文字知识图谱构建及应用

安 波
（中国社会科学院民族学与人类学研究所，北京 100081）

摘要：文字是文化和知识传播最重要的载体，是族群文明的重要标志，对于民族研究和语言研究具有重要价值。由于年代久远或使用人口较少，许多古汉字或少数民族文字已经濒临灭绝，严重影响了语言文字的多样性和语言发展规律的研究。为了充分保护和挖掘这些文字的价值，本文基于中华字库工程成果数据设计了一个大规模中华文字知识图谱，其涵盖楷书汉字及多种古汉字和少数民族文字。该图谱为研究中华文字的发展、汉藏同源、民族共同体等具有重要意义。为了更好地展示文字知识图谱的作用，构建了一个面向文字知识的智能问答系统，利用文字知识图谱的数据满足用户以自然语言方式进行查询的需求。

关键词：知识图谱；中华字库；智能问答；字际关系；关系抽取
DOI：10.3772/j.issn.1674−1544.2022.01.009 **CSTR：**15994.14.issn.1674−1544.2022.01.009
中图分类号：TP391 **文献标识码：**A

Construction and Application of Character Knowledge Graph

AN Bo
(Institute of Ethnology and Anthropology, Chinese Academy of Social Sciences, Beijing 100081)

Abstract: Characters are one the most important carrier of culture and knowledge dissemination, an important symbol of ethnic civilization, and are of great value to national and language research. Many ancient Chinese characters and minority characters are on the verge of extinction due to their long history or small population, which has seriously affected the diversity of languages and the study of the law of language development. In order to fully protect and mine the value of these characters, this paper designs and implements a large-scale characters knowledge graph based on the achievement data of Chinese Character Library Project, including regular script Chinese characters, ancient Chinese characters and minority characters. The knowledge graph is of great significance for the study of the development of Chinese characters, Sino-Tibetan, national community and other topics. In order to show the function of text knowledge graph, this paper implements an intelligent question answering system for text knowledge, which can use the data of text knowledge atlas to meet the query needs of users.

Keywords: knowledge graph, Chinese Character Library Project, question answering, word relationship, relation extraction

作者简介：安波（1986—），男，中国社会科学院民族学与人类学研究所副研究员，研究方向为民族语言处理、自然语言处理、机器学习。

基金项目：国家自然科学基金项目“知识增强的中文复述识别关键技术研究”（62076233）；国家新闻出版重大科技工程项目“中华字库工程应用平台研发”（0610−1041BJNF2328/23）。
收稿时间：2021 年 11 月 1 日。

0　引言

数千年来中华大地上诞生了很多不同的语言和文字，有些文字目前还在广泛地使用，如汉字、藏文等，也有些文字已经很少有人使用或濒临灭绝，如西夏文、金文等。这些文字对于研究中华民族的历史、文化、传统习俗和语言发展的规律具有重要价值。为了更好地保护和研究中华民族文字，且保护与挖掘这些文字的价值，近年来国家立项了多个与文字相关的重大文化科技工程，如“中华字库工程”“语保工程”[1]等。其中，中华字库工程是目前最为系统和完整地进行中华民族文字整理工作的工程。“中华字库工程”[2]汇集了包括中国科学院软件研究所、中版集团数字传媒有限公司、首都师范大学、北京中标中易电子信息技术有限公司等在内的国内从事文字处理的科研机构和公司，尝试利用信息化、数字化手段进行原始数据的采集、整理，建立准确的字际关系，形成标准的国际编码、字体文件、输入法、字符属性数据集等，并利用云计算技术将采集的成果数据对外提供字型云服务、属性云服务和输入法云服务，且基于这 3 种服务提供成果字库和其他数据的输入、显示和属性查询服务。“中华字库工程”的研发建设，为研究中华文字的产生、发展和演变提供数据基础，是提高中国文化“软实力”的一项重要举措，对于中华文明的普及与传播具有重大的战略意义。目前，已经收集整理了包含楷书汉字、古汉字、少数民族语言等语种并针对每种文字形成了对应的字体文件、编码空间、字图，并针对每个字符收集整理了相对应的属性信息。迄今已有一些在线的文字资源库可以使用，如“在线新华字典”等。然而，中华字库工程的数据存储采用的是传统文件系统（字体、输入法、字图）、关系型数据库（属性），不同类型的数据之间是隔离的，没有建立合理的关联，致使这些文字资源主要以现代汉字为主，缺少古汉字、少数民族文字的相关内容。因此无法很好地进行数据融合、查询、挖掘与推理等工作，不利于对文字整理成果数据的深入研究与应用。

知识图谱（Knowledge Graph）[3] 是一种语义网络（Semantic Network）[4]，最早是谷歌于 2012 年提出的且应用于搜索引擎[5]、推荐[6]等，取得了很好的效果。知识图谱使用结构化的网络结构描述概念、实体（Entity）及关系（Relation），实体之间通过关系建立关联，形成图状结构的知识结构，更加接近人类对世界的认知形式，提供了一种更好的组织、管理和使用知识的方式。知识图谱技术是指知识图谱的构建与应用的技术，融合了认知计算[7]、表示学习[8]、查询与推理等技术的交叉方向。知识图谱的构建通常包括本体（Ontology）[9]设计和知识抽取两个阶段。其中，本体定义了知识图谱的实体、关系的类型等信息，限定了知识图谱的范围，本体设计是由领域专家与知识图谱专家人工设计的。而知识抽取则主要基于模型、规则等计算机技术实现。知识抽取常见的任务包括知识表示与建模、知识表示学习、实体识别与链接[10]、实体关系识别[11]、事件抽取[12]等。为了进一步提升知识图谱的覆盖率，知识补全和知识融合[13−15]也是常用的技术。典型的知识图谱的应用有知识查询与推理、智能搜索、知识问答、对话系统等。

鉴于知识图谱能够支持知识的快速查询与推理，本文将针对中华字库工程研发建设中存在的问题，采用知识图谱的方式对字库工程的成果数据进行组织，挖掘数据中的知识，形成中华字库成果数据知识图谱，利用图数据库进行数据的存储，并基于实体识别、关系抽取和三元组抽取等步骤，构建一个包含 60 万个实体、1 000 万个三元组的大规模文字知识图谱。在该文字知识图谱的基础上，构建智能问答系统[16]，回答用户以自然语言形式提出的问题。

1　中华文字知识图谱构建

1.1　文字数据

通过前期积累，整理形成了超过 50 万字的成果字符，以及这些字符对应的编码、字形、输入法、扫描图档、属性等信息。在很大程度上满

足了新闻出版行业的文字标准化的迫切需求，加快了国家尤其是民族地区的信息化进程，增强了国家文化“软实力”。

然而，目前这些数据以字体文件、文本、关系数据库等不同的形式进行存储，数据之间没有建立关联，限制了成果数据的进一步挖掘和应用。针对该问题，本文提出使用更为先进的知识图谱的方式进行数据的重新组织和存储，建立数据之间的关联，为后续的字库数据挖掘和应用提供支撑。

1.2　知识图谱本体设计

本体是知识图谱的知识表示的基础，是知识图谱的抽象概念集合，是某个领域的概念框架，如“人”“地”“事件”等。知识图谱是在本体基础上的具象，在本体中通常包含实体概念和关系概念。如在人物知识图谱中，“人”是一个实体概念，“姚明”则是一个具体的实体。“妻子”是一个关系的概念，三元组（姚明，妻子，叶丽）则是关系“妻子”的一个具体实例。知识图谱最终是由三元组构成。三元组包含头部实体、关系和尾部实体3个元素。在前面的例子中，“姚明”为头部实体，“妻子”为关系，“叶丽”为尾部实体。通过这些三元组可以构建实体之间的相互关系，支持查询、推理等操作。

中华文字知识图谱主要涉及字符、语种、地域和文献四类实体。在中华文字知识图谱中对关系和属性不做严格地进行区分，均采用三元组的方式进行存储。如字符编码属性和字符之间异体关系均采用三元组进行存储。关系和属性在该知识图谱中不进行区分，均以关系的形式进行存储和使用。目前，中华文字知识图谱中的主要关系类型如表1所示。同时，中华文字知识图谱不仅针对现代汉语，而且针对古汉语、傣文、西夏文等少数民族文字。中华文字知识图谱包含的部分语种见表2。

1.3　知识图谱构建

基于上述本体定义的实体和关系类型，本文设计信息抽取系统实现从文本、图片、字体、数据库等信息中抽取文字对应的实体、关系和三元组构建知识图谱。文字知识图谱抽取的整体框架如图1所示。该系统主要包括数据预处理、实体抽取、关系抽取、实体链接和三元组抽取。主要对不同模态的数据进行数据处理，形成能够进行统一信息抽取的数据形式。然后再进行实体和关系的识别。同时，由于相同实体在不同的数据中有不同的模态和形式，需要对这些实体进行链

表1　中华字库关系/属性列表

类别	属性	属性数值类型	定义	示例
字形属性	总笔画数	数值	构成一个汉字或汉字部件的笔画的数目	如字符“部”的笔画数为：10
	全笔顺	文本	一个汉字按照书写时笔画的次序和方向确定的起笔至末笔的笔形序列	如“好”的笔顺：㇇丿一フ亅一
	字形	图像	构成方块汉字的二维图形	部
	字体文件	文本	该字符所在的字体文件	
	汉字描写序列	文本	对汉字字形结构的描写	如字符“好”的描写为：⿰女子
	康熙部首序号	数值	《康熙部首表》规定的主部首的序号	如“一”的序号为：1.0
	康熙主部首	文本	标注康熙部首的主部首	如“吃”的主部首为：“口”
	康熙附形部首	文本	标注康熙主部首的附形部首	如“土”的附形部首为：⼟
	康熙部首外笔画数	数值	部首外笔画数是整字总笔画数除部首实际笔画数之外的笔画数	如字符“好”的为：2

续表

类别	属性	属性数值类型	定义	示例
字形属性	康熙部首外笔顺	文本	部首外笔顺是整字全笔顺除部首实际笔顺之外的笔顺	如“好”的为： 
	规范主部首	文本	同一部首不同写法中具有代表性的书写形式	如“草”的主部首为：“艹”
	规范部首序号	数值	规范部首对应的序号	如“宀”的序号为：5.0
	规范附形部首	文本	附属于主部首的书写形式，有增繁、简化、变形和从属等多种形式	如“艹”的规范附形部首为：“艸”
	规范部首外笔画数	文本	部首外笔画数是整字总笔画数除部首实际笔画数之外的笔画数	如字符“好”的为：2
	规范部首外笔顺	文本	部首外笔顺是整字全笔顺除部首实际笔顺之外的笔顺	如“空”的规范部首外笔顺为：
读音属性	汉语拼音	文本	有规范的汉语拼音	如“空”的拼音：“kòng”
字义属性	释义	文本	对汉字的意义或用法的解释	如“乘”的释义为同“乘”，乘坐
	例证	文本	用于证明字形的出处、来源及音义的正确性、可靠性的文本证据	古樂府有飲馬長城窟行
	例证出处	文本	例证出处应包含正题名和所在页面信息	《古文苑 000408−000322 ）
关系属性	异体	文本	两个字符音义相同，但是字形不同	㚔与“侫”
	正误		一个字是另外一个字的错误用法	奮与奮
	通用		两个字在一定情况下可以通用	“胃”与其关系字“謂”之间存在通用关系
	繁体		一个字的繁体形式	鰛与鰪
	简体	文本	一个字的简体形式	简与簡
编码属性	字符编码	文本	字符对应的Unicode编码	如  对应的编码是 090002
	编码平面	数字	字符对应的编码平面	如  的编码平面为 09
语言属性	语种	文本	该字符所属的语种	属于甲骨文
	地域	文本	使用该语种的地理区域	
	时代	文本	该语种主要的使用年代	“宴台碑”属于正大元年（1224）

表 2　中华文字知识图谱包含的部分语种

编号	类型	语种	编号	类型	语种
1	汉字	楷书汉字	12	少数民族文字	印度式文
2		甲骨文	13		拉丁变体
3		金文	14		纳西东巴文
4		楚简帛书	15		契丹小字
5	少数民族文字	傣文	16		哥巴文
6		阿拉伯系文	17		水族
7		栗特系文	18		八思八文
8		南方仿汉文	19		突厥文
9		西夏文	20		傈僳文
10		女真文	21		彝文
11		契丹大字			

接，实现实体的归一化处理。如利用字符编码将文本形式、图片形式和字体文件中的同一字符关联起来，便于后续的三元组抽取。

利用上述信息抽取系统，本文抽取了超过60万个实体和超1 000万个三元组，构成了当前最大规模的中华文字知识图谱。本文利用Neo4j构建三元组数据，形成可视化的数据查询接口，具体的示例如图2所示。

2　基于文字知识图谱的智能问答

知识图谱将知识通过关系串联起来，形成网状结构，能够支撑很多智能应用，典型的应用主要包括智能问答、智能客服、推荐系统等。本文在文字知识图谱的基础上构建了一个智能问答系统，实现了在文字知识图谱上的智能问答，并实现了支持自然语言的方式进行文字相关关系/属性的查询。该系统的整体架构如图3所示。主要包括实体抽取、意图识别、知识图谱查询与推理和回复生成4个模块。本文基于这4个模块，利用中华文字知识图谱实现一个较为完整的智能问答系统，以完成针对该知识图谱的自动查询，满足该知识图谱的基本查询需求。

2.1　实体抽取

实体抽取模块的主要任务是从用户提出的问题中准确地识别实体词。在文字知识图谱中，实体词主要为字符、语种、地域和文献。本文基于Bert+CRF[17]（预训练语言模型+条件随机场）实现实体识别。该模型将实体识别转换为序列标注问题，通过给问题中的每个字预测不同的标签实现对实体的识别。具体的例子如图4所示，其中“S”表示为一个实体字，“O”表示为其他字符。该模型能够基于大规模预训练语言模型实现。预训练语言模型的微调机制使得该模型具有更好的鲁棒性，所需要的标注数据也更少，尤其适合于文字知识图谱这种缺少标注数据的模型。

2.2　意图识别

意图识别是在实体识别的基础上，判断用户的意图。如图4中的问题对应的意图是“异体字”。本文将意图识别建模为一个分类问题，将

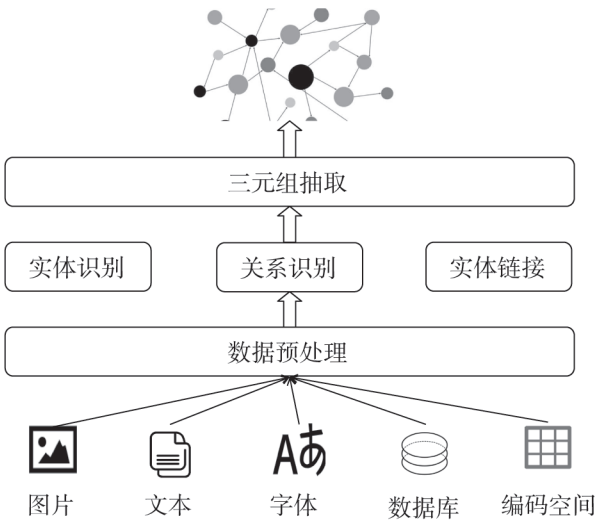

图1　文字知识图谱构建流程

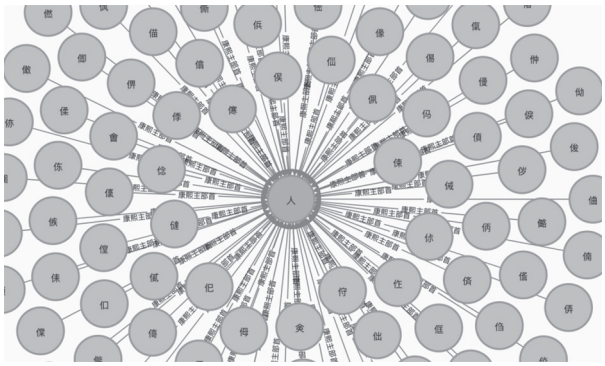

图2　以“人”作为部首的字符

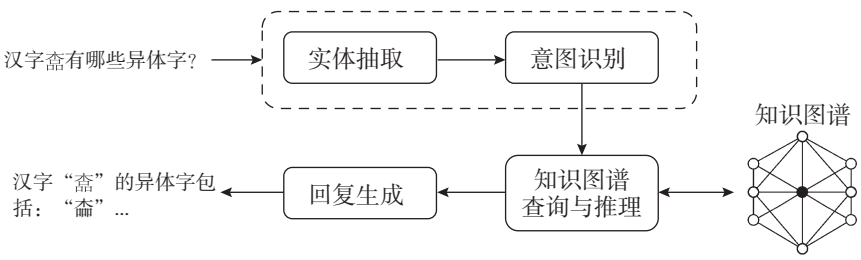

图3　基于文字知识图谱的智能问答

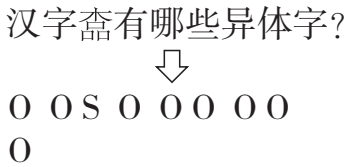

图 4　实体识别示意

用户的问题分类为不同的关系/属性的查询需求。本文利用Bert+FC[18]（预训练语言模型+全连接分类器）作为文本分类模型，用于实现意图识别。

2.3　知识图谱查询与推理

当识别了用户问题中的核心实体和意图后，可以生成知识图谱的查询命令，实现单跳和多跳的查询。所谓“单跳”指的是通过一次关系查询操作即可满足用户的需求。如图 4 的问题可以通过查询目标字的所有异体字来满足用户查询。“多跳”是那些不能通过单一的关系来满足用户需求的问题，如问题“字符顸对应语种有多少字?”需要通过“语种”和“字符列表”两个关系来满足用户查询。在多跳场景下识别用户的意图后与知识图谱中的关系组合进行相似度计算，最终得到查询的关系组合。本文基于SentenceBert[19]模型实现意图与关系组合的相似度计算。该模型在大规模语言模型的基础上计算句子/短语/词汇之间的相似度，能够充分利用上下文信息。

2.4　回复生成

基于上述模块，系统完成对用户问题的理解和知识的查询与推理，得到用户所需要的信息。为了能够以更自然的交互方式与用户进行问答交互，本文实现了基于规则结合文本生成模型T5[18]进行回复生成。

3　结语

本文在中华字库成果数据的基础，采用知识图谱技术和数据组织方式构建了一个大规模的中华文字知识图谱。文字知识图谱涵盖了楷书汉字、古汉字、少数民族文字等多种语言文字，包含了字形、字音、字义、编码空间等多种模态和类型的属性数据。本文采用实体抽取、关系识别、三元组抽取等技术构建了知识图谱。并利用深度学习在文字知识图谱的基础上实现了一个智能问答系统，满足了用户对文字知识查询的需求。

参考文献

[1] 丁石庆.北方人口较少民族语言词汇缺失现象探究:以语保工程民语资源调查材料为例[J]. 民族语文, 2020(3): 8.

[2] 孟忻. “中华字库”工程——中华民族有史以来规模最大的汉字及少数民族文字整理工作[J]. 中国索引, 2013(1): 2.

[3] 刘峤, 李杨, 段宏, 等. 知识图谱构建技术综述[J]. 计算机研究与发展, 2016, 53(3): 582.

[4] 李洁, 丁颖. 语义网, 语义网格和语义网络[J]. 计算机与现代化, 2007 (7): 38−41.

[5] 邱均平, 胡文君, 罗力.基于知识图谱的国际网络搜索引擎研究现状与前沿分析[J]. 图书情报工作, 2010, 54(24): 89.

[6] 常亮, 张伟涛, 古天龙, 等. 知识图谱的推荐系统综述[J]. 智能系统学报, 2019, 14(2): 207−216.

[7] 陈伟宏, 安吉尧, 李仁发, 等. 深度学习认知计算综述[J]. 自动化学报, 2017, 43(11): 1886−1897.

[8] 刘知远, 孙茂松, 林衍凯, 等. 知识表示学习研究进展[J]. 计算机研究与发展, 2016, 53(2): 247.

[9] 张德政, 谢永红, 李曼, 等. 基于本体的中医知识图谱构建[J]. 情报工程, 2017, 3(1): 35−42.

[10] 刘浏, 王东波. 命名实体识别研究综述[J]. 情报学报, 2018, 37(3): 329−340.

[11] 鄂海红, 张文静, 肖思琪, 等. 深度学习实体关系抽取研究综述[J]. Journal of software, 2019, 30(6): 1793−1818.

[12] 高强, 游宏梁. 事件抽取技术研究综述[J]. 情报理论与实践, 2013, 36(4): 114−117.

[13] 丁建辉, 贾维嘉. 知识图谱补全算法综述[J]. 收藏, 2018(1): 56−62.

[14] 杨飘, 董文永. 基于 BERT 嵌入的中文命名实体识别方法[J]. 计算机工程, 2020, 46(4): 40−45.

[14] 刘爽, 杨辉, 李佳宜, 等.面向多数据源的少数民族文化知识图谱构建[J].计算机技术与发展, 2021, 31(8): 191−197, 203.

[15] 刘烨宸, 李华昱.领域知识图谱研究综述[J].智能系统学报:计算机系统应用, 2020, 29(6): 1−12.

[16] 张云中, 祝蕊.面向知识问答系统的图情学术领域知识图谱构建: 多源数据整合视角[J].情报科学, 2021, 39(5): 115−123.
[17] 严佩敏, 唐婉琪. 基于改进 BERT 的中文文本分类[J]. 工业控制计算机, 2020, 33(7): 108−110.
[18] REIMERS N, GUREVYCH I. Sentence−bert: sentence embeddings using siamese bert−networks[J]. arXiv preprint arXiv: 1908.10084, 2019.
[19] DONG L, YANG N, WANG W, et al. Unified language model pre−training for natural language understanding and generation[J]. arXiv preprint arXiv: 1905.03197, 2019.

（上接第55页）

LAMOST二期巡天即将完成，LAMOST的观测数据管理和开放共享将进入崭新阶段。未来LAMOST数据团队将继续优化数据质量，提升数据的可追溯性，规范数据入库过程，完善数据发布系统，进行更深更广的国际化推广，打造国际权威的科学数据库系统，并继续推动LAMOST巡天数据在可视化与可视分析、人工智能、机器学习、科普教育等领域的应用。同时，国家天文科学数据中心将以国际化先进理念为指导，打造科学平台，实现数据与科研要素的深度融合，对天文数据的开放共享进行全新探索，推动科研模式的变革[11]。

致谢

郭守敬望远镜（Large Sky Area Multi−Object Fiber Spectroscopic Telescope, LAMOST）是中国科学院建设的国家重大科学项目。该项目资金由国家发展和改革委员会提供。LAMOST由中国科学院国家天文台运营和管理。本文得到了中国虚拟天文台、国家天文科学数据中心、中国科学院科学数据中心体系提供的数据资源和技术支持。感谢国家天文台—阿里云天文大数据联合研究中心对本项工作的支持。

参考文献

[1] CUI X Q , ZHAO Y H , CHU Y Q , et al. The Large Sky Area Multi−Object Fiber Spectroscopic Telescope (LAMOST)[J]. 云南天文台台刊, 1999, 12(S1): 257−260.
[2] 张海龙, 冶鑫晨, 李慧娟, 等.天文数据检索与发布综述[J].天文研究与技术, 2017, 14(2): 212−228.
[3] 崔辰州, 赵永恒, 赵刚, 等. 虚拟天文台的技术进展[J]. 天文学进展, 2002, 20(4): 302−311.
[4] LUO Ali, ZHANG Haotong, ZHAO Yongheng, et al. Data release of the LAMOST pilot survey[J]. Research in astronomy and astrophysics, 2012(9): 1243−1246.
[5] The Large Sky Area Multi−Object Fiber Spectroscopic Telescope (LAMOST)[N]. Inquiries of Heaven, 2012−08−24(2).
[6] LUO A , ZHAO Y , ZHAO G , et al. The first data release (DR1) of the LAMOST regular survey[J]. Research in astronomy and astrophysics, 2015, 15(8): 1095−1124.
[7] 崔辰州, 赵永恒. 中国虚拟天文台研发策略与重点[J]. 天文研究与技术—国家天文台台刊, 2004(3): 203−209.
[8] HE B, FAN D, CUI C, et al. The LAMOST data archive and data release[C]. San Francisco: Astronomical Society of the Pacific, 2016(512):153.
[9] SZALAY A, GRAY J, THAKAR A, et al. The SDSS sky server, public access to the sloan digital sky server data[C]. the 2002 ACM SIGMOD international conference. ACM, 2002.
[10] WILKINSON M D, DUMONTIER M, AALBERSBERG I J , et al. The FAIR guiding principles for scientific data management and stewardship[J]. Scientific data, 2016, 3(160018 (2016)): 167−172.
[11] CUI C, TAO Y, LI C, et al. Towards an astronomical science platform: experiences and lessons learned from Chinese virtual observatory[J]. Astronomy and computing, 2020(32):1−8.