doi:10.3772/j.issn.2095-915x.2023.05.007

英汉双语富媒体知识图谱构建工程研究
——以 CNS 英文期刊为例

开放科学（资源服务）标识码（OSID）

韦向峰[1,2] 缪建明[3] 张全[1] 袁毅[1]

1. 中国科学院声学研究所 北京 100190；
2. 富媒体数字出版内容组织与知识服务重点实验室 北京 100038；
3. 中国兵器工业信息中心 北京 100089

摘要：[目的 / 意义] 研究自动构建英汉双语富媒体知识图谱的方法和过程，为跨语言多模态知识图谱的自动构建提供借鉴参考，对及时获取最新英文科研成果、科技情报监测等具有重要意义。[方法 / 过程] 采用自顶向下和自底向上相结合的方法，先从顶层设计要抽取的主要实体、属性和关系，从底层非结构化文本数据进行分析抽取细粒度的实体和属性，对有歧义实体和跨语言实体进行实体对齐，对跨媒体的实体进行实体链接，用图数据库实现知识图谱的存储及应用。[局限] 未来需进一步提高细粒度实体的抽取正确率，对音视频媒体进行特征提取和内容自动识别。[结果 / 结论] 以 CNS（*Cell*、*Nature*、*Science*）等英文科技期刊网站为例，通过数据抓取、实体抽取、属性抽取、知识融合、跨媒体链接等过程，实现了英汉双语富媒体知识图谱的构建、存储和可视化展示。

关键词：富媒体；知识图谱；实体抽取；实体对齐；语步识别

中图分类号：G35; TP391

Research on the Construction of English-Chinese Bilingual Rich Media Knowledge Graph: A Case Study of *CNS* English Journal

WEI Xiangfeng[1,2] MIAO Jianming[3] ZHANG Quan[1] YUAN Yi[1]

1. Institute of Acoustics, Chinese Academy of Science, Beijing 100190, China;
2. The Key Laboratory of Rich-Media Knowledge Organization and Service of Digital Publishing Content, Beijing 100038, China;
3. Information Center of China North Industries Group Corporation Limited, Beijing 100089, China

Abstract: [Objective/Significance] It is of great significance for scientific and technological information monitoring and

基金项目 2022 年富媒体数字出版内容组织与知识服务重点实验室开放基金“基于英文科技出版物的跨语言富媒体知识工程研究”（ZD2022-10/01）。

作者简介 韦向峰（1976-），博士，副研究员，研究方向为人工智能、语音信号与信息处理、大数据与知识组织，Email：wxf@mail.ioa.ac.cn；缪建明（1977-），博士，研究员，研究方向为智能物流技术、智能评估技术；张全（1968-），博士，研究员，研究方向为人工智能、语义分析与处理；袁毅（1967-），学士，高级工程师，研究方向为计算机软件、计算机应用。

引用格式 韦向峰 , 缪建明 , 张全 , 等 . 英汉双语富媒体知识图谱构建工程研究——以 CNS 英文期刊为例 [J]. 情报工程 , 2023, 9(5): 84-96.

obtaining the latest English scientific research results in time, with researching the method and process of automatically constructing the English-Chinese rich media knowledge graph. It is also a meaningful experience for constructing cross-language and cross-media knowledge graph. [Methods/Processes] The approach that combines top-down and bottom-up methods is employed, starting with top-level design for extracting primary entities, attributes, and relationships. For fine-grained entities and attributes, analysis and extraction are performed from the bottom-up analyzing unstructured textual data. Ambiguous entities and cross-lingual entities require entity alignment, while cross-media entities require entity linking. By using a graph database, teh storage and its application of the knowledge graph can be implemented. [Limitations] Future works include further improving the accuracy of fine-grained entity extraction, extracting features and automatically recognizing content for audio and video media. [Results/Conclusions] Taking CNS (*Cell*, *Nature*, *Science*) and other English scientific and technological journal websites as an example, this paper successfully constructed a bilingual English-Chinese multimedia knowledge graph through data scraping, entity extraction, attribute extraction, knowledge fusion, cross-media linking.

Keywords: Rich media; knowledge graph; entity extraction; entity alignment; moves recognition

引言

知识图谱本质上是一种具有有向图结构的语义网络知识库，其中图的结点代表实体或概念，图的边代表实体或概念之间的各种语义关系[1]。知识图谱作为一种形式化的知识表示方法，具有结构化、图形化、可推理等优点，被广泛应用于搜索引擎、自然语言处理、情报分析、智能客服等领域。目前通用领域有 DBpedia[2]、YAGO[3]、FreeBase[4] 等大规模百科知识类英文知识图谱，中文通用知识图谱有 CN-DBpedia[5]、Zhishi.me[6] 以及 OpenKG.CN 平台[7]。但是以百科知识为基础的知识图谱无法满足专业化知识推理、精细化应用场景和特定领域需求，因此垂直领域的专用知识图谱获得了广泛研究和发展，例如学术文献领域的知识图谱 SciGraph[8]、OAG（Open Academic Graph）[9]、AMiner[10]、AceKG[11]。随着知识图谱的发展以及智能时代的来临，知识图谱需要融合不同的语言、包含更多模态或媒体的知识内容，例如文本、图片、音频、视频等富媒体。如何自动构建跨语言、跨模态的垂直领域知识图谱，成为知识图谱自动构建技术研究的热点之一。

知识图谱的自动构建是一个复杂过程，涉及从不同数据源中抽取、整合和表示知识。一般而言，首先需要从网页或结构化数据源进行数据采集与抽取，从中提取实体、关系和属性等信息；接着对从不同源头获得的数据进行数据清洗，包括去除重复项、修复拼写错误、统一格式等；然后使用自然语言处理等技术进行实体抽取、属性抽取、关系抽取等，获取知识图谱的实体、属性和关系；之后通过实体对齐、关系合并等技术对知识图谱进行知识融合，以确保知识图谱的一致性；最后需要进行知识表示和知识存储，一般采用三元组表示知识图谱，使用图数据库存储知识图谱。

本文将以构建面向世界一流英文科技期刊的跨语言多模态知识图谱为例，探讨如何从单一语种的英文科技期刊网站中自动获取文本、图片、音视频等富媒体数据信息并翻译为中文，同时探索建立生命科学、医学、化学等领域的细粒度跨语言富媒体知识图谱构建技术，实现富媒体实体关系的抽取、关联、跨语言映射和存储。本文的跨语言多模态知识图谱技术和方

doi:10.3772/j.issn.2095-915x.2023.05.007

法，可为其他垂直领域的知识图谱构建提供参考和借鉴；所构建的英汉双语富媒体知识图谱可为下游任务如跨语言文献信息检索和推荐、科学研究热点分析、科技情报监测等提供重要支撑。

1 构建方法

知识图谱的构建方法可以分为自顶向下和自底向上两种方法[12]。自顶向下方法一般是借助结构化的网站数据源或者领域专家的先验知识来构建知识图谱，通常从事先定义的领域本体开始，包括实体类别、属性和关系等，通过规则和规范来指导实体和关系的抽取和构建；自底向上方法一般是通过自动化算法和技术从大规模的非结构化数据中抽取实体和关系，然后逐步构建知识图谱，通常包括实体抽取、关系抽取和实体链接等步骤。

本文采取自顶向下和自底向上相结合的方法。首先根据 CNS 等英文科技期刊网站的结构特点设计了期刊、科技论文、论文作者、科研机构、图片、音频、视频等实体，并设计了各个实体的属性和它们之间的关系，例如科技论文的摘要、发表日期等属性，这些实体关系的设计以属性关系为主；其次，对于非结构化的论文摘要文本内容，采用自动语步识别技术获取摘要的背景、方法、结果和结论等语步，作为科技论文实体的细粒度属性；第三，从非结构化的论文摘要文本中抽取学科领域的专业术语等实体，这些实体之间的关系以实体共现的语句谓词为主；第四，将英文知识图谱中的实体、属性和关系翻译为中文，然后进行单语言知识图谱的实体消歧、跨语言知识图谱的实体对齐；第五，将图片、音频和视频等富媒体信息内容进行实体链接，链接到科技论文、主题等实体；最后，把获得的跨语言富媒体知识图谱存储到图数据库中，以便进行后续的知识图谱检索和相关应用。

对于知识图谱的构建而言，一般需要经过实体抽取、属性抽取、关系抽取等知识抽取的处理，然后进行实体对齐、属性对齐等知识融合的处理。对于跨语言和跨模态的知识图谱，还需要进行跨语言的实体对齐、跨模态的实体属性映射等处理。

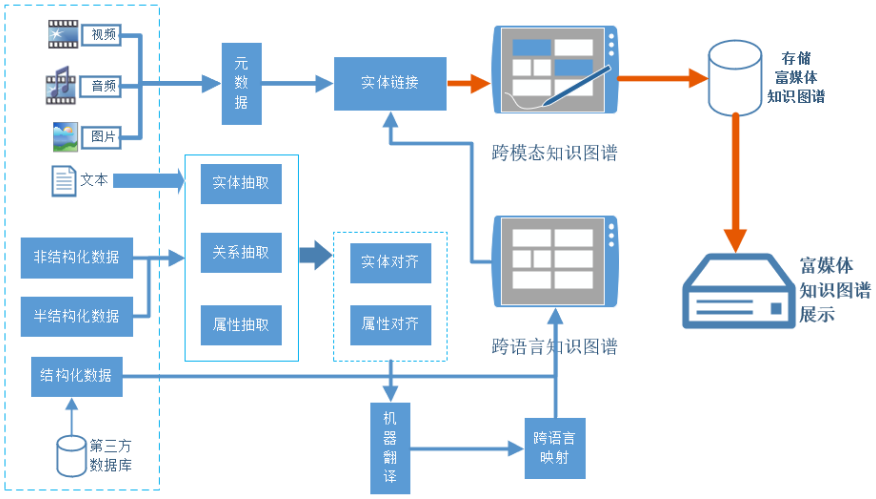

图 1 英汉双语富媒体知识图谱自动获取系统框架

如图 1 所示，本文的英汉双语富媒体知识图谱构建工作主要包括确定数据来源、实体抽取、属性抽取、实体对齐、跨语言映射、跨媒体链接等处理步骤或阶段，最后对所构建的多模态跨语言知识图谱进行存储及应用展示。

2 数据来源和实体设计

本文的数据来源是世界一流的英文科技期刊网站，主要以三大顶刊 *Cell*、*Nature*、*Science*（CNS）及其子刊为主，同时包括生命科学、医学、化学等学科的其他知名英文科技期刊（见表 1）。利用 Python 编写的软件程序，可自动获取这些网站中期刊、科技论文、论文作者、科研机构等实体信息，以及相关的图片、音频和视频实体信息。

本文设计的主要实体及其属性如表 2 所示。在本文设计的知识图谱实体、属性和关系中，主题既是实体也是属性，实体之间通过主题或属性形成关联关系。主题分为生命科学、化学、医学、综合等学科类别，除主题外知识图谱中的其他实体都可以归属于某个学科类别的主题。其中，科技论文实体处于核心重要位置，科技论文不仅关联文本媒体形式的期刊、论文作者、科研机构，而且关联图片媒体、音频媒体和视频媒体。对于音频媒体内容，主要通过 *Scientific American*（科学美国人）杂志中

表 1　数据来源的主要期刊信息

英文名称	中文名称	网址	主题
Cell	《细胞》	https://www.cell.com/	综合
Nature	《自然》	https://www.nature.com/	综合
Science	《科学》	https://www.science.org/	综合
Cell Metabolism	《细胞 - 代谢》	https://www.cell.com/cell-metabolism/home	生命科学
Cell Stem Cell	《细胞 - 干细胞》	https://www.cell.com/cell-stem-cell/home	生命科学
Immunity	《免疫》	https://www.cell.com/immunity/home	生命科学
Cancer Cell	《癌细胞》	https://www.cell.com/cancer-cell/home	生命科学
Nature Immunology	《自然 - 免疫学》	https://www.nature.com/ni/	生命科学
Nature Genetics	《自然 - 遗传学》	https://www.nature.com/ng/	生命科学
Nature Methods	《自然 - 方法学》	https://www.nature.com/nmeth/	生命科学
Nature Cell Biology	《自然 - 细胞生物学》	https://www.nature.com/ncb/	生命科学
Nature Biotechnology	《自然 - 生物技术》	https://www.nature.com/nbt/	生命科学
Nature Neuroscience	《自然 - 神经科学》	https://www.nature.com/neuro/	生命科学
Nature Medicine	《自然 - 医学》	https://www.nature.com/nm/	医学
JAMA Network	《美国医学会杂志》	https://jamanetwork.com/	医学
The New England Journal of Medicine	《新英格兰医学杂志》	https://www.nejm.org/	医学
The British Medical Journal	《英国医学会杂志》	https://www.bmj.com/	医学
The Lancet	《柳叶刀》	https://www.thelancet.com/	医学
Nature Chemistry	《自然 - 化学》	https://www.nature.com/nchem/	化学
Journal of the American Chemical Society	《美国化学会志》	https://pubs.acs.org/journal/jacsat	化学
Angewandte Chemie	《德国应用化学》	https://onlinelibrary.wiley.com/journal/15213773	化学

doi:10.3772/j.issn.2095-915x.2023.05.007

表 2　主要的实体和属性

实体	属性
主题	主要用于定义实体间的关系，例如所属学科、领域等
期刊	主题、期刊名称、网址、期刊影响因子、语种、刊名缩写等
科技论文	主题、DOI、标题、摘要、发表期刊、出版日期、期、卷等
论文作者	主题、姓名、邮箱、关联的科研机构、ORCID 等
科研机构	主题、机构名称、机构简称、地址等
图片	关联的科技论文、主题、图片文件地址、文件大小、像素、语义标签、文本描述等
音频	关联的科技论文、主题、音频文件地址、文件大小、时长、格式、比特率、采样频率、文本描述等
视频	关联的科技论文、主题、视频文件地址、文件大小、时长、格式、帧速率、帧高度、帧宽度、分辨率、文本描述等

的播客网站（https://www.scientificamerican.com/podcasts/）进行自动获取。利用数据获取与解析的自动化工具，本文从 22 个英文期刊网站中获取了期刊的历史科技论文及其相关属性，其中科技论文共 13135 篇，科技论文中的图片共 8738 张，科技论文中的关联视频共 911 个，音频媒体内容（包括音频文件及其对应的文本内容）共 4638 个。在富媒体知识图谱中音频实体可以通过主题或音频文本内容中出现的科技期刊名称与科技论文实体进行关联。

3　实体抽取

实体抽取是知识图谱构建的关键环节之一，其主要任务是从非结构化的文本中自动识别出各种实体，并将其标注出来。对通用领域而言，实体抽取可以从文本中识别出特定名称的实体，如人名、地名、机构名等；对垂直领域而言，实体抽取可以从文本中识别出具有特定专业含义的术语，例如医学领域的疾病名称、药品名称、医学检查方法等。对于表 2 中论文作者、科研机构等大多数实体，很容易通过半结构化的网页信息内容抽取得到。而在科技论文的非结构化摘要文本中，包含有各个学科领域的众多术语或实体，如何从摘要文本中自动识别抽取得到相关学科领域的术语实体是在构建知识图谱时需要研究解决的一个问题。

由于科技论文涉及众多细分的学科和领域，各学科领域的专业术语的特点和规律并不相同，本文仅仅讨论针对生物医学领域的化学药物名称的实体抽取。生物医学领域的化学药物名称有其自身的特殊性，这使得化学药物名称的实体抽取正确率远低于通用领域的实体抽取正确率。这些特殊性在于以下几个方面：（1）名称往往很长，例如“sodium dodecyl sulphate polyacrylamide”。（2）命名方式多样，没有统一标准。有的采用国际理论和应用化学联合会制定的命名方式如“8-O-trans-cinnamoyl caryoptoside”，有的采用惯用名如“captafol”，有的采用简称如“PCAHs”等。（3）歧义性缩写，化学药物名称经常出现缩写，且缩写没有统一规律。（4）不断出现的新的化学药物名称，仅仅依靠词典规则方法难以将其全部覆盖。

实体抽取的方法可以分为：基于词典的方法、基于规则的方法、基于统计学习的方法和基于深度学习的方法。就知识图谱实体抽取的效果而言，基于深度学习的方法要优于其他方

法。目前基于 Bi-LSTM-CRF 模型的深度学习方法在化学药物语料库上取得的实验结果比传统的基于 CRF 的统计机器学习方法的结果要好，成为了主流的化学药物名称实体抽取方法。例如，一种基于 CNN-Bi-LSTM-CRF 模型用于生物医学领域的实体识别方法 [13]，先利用 CNN 学习单词字符级向量，然后使用 Bi-LSTM-CRF 模型进行实体识别，在 BioCreative II GM 和 JNLPBA2004 生物医学语料上取得了较好的结果，但是仍有进一步改进的空间。因为这些基于 Bi-LSTM-CRF 模型的深度学习方法都是在句子内部进行实体抽取，容易出现抽取得到的实体不一致的问题，也就是说，同一篇文档中提及的相同实体由于上下文不同可能会被标注成不同的实体。

为了缓解这种文档内抽取实体不一致的问题，本文将注意力机制引入到 Bi-LSTM-CRF 模型中，将文档作为模型的输入单元，通过注意力机制来捕获文档全局信息，使同一篇文档不同句子中的相关词被视为相互依赖的标签，从而进一步提升深度学习模型抽取化学药物名称实体的效果。首先定义输入文档为 $\mathbf{D}=\{S_1, S_2, \ldots, S_i, \ldots, S_m\}$，由 m 个句子组成，一个句子定义为 $\mathbf{S}=\{w_1,w_2, \ldots, w_i, \ldots, w_n\}$，由 n 个词语组成。文档需要经过嵌入层（词向量）和一个 Bi-LSTM 层，然后进入到一个新的注意力层，以捕获文档级别相关词语的依赖信息。在注意力层引入一个注意力矩阵 A 来计算当前目标词语和文档中所有词语的相似度得分。注意力矩阵 A 中的权重值是第 t 个词 w_t 在文档全文范围内对应第 i 个词 w_i 表所分配的注意力权值，利用公式（1）进行计算。

$$\alpha_{ti}=\frac{exp(score(w_t,w_i))}{\sum_{k=1}^{n}exp(score(w_t,w_k))} \quad (1)$$

其中，$score(w_t,w_i)$ 是词语 w_t 和词语 w_i 之间的相似度得分函数，可通过词语对应的词向量之间的欧式距离、余弦距离或曼哈顿距离计算得到。为了获取文档级信息并学习注意力权值高的词语信息，将得到的注意力权值对 Bi-LSTM 层的输出进行加权求和得到文档全局向量。然后，将此全局向量和 Bi-LSTM 层的输出进行拼接，使用 Tanh 函数作为激活函数，得到注意力层的输出。最后，使用 CRF 层来评估输出标签之间的依赖关系，使用 Softmax 函数计算文档 **D** 中标签序列的条件概率。经过标注语料库的训练后，基于注意力机制的 Bi-LSTM-CRF 模型可以对生物医学领域的化学药物名称进行文档级的标签标注，实现化学药物名称的实体抽取。

本文采用 BioCreative 评测发布的 CDR 语料库作为训练数据集和测试数据集。CDR 原始的语料库包含 1500 个 PubMed 论文摘要，分为训练集（摘要 500 篇）、开发集（摘要 500 篇）、测试集（摘要 500 篇）。本文把原始语料库的训练集和开发集合并作为本文的训练数据集，把原始语料库的测试集直接作为本文的测试数据集。训练基于注意力机制的 Bi-LSTM-CRF 模型时的一些超参数设置如下：词向量维度设为 50，字符向量维度设为 25，字符级 Bi-LSTM 神经单元大小设为 25，词语级 Bi-LSTM 神经单元大小设为 100，学习率设为 0.001，优化函数采用随机梯度下降函数 SGD。实验结果表明，基于注意力机制的 Bi-LSTM-CRF 模型比一般的基于句子的 BiLSTM-CRF 模型性能要高 0.7 个百

doi:10.3772/j.issn.2095-915x.2023.05.007

分点，而且可以减少实体抽取中的不一致错误。

4 属性抽取

对文本媒体的实体属性而言，属性抽取是指从文本或数据中识别和提取实体的属性信息的过程。属性抽取的目标是从结构化或非结构化的数据中自动识别和提取实体的属性。表 2 中的大多数文本媒体实体的属性，可通过英文期刊网站内容的半结构化文本内容信息获取，使用规则和自动化程序的方法获得科技论文的标题、科技论文的摘要、科技论文的发表日期、论文作者的电子邮箱等属性。

对于科技论文实体的非结构化的摘要文本属性，可以进一步细化为背景、方法、结果和结论等语步属性，这样可以很方便地检索和查询科技论文在做哪方面的研究、使用了什么方法、得到了什么结论，丰富整个知识图谱的属性粒度，提高论文信息获取和阅读的效率。为了获取和建立更细粒度的属性知识，需要对非结构化的论文摘要文本内容数据进行语步自动识别。语步是指为实现语篇整体交际目的，语篇中具有某种特定交际功能的部分[14]。例如，科技论文中的摘要文本可以细分为背景、方法、结果、结论等语步，目的是向读者简要全面地介绍整个论文的概貌（为什么做、如何做、做的结果）。有的英文科技期刊要求论文作者自己把论文摘要划分出语步，如 *The NEW ENGLAND JOURNAL of MEDICINE*（新英格兰医学杂志）网站中的论文就给出了论文摘要的背景、方法、结果和结论的固定格式及文本内容。但大多数科技论文的摘要文本只是一段文字文本，没有专门划分出语步，需要利用深度学习等模型算法自动识别出科技论文摘要文本的语步。

科技论文摘要的语步属性抽取本质上就是文本自动分类，其主要方法有基于规则的方法、基于浅层机器学习模型的方法和基于深度学习模型的方法。基于深度学习模型的语步识别可充分利用句子潜在语言学特征，无论是在通用性还是效果方面均有所改善，是目前较为高效且主流的摘要语步识别与分类方式，主要采用的模型有长短期记忆网络（LSTM）、双向 LSTM、BERT 等模型。本文的语步自动识别模型是在基于 BERT 模型的 SciBERT[15] 的基础上构建得到，基本参数采用了 SciBERT 的预训练参数。SciBERT 模型的训练语料来自文献检索网站 Semantic Scholar 的随机采样的论文全文，共计 114 万篇论文（其中 18% 来自计算机领域，82% 来自生物医学领域）。模型的深度神经网络（Transfomer 模块）层数为 12，隐藏状态的尺寸为 768，自注意力头数为 12，共有 1.1 亿个参数。本文使用《新英格兰医学杂志》期刊网站获得的论文摘要进行训练，并对分类器以及 BERT 模型的最后两层网络的参数进行优化，实现损失函数的最小化。从《新英格兰医学杂志》期刊网站获得的论文 615 篇，按照 8:2 比例（即 80% 共 492 篇摘要文本作为训练样本，20% 共 123 篇摘要文本作为测试样本 ）进行 SciBERT 模型和本文优化模型的实验，得到如表 3 所示的实验结果。

表 3 摘要文本语步自动识别的结果

模型	语步	准确率	召回率	F1 值
SciBERT	背景	76.15%	74.81%	75.48%
	方法	68.24%	65.63%	66.94%
	结果	73.36%	70.05%	71.71%
	结论	75.83%	71.58%	73.71%
优化模型	背景	80.37%	79.62%	80.00%
	方法	78.52%	76.91%	77.72%
	结果	79.43%	78.25%	78.84%
	结论	85.26%	83.96%	84.61%

5 知识融合

知识融合是指将知识图谱中不同来源和不同结构的知识数据进行整合和合并，确保知识图谱中的数据一致性和准确性。本文的知识融合主要是实体对齐，主要探讨科技论文中抽取到的论文作者的实体对齐问题，以及分别从英文文本和中文文本抽取得到的实体的跨语言实体对齐问题。

5.1 作者实体对齐

本文中的作者实体对齐是需要判断来自不同科技论文的同名同姓的作者是否是真实物理世界中的同一个人。科技论文中产生同名同姓作者的原因主要有：（1）同一个人在同一个科技期刊或者在不同的科技期刊发表了不同的论文；（2）两个或者多个不同的人具有相同的姓名，并且在同一个科技期刊或者在不同的科技期刊发表了不同的论文；（3）由于英文的姓和名排列顺序与中文的排列顺序不一致，或者不同期刊对论文作者姓和名排列顺序的不一致，导致姓和名颠倒的两个不同姓名的论文作者可能指向同一个人。本文共获得论文作者 165586 个（未进行去重处理），去重处理后获得 124998 个；经过去重处理后获得的科研机构实体共 97238 个。

从 CNS 等科技期刊网站的文本内容中可以提取出论文作者的隶属单位、邮箱或者 ORCID（Open Researcher and Contributor ID，即开放学术出版物及学术产出的作者标识符，网站：https://orcid.org/）。其中 ORCID 是全世界范围内唯一 16 位身份识别码，相当于科研工作者在学术领域的身份证。因此，如果文本媒体的知识图谱中已经存在和待构建的论文作者实体姓名相同的论文作者实体，第一步先判断已经存在的论文作者和待构建的论文作者这两个实体的 ORCID 是否相同，若相同则视为同一个实体，否则进行下一步的判断；第二步，判断这两个实体的电子邮箱地址是否相同，若相同则视为同一个实体，否则进行下一步的判断；第三步，判断这两个实体的隶属单位是否相似或相同，若相同或者相似度达到一定阈值则视为同一个实体（相似度采用编辑距离进行计算），否则进行下一步的判断；第四步，查找知识图谱中是否有与待构建的论文作者的姓和名顺序颠倒的论文作者实体，如果有则执行第一步到第三步的处理步骤。本文从 13135 篇科技论文中抽样得到 2974 篇科技论文，作者总数为 50982（不考虑姓名是否相同），姓名唯一的作者有 36085 个，因此作者重名率为 29.22%。对重名的论文作者进行实体对齐处理后，其实体对齐结果如表 4 所示。

表 4 论文作者实体对齐处理结果的准确率

对齐方法	作者数量	占比	准确率
ORCID	609	4.09%	100%
电子邮箱地址	738	4.95%	100%
隶属单位	13100	87.94	90%
姓名颠倒	63	0.42%	85%
无法判定	387	2.6%	——
合计	14897	100%	93.75%（平均）

5.2 跨语言实体对齐

从英文科技期刊网站中获取的实体文本数据都是英文实体或属性，需要将这些英文实体或属性的内容翻译为中文。自动化的文本机器翻译技术可以将英文文本内容自动翻译为中文

doi:10.3772/j.issn.2095-915x.2023.05.007

文本内容，且具有较高的可懂率，这不仅有助于科研人员方便快捷地获取英文科技论文信息，也有助于英汉双语富媒体知识图谱的构建。本文使用成熟的机器翻译软件将科技论文的标题、摘要等文本数据从英文翻译为中文，从而得到中文知识图谱中的科技论文属性信息。利用百度翻译开放平台，通过 Python 程序语言调用其 API 接口实现了从英语到汉语的文本翻译。使用本文第 3 节介绍的实体抽取方法可以分别抽取得到中文文本和英文文本中的各种实体，包括通用领域的一般实体和学科领域的术语实体。使用跨语言实体对齐技术将这些实体进行跨语言对齐，不仅可以丰富知识图谱中实体之间的链接关系，也可以为知识图谱的应用如跨语言文献检索提供必要的知识基础。

跨语言的实体对齐与单语言的实体对齐是相似的，都可以采用基于知识表示的 TransE[16]、TransH[17] 和 TransR[18] 等模型。这些模型主要是将知识图谱嵌入到词向量空间，就可以对实体的相似度进行计算和处理。例如 TransE 模型的基本思想是：考虑到知识图谱由三元组（$\boldsymbol{h}$，$\boldsymbol{l}$，$\boldsymbol{t}$）组成，$\boldsymbol{h}$ 和 $\boldsymbol{t}$ 分别表示不同的实体，$\boldsymbol{l}$ 用于表示两个实体间的关系，因此可以将头实体 $\boldsymbol{h}$ 的向量和关系 $\boldsymbol{l}$ 的向量进行 L2 范式计算，所得结果应近似于尾实体 $\boldsymbol{t}$ 的向量。这种训练学习的方法将知识图谱的图表示中实体间的关系考虑在内，因此所得到的实体向量表示也包含了实体间的关系信息。TransE 模型的训练流程如下：（a）根据所设置的维度超参数，对实体和关系的向量进行随机初始化。具体方法是在均匀分布 $\left(-\frac{6}{\sqrt{k}},\frac{6}{\sqrt{k}}\right)$ 中随机采样，其中 k 是向量的维度，然后对实体和关系的随机初始化结果进行归一化，即除以向量的 L2 范数。（b）根据所设置的 Batch 超参数 b，从训练集 S=（$\boldsymbol{h}$，$\boldsymbol{l}$，$\boldsymbol{t}$）中，构成正样本。针对每一个正样本，均替换其中的头实体 h 或尾实体 t，构成负样本集 S=（$\boldsymbol{h'}$，$\boldsymbol{l}$，$\boldsymbol{t'}$）。（c）根据 TransE 模型的损失函数，依次选取正负样本对模型进行训练。损失函数如公式（2）所示，其中 $\gamma>0$，是一个边界超参数。若希望降低 f_{loss} 的值，则需要降低正样本的距离 d（$\boldsymbol{h}$，$\boldsymbol{l}$，$\boldsymbol{t}$）并增加负样本的距离 d（$\boldsymbol{h'}$，$\boldsymbol{l}$，$\boldsymbol{t'}$）。距离函数 d（$\boldsymbol{h}$，$\boldsymbol{l}$，$\boldsymbol{t}$）为 L2 范数，即公式（3）。

$$f_{loss}=\left[\gamma+d\left(h,l,t\right)-d\left(h',l,t'\right)\right] \quad (2)$$

$$f_{score}=\left\|h+l-t\right\|_2 \quad (3)$$

为了检验 TransE、TransH 和 TransR 在跨语言知识图谱中的实体对齐能力，本文选择公开数据集 DBpedia 的英文、中文两种语言的知识图谱作为训练数据集和测试数据集（具体数量见表 5）。DBpedia 中的一部分跨语言三元组之间已经建立了跨语言链接，具有英文数据和中文数据之间的对齐集。训练时的一些超参数配置如下：向量空间维数 m, n = 75、学习率 λ= 0.01、训练次数 epoch=400。每次更新参数后，正则化实体和关系的嵌入向量的 L2 范数为 1。TransE、TransH 和 TransR 模型的训练结束后，使用测试集对各个模型进行测试，测试时使用 Hits@10 作为评价指标。对测试集中的每一个实体对（$\boldsymbol{e}_1$，$\boldsymbol{e}_2$），其中 $\boldsymbol{e}_1$ 为英文知识图谱中的实体，$\boldsymbol{e}_2$ 为中文知识图谱中的实体。对于每一个 $\boldsymbol{e}_1$，在中文知识图谱中寻找与之相似度最高的 10 个实体 $\{ee_1, ee_2, \cdots, ee_{10}\}$，那么 $e_2 \in \{ee_1, ee_2, \cdots, ee_{10}\}$ 的平均比例即为从英文

知识图谱对齐到中文知识图谱的 Hits@10。

表 5 跨语言实体对齐处理结果的准确率

对齐方法	对齐方向	三元组对数量	实体数量	Hits@10
TransE	英 - 中	67310	7865	71.35%
TransH	英 - 中	67310	7865	76.82%
TransR	英 - 中	67310	7865	73.91%

实验数据集中，英文知识图谱到中文知识图谱的对齐三元组对个数为 67310 个，实体对齐对的个数为 7865 个。数据集按照 8:2 的比例分为训练数据集和测试数据集，TransE、TransH 和 TransR 模型经过训练后，在测试数据集上的实验结果如表 5 所示。从表 5 可以看出，TransH 模型的实体对齐效果较好。因此，本文使用 TransH 模型对从英文科技期刊获得的跨语言知识富媒体图谱中的英文文本实体到中文文本实体的对齐进行了抽样测试，采样了 1000 个英 - 中实体对进行了跨语言的实体对齐测试，Hits@10 准确率为 77.26%。

6 跨媒体实体链接

除了文本媒体形式的实体之外，本文还自动获取了与科技论文实体相关的图片、音频、视频等媒体形式的实体。科技论文实体与这些富媒体实体之间需要进行链接，建立实体之间的关系，以便进行知识图谱的检索、可视化和下游应用。同时，图片、音频和视频等富媒体实体自身还具有一些用于描述这些实体的属性，也需要进行属性的抽取。

图片媒体数据是根据科技论文的 URL 地址去获取的，可以很方便地将图片实体关联到科技论文实体（把科技论文的 URL 地址作为唯一 ID）。图片实体的属性除了所关联的科技论文之外，还有自身对应的文件路径、文件大小、尺寸大小（按像素计算的长 x 高）、对应的文本描述等属性。视频媒体数据与图片媒体类似，是根据科技论文的 URL 地址去获取的，因此可以很方便地将视频实体关联到科技论文实体。为了方便后续的数据处理和应用，还需要提取视频实体的属性数据，包括视频文件路径、视频文件大小、视频格式、时长、帧宽度、帧高度、帧速率、数据速率、总比特率、宽高比等。

音频媒体数据主要来自 *Scientific American* 杂志的播客频道，其内容是介绍科学知识或学术期刊上的一篇论文，音频长度一般不超过 60 秒，涵盖了生命科学、天文学、物理学、地球科学等多个领域的前沿研究成果和科技进展。该播客频道网站还提供了音频对应的转写文本，可用于对照阅读或文本分析。为了将音频实体关联到科技论文实体，本文利用主题作为中间实体媒介或者音频实体与科技论文实体之间的主题关系，把相同主题的音频实体和科技论文实体进行实体关系的链接。科技论文的主题分为生命科学、化学、医学、综合类等，音频的主题提供用户指定主题和自动标注主题两种方式。自动标注主题采用基于 LDA（Latent Dirichlet Allocation）模型的主题分析方法，该方法主要利用主题词语在文本中的分布，获得“主题 - 词语”分布和“文本 - 主题”分布，然后通过极大似然估计方法确定文本对应的主题。此外，对于音频实体本身，还需要提取音频实体对应的文件路径、文件大小、采样频率、比特率、音频时长等属性。

doi:10.3772/j.issn.2095-915x.2023.05.007

7 知识图谱存储及应用

知识图谱的形式化表达一般归结为“< 实体 1，关系，实体 2>”和“< 实体，属性，属性值 >”等三元组。这些三元组数据可以用关系数据库、语义网 RDF 描述语言、图数据库等进行存储。本文采用图数据库 Neo4j 进行知识图谱的存储和可视化。首先，需要创建科技论文实体，其属性包括主题、URL、DOI、标题、作者、摘要、关键词、来源期刊、来源期刊的卷期、发表日期、起始页码、结束页码等。其次，创建论文作者实体，其属性包括关联的论文 URL、姓名、科研机构、电子邮件、ORCID 等。第三，创建其他实体，如图片实体、音频实体、视频实体、期刊实体，以及这些实体的属性，例如图片或音视频对应的富媒体文件的存储路径。第四，建立实体之间的关系，主要有科技论文与论文作者的关系（HAS_AUTHOR）、科技论文与图片的关系（HAS_PICTURE）、科技论文与音频的关系（通过主题关联）、科技论文与视频的关系（HAS_VIDEO）、科技论文与期刊的关系（HAS_JOURNAL）。最后，利用 Neo4j 的数据库服务和查询界面实现整个知识图谱的可视化查询。

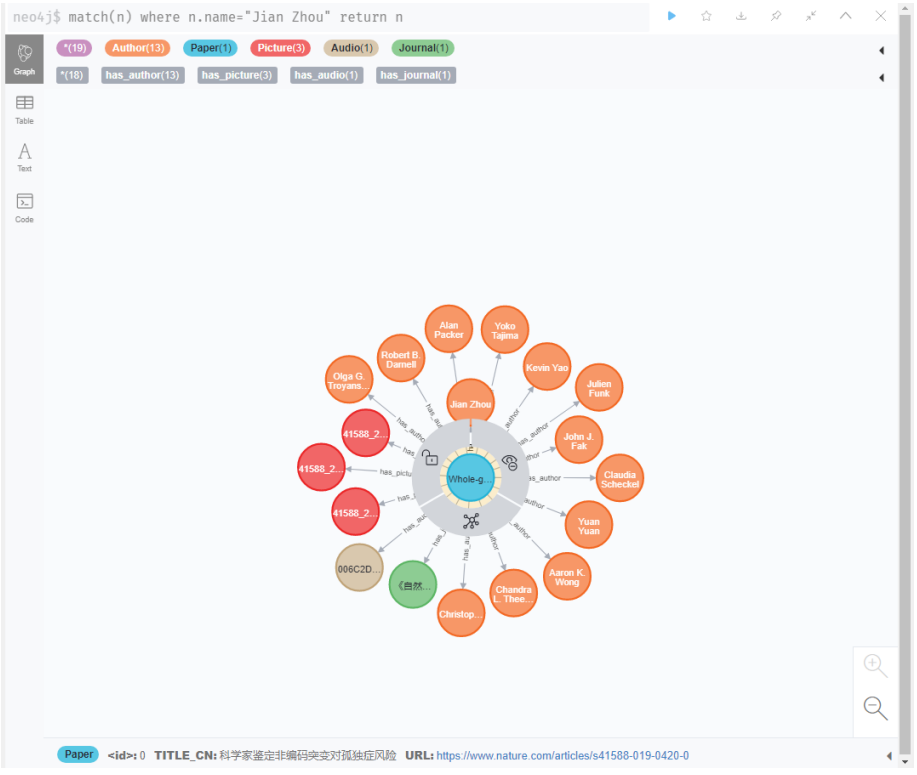

图 2 知识图谱的可视化查询界面

图 2 为在 Neo4j 的查询界面中查询姓名为“Jian Zhou”的作者后返回的知识图谱中的实体及其关系，中间的核心实体为科技论文实体，周围的关联节点为其属性或者具有关联关系的论文作者、图片、音频、视频或期刊等实体。

使用本文的面向 CNS 英文期刊的数据获取与解析、英语文本自动翻译、图片和音视频数

据处理等技术，还可以实现英语科技期刊科技论文网页实时自动转换为相对应的汉语知识图谱内容结果。如图 3 所示，输入一个英语科技期刊的科技论文网页地址，点击“获得结果”按钮后可以实时获取该科技论文对应的汉语富媒体知识图谱网页内容。其中，汉语的标题内容和摘要内容均为使用机器自动翻译得到的结果。

图 3 英文科技网站论文富媒体知识图谱的自动生成（示例）

8 结语

本文使用自顶向下和自底向上相结合的方法来构建英汉双语富媒体知识图谱，以 CNS 英文科技期刊网站的学术文献信息为例，首先在顶层设计了主题、期刊、科技论文、论文作者、科研机构、图片、音频和视频等实体及其属性，然后从半结构化的英文科技期刊网站内容中获取了这些实体和属性。针对非结构化的论文摘要或全文文本数据，通过深度学习模型从文本数据中自动抽取出学科领域的术语实体以及通用领域的实体，并将论文摘要细分为背景、方法、结果和结论等语步，实现对知识图谱的实体抽取和属性抽取。在知识图谱构建过程中，对同名作者实体、跨语言的实体，采用规则方法和 TransH 模型方法进行了实体对齐，以确保知识图谱中实体的一致性和知识的准确性。对图片媒体、音频媒体和视频媒体，提取了它们的属性并和科技论文实体进行了跨媒体的链接。最后，用图数据库 Neo4j 对知识图谱中的实体和关系进行存储，实现可视化查询和跨语言知识图谱获取等应用。本文所构建的英汉双语富媒体知识图谱可以为下游任务如跨语言文献信息检索和推荐、科学研究热点分析、科技情报监测等提供重要的数据支撑。

在构建英汉双语富媒体知识图谱的过程中，自动从文本媒体数据中抽取学科领域的术语实体的准确率并不是很高，且在领域通用性和适用性上还有待提升。在图片、音频和视频媒体数据的处理上，没有进一步地提取它们的自身特征作为属性，未来可以使用卷积神经网络（CNN）、梅尔倒谱系数（MFCC）、视频特征编码等技术进行属性自动抽取。希望本文的探索研究能为多模态跨语言知识图谱的构建提供参考，为科技文献知识图谱的进一步应用建立更加坚实的基础。

doi:10.3772/j.issn.2095-915x.2023.05.007

参 考 文 献

[1] 漆桂林 , 高桓 , 吴天星 . 知识图谱研究进展 [J]. 情报工程 , 2017, 3(1): 4-25.

[2] LEHMANN J, ISELE R, JAKOB M, et al. DBpedia – A large-scale, multilingual knowledge base extracted from Wikipedia[J]. Semantic Web, 2015, 6(2): 167-195.

[3] REBELE T, SUCHANEK F, HOFFART J, et al. YAGO: A multilingual knowledge base from Wikipedia, Wordnet, and Geonames [C]//The Semantic Web–ISWC 2016: 15th International Semantic Web Conference, Kobe, Japan, October 17–21, 2016, Proceedings, Part II 15. Springer International Publishing, 2016: 177-185.

[4] BOLLACKER K, EVANS C, PARITOSH P, et al. Freebase: a collaboratively created graph database for structuring human knowledge [C]//Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. 2008: 1247-1250.

[5] XU B, XU Y, LIANG J, et al. CN-DBpedia: A never-ending Chinese knowledge extraction system [C]// International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems. Cham: Springer International Publishing, 2017: 428-438.

[6] NIU X, SUN X, WANG H, et al. Zhishi. me-weaving Chinese linking open data [C]//The Semantic Web–ISWC 2011: 10th International Semantic Web Conference, Bonn, Germany, October 23-27, 2011, Proceedings, Part II 10. Springer Berlin Heidelberg, 2011: 205-220.

[7] 叶宏彬 , 张宁豫 , 陈华钧 , 等 .OpenConcepts: 一个开放的细粒度中文概念知识图谱 [J]. 中文信息学报 , 2023, 37(1): 46-53.

[8] HENNING S. Springer Nature SciGraph: Supporting open science and the wider understanding of research [EB/OL]. (2017-03-09) [2022-08-05]. https://www.springer.com/gp/about-springer/media/press-releases/corporate/springer-nature-scigraph--supporting-open-science-and-the-wider-understanding-of-research/12129600.

[9] SINHA A, SHEN Z, SONG Y, et al. An overview of Microsoft Academic Service (MAS) and applications [C]//Proceedings of the 24th International Conference on World Wide Web. 2015: 243-246.

[10] TANG J, ZHANG J, YAO L, et al. ArnetMiner: Extraction and mining of academic social networks [C]//Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2008: 990-998.

[11] WANG R, YAN Y, WANG J, et al. AceKG: A large-scale knowledge graph for academic data mining [C]//Proceedings of the 27th ACM International Conference on Information and Knowledge Management. 2018: 1487-1490.

[12] 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.

[13] 李丽双 , 郭元凯 . 基于 CNN-BiLSTM-CRF 模型的生物医学命名实体识别 [J]. 中文信息学报 , 2018, 32(1): 116-122.

[14] 郭航程 , 何彦青 , 兰天 , 等 . 基于 Paragraph-BERT-CRF 的科技论文摘要语步功能信息识别方法研究 [J]. 数据分析与知识发现 , 2022, 6(2/3): 298-307.

[15] BELTAGY I, LO K, COHAN A. SciBERT: a pretrained language model for scientific text [EB/OL]. (2019-09-10) [2022-08-05]. https://arxiv.org/abs/1903.10676.

[16] BORDES A, Usunier N, Garcia-Duran A, et al. Translating embeddings for modeling multi-relational data [C]//Proceedings of Neural Information Processing Systems (NIPS), 2013: 1-9.

[17] WANG Z, ZHANG J, FENG J, et al. Knowledge graph embedding by translating on hyperplanes [C]// Proceedings of AAAI, 2014: 1112-1119.

[18] LIN Y, LIU Z, SUN M, et al. Learning entity and relation embeddings for knowledge graph completion [C]//Proceedings of AAAI, 2015: 2181-2187.