中滑动窗口相似性检测进行视频分段的方法，接下来对本研究中小节分割算法的工作原理进行详细介绍。

滑动窗口相似性测量部分的输入为整体文本数据，输出为文本片段适合分割的时间节点。工作原理如下：首先，以特定步长的窗口在文本上移动，其次，查阅知识实体词典，对每个窗口中捕捉知识实体，并计算窗口中包含知识实体平均的词向量，该词向量可采用

---

[74] Galanopoulos, D., & Mezaris, V. (2019, January). Temporal lecture video fragmentation using word embeddings. In International Conference on Multimedia Modeling (pp. 254-265). Springer, Cham.

[75] new-MOVING-lecture-video-fragmentation-technologies-in-videolectures-net-platform. (2019). Retrieved 9 May 2021, from http://moving-project.eu/2019/01/21/new-moving-lecture-video-fragmentation-technologies-in-videolectures-net-platform/

Word2vec 进行表征；第三，可计算某一点处相邻两个窗口的平均词向量相似度，并映射为函数 y，则 y 代表两个相邻窗口之间的相似度分数。最后，可作出 y 的图像并获得一条曲线（如图），这个图像中最深的波谷被判定为视频分节边界的候选点，山谷的深度是由波谷左右两边最靠近的波峰来确定的，设 val 为局部极小值，即波谷对应的 y 值，$Peak_1$为该点左边最靠近的波峰的峰值，$Peak_2$为该点右边最靠近的波峰的峰值，则该波谷的深度为：

$$Depth = （Peak_1 - val）+ （Peak_2 - val）$$ 式（3-6）

则$Depth$表征了在特定时间邻近窗口中知识实体的语义变化程度，可设定阈值 Thr，判定当$Depth$大于给定的Thr时，即可将该点视为视频小节分割点。

4 研究设计

本研究以教学视频转录的文本记录作为数据源，设计并实现了一套基于自然语言处理技术、采用了深度学习、统计学、规则方法的知识图谱半自动化构建方案。本节介绍了研究目标及构建流程图，进一步详细介绍了每个环节中的研究内容，说明了使用的数据来源和技术框架。

4.1 研究目标及构建思路

本研究的主要目标是，设计并实现一套知识图谱半自动化构建方案。该方案注重成本效益和提取结果的有效性，利用深度学习、统计学和基于规则的自然语言处理技术来挖掘文本记录，从而半自动化构建并可视化地呈现知识图谱。该项研究利用了《教育大数据》课程的文本记录进行示例。

据目前所能查到的中英文文献中，这项研究是第一个为中文课程内容构建知识图谱所做的尝试，也是第一个采用半自动化方法构建中文课程内容知识图谱的研究。此外，这项研究是第二个尝试构建内容知识图谱的研究工作。

该知识图谱构建方案的有三个主要目标：

（1）能够正确反映该文本中涵盖的知识体系，包括课程主题、子主题以及包含知识概念之间的关系。

（2）设计并实现具有成本效益的知识图谱构建方案，降低对人工尤其是专家的依赖，以降低知识图谱构建的人力成本和时间成本，为频繁创建和更新知识图谱创造条件。

（3）能够结构化教学过程中演讲者的文本数据，为更多教育技术应用提供结构化的数据基础。例如教学内容知识图谱能应用于开发能够帮助学生回顾课堂内容、完善知识结构的学习管理系统；应用于教师整理授课知识点并在此基础上改良并更新课程设计，改良教学质量；协助教学管理者进行教学管理、评估一线教学质量、将教学培养知识体系和社会需求进行匹配，改进更加符合社会需求的课程方案。

在研究目标的基础上，结合已有的文献基础，研究思路如图 13 所示。

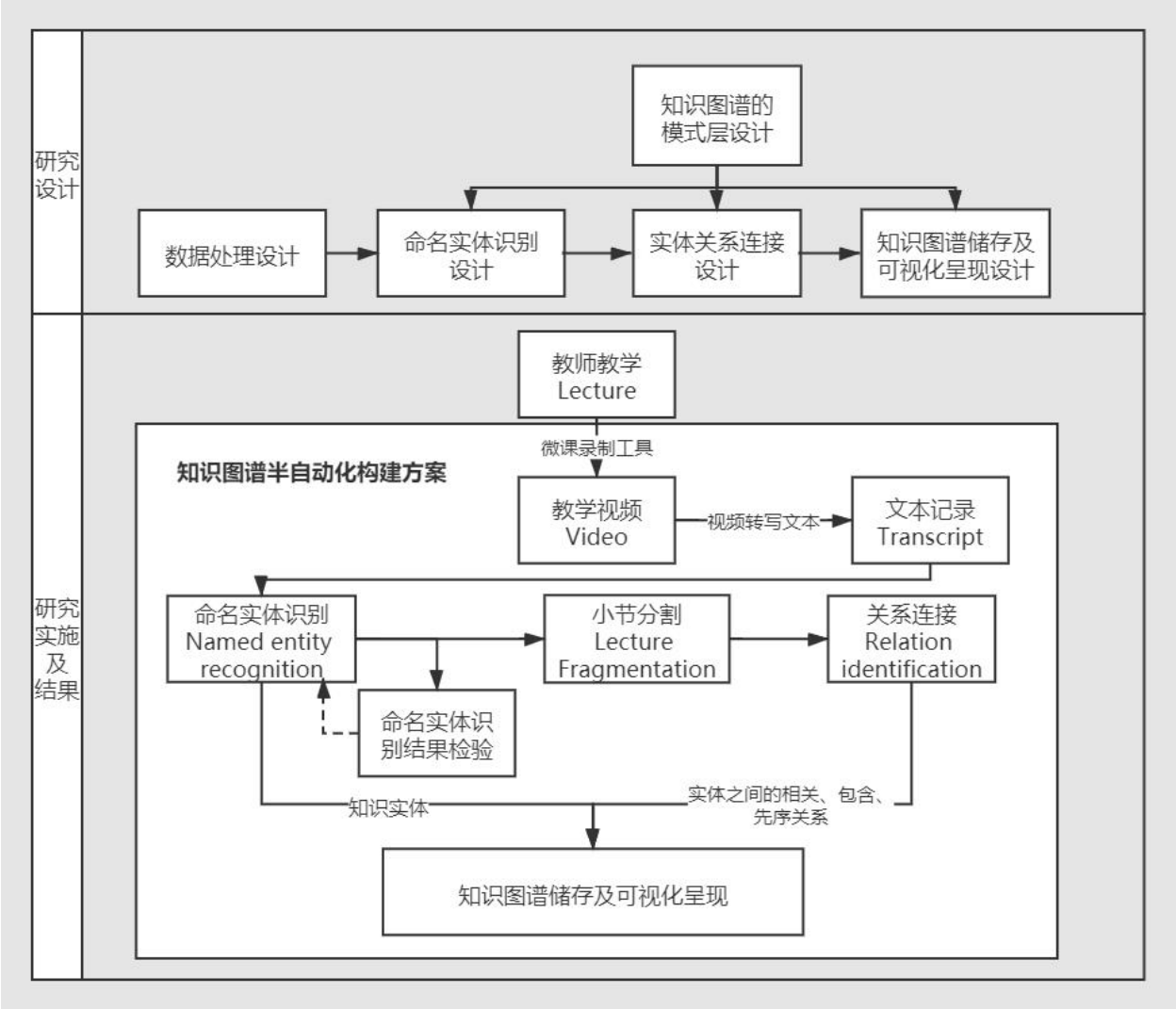


图 13 知识图谱半自动化构建方案设计图

第一步：知识图谱的模式层设计。本研究将采用自上而下和自下而上相结合的知识图谱构建方法，首先设计所构建知识图谱的模式层，目的是能够在一定程度上抽象表征文本知识图谱的实体和实体之间的关系；

第二步：数据处理设计。将教师教学过程用微课录制工具录制后，利用视频转写文本技术获得视频（音频）的文本记录，并且在后续的自动化构建步骤中仅使用文本记录将具有成本效益；

第三步：命名实体识别设计。采用基于信息论及深度学习方法对文本数据进行命名实体识别并检验，得到实体词典，实现能半自动识别文本记录中知识实体的程序；

第四步：实体关系连接设计。分成两个部分：①小节分割。采用基于词向量及统计学方法对演讲文本进行小节分割，得到被分割成多个小节的文本，实现能够将演讲文本小节分割的程序，为关系连接做准备；②关系连接，结合①操作得到的小节分割，采用基于规则的方法对第三步得到的知识实体进行实体关系抽取，得到实体之间的关系；

第五步：知识图谱储存与可视化呈现设计。利用 Python 程序把上述实体词典和实体之

间的关系储存到 Neo4j 图数据库中，并可视化呈现。

4.2 构建方案设计

构建方案主要包括五个模块：知识图谱的模式层设计、数据处理设计、命名实体识别设计、实体关系连接设计、知识图谱储存及可视化呈现设计。

4.2.1 模式层设计

进行知识图谱模式层的设计需要定义知识图谱构建方法、实体、实体间关系，这项研究将知识图谱中的节点称为知识实体，连边称为实体关系，具体设计如下：

(1) 知识图谱构建方法

由于本研究目标是从文本记录中挖掘知识体系，属于没有明确定义结构的信息。因此本研究在定义知识实体的部分采用自下而上的方法，在关系抽取的部分关注知识实体关系的教育学意义，因此采用自上而下的方法。

(2) 知识实体定义

根据构建目标，本研究构建的知识图谱需要尽可能大范围地覆盖文本所涵盖的主题词汇。因此，本研究中定义知识实体为能够包含一定含义的词，且这个词需要能够和该句话的关键信息有联系，其表征为字或词语。并且在挖掘知识实体的过程中采用自下而上的方法，即通过人工标注数据集中的知识实体。

(3) 实体关系定义

本研究构建的知识图谱属于教育领域的知识图谱，因此在定义实体关系时需要考虑实体关系的教育目的性和教育学意义。

在教育领域，带有教育目的的实体关系更多的关注，主要是这四种：包含关系、递进关系、因果关系、先序关系。其中先序关系和包含关系是大多数研究者更为关注的。另外，在构建知识图谱时，可以考虑知识之间的隐性关系，即两个词之间在人类知识库里客观的相关关系。

综上，本研究定义实体关系为包含关系、先序关系以及相关关系这三种带有教育学意义的关系。

4.2.2 数据处理设计

本研究所涉及的数据处理包括四步。

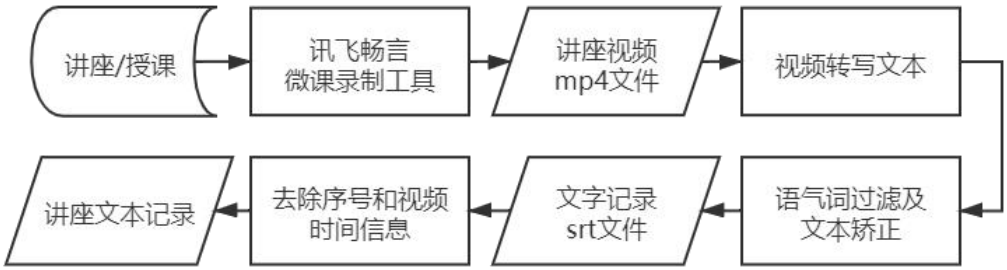


图 14 数据处理流程设计图

第一步，录制，利用讯飞畅言微课录制工具对教学过程进行录制，获得质量较好的课程微课 mp4 文件；

第二步，转录，采用网易见外工作台的视频转写文本技术进行转录，将 mp4 文件上传到见外在线工作台上，工作台自动转录生成质量较好的文本；

第三步，语气词过滤及文本矫正，在工作台上使用常用语气词过滤功能，并矫正转录错误的词语，以保证文本数据的质量，过滤后导出文本记录 srt 文件，过滤的常用语气词 19 个常用语气词；

第四步，清洗，将文本记录用 xlsx 文件格式打开，去除其中的序号和视频时间这些无关信息，得到纯文本的文本记录。

4.2.3 命名实体识别设计

命名实体识别是知识图谱构建中的重要环节，该环节的输入是经过数据处理后的文本记录，输出是对目标数据集中抽取出的知识实体和训练好的模型。

本研究将知识图谱的命名实体称为知识实体，这样的知识实体通常为一个词或者短语来表征所涵盖的内容，例如“大数据”。

本研究中的命名实体识别流程分为三步，如下图所示；

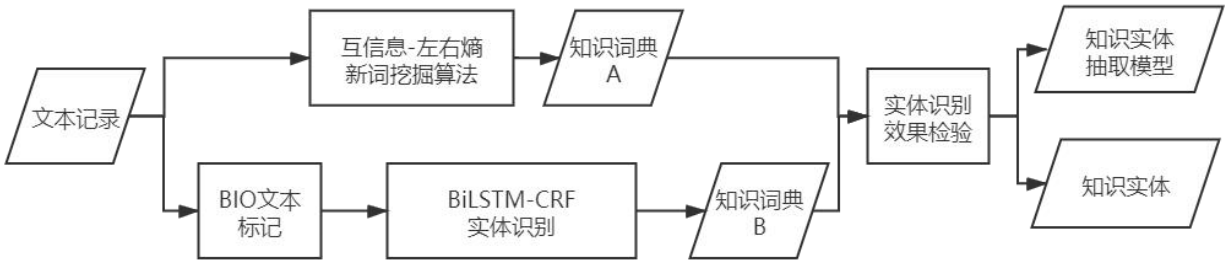


图 15 知识实体识别流程设计图

(1) **第一步：采用互信息-左右熵的新词挖掘算法处理文本记录获得词典 A；**

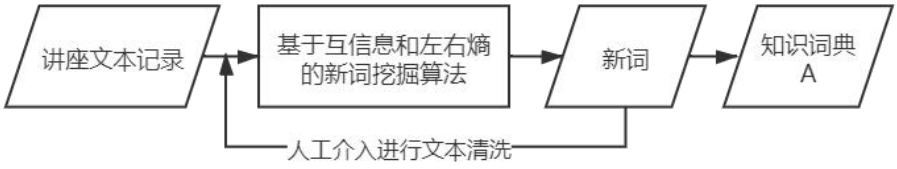


图 16 基于互信息和左右熵的新词挖掘算法实验设计图

这项研究将采用基于信息论的算法，通过人机协同的方式从文本记录中初步得到知识词典 A。

首先利用互信息-左右熵算法挖掘出不超过 300 个新词，接着人工对挖掘出的新词进行筛选，将不属于知识实体的新词放入 stopwords.txt 停用词表中，进行多轮迭代，直到挖掘得到的新词大部分为知识实体可停止迭代，将这些新词构成知识词典 A。

(2) 第二步：基于 BiLSTM-CRF 实体识别算法训练命名实体识别模型并获得知识词典 B

图 17 基于监督学习的命名实体识别实验设计图

这项研究进一步将命名实体识别任务看作序列标记任务，通过有监督的深度学习方法训练模型并得到知识词典。

首先对文本进行 BIO 文本序列标记，这是通过人工对数据标注知识实体，获得训练集和测试集；接着使用 BiLSTM-CRF 实体识别算法训练模型并保存模型，最后将该模型作用在目标数据集上获得知识实体词典 B；

(3) 第三步：实体识别效果检验

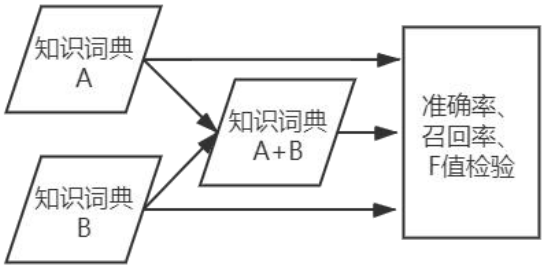


图 18 实体识别效果检验设计图

输入词典 A，词典 B，词典 A+词典 B 共三种词典，计算准确率、召回率、F 值并进行检验。若三个指标都超过 0.75，则表示模型训练可接受，若三个指标均高于 0.8 则说明实

体识别效果较好，选取其中效果最好的词典对应的方法作为抽取知识图谱的方法，并将词典保存下来便于后续构建知识图谱。

4.2.4 实体关系连接设计

在实体关系连接的环节中，输入是经过命名实体识别得到的知识实体，输出是知识实体之间的关系。

本研究关注知识实体之间的包含关系、先序关系和相关关系这三种关系。这项研究先定义三种关系的判断规则，再进行基于规则的关系识别。

由于演讲文本大多数情况下不具有明显的大纲结构，也不会具有如“第一部分”“第二部分”等明显的小节分隔的示意词，因此这项研究利用滑动窗口计算句子之间的相似度并对演讲文本进行小节分割，进而利用小节分割结果来识别知识实体的特定关系。关系连接的具体流程分为四步，如下图所示：

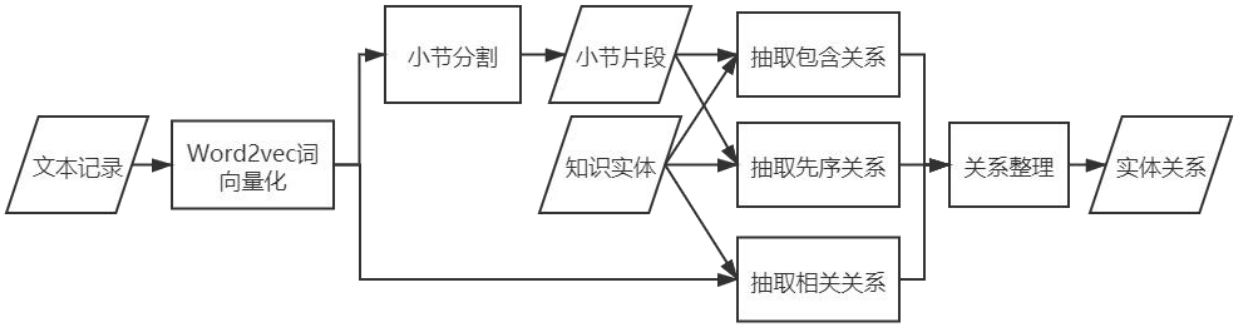


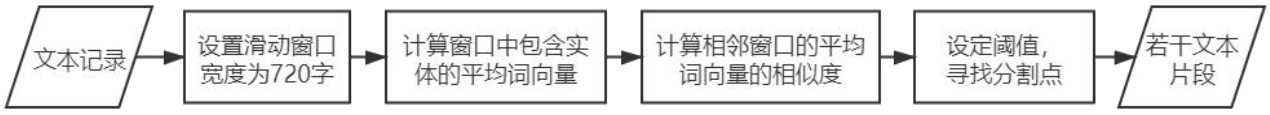
图 19 实体关系连接流程设计图

(1) 第一步：抽取相关关系

利用 Word2vec 词向量化知识实体，得到每个知识实体的词向量，计算此向量与其他知识实体的向量夹角余弦值，并取相似度最高的 3 个知识实体，标注它们之间的相关关系；

(2) 第二步：小节分割

由于连续的演讲文本不利于实体关系的识别，本研究把长演讲文本分割成多个小节片段，每小节是围绕某个主题展开的阐述。具体操作分成 4 步：



① 设置一个滑动窗口，按照一定步长计算滑动窗口的句向量；

② 计算相邻窗口之间的相似度并映射成 y 函数，则 y 函数将形成一条连续曲线；

③ y 函数的极小值（即曲线上的波谷）代表该点处左右两边的文本相似度小，即可认为是可进行小节分割的地方；

④ 设定一个阈值，判断极小值的波谷深度超过阈值就进行小节分割，得到多个小节片段集合；

(3) 第三步：抽取包含关系和先序关系

在本研究中，包含关系定义为：两个实体之间具有层级结构，也通常指一个知识实体涵盖了另一个知识实体。先序关系定义为：在同一层级的知识概念中，先出现的概念先序于后出现的概念。主题词意为某一文段中词频最高的三个知识实体。抽取具体规则如下：

包含关系包括：

① 规则一：同一演讲内，该讲的主题词包含各小节的小节主题词，如第二讲主题词包括第二讲中各小节的小节主题词；

② 规则二：同一小节中，小节主题词包含该小节中所有知识实体，如第 1 小节主题词包括了第一小节中出现的所有知识实体；

先序关系包括：

① 规则一：按照课程的先后顺序，靠前的演讲主题先序于靠后的演讲主题，如第一讲主题词先序于第二讲主题词；

② 规则二：在同一演讲内，前一小节的主题词先序于后面小节的主题词，如第 1 小节主题词先序于第二小节主题词；

(4) 第四步：整理前序抽取出的关系

定义整理规则如下：

① 法则一：如果一个实体 A 与自己本身出现包含、先序、相关关系，则删除这些关系；

② 法则二：如果两个实体 A 与 B 之间属于包含关系或先序关系，且又属于相关关系；则删除 A 与 B 的相关关系。

③ 法则三：如果两个实体 A 与 B 之间存在 A 包含 B 且 B 包含 A 或者 A 先序于 B 或者 B 先序于 A 的关系，则删除 A 与 B 之间的所有包含和先序关系，将 A 与 B 判定为相关关系。

4.2.5 知识图谱储存及可视化呈现设计

构建知识图谱的原材料准备好之后，需要将上述得到的知识实体和实体关系，通过 python 储存到图数据库 Neo4j 文件中，在 Neo4j Community Server 上即可查看可视化的知识图谱。此过程可分为四步，如下图所示：

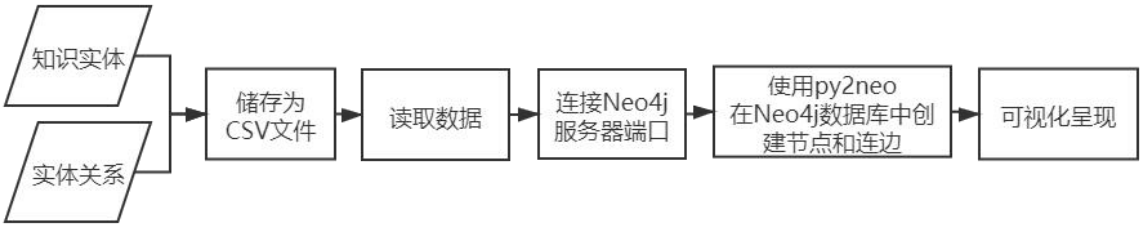


图 20 知识图谱储存及可视化呈现流程设计图

(1) 第一步：读取知识实体和关系；
(2) 第二步：启动 Neo4j 服务器并连接端口；
(3) 第三步：创建节点和连边；
(4) 第四步：访问 Neo4j 图数据库进行可视化呈现；

4.3 数据来源

本研究旨在设计并实现知识图谱半自动化构建方案，对文本的内容主题及类型并无要求，只需文本质量较高且易得即可。因此本研究使用讯飞畅言微课录制工具记录了北京师范大学公共选修课《教育大数据》，共获得三讲教学视频，全长为 8 课时。平均每课时视频长度为 44.01 分钟，共计 352.09 分钟。

4.4 构建技术框架说明

知识图谱半自动构建方案的技术主要发生在四个模块：数据处理、命名实体识别、实体关系连接、知识图谱储存及可视化呈现，展示如图所示：

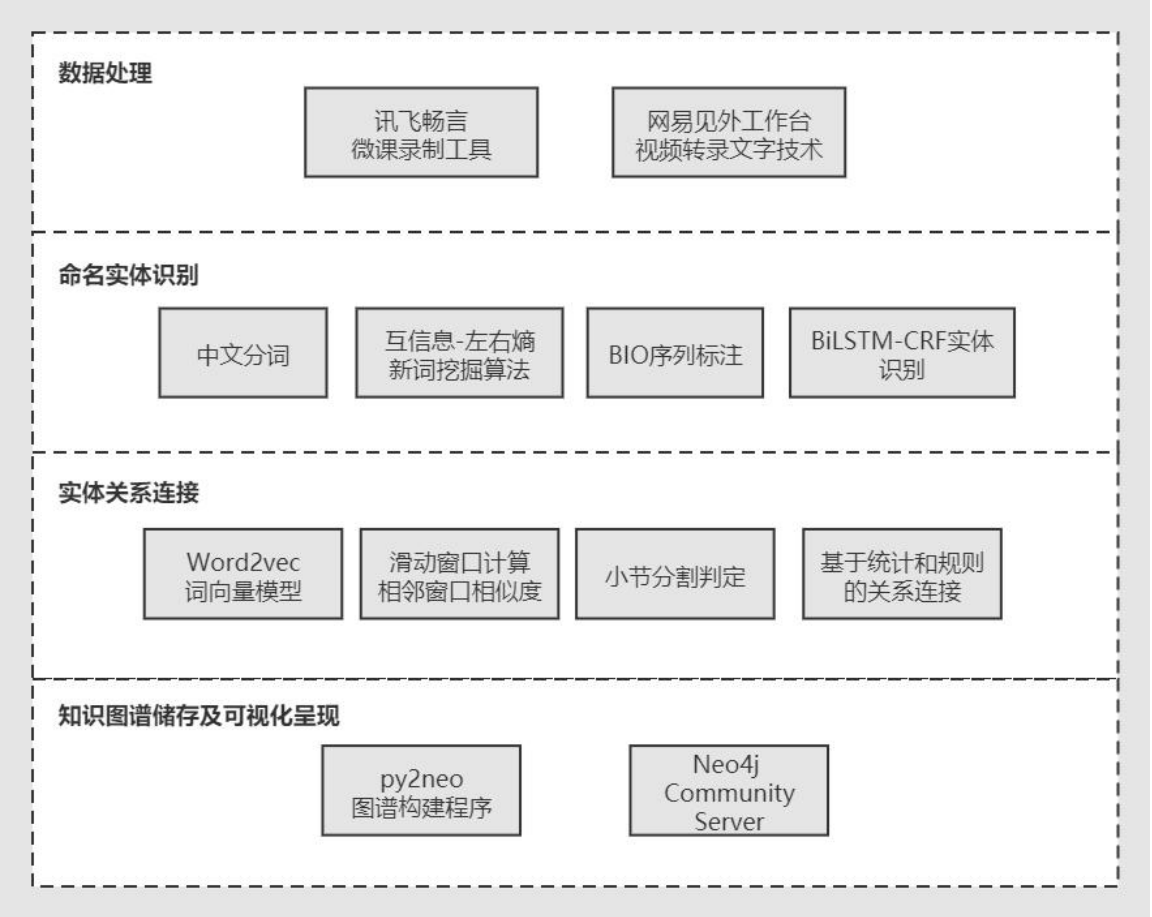


图 21 知识图谱半自动化构建技术框架图

5　知识图谱的半自动化构建实现与检验

根据设计方案，本章以《教育大数据》为数据源，进行了数据处理、命名实体识别及检验、基于统计方法和规则的实体关系抽取以及知识图谱的储存及可视化呈现。

5.1 数据处理

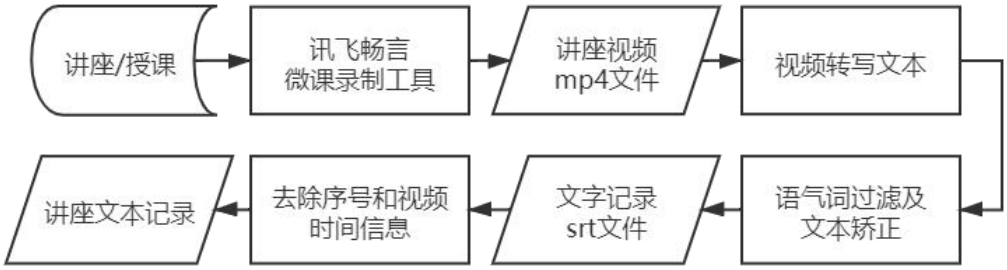


图 22 数据处理流程图

第一步，录制，利用讯飞畅言微课录制工具对教学过程进行录制，获得质量较好的课程微课 mp4 文件。总计获得三讲教学视频 mp4 文件，全长为 8 课时。平均每课时视频长度为 44.01 分钟，共计 352.09 分钟。

第二步，转录，采用网易见外工作台的视频转写文本技术进行转录。将 mp4 文件逐一上传到见外在线工作台，工作台将自动转录视频并在线显示转录文本。

图 23 网易见外工作台工作界面

第三步，语气词过滤及文本修正，在工作台上使用常用语气词过滤功能并修正转录错误的词语，以保证文本数据的质量，过滤后导出文本记录 srt 文件。

第四步，清洗，将文本记录用 xlsx 文件格式打开，去除其中的序号和视频时间这些无关信息，得到纯文本的文本记录。得到数据集共计 79693 字。从三讲课程中随机选择的一讲课程，作为目标数据集，在深度学习中作为测试集，该讲课程共计 16670 字。

5.2 命名实体识别及效果检验

本小节以文本记录作为输入，从文本记录中使用基于左右熵和互信息的新词挖掘算法和基于深度学习的命名实体识别算法提取知识实体形成知识词典，并且对知识实体的提取效果进行检验，采用效果最好的知识实体抽取方法作为知识图谱半自动化构建方案中的一部分。该环节主要分成三个步骤：

(1) 基于互信息和左右熵的新词挖掘算法

(2) 基于深度学习的命名实体识别方法

(3) 实体识别效果检验

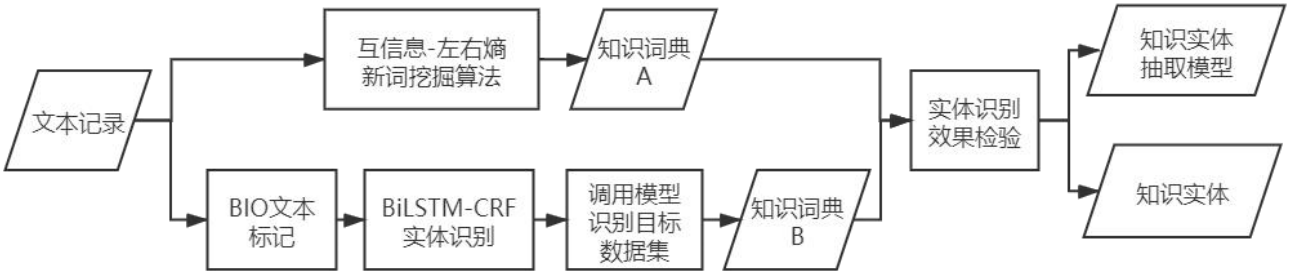


图 24 知识实体识别流程图

5.2.1 基于互信息和左右熵的新词挖掘

这项研究采用基于互信息和左右熵的新词挖掘算法从文本记录中抽取新词，并输出为知识词典 A，在这个过程中需要人工介入以提高新词挖掘效果。

数据集说明：目标数据集共计 16412 字。

实验配置：该算法采用 Python 程序对 Github 项目（zhanzecheng, 2018）[76]进行实现，设置参数如下：每次抽取的新词个数最多为 300 个，设置候选词最大长度不超过 10。

实验流程：

[76] zhanzecheng/Chinese_segment_augment. (2018). Retrieved 27 April 2021, from https://github.com/zhanzecheng/Chinese_segment_augment

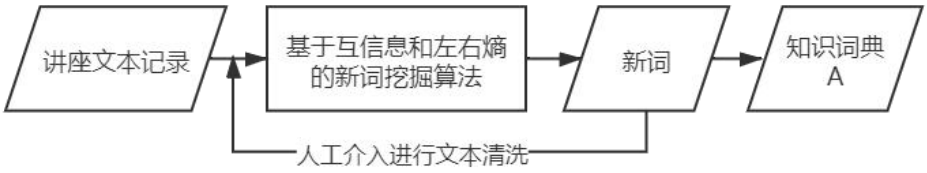


图 25 基于互信息和左右熵的新词挖掘算法实验

在首次实验时，对文本记录采用该算法挖掘得到 300 个新词，此时需要人工介入对抽取得到的新词进行判断。若某词属于应去除的词汇类型（见文本清洗中的表 2 无关词示例），则将其放入 stopwords.txt 文件中。

再次进行新词抽取，算法则会挖掘出不包含停用词而得到 300 个新词。据此进行数轮人机协同的迭代过程后，直到新词中出现的词大部分为具有含义的知识实体，则可将抽出来的词存入知识词典 A。

实验结果：

以下是对目标文本的处理结果，分为文本清洗和新词挖掘两部分。

(1) 文本清洗

在每一轮的新词挖掘结果中，我们能发现当中有一部分词是人们在演讲时频繁使用的与演讲主题无关的“惯用词”，如一段文本“我们能够看到”中的“能够”“看到”都是演讲者习惯添加的词汇，同时它们容易被识别为新词，因此需要在新词挖掘后通过人工介入去除这些词汇。通过对目标文本记录的挖掘和分析，对需要去除的词汇展示如下：

去除词汇类型	示例
无关动词	能够，知道，遇到，学了，可以看一看，我用，我要，我举，我来，提完，看一看哈，可以看得出来，记到小数点，光线调暗，算不动，不太清楚，把认出来，这张图用，涂红，光线调暗，没那么准，没配
连词助词	比方说，等等等等，第四个哈是，从此以后，再也不会，巴拉巴拉，第一个，第四个，不谈
时间名词	前两天，从此以后，上百年上千年，上个世纪
无关名词	含义
错误新词	电子邮箱你们，同学 20

表 4 无关词示例

(2) 新词挖掘结果

通过多轮迭代，最终从目标文本记录中挖掘出的 122 个新词如下：

表 5 利用新词挖掘算法从目标文本中获得的新词

新词列表
人工智能，场景，大数据，90 年代，卷积神经网络，16 位，人工智能应用，感知智能，结构化数据，2000 年，80 年代，运算智能，自然语言处理，无人驾驶，类脑计算，50 年代，数据体量，人机协同，星球大战，感知认知，算法层面，因果律判别，马斯克，认知神经科学，诊断新冠，战胜卡斯帕罗夫，Alpha，70 年代，深度神经网络，无人驾驶模式，数据确权，神经网络，因果律发现，核酸检测，自然语言理解，辩证唯物主义，角度，经典套路，围棋，虚拟现实，VR，德智体美劳，顺丰，圆通，新冠，无症状，模式识别，ibm 深蓝，无人驾驶购物车，学习动机，学习策略，无人区，五育并举，脑力劳动，主观题批阅，脑机接口，广告宣传，购物车小车，google，486cpu，教学法，教学软件，ASCLL，政治课，编程，200 字，世界冠军，终极版，数据，电子邮箱，体量，人脸识别，因果律发现，确权问题，表达方式，顶峰，判别，因果，性别，政治局，潘云鹤，习总，亮度，顶峰，政治课，因果，咳嗽，人力资源强国，ASCLL 码，疫情，健康码，体力劳动，新一代，学习者建模，阿发狗，主观题，作文，收件箱，比尔盖茨，历史，中学语文，算力，准确率，现代化 2035，图灵奖，特斯拉，科幻片，人工，李世石，教学设备，机器，北太平庄，刷题，中叶，区块链，美团，计算机，信息科学，脑科学，肺部造影，神经科学，程序员

5.2.2 基于深度学习的命名实体识别

除了新词挖掘算法，这项研究还采用了基于深度学习技术进行命名实体识别，以实现能够对目标文本记录自动化抽取知识实体的效果。

本研究把命名实体识别任务视为序列标记任务，即通过训练好的自动化程序，实现对目标文本语句标注出对应的 BIO 序列标签，即以 B 标注知识实体的第一个字，I 标注知识实体的除第一字以外的字，O 标注除了知识实体以外的文本。例如“从大数据到人工智能”的标注结果应为“OBIIOBIII”。

数据集说明：对所有数据集文本记录进行人工标记，训练集共 63023 字，测试集共 16670 字，比例约为 8：2，符合深度学习训练比例。

实验配置：本研究利用 Github 上 luopeixiang 的 named_entity_recognition 项目训练 BiLSTM-CRF 命名实体识别模型。

实验流程：

图 26 基于 BiLSTM-CRF 命名实体识别实验流程

首先对文本进行 BIO 文本序列标记，通过人工对数据标注知识实体，获得训练集和测试集；接着使用 BiLSTM-CRF 实体识别算法训练模型并保存模型，最后将该模型作用在目标数据集上获得知识实体词典 B；

(1) BIO 文本标记

通过人工对文本进行 BIO 序列标记。

实验结果：

数据集描述性统计信息如表 6 所示。

表 6 数据集描述性统计信息

数据集	总字数	B	I	O
训练集	63023	3822	8406	50743
测试集	16670	1395	2549	12703

(2) BiLSTM-CRF 实体识别

本研究利用 Github 上 luopeixiang 的 named_entity_recognition 项目训练 BiLSTM-CRF 命名实体识别模型。

神经网络训练参数为：

表 7 神经网络训练主要参数

参数	值
Batch size	64
Epoch	30
隐藏状态维数	128
向量维度	128
学习率	0.001
dropout	0.5

实验结果：

训练完毕的 BiLSTM-CRF 模型在测试集上的效果如表 6，模型的准确率、召回率及 F 值都在 0.75 以上表示训练的模型可接受。

表 8 BiLSTM-CRF 训练模型效果

准确率	召回率	F 值
0.7523	0.7953	0.7506

(3) 调用模型识别目标数据集中的实体，获得知识词典 B

本研究进一步调用上述训练的 BiLSTM-CRF 模型对目标数据集进行处理。

实验结果：

去重后获得 398 个知识实体，表 10 呈现了部分识别结果：

表 10 词向量-BiLSTM-CRF 实体识别结果（部分）

文本	序列标注结果	知识实体识别
让大家呢对大数据从数据采集数据质量分析数据清洗和存储对大家做了比较系统全方位介绍	OOOOOBIIOBIBIBIBIBIBIBIOBIOOOOOOOOBIBIIOO	“大数据”“数据”“采集”“质量”“分析”“清洗”“存储”“系统”“全方位”
尤其涉及到了一些基本机器学习深度学习一些算法	OOOOOOOOOOBIII BIIIOOBI	“机器学习”“深度学习”“算法”
以及基于复杂网络应用	OOOOBIIIBI	“复杂网络”“应用”
90 年代末人工智能曾经很火了	BIIIOBIIIOOOOO	“90 年代”“人工智能”

5.2.3 实体识别效果检验

本研究进一步对词典 A，词典 B 以及词典 A+词典 B 进行准确率、召回率、F 值的检测。输入词典 A，词典 B，词典 A+词典 B 共三种词典，计算准确率、召回率、F 值并进行检验。选取其中效果最好的词典及对应的方法作为抽取知识图谱的方法。

数据集说明：词典 A 共计 122 个词，词典 B 去重后共计 398 个词。

实验配置：使用 python 对词典进行合并。

实验流程：

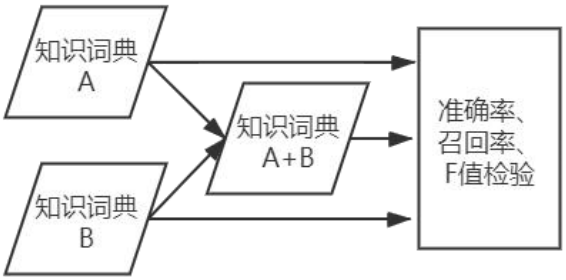


图 27 实体识别效果检验流程

首先对词典 A 进行检验。由于新词挖掘算法是人机协同的方式进行的，无法得到也没有必要从数据集中获得所有可能得到的新词，故无法获得所有的假阴例和真阴例，因此无法计算该方法的召回率和 F 值。

第二步对词典 B 进行检验。由于目标数据集与模型训练的测试集相同，因此实体识别效果可使用模型在测试集上的效果进行表征。

第三步对词典 A 和词典 B 进行合并去重，并检验其准确率。

实验结果：

(1) 词典 A

在新词挖掘环节中获得词典 A 共计 122 个新词，其中有 65 个为目标数据集中真正的知识实体，57 个新词非知识实体。故词典 A 的准确率为挖掘得到的新词中真正的知识实体所占比例，即 0.5328。

(2) 词典 B

由于目标数据集与模型训练的测试集相同，因此实体识别效果可使用模型在测试集上的效果进行表征。

(3) 词典 A+词典 B

词典 B 为实体识别模型作用在目标数据集上得到的知识实体词典，与词典 A 合并后进行检验。

得到三种词典的准确率、召回率和 F 值如表 11 所示。

表 11 实体识别效果检测结果

词典序号	准确率	召回率	F 值
词典 A	0.5328	-	-
词典 B	0.7523	0.7953	0.7506

词典 B 的有监督的学习算法是可以接受的，同时词典 A 能够补充一部分模型训练中没有抽取出的实体词汇，提高了召回率。以上结果说明基于互信息和左右熵的新词挖掘算法结合 BiLSTM-CRF 训练模型能够有效地对文本数据中的知识实体进行命名实体识别。

5.3 基于统计方法和规则方法的实体关系连接

本节使用基于词向量的相关关系抽取方法、基于滑动区间计算句向量的小节分割算法、基于规则的包含关系和先序关系的抽取方法对实体关系进行连接，关系连接实验流程如下：

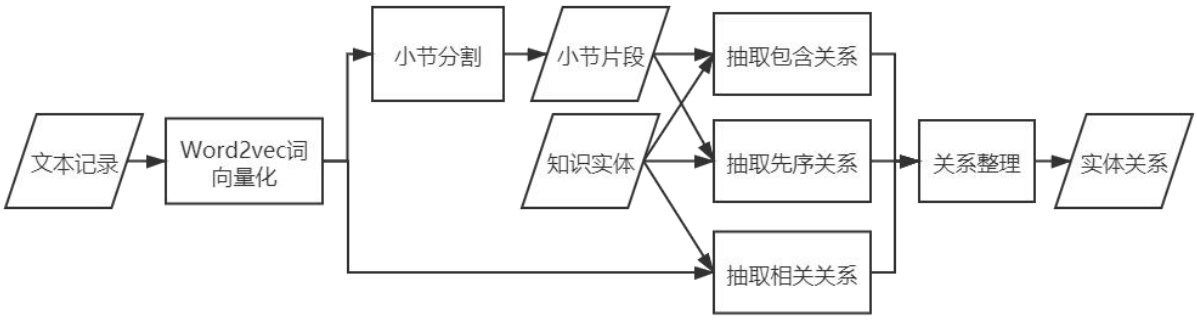


图 28 关系连接实验流程图

5.3.1 基于词向量的相关关系抽取实验与结果

这项研究使用预训练的 Word2vec 词向量模型生成每个知识实体的词向量，进一步计算每个知识实体之间的相似度，以实体词向量之间夹角的余弦值作为词之间的相似度指标，对每个实体只保留与其相似度超过 0.6 且相似度最高的前三个实体标注相关关系。

数据集说明：目标数据集中实体共 413 个，储存为 CSV 文件。

实验配置：该算法使用 gensim 包实现，使用腾讯 AI Lab 的 800 万中文语料预训练模型文件，其中每个中文词的向量维数为 200 维。

规则说明：由于演讲中存在一些词汇为组合词不存在在预训练的词向量模型中，如“数据清洗”等。因此使用 jieba.lcut 将该词切分，再计算每一个词向量的平均值作为其组合词的词向量。例如，“数据清洗”能够经过 jiebe.lcut 切分为“数据”和“清洗”，则整个词“数据清洗”的词向量计算为“数据”和“清洗”两个词向量的平均值。

实验结果：

本研究对目标实体数据集采用 Word2vec 模型共获得 1539 个相关关系，下表为具有相关关系的词（部分）。其中一些绝大多数抽取出的相关关系都存在较强的相关意义，如（讲座，知识服务）、（讲座，教学）、（大数据，建模）、（数据清洗，难处理数据）、（数据建模，采集数据）、（深度学习，神经网络）、（全方位，科学化）、（教育，智慧教育）、（教育，学校教育）、（适应，复杂场景）等。

相关关系（部分）
（讲座，知识服务），（讲座，教学），（讲座，学习服务），（大数据，建模），（报告，调查），（报告，数据），（数据清洗，难处理数据），（数据清洗，数据建模），（数据清洗，数据），（数据建模，数据），（数据建模，数据确权），（数据建模，采集数据），（应用，人工智能应用），（应用，应用场景），（应用，典型应用），（全方位，智能教育），（全方位，科学化），（深度学习，深度神经网络），（深度学习，学习能力），（深度学习，学习），（算法，分类算法），（算法，运算智能），（原理，机理），（原理，方法），（中叶，上个世纪），（中叶，工业革命），（中叶，发展），（教育，学校教育），（教育，智慧教育），（教育，教育服务），（信息化，智能化学习），（信息化，教育现代化），（信息化，智能教育），（融合，结合场景），（融合，混合增强智能），（融合，混合智能），（变革，人工智能时代），（变革，浪潮），（变革，智能化学习），（适应，适应性），（适应，增强学习），（适应，复杂场景）.....

5.3.2 基于滑动区间的小节分割

这项研究通过设置滑动窗口并计算相邻窗口的句向量相似度，找到相似度较低的若干个可分割点，进而从文本记录中分割小节，输出为小节片段进行储存。

数据集说明：目标文本数据共计 16412 字。

实验配置：该算法使用 gensim 包中 Word2vec 实现，使用腾讯 AI Lab 的 800 万中文语料预训练模型文件，其中每个中文词的向量维数为 200 维。经过多次调试设定滑动窗口宽度为 720 字，步长为窗口宽度的 1/6，即 120 字，极小值深度阈值为 0.1。

实验流程：

图 29 文本小节分割流程图

实验结果：

本研究使用基于滑动窗口对目标数据集进行小节分割，对于文本中的特定字，其相邻窗口中知识实体集合的相似度如下图。图中横轴为字的位置索引，纵轴为该字相邻滑动窗口中的知识实体集合的相似度。

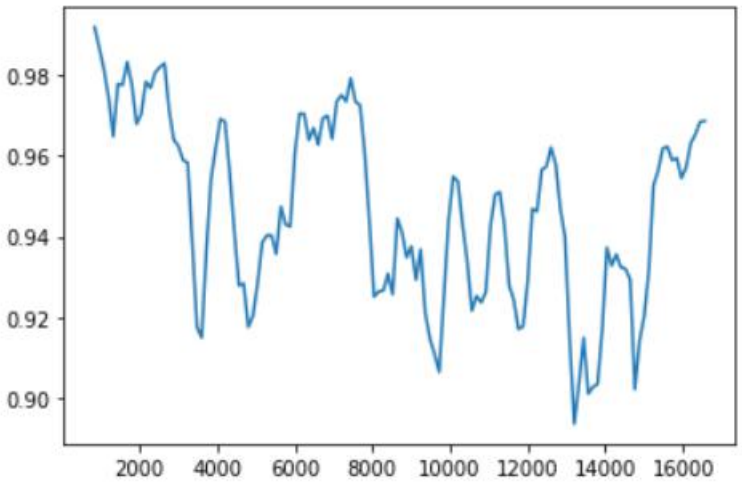


图 30 每个字相邻窗口中知识实体的相似度

共分割得到小节片段 24 个文本片段，部分小节文本如下：

序号	小节文本
1	ok 咱们上课今天是门课最后一次课前面呢我们用了十次讲座实际上是十个方方面面关于大数据领域报告让大家呢对大数据从数据采集数据质量分析数据清洗和存储数据建模以及相应数据应用对大家做了比较系统全方位介绍课程呢不可避免有些是大家听起来会比较难尤其涉及到了一些基本机器学习深度学习一些算法基本介绍但实际上我们并没有涉及到非常深入算法更多时候呢实际上是把算法一些基本原理以及基本一些应用给大家做了介绍那么在前面几周课程当中大家也应该都听到了关于大数据应用于我们教育教学场景学习场景教育教学管理场景区域教育治理场景以及基于复杂网络应用是我们对它应用层面系统介绍那么今天呢我们不再具体去讨论大数据我们看一看哈大数据在往下进一步应用时候呢我们现在关心一些话题实际上今天课程主题呢如何从大数据走向了我们人工智能两个概念呢它实际上是一脉相承概念人工智能概念提出呢大家都知道实际上在上个世纪中叶就开始谈人工智能但是都没有做起来在上个世纪 90 年代末人工智能曾经很火了一致但是后来呢又逐渐走向了销声匿迹到最近十年人工智能成为了我们热点从大角度来说话呢人工智能应用呢是我们大数据加上了算力再加上了算法我们三者有机了结合形成了我们人工智能但从小角度来说话回归到我们教育场景实际上我们现在用还是比较比较少比较弱也是比较 low 相应呢强人工智能应用进入到了我们教育领域实际上还有很长一段路要走但不管怎么说我们今天课呢重点我们看一看一类技术将来到我们教育当中会怎么样去
2	讨论咱们教育当中讨论呢借用一张图实际上去概括去说明智慧教育我们国内发明了很多一些词哈从我们教育信息化开始些年我们在谈信息化 20 逐渐我们在谈智慧教育智慧学习最近我们在谈智能教育从去年时候我们开了第一届全球智能教育大会今年

	12 月份也其实下周下下周第二届全球人工智能教育大会将在北师大开线上线下融合方式我们来开一场会届时呢应该是教育部保存部长也会到北师大来一起来开会会去讨论问题呢当然我们大部分会都是线上哈会主题呢实际上去解决了问题么多些技术进入到了教育场景了以后呢到底对我们教育会形成样变革我们教育要想适应一种变革我们到底要做一些各位同学你们要做一些我们作为老师来说话我们要做一些做一系列些工作到底是对于我们所谓人力资源强国建设到底会带来样影响现在很多人都在讨论和研究话题那么话题你看我们在谈包括虚拟现实 VR 引入到了我们教育场景当中会带来样一些应用我们云计算我们互联网进入到了教育领域当中我们会带来什么样一些应用我们知识计算我们信息化以及我们大数据一起进入到了教育场景当中我这里概括起来一句话他我们唯一希望是给每孩子包括各位在内每同学去提供最适合你教育服务最适合你最灵活最个性化最适合了你发展但现在问题是呢东西是适合你你们不太清楚我们也不太清楚我们就希望是利用大数据通过数据挖掘去挖掘出来一些东西能够体现出来是你一些特征我一些特征从我这样一个老师我适合用什么样一种教学方式来提

5.3.3 基于规则的包含关系和先序关系抽取实验与关系整理结果

这项研究通过小节片段之间形成的结构关系，利用词频统计获得每个片段词频最高的前三个词作为该片段的主题词，并基于规则提取片段与片段之间、片段内实体包含和先序关系并进行储存。

数据集说明：目标数据集文本片段共 24 个，实体共 413 个。

实验规则说明：

(1) 主题词：某文本范围中的主题词为该范围内词频最高的三个词。

(2) 包含关系抽取规则说明：

① 规则一：同一演讲内，该讲的主题词包含各小节的小节主题词，如第二讲主题词包括第二讲中各小节的小节主题词；

② 规则二：同一小节中，小节主题词包含该小节中所有知识实体，如第 1 小节主题词包括了第一小节中出现的所有知识实体；

(3) 先序关系抽取规则说明：

① 规则一：按照演讲先后顺序，靠前的演讲主题先序于靠后的演讲主题，如第一讲主题词先序于第二讲主题词；

② 规则二：在同一演讲内，前一小节的主题词先序于后面小节的主题词，如第 1 小节主题词先序于第二小节主题词；

(4) 关系整理规则说明：

① 法则一：如果一个实体 A 与自己本身出现包含、先序、相关关系，则删除这些关

系；

② 法则二：如果两个实体 A 与 B 之间属于包含关系或先序关系，且又属于相关关系；则删除 A 与 B 的相关关系。

③ 法则三：如果两个实体 A 与 B 之间存在 A 包含 B 且 B 包含 A 或者 A 先序于 B 或者 B 先序于 A 的关系，则删除 A 与 B 之间的所有包含和先序关系，将 A 与 B 判定为相关关系。

实验流程：

步骤一：读入先前采用 Word2vec 词向量模型抽取的相关关系集合 $R = \{R_1, R_2, \ldots\}$；

步骤二：遍历全部文本记录和小节文本，计算各讲以及各小节知识实体的词频，选取在各讲及各小节中词频最高的若干个知识实体作为主题词，获得各讲主题词集合 $L = \{L_1, L_2, \ldots\}$和各小节主题词集合$T_{L_i} = \{T_{L_1}, T_{L_2}, \ldots\}$，并从中抽取和包含关系集合 $C = \{C_1, C_2, \ldots\}$和先序关系集合 $P = \{P_1, P_2, \ldots\}$；

步骤三：若集合 R、集合 P、集合 C 中出现实体 A 与自身出现包含、先序、相关关系，则删除关系；

步骤四：若存在实体 A 与 B 之间属于包含关系或先序关系且又属于相关关系；则删除 A 与 B 的相关关系；

步骤五：若两个实体 A 与 B 之间存在 A 包含 B 且 B 包含 A 或者 A 先序于 B 或者 B 先序于 A 的关系，则删除 A 与 B 之间的所有包含和先序关系，将 A 与 B 判定为相关关系加入到相关关系集合 R 中。

实验结果：

经过关系整理，共获得 1176 个包含关系和 4 个先序关系。

包含关系部分列出如下表，每个元组表示前面的词包含后面的词。可以看到表中有一些元组具有比较强的可解释性。例如“教育”主题包含了“智慧学习”以及“人工智能教育大会”，说明在演讲人在围绕“教育”为主题的阐述中，“智慧学习”以及“人工智能教育大会”作为知识实体在小节中出现。可认为在知识图谱中可形成“教育”知识实体包含“智慧学习”以及“人工智能教育大会”的关系，有利于表征演讲人组织的知识结构。

同时，有一些元组的可解释性较差。例如“（算法, 健康码）”、“（算法, 疫情）”，客观世界的知识结构无法直接认可“算法”知识实体包含了“健康码”和“疫情”这两个知识实体。但通过回溯文本数据可以发现，演讲人在该片段中的主题是谈论现代的信息化环境提高了数据的采集效率从而能够使原有的算法能够更高质量地提供合适的服务，并使用了疫情中的健康码案例来加以来说明。而案例可属于知识结构中的一部分，因此可认为该片段主题包含了“健康码”“疫情”这两个知识实体。同时该片段的主题词由三个词组成，其中“算法”包含在内，因此可以认为“算法”知识实体间接地包含“健康码”和“疫情”。这个案例说明了在此结构定义下的包含关系丰富了原有包含关系的内涵。

表 12 从目标数据集中提取的包含关系和先序关系（部分）

包含关系（部分）
（大数据, 数据清洗），（大数据, 数据建模），（大数据, 应用），（大数据, 原理），（大数据, 教学），（大数据, 管理），（大数据, 教育治理），（算法, 深度学习），（算法, 原理），（算法, 教育治理），（教育, 热点），（教育, 有机结合），（教育, 智慧教育），（教育, 教育信息化），（教育, 信息化），（教育, 智慧学习），（教育, 智能教育），（教育, 人工智能教育大会），（教育, 融合），（教育, 教育部），（教育, 适应），（教育, 个性化），（教育, 教学方式），（教育, 建模），（教育, 适应性），（大数据, 云计算），（大数据, 互联网），（大数据, 知识计算），（大数据, 唯一希望），（大数据, 教育服务），（大数据, 灵活），（大数据, 个性化），（大数据, 特征），（大数据, 建模），（算法, 无监督），（算法, 卷积神经网络），（算法, 上个世纪），（算法, 数据），（算法, 健康码），（算法, 疫情），（算法, 人脸识别），（算法, 图像识别），（算法, 语音识别），（算法, 畅言），（算法, 机器学习），（算法, 具体场景），（算法, 智能化学习），（算法, 服务），（算法, 结合场景），（算法, 测量机器），（算法, 分类算法），（大数据, 相关），（大数据, 挖掘），（大数据, 因果），（大数据, 因果律），（大数据, 干预），（大数据, 要素），（大数据, 测量），（大数据, 教育学），（大数据, 教育要素），（大数据, 评价），（大数据, 预测），（大数据, 情绪状态），（大数据, 兴趣状态），（大数据, 动机状态），（大数据, 典型），（大数据, 智能），（大数据, 学习服务），（大数据, 问题场景），（大数据, 有限），（大数据, 领域专家），（批阅, 人工智能），（批阅, 工业革命），（批阅, 人类），（批阅, 体力劳动），（批阅, 解放），（批阅, 脑力劳动），（批阅, 利害性考试），（作文, 脑力劳动），（作文, 教育场景），（作文, 工作），（作文, 批阅），（作文, 阅读），（作文, 文献），（作文, 产品），（作文, 程度），（作文, 利害性考试），（作文, 技术），（人工智能, 工业革命），（人工智能, 体力劳动），（人工智能, 解放），（人工智能, 信息革命），（人工智能, 互联网），（人工智能, 自动化）.....

从目标数据集提取的先序关系如下表。先序关系数量较少且较难进行解释。

表 13 从目标数据集中提取的先序关系

先序关系
（应用，教育），（大数据，教育），（大数据，作文），（大数据，批阅）

5.4 知识图谱储存及可视化呈现

本节将上述得到的知识实体和实体关系，储存为 CSV 文件。通过 python 中的 py2neo 储存到图数据库 Neo4j 中，并通过 Neo4j Community Server 的本地服务打开浏览器输入 http://localhost:7474/browser/，即可查看知识图谱。若需要进一步使用该图谱，则可以使用 Neo4j 语言 Cypher 查询相关节点和连边。

输入说明：知识词典 A+C、相关关系集合 R、先序关系集合 P、包含关系集合 C 储存的 CSV 文件。

实验配置：该环节由 python 程序完成，使用第三方包 py2neo 2021.1.1 版本进行。

实验流程：

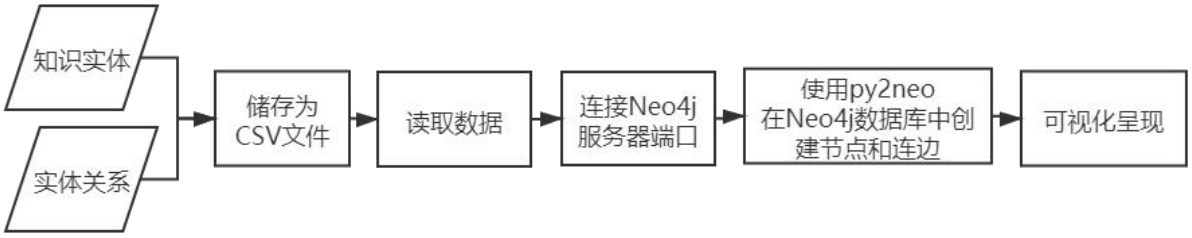


图 31 知识图谱储存及可视化呈现流程图

(5) 第一步：使用 pandas 从 CSV 文件中读取知识实体和关系；

(6) 第二步：启动 Neo4j 服务器并使用 py2neo 连接端口“http://localhost:7474”；

① 在本地调用 Neo4j 服务，在 cmd 窗口输入 neo4j.bat console 启动 Neo4j 服务器；

② 使用 py2neo 连接“http://localhost:7474”；

(7) 第三步：创建节点和连边；

① 创建两个节点实例 node_1 和 node_2，定义其标签（label）为 Knowledge，定义属性 name 为知识实体名称；

② 将节点实例 node_1 和 node_2 分别融合（merge）到图中，程序能够自动检测是否存在相同标签和属性的名称节点，若存在则不会新建节点，若不存在则新建节点；

③ 建立连边关系（node_1，实体关系，node_2），同样融合到图中，程序能够自动检测是否存在相同起始节点、终止节点和连边关系，若存在则不会新建关系，若不存在则新建关系；

(8) 第四步：打开 Neo4j 图数据库进行可视化呈现；

经过上述步骤成功将知识图谱储存到 Neo4j 图数据库中，可以从本地浏览器中输入“http://localhost:7474”访问该数据库。

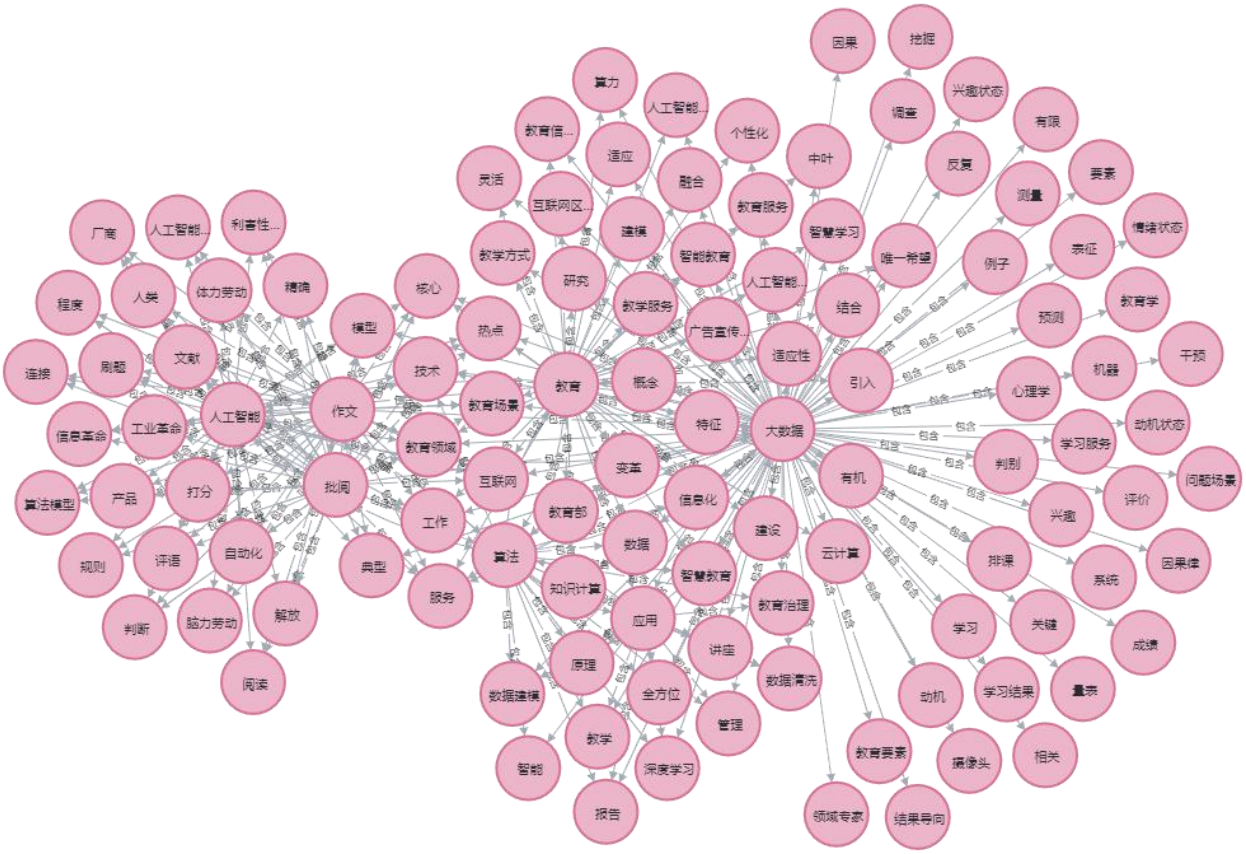


图 32 在 Neo4j 图数据库中呈现的知识图谱（部分数据）

6 总结与展望

6.1 全文工作总结

6.1.1 研究成果

在这项研究工作中，从文本数据中自动化地提取知识图谱的问题得到了一定程度的解决。以往从文本这类非结构化数据中抽取知识图谱主要依赖专家进行人工处理，这种方法不仅费时费力，而且随着文本信息的快速增多，人工处理已经逐渐变得不可行。而在教育领域中文本数据大量存在，缺乏对文本信息的自然语言理解会阻碍人工智能技术在教育领域的应用。

本研究以教学视频转录的文本记录作为数据源，设计并实现了一套基于自然语言处理技术、采用了深度学习、统计学、规则方法的知识图谱半自动化构建方案。该方案注重了构建成本效益并检验了提取结果的有效性。不仅降低了从文本中构建知识图谱的成本，还能为更多教育技术应用提供结构化的数据基础。

在数据处理环节，本研究设计这个环节以保证视（音）频转录文本的质量，并且使用第三方平台进行视频转录和语气词过滤，这种方法可以轻松迁移以处理其他的音视频数据。

在命名实体识别及效果检验环节，这项研究采用了人机协同和深度学习的半自动化和自动化的构建方法，利用了基于互信息和左右熵的新词挖掘算法以及 BiLSTM-CRF 进行命名实体识别。

在实体关系连接环节中，这项研究采用滑动区间相似度测量的方法对连续的讲座文本进行小节分割，再运用了统计学方法和规则方法抽取实体间的相关关系、包含关系和先序关系。

在知识图谱储存及可视化呈现环节，这项研究将抽取出来的知识实体和实体关系通过 Python 程序储存到 Neo4j 图数据库中，并在浏览器中可进行可视化呈现，方便浏览和传输。

其次，本研究选用《教育大数据》课程中的文本数据，通过上述方案半自动化地构建了知识图谱，抽取并检验目标数据集中的 413 个知识实体抽取的准确率、召回率和 F 值，验证了命名实体识别方案的可行性。通过文本小节分割环节将目标文本切分为 24 个小节，并通过关系抽取获得了 1539 个相关关系、1176 个包含关系、4 个先序关系，构建了《教育大数据》课程的知识图谱。通过 python 储存在 Neo4j 图数据库中，进行可视化呈现。

6.1.2 对知识图谱构建方案的反思

本节对这项研究提出的知识图谱半自动化构建方案和结果进行反思，这项研究经过实证，认为以文本数据为数据源进行半自动化构建知识图谱时需要注意以下几个问题：

(1) 数据采集和处理的问题。

由于知识图谱的构建对于文字的准确性要求较高，因此在视频（音频）采集、视频转录成文本的过程中也需要有较高的准确度，才能构建出可用的知识图谱。在本研究数据处理环节中，进行语音转录文本时需要反复对照原音频，才能矫正一些转录出错的文本，会是不得不付出的时间成本。

此外由于笔者对训练模型的了解不足，在一开始进行数据处理时，去除了语音转文本中的标点符号以及空格，因此在模型训练过程中缺失了一部分的位置信息，导致训练效果并没有达到理想状态，这在未来的研究中会进一步解决。

(2) 自动化程度和实体识别以及关系连接准确度之间的平衡的问题。

由于文本数据量的快速增多，自动化的处理机制成为了据必要的处理手段和要求。但目前不论是半自动化构建知识图谱还是自动化知识图谱构建还远远没有达到预期效果，直接应用在教育领域也是具有风险的。而且由于教育领域内的风险往往是不可逆的，因此在构建教育领域的知识图谱时，我们必须谨慎考虑图谱中的知识实体以及实体关系的准确性，同时尽可能采取可自动化的方案，以提高教育数据的处理质量和数量，来推动教育与学习中智能化技术解决问题的进展。

(3) 实体关系连接规则定义的主观偏误的问题。

在基于统计方法和规则方法的实体关系抽取部分，本研究抽取出了知识实体之间的相关关系、包含关系和先序关系。由于关系抽取规则主要依靠人工定义规则，因此存在一定的主观偏误。例如在本研究中抽取的知识实体之间的相关关系和包含关系数量较多且效果较好，而基于定义规则抽取出的先序关系较少且较难解释。可能是由于本研究定义的先序关系规则较弱，因此抽取先序关系的效果不好。在未来的研究中可尝试引入外部词典或采用概率关联等算法进行先序关系抽取。

6.1.3 研究价值

本研究设计并实现了一套半自动化抽取知识图谱的方案，此研究的研究价值在于：

(1) 理论价值：

在这项研究工作中，从文本数据中自动化地提取知识图谱的问题得到了一定程度的解决。以往从文本这类非结构化数据中抽取知识图谱主要依赖专家进行人工处理，这种方法不仅费时费力，而且随着文本信息的快速增多，人工处理已经逐渐变得不可行。而在教育领域中文本数据大量存在，缺乏对文本信息的自然语言理解会阻碍人工智能技术在教育领域的应用。

这项研究是第一个为中文课程内容构建知识图谱所做的尝试，也是第一个采用半自动化方法构建中文课程内容知识图谱的研究。我们尝试使用多种半自动化、自动化技术手段从非结构化的教学过程的文本记录中抽取了课程涵盖的知识体系，尝试了一种新的非结构化数据处理的可能性，为该领域做出了新的探索。

(2) 实践价值：

本研究提出的半自动化构建方案是具有成本效益的，使用的工具都具有易得性。同时，该方案具有较低的内容依赖性，可迁移到其他文本数据的知识图谱构建中。可利用该方案节省构建知识图谱的时间成本和人力成本，从而能够让使用者专注于知识图谱的应用设计，如教学管理、教学质量评估等，从而促进教育大数据的利用以解决教育与学习中存在的问题。

6.1.4 研究局限与不足

这项研究提出的知识图谱构建方案还有一定的局限性。

首先，在数据处理和命名实体识别方面，由于一开始对文本的深度学习模型训练方法的不熟悉，在前期数据处理的过程中没有选择带有标点符号的语音转文本工具，这样的做法导致在训练模型时缺少了字词出现在句子中的位置信息。而重新转录文本进行校对并进一步人工标注 BIO 序列的工程量巨大，由于时间限制，笔者只能在训练模型时尝试以字数为步长增加断句标识，以随机挑选的十个句子的平均长度作为基准，上下浮动五个字的范围对文本数据进行断句调试。

其次，在文本的小节分割环节中仍然存在不足，在目前的小节分割环节中，滑动窗口的长度参数是根据人工多次调试得到的一个参数。由于随着窗口长度数值的增大，分割性能会逐渐变好之后逐渐变差，因为当窗口长度增大时一个窗口中能够包含的线索平均数量也就增多了，能够与其他窗口的区分也变大了；而最佳窗口有一个上限，当一个窗口过大时，由于窗口中囊括的主题词太多因此难以和其他小节进行区分。因此通过人工判断分割性能是否变好，可以粗略地获得该参数。但在未来要更精确地应用小节分割短发，还需要进一步对滑动窗口参数进行指标判定。

最后，在实体关系连接方面，第一是该知识图谱抓取的三种实体关系只是知识关系中的一部分，由于使用实体之间的相关、包含、先序来概括所有知识实体之间的关系，这样的定义方式疏漏了一些句子表达的含义，如“从大数据走向人工智能”，“走向”的意味并没有被系统捕捉到，因此在这样的构建方案下，一些知识之间的含义会被疏漏。第二是实体先序关系提取规则存在不足。本研究中先序关系的判定依据是讲座文本中的前后顺序。但这样假设了演讲人会根据先序关系对演讲内容进行组织。这个假设较为粗浅，在未来还可继续完善，例如使用概率关联规则挖掘 Apriori 算法进行先序关系的挖掘等。

6.2 未来工作展望

因为研究时间受限以及笔者研究水平有限，目前的讲座知识图谱还具有进一步改善的空间，在未来的研究中希望能够继续完善。

第一，在数据处理部分，尽可能获取带有正确断句信息的文本数据进行模型训练。

第二，在命名实体识别部分，可进一步考虑文本信息中的代词进行，完善文本信息抽取；另外可以考虑引入外部知识库和词典进行识别。

第三，完善关系连接方法，可进一步考虑采用自下而上的关系连接方法，尽可能还原文本信息中原来的实体关系。

第四，增加知识融合、实体消歧等环节进一步提高知识图谱的可用性。

参考文献

[1]. 陈新亚, & 李艳. (2020). 《2020 地平线报告: 教与学版》 的解读及思考--疫情之下高等教育面临的挑战与变革. 远程教育杂志, 38(2), 3-16.

[2]. Feldstein, M., & Hill, P. (2016). Personalized learning: What it really is and why it really matters. Educause review, 51(2), 24-35.

[3]. 朱艳茹, 范亚芹, & 赵洋. (2018). 基于知识图谱的自适应学习系统知识模型构建. 吉林大学学报: 信息科学版, 36(3), 345-350.

[4]. 祁晓慧.(2020).多模态课程知识图谱构建与应用研究(硕士学位论文,吉林大学).

[5]. Shi, W., Liu, X., Gong, X., Niu, X., Wang, X., Jing, S., ... & Luo, J. (2019, November). Review on Development of Smart Education. In 2019 IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI) (pp. 157-162). IEEE.

[6]. Rizun, M. (2019). Knowledge graph application in education: a literature review. Acta Universitatis Lodziensis. Folia Oeconomica, 3(342), 7-19.

[7]. 朱福军. (2018). 基于学习元的领域知识图谱自动构建研究 (Master's thesis, 四川师范大学).

[8]. 朱艳茹. (2018). 基于知识图谱的自适应学习系统设计与实现 (Master's thesis, 吉林大学).

[9]. Wang, S. (2017). Knowledge Graph Creation from Structure Knowledge.

[10]. Basu, S., Yu, Y., Singh, V. K., & Zimmermann, R. (2016, January). Videopedia: Lecture video recommendation for educational blogs using topic modeling. In International Conference on Multimedia Modeling (pp. 238-250). Springer, Cham.

[11]. 侯俊萌. (2017). 基于 MOOC 的高等教育知识图谱的构建 (Doctoral dissertation, 北京: 北京邮电大学).

[12]. Ilkou, E., & Signer, B. (2020). A Technology-enhanced Smart Learning Environment based on the Combination of Knowledge Graphs and Learning Paths. In CSEDU (2) (pp. 461-468).

[13]. Signer, B., & Norrie, M. C. (2007, November). As we may link: a general metamodel for hypermedia systems. In International Conference on Conceptual Modeling (pp. 359-374). Springer, Berlin, Heidelberg.

[14]. Chen, P., Lu, Y., Zheng, V. W., Chen, X., & Li, X. (2018, June). An automatic knowledge graph construction system for K-12 education. In Proceedings of the Fifth Annual ACM Conference on Learning at Scale (pp. 1-4).

[15]. 蒋琪. (2020). 基于 MOOC 的交互式领域知识地图半自动化构建研究 (Master).

Beijing Normal University.

[16]. Hodge, G. (2000). Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files. Digital Library Federation, Council on Library and Information Resources, 1755 Massachusetts Ave., NW, Suite 500, Washington, DC 20036.

[17]. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., & Taylor, J. (2008, June). Freebase: a collaboratively created graph database for structuring human knowledge. In Proceedings of the 2008 ACM SIGMOD international conference on Management of data (pp. 1247-1250).

[18]. Wu, W., Li, H., Wang, H., & Zhu, K. Q. (2012, May). Probase: A probabilistic taxonomy for text understanding. In Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data (pp. 481-492).

[19]. Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., ... & Zhang, W. (2014, August). Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 601-610).

[20]. Zheng, Q., Liu, J., Zeng, H., Guo, Z., Wu, B., & Wei, B. (2019). Knowledge forest: A novel model to organize knowledge fragments. arXiv preprint arXiv:1912.06825.

[21]. Doignon, J. P., & Falmagne, J. C. (1985). Spaces for the assessment of knowledge. International journal of man-machine studies, 23(2), 175-196.

[22]. Sun, L., Cheng, R., Cheung, D. W., & Cheng, J. (2010, July). Mining uncertain data with probabilistic guarantees. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 273-282).

[23]. Chen, P., Lu, Y., Zheng, V. W., Chen, X., & Yang, B. (2018). KnowEdu: a system to construct knowledge graph for education. Ieee Access, 6, 31553-31563.

[24]. Liang, C., Ye, J., Wu, Z., Pursel, B., & Giles, C. L. (2017, February). Recovering Concept Prerequisite Relations from University Course Dependencies. In AAAI (pp. 4786-4791).

[25]. Liu, H., Ma, W., Yang, Y., & Carbonell, J. (2016). Learning concept graphs from online educational data. Journal of Artificial Intelligence Research, 55, 1059-1090.

[26]. Wang, S., Liang, C., Wu, Z., Williams, K., Pursel, B., Brautigam, B., ... & Giles, C. L. (2015, September). Concept hierarchy extraction from textbooks. In Proceedings of the 2015 ACM Symposium on Document Engineering (pp. 147-156).

[27]. Chaplot, D. S., Yang, Y., Carbonell, J., & Koedinger, K. R. (2016). Data-Driven Automated Induction of Prerequisite Structure Graphs. International Educational Data Mining Society.

[28]. Zhu, Y., Cao, X., Bian, Y., & Wu, J. (2014, September). CKGHV: a comprehensive

knowledge graph for history visualization. In IEEE/ACM Joint Conference on Digital Libraries (pp. 437-438). IEEE.

[29]. Sun, K., Liu, Y., Guo, Z., & Wang, C. (2016). Visualization for Knowledge Graph Based on Education Data. International Journal of Software and Informatics, 10(3).

[30]. Zhao, T., Huang, Y., Yang, S., Luo, Y., Feng, J., Wang, Y., ... & Zhu, F. (2019, April). Mathgraph: A knowledge graph for automatically solving mathematical exercises. In International Conference on Database Systems for Advanced Applications (pp. 760-776). Springer, Cham.

[31]. Zhu, Y., Cao, X., Bian, Y., & Wu, J. (2014, September). CKGHV: a comprehensive knowledge graph for history visualization. In IEEE/ACM Joint Conference on Digital Libraries (pp. 437-438). IEEE.

[32]. Zhao, T., Huang, Y., Yang, S., Luo, Y., Feng, J., Wang, Y., ... & Zhu, F. (2019, April). Mathgraph: A knowledge graph for automatically solving mathematical exercises. In International Conference on Database Systems for Advanced Applications (pp. 760-776). Springer, Cham.

[33]. Shanmukhaa, G. S., Nandita, S. K., & Kiran, M. V. K. (2020, March). Construction of Knowledge Graphs for video lectures. In 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS) (pp. 127-131). IEEE.

[34]. Rizun, M. (2019). Knowledge graph application in education: a literature review. Acta Universitatis Lodziensis. Folia Oeconomica, 3(342), 7-19.

[35]. Ehrlinger, L., & Wöß, W. (2016). Towards a Definition of Knowledge Graphs. SEMANTiCS (Posters, Demos, SuCCESS), 48, 1-4.

[36]. Paulheim, H. (2017). Knowledge graph refinement: A survey of approaches and evaluation methods. Semantic web, 8(3), 489-508.

[37]. CS 520: Knowledge Graphs. (2020). Retrieved 10 December 2020, from https://web.stanford.edu/class/cs520/

[38]. Qingjie, L., Lingyu, X., Jie, Y., Lei, W., Yunlan, X., Suixiang, S., & Yang, L. (2014, September). Research on domain knowledge graph based on the large scale online knowledge fragment. In 2014 IEEE Workshop on Advanced Research and Technology in Industry Applications (WARTIA) (pp. 312-315). IEEE.

[39]. 肖仰华. (2020). 领域知识图谱落地实践中的问题与对策_实体. Retrieved 10 December 2020, from https://www.sohu.com/a/280006592_100099320

[40]. Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., ... & Zhang, W. (2014, August). Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge

discovery and data mining (pp. 601-610).

[41]. Wu, W., Li, H., Wang, H., & Zhu, K. Q. (2012, May). Probase: A probabilistic taxonomy for text understanding.

[42]. Frontini, F., Del Gratta, R., & Monachini, M. (2013, December). GeoDomainWordNet: Linking the geonames ontology to WordNet. In Language and Technology Conference (pp. 229-242). Springer, Cham.

[43]. 刘烨宸, & 李华昱. (2020). 领域知识图谱研究综述. 计算机系统应用, 29(6), 1-12.

[44]. 王昊奋. (2019). 知识图谱 方法、实践与应用. 北京: 电子工业出版社.

[45]. 刘峤, 李杨, 段宏, 刘瑶, & 秦志光. (2016). 知识图谱构建技术综述. 计算机研究与发展, 53(3), 582.

[46]. Yan, J., Wang, C., Cheng, W., Gao, M., & Zhou, A. (2018). A retrospective of knowledge graphs. Frontiers of Computer Science, 12(1), 55-74.

[47]. Leaman, R., Wei, C. H., & Lu, Z. (2015). tmChem: a high performance approach for chemical named entity recognition and normalization. Journal of cheminformatics, 7(S1), S3.

[48]. Han, A. L. F., Wong, D. F., & Chao, L. S. (2013, June). Chinese named entity recognition with conditional random fields in the light of Chinese characteristics. In Intelligent Information Systems Symposium (pp. 57-68). Springer, Berlin, Heidelberg.

[49]. Whitelaw, C., Kehlenbeck, A., Petrovic, N., & Ungar, L. (2008, October). Web-scale named entity recognition. In Proceedings of the 17th ACM conference on Information and knowledge management (pp. 123-132).

[50]. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural architectures for named entity recognition. arXiv preprint arXiv:1603.01360.

[51]. Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:1508.01991.

[52]. 苏丰龙, 谢庆华, 邱继远, & 岳振军. (2016). 基于深度学习的领域实体属性词聚类抽取研究. 微型机与应用, 35(1), 53-55.

[53]. Feng, J., Huang, M., Zhao, L., Yang, Y., & Zhu, X. (2018). Reinforcement learning for relation classification from noisy data. arXiv preprint arXiv:1808.08013.

[54]. Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. Lingvisticae Investigationes, 30(1), 3-26.

[55]. Ratnaparkhi, A. (1996). A maximum entropy model for part-of-speech tagging. In Conference on empirical methods in natural language processing.

[56]. Collins, M., & Duffy, N. (2002, July). New ranking algorithms for parsing and tagging: Kernels over discrete structures, and the voted perceptron. In Proceedings of the 40th

Annual Meeting of the Association for Computational Linguistics (pp. 263-270).

[57]. Knoke, D., Burke, P. J., & Burke, P. J. (1980). Log-linear models (Vol. 20). Sage.

[58]. 刘克彬, 李芳, 刘磊, & 韩颖. (2007). 基于核函数中文关系自动抽取系统的实现. 计算机研究与发展, 44(8), 1406-1411

[59]. Carlson, A., Betteridge, J., Wang, R. C., Hruschka Jr, E. R., & Mitchell, T. M. (2010, February). Coupled semi-supervised learning for information extraction. In Proceedings of the third ACM international conference on Web search and data mining (pp. 101-110).

[60]. Liwei, C., Yansong, F., & Dongyan, Z. (2013). Extracting relations from the web via weakly supervised learning. Journal of Computer Research and Development, 50(9), 1825.

[61]. Zhang, Y., & Zhou, J. (2000, October). A trainable method for extracting Chinese entity names and their relations. In Second Chinese Language Processing Workshop (pp. 66-72).

[62]. 互联网时代的社会语言学：基于SNS的文本数据挖掘 | Matrix67: The Aha Moments. (2012). Retrieved 27 April 2021, from http://www.matrix67.com/blog/archives/5044

[63]. Mutual information. Retrieved 27 April 2021, from https://nlp.stanford.edu/IR-book/html/htmledition/mutual-information-1.html

[64]. 反作弊基于左右信息熵和互信息的新词挖掘. (2017). Retrieved 27 April 2021, from https://zhuanlan.zhihu.com/p/25499358

[65]. 谢腾, 杨俊安, & 刘辉. (2020). 基于 BERT-BiLSTM-CRF 模型的中文实体识别. 计算机系统应用, 29(7), 48-55.

[66]. Lafferty, J., McCallum, A., & Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

[67]. McCallum, A., & Li, W. (2003). Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons.

[68]. luopeixiang/named_entity_recognition. (2019). Retrieved 1 May 2021, from https://github.com/luopeixiang/named_entity_recognition

[69]. Galanopoulos, D., & Mezaris, V. (2019, January). Temporal lecture video fragmentation using word embeddings. In International Conference on Multimedia Modeling (pp. 254-265). Springer, Cham.

[70]. new-MOVING-lecture-video-fragmentation-technologies-in-videolectures-net-platform. (2019). Retrieved 9 May 2021, from http://moving-project.eu/2019/01/21/new-moving-lecture-video-fragmentation-technologies-in-videolectures-net-platform/

[71]. zhanzecheng/Chinese_segment_augment. (2018). Retrieved 27 April 2021, from https://github.com/zhanzecheng/Chinese_segment_augment

致　谢

感谢读者看到这里。

在此论文完成之际，我也即将告别本科学习阶段，迈入新的征程。四年时光匆匆，想做的事很多，收获的美好也很多。在此我想向所有帮助过关心过我的老师、同学、朋友以及家人献上最真挚的感谢。

感谢我的本科生导师张婧婧教授对我四年学业和生活的关心和支持，是婧婧老师带领我尝试做研究、撰写报告，婧婧老师的科学家精神一直指引着我、感染着我。感谢论文指导老师郑勤华教授给予本研究工作的倾心指导和帮助，是郑老师让我看到了教育科技项目落地实践的伟大意义。感谢班主任张志祯老师于我在疫情期间在国外交换学习生活的温暖关怀。感谢丁道勇老师的教育学课程启迪，感谢陈桄老师在我选考发展心理学的鼓励，感谢美凤老师在我提前修读课程的耐心辅导，感谢孙洪涛老师在我尝试产品设计的指引，感谢李芒老师和我分享的知识地图，感谢张进宝老师在我探索学龄前儿童计算思维项目的支持，感谢卢宇老师和李艳燕老师激发了我对人工智能教育应用的兴趣，感谢吴娟老师和李爽老师的帮助，不然在大四这学期我差一点修不完规定课程无法按时毕业。还要感谢所有任课老师对我的耐心和悉心指导。在此，还要感谢 SDSU 王敏娟教授给我在美国生活和学习上的帮助，我很幸运能够认识敏娟老师，感受到她带给我以及许多学生的温暖和光芒。

感谢这一路遇到的师兄师姐们！感谢君磊师兄、蒋琪师姐对这篇论文的帮助和建议，感谢玻玻师姐、晓杰师姐、艺萌师姐的一路帮助；感谢在满天星丹妮姐和孙尧姐的培养与鼓励，还要感谢怀波师兄、欣竹师姐、嘉灵师姐、陈佑师兄好多师兄师姐的建议和帮助。

感谢四年来一路同行的朋友们！感谢一起做课程项目的陈蕊、颢琳、王雯好多同学，一起熬大夜做数学建模的梦园和 Chiaming，感谢听我讲论文问题的李知航、詹子贤和弟弟昀翔；感谢一起备考经济学双学位的友友们；感谢室友们陪伴四年一起度过了大学这个如此重要的阶段；感谢我的好朋友可昕和 lulu，和你们在一起的时光是我大学四年里最最宝贵的回忆；感谢我在 San Diego 遇到的虞瑶和冰冰还有室友 Kaylee 和 Andres，和你们一起学习做饭爬山，一起度过了无数个在美国的夜晚，那将成为我最难忘的大学回忆。

感谢爸爸妈妈以及家人一直以来对我的支持和鼓励，是你们每一次在登机口送别我、看我远行，你们是我最温暖的港湾；感谢我的猫提提在去年出现在我的生活里，带给我很多很多的快乐；感谢爱我的人们和我爱的人们，给予我生活里点点滴滴的温暖。

感谢自己，在每一个艰难的时光里都走过来了。也希望我能继续拥有勇气，无畏地走向前方。

简靖琳

2021 年 5 月