



文献 CSTR:

32001.14.11-6035.csd.2025.0024.zh



文献 DOI:

10.11922/11-6035.csd.2025.0024.zh

数据 DOI:

10.57760/sciencedb.j00001.01301

文献分类: 计算机科学技术

收稿日期: 2025-03-08

录用日期: 2025-05-30

发表日期: 2026-02-05

* 论文通信作者

胡明昊: huminghao16@gmail.com

白晓颖: baixy@aibd.ac.cn

基于科技信息的多模态实体链接数据集 SI-MEL

王芳¹, 王少煌², 胡明昊^{3*}, 耿国桐³, 白晓颖^{3*}

1. 北京大学, 北京 100871

2. 新疆大学, 乌鲁木齐 830046

3. 军事科学院军事科学信息研究中心, 北京 100142

摘要: 实体链接是自然语言处理中的核心任务之一, 广泛应用于多种下游任务。在科技信息领域, 由于缺乏大规模、高质量的公开数据集, 现有模型在准确性上仍存在明显差距, 尤其是在处理专业术语和缩略语等短语时, 模型难以有效捕捉语义关联, 导致链接效果不理想。为解决这一问题, 本研究基于 Wikipedia 的科技信息相关文章, 设计数据处理框架, 包括原始数据获取 (文章爬取与数据清洗)、候选集生成 (提及提取与检索、图片爬虫、相似度计算) 及最终数据集 (人工验证)。通过该框架, 构建了 SI-MEL (Science Information - Multimodal Entity Linking) 数据集, 涵盖组织 (1415 条)、人物 (4932 条) 和专业 (8693 条) 三个主题。实验结果表明, 通过不同的模型输入扩充策略, 数据集的结构化和多样性显著增强了链接效果。SI-MEL 的发布为科技信息领域的多模态实体链接提供了重要的研究资源, 具有推动该领域发展的重要意义。

关键词: 科技信息; 多模态; 实体链接; 数据集

数据库 (集) 基本信息简介

数据库 (集) 名称	基于科技信息的多模态实体链接数据集 SI-MEL
数据通信作者	胡明昊 (huminghao16@gmail.com); 白晓颖 (baixy@aibd.ac.cn)
数据作者	王芳、胡明昊、白晓颖
地理区域	全球
数据量	2.49 GB
数据格式	*.json
数据服务系统网址	https://doi.org/10.57760/sciencedb.j00001.01301
基金项目	国家自然科学基金 (62476283、62376284)
数据库 (集) 组成	数据集包含15040条科技信息实例, 包含组织、人物、专业等主题。数据实例主要有名称、数据Qid、中英文属性、视觉信息、候选实体集。

引言

实体链接 (Entity Linking) 是指通过计算机技术将文本中的提及 (Mention) 与知识图谱中相应的实体进行链接的过程^[1], 也是支撑自然语言处理 (Natural Language Processing) 领域各个下游任务的重要技术。实体链接已经取得了长足的

发展，从早期基于规则的方法，到中期基于统计学习的方法，再到目前主流的基于深度学习的方法。值得注意的是，基于深度学习的实体链接方法在模型训练阶段高度依赖大规模高质量的数据资源。在语料资源丰富的领域（如人物、新闻及电影等），目前已构建了多个多模态实体链接数据集，包括 Wikiperson、WikiDiverse、MEL 等代表性数据集。这些数据集的详细内容如表 1 所示。

表 1 多模态实体链接数据集的简要情况

Table 1 An overview of the multi-modal entity linking dataset

数据集名称	来源	主题	大小	构建方式
Snap ^[2]	Freebase	Social Media	12K captions	人工标注
Twitter ^[3]	Twitter users	Social Media	4M tweets	半自动标注
Movie ^[4]	Wikipedia	Movies	1K reviews	
Weibo-MEL ^[5]	Weibo	Social Media	25K samples	
Wikidata-MEL ^[5]	Wikidata	News	18K samples	
Richpedia-MEL ^[5]	Richpedia	encyclopedia	17K samples	
WikiDiverse ^[6]	Wikipedia	News	8K captions	人工标注
WikiPerson ^[7]	Wikipedia	Person	50K images	人工标注

经调查发现，面向科技信息领域的多模态实体链接数据集在公开可用资源方面存在显著不足。与人物、新闻及电影等领域的多模态实体链接数据集相比，科技信息领域的实体链接模型性能相对较差，在一定程度上制约了多模态实体链接技术的进一步发展。具体而言，在实际链接过程中，实体表达形式较为简洁，但可能蕴含丰富的深层语义信息。然而，现有实体链接模型难以有效捕捉和解析此类语义信息，如图 1 所示。由于实体表达形式的高度简化和“一词多义”现象，模型对实体语义的理解能力受限，降低了链接准确率。因此，面向科技信息构建含有实体的上下文、属性以及视觉等不同粒度语义信息的多模态实体链接数据集，为实体链接模型提供必要的补充语义信息。

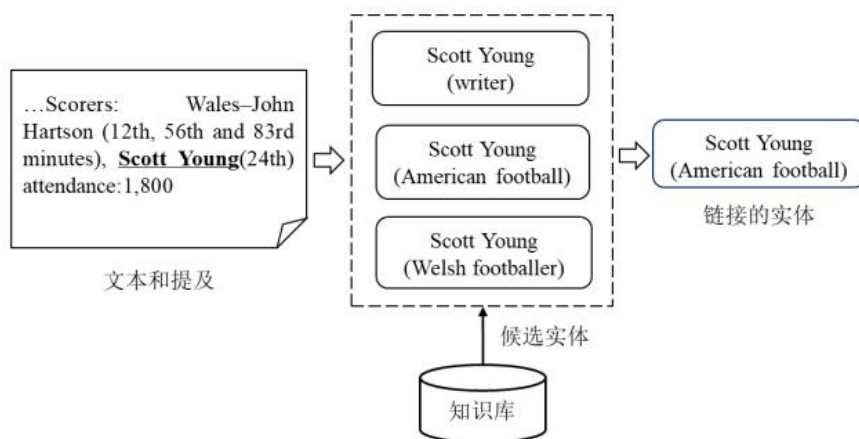


图 1 实体链接主要流程示意图

Figure 1 The main workflow of entity linking

本研究基于大规模科技信息的原始文章，采用多模态实体链接数据集的系统化构建流程，构建了基于科技信息领域的多模态实体链接数据集。本数据集遵循统一的标准规范进行结构化存储，涵

盖实体的标签、属性信息（包括描述、别名等）、视觉信息以及候选实体集合等。本数据集的构建不仅为自然语言处理的下游任务提供了高质量的资源支持，还可作为外部知识库，为大语言模型的训练与优化提供重要的语义补充，从而进一步提升模型在科技信息领域的理解与应用能力。

1 数据采集和处理方法

1.1 数据采集

科技信息的原始语料来源于 Wikipedia (<https://www.wikipedia.org/>) 页面，通过爬虫技术获取涵盖组织、人物和专业等主题的所有相关文章内容，并对其进行清洗与处理。具体处理步骤包括：1) 数据筛查与去重。通过对标题和正文内容的哈希值进行比对，识别并删除重复的条目。此外，使用算法检查内容相似度，剔除内容雷同的数据，以保证数据的多样性和覆盖度；2) 标题有效性筛选。通过正则表达式和自然语言处理技术，对标题进行严格筛选，删除标题为单字符、特殊符号或乱码的数据，以确保标题的规范性和有效性；3) Wikidata 跳转链接验证。通过遍历所有数据，验证其是否包含有效的 Wikidata (<https://www.wikidata.org/>) 跳转链接。缺乏有效链接的数据将被剔除，因为 Wikidata 链接为每个数据条目提供了进一步的知识和结构化信息，有助于提升数据的可信度和语义关联性。最终获得科技信息的原始数据，共计 27650 个 JSON 文件，原始数据大小 557 MB，包含标题和正文内容。

1.2 数据处理

为保障数据准确性，采用图 2 所示的三阶段处理流程。首先，基于原始数据中 Wikipedia 页面的超链接定向跳转至 Wikidata，以提取对应的属性信息。接着，使用 Wikidata 检索，选取前 100 个搜索结果的关键属性（如 QID、名称及描述等）构建初始候选实体集合。最后，构建多源图像爬取框架（包括 Wikidata、Wikipedia 及 Wikimedia 等），利用预训练模型 CLIP (ViT-B-32) [8] 的图文多模态相似度计算，筛选文本描述与图像信息均完备的候选实体集合。

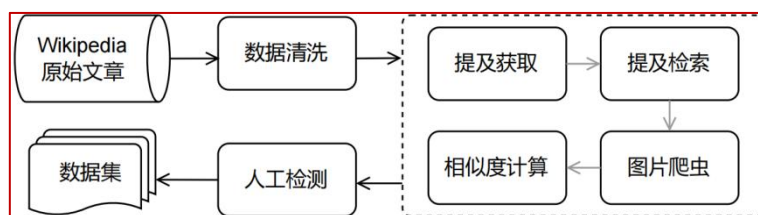


图 2 数据采集及处理方法

Figure 2 Data collection and processing workflow

2 数据样本描述

2.1 数据结构

SI-MEL 数据集的总大小为 2.49 GB，涵盖了组织、人物和专业三个主要主题，数据集详情如表 2 所示，详细列出了每个子主题的实例数、文件大小、图片数量及总数据量。

表 2 数据集基本情况

Table 2 Basic information of the dataset

主题	子主题	实例数	文件大小	图片数量	总大小
组织	机构	462	1 MB	1848	2.49 GB
	高校	953	2.1 MB	3812	
人物	科研	2096	4.5M	8384	
	政治	2836	6.2 MB	11344	
专业	软件	3189	7.1 MB	12756	
	硬件	5504	12.3 MB	22016	

在组织类别中，子主题包括“机构”和“高校”。“机构”类别包含 462 个实例，文件大小为 1 MB，图像数量为 1848 张；而“高校”类别则包含 953 个实例，文件大小为 2.1 MB，图像数量为 3812 张。在人物类中，子主题包括“科研”和“政治”。“科研”类别包含 2096 个实例，文件大小为 4.5 MB，图像数量为 8384 张；“政治”类别则包含 2836 个实例，文件大小为 6.2 MB，图像数量为 11344 张。在专业类中，子主题包括“软件”和“硬件”。“软件”类别的实例数量为 3189，文件大小为 7.1 MB，图像数量为 12756 张；而“硬件”类别包含 5504 个实例，文件大小为 12.3 MB，图像数量为 22016 张。

各子主题的详细信息为后续的数据分析和模型训练提供了扎实的基础，尤其是在数据量、文件大小及图像数量上的分布，能够为训练过程中的多样性和数据丰富性提供有力支持。

2.2 数据实例

每条数据以字典形式存储，共有 6 个项，对应文件中的每一个关键词分别为提及名称、数据 Qid、中英文属性信息、图片信息、候选实体集合以及数据难度。数据样本如图 3 所示。

其中，“提及名称”表示实体在现实世界中的实际名称；“Qid”是 Wikidata 为每个实体分配的唯一标识符。数据构建过程中，主要采集英文相关属性信息，但在爬取过程中，如果发现完整的中文信息，也会将其保留。图片信息则表示与实体相关的视觉数据，作为该实体的多模态表示之一。候选集经过检索和筛选，确保其包含高质量的实体候选，满足多模态实体链接任务的需求。数据难度指标用于评估候选集中的实体是否包含正确答案。

3 数据质量控制和评估

3.1 质量控制

基于 Wikipedia (<https://www.wikipedia.org/>) 页面获取科技信息初始数据时，首先对原始文本进行去噪处理，网页信息会出现不完全符合规范的内容，含有部分空、乱码等页面，通过语义匹配将这些无用页面去除，获得原始页面信息；接着对原始页面信息进行对照检查，去除图片乱码以及单字符标题的页面，获取科技信息的原始数据。

在数据处理过程中，首先，通过解析 Wikipedia 页面内嵌的 Wikidata 超链接，利用爬虫代码精准提取实体属性信息；接着，利用搜索引擎 Wikidata 检索提及，对每个提及返回的 Top-100 搜索条目进行关键属性（QID 唯一标识符、实体名称、多语言描述等）的提取，形成初始候选实体集合；最

后，整合 Wikidata、Wikipedia 及 Wikimedia 的跨模态数据资源，通过爬虫技术获取图片信息，将候选实体的文本描述与对应的图像输入 CLIP（ViT-B-32）模型，利用模型的文本-图像对比学习机制，评估文本和图像之间的相似性。在此基础上，设定了超参数相似度阈值 s ，最终筛选出图文描述匹配度高于阈值的优质候选实体集合。这一过程有助于去除无关或低质量的实体，从而确保数据集中包含的实体与实际任务高度相关，提高了数据集的质量和有效性。

```
{
  "entity": "Rafael Landívar University",
  "Qid": "Q2496136",
  "attribute": {
    "en": {
      "label": "Rafael Landívar University",
      "description": "Catholic university if Guatemala",
      "aliases": ["Rafael Landívar University"],
      "context": "Rafael Landívar University (Universidad Rafael Landívar) is a private Catholic coeducational higher education institution run by the Society of Jesus in Vista Hermosa III Guatemala... Many recent Guatemalan presidents have studied at the university, including Ramiro de León Carpio, Álvaro Arzú, and Óscar Berger.\n",
      "Type": "Private, Catholic coeducational higher education institution",
      "zh": {
        "标签": "拉法耶兰帝瓦大学",
        "描述": "危地马拉的大学",
        "别名": ["拉法耶蘭帝瓦大學", "拉斐尔兰迪瓦尔大学"],
        "上下文": "N/A",
        "attribute": "N/A"
      }
    },
    "image": ["Landivar Campus Central.jpg",
      "https://commons.wikimedia.org/wiki/Special:FilePath/Landivar%20Campus%20Central.jpg",
      "/content/drive/MyDrive/Multi-Dataset-2025/college_candidates_images/Rafael Landívar University-Q2496136/Q2496136.jpg"],
    "candidates": [
      {
        "Qid": "Q2496136",
        "name": "Rafael Landívar University",
        "description": "Catholic university if Guatemala",
        "image": ["Landivar Campus Central.jpg", "https://commons.wikimedia.org/wiki/Special:FilePath/Landivar%20Campus%20Central.jpg", "/content/drive/MyDrive/Multi-Dataset-2025/college_candidates_images/Rafael Landívar University-Q2496136/Q2496136.jpg"]
      },
      {
        "Qid": "Q16424000",
        "name": "Rafael",
        "description": "male given name",
        "image": ["General Rafael Urdaneta Bridge view from the lake to Cabimas side.jpg", "https://upload.wikimedia.org/wiki/File:General Rafael Urdaneta Bridge view from the lake to Cabimas side.jpg", "/content/drive/MyDrive/Multi-Dataset-2025/college_candidates_images/Rafael Landívar University-Q2496136/Q16424000.jpg"]
      },
      {
        "Qid": "Q37478521",
        "name": "Rafael",
        "description": "family name",
        "image": ["General Rafael Urdaneta Bridge view from the lake to Cabimas side.jpg", "https://upload.wikimedia.org/wiki/File:General Rafael Urdaneta Bridge view from the lake to Cabimas side.jpg", "/content/drive/MyDrive/Multi-Dataset-2025/college_candidates_images/Rafael Landívar University-Q2496136/Q37478521.jpg"]
      }
    ],
    "difficulty": 0
  }
}
```

图3 数据样本示例图

Figure 3 Examples of data samples

3.2 效果评估

为了验证本数据集在提升科技信息多模态实体链接效果方面的有效性，选择了多个具有代表性的基线模型，包括 ALIGN^[9]、BLIP-2^[10]、CLIP^[8]、FLAVA^[11]、OWL-ViT^[12]、SigLIP^[13]、Qwen-VL^[14]和 MiniCPM-V^[15]。这些模型涵盖了多种最新的视觉-语言预训练架构，能够较好地反映当前多模态实体链接领域的主流研究进展。

在实验设计方面，为了评估不同输入信息扩充策略对模型性能的影响，设定四种不同的输入信息扩充策略。1) 无扩充（对照组）。基线模型仅使用原始的文本或图像输入数据进行评估，没有任何额外的语义信息扩充，主要用于衡量模型在标准输入条件下的基本性能表现。2) 上下文扩充。为每个实体添加上下文信息，并将其与原始文本一起作为模型的输入。3) 实体描述扩充。为每个实体添加具体描述信息，并将其与原始文本一起作为模型的输入。4) 上下文与实体描述联合扩充。为每个主题同时使用上下文和实体描述进行扩充。通过这些不同的实验设置，能够全面评估每种扩充策略对模型性能的提升效果。

在评测指标方面，使用了准确率（ACC）作为主要评估标准，以量化不同扩充方式对多模态实体链接任务的影响。此外，为了确保实验的公平性，对各模型的超参数进行了统一设置。例如，文本长度均为 512，并在相同的硬件条件下（A100 42G）进行评估。

从表 3 中的实验结果可知，与不使用任何扩充方法的基线模型相比，使用上下文、描述进行语义信息扩充的方法，可以有效提升多模态实体链接的性能。具体而言，使用语义信息扩充的模型，与基线模型相比 ACC 值最高提升了 7.5%，表明无论使用上下文还是实体描述作为辅助信息进行信息扩充，都可以在一定程度上提高链接准确率。

表 3 SI-MEL 数据集上的主要实验结果

Table 3 Primary experiment results on SI-MEL

任务	ALIGN	BLIP-2	CLIP	FLAVA	OWL-ViT	SigLIP
实体提及	0.681	0.724	0.755	0.663	0.492	0.717
提及+上下文	0.682	0.736	0.756	0.663	0.492	0.718
提及+描述	0.683	0.743	0.767	0.665	0.493	0.720
提及+上下文+描述	0.701	0.751	0.776	0.669	0.492	0.723

此外，对比表 3 和表 4 中的实验结果可知，小规模预训练基线模型（ALIGN、BLIP-2、CLIP、FLAVA、OWL-ViT、SigLIP）在实体链接任务上的表现优于大规模多模态基线模型（Qwen-VL、MiniCPM-V）。这是由于小规模模型的训练语料的数据适配性与目标函数的针对性，使其对该任务的适应性更强，从而取得更优效果。最后，Qwen-VL 和 MiniCPM-V 采用多轮对话机制（表格中用*表示），相较于单轮对话，能够显著提升多模态实体链接的性能。

表 4 SI-MEL 数据集在大模型上的实验结果

Table 4 The experiment results of LLMs on SI-MEL

任务	单轮对话		多轮对话	
	Qwen-VL	MiniCPM-V	Qwen-VL*	MiniCPM-V*
实体提及	0.466	0.543	0.378	0.393
提及+上下文	0.468	0.551	0.381	0.419
提及+描述	0.523	0.558	0.396	0.467
提及+上下文+描述	0.524	0.562	0.397	0.468

4 数据价值

本数据集专注于科技信息领域，旨在提供包含实体上下文、属性、视觉信息等不同粒度语义信息的多模态实体链接语料。它为实体链接模型提供了所需的丰富多模态数据，支持模型在理解和链接科技领域实体时的性能提升。当前，科技信息领域的多模态实体链接数据集仍然相对匮乏，这在一定程度上限制了实体链接模型在该领域的应用和发展，导致模型对实体语义的理解不足，进而影响了链接准确率。本数据集的构建，填补了这一空白，为科技信息领域的实体链接研究提供了高质量的训练和评测数据，帮助提升模型在科技领域实体的语义识别能力，推动多模态实体链接技术的进

步。

此外，本数据集不仅在实体链接领域具有重要的应用价值，还可广泛应用于自然语言处理的其他下游任务。例如，可用于支持多模态实体链接、知识图谱构建、科技文献的检索、智能问答系统的构建等场景，为相关领域的研究提供强有力的支持。更重要的是，作为外部知识库，能够为大型语言模型提供辅助信息，从而增强其在科技领域的应用效果。综上所述，本数据集在提升科技领域实体语义理解、推动多模态实体链接技术发展以及促进相关应用的实际落地等方面，具有重要的研究价值和实际意义。

致 谢

本数据集所使用的数据来源是 Wikipedia、Wikidata 及 Wikimedia 等公开网站。本数据只用于科学研究，不用于商业性目的。特别感谢参与本数据集工作的专业人员。

作者分工职责

王芳（1997—），女，山东省济南市人，博士研究生，研究方向为人工智能。主要承担工作：数据采集，数据集构造，数据校对，论文撰写。

王少煌（1997—），男，山西省太原市人，硕士研究生，研究方向为自然语言处理、大语言模型领域微调技术。主要承担工作：数据校对。

胡明昊（1990—），男，四川省成都市人，博士，高级工程师，研究方向为人工智能。主要承担工作：数据集选择与采集指导，数据集构造方法的指导和论文写作指导。

白晓颖（1973—），女，辽宁省大连市人，博士，研究员，研究方向为软件工程。主要承担工作：数据集构造方法和论文写作指导，数据集质量控制与综合管理。

耿国桐（1975—），男，山东省淄博市人，博士，研究员，研究方向为信息处理。主要承担工作：论文写作指导。

参考文献

- [1] FENSEL D, ŞİMŞEK U, ANGELE K, et al. Introduction: what is a knowledge graph? [M]//Knowledge Graphs: Methodology, Tools and Selected Use Cases. Cham: Springer International Publishing, 2020: : 1 – 10. DOI:10.1007/978-3-030-37439-6_1.
- [2] MOON S, NEVES L, CARVALHO V. Multimodal named entity disambiguation for noisy social media posts[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia. Stroudsburg, PA, USA: ACL, 2018: 2000-2008.. DOI:10.18653/v1/p18-1186.
- [3] ADJALI O, BESANÇON R, FERRET O, et al. Building a multimodal entity linking dataset from tweets[C]//Proceedings of the Twelfth Language Resources and Evaluation Conference. 2020: 4285-4292.
- [4] GAN J R, LUO J C, WANG H W, et al. Multimodal entity linking: a new dataset and a baseline[C]//Proceedings of the 29th ACM International Conference on Multimedia. Virtual Event, China. ACM, 2021: 993 – 1001. DOI:10.1145/3474085.3475400.

- [5] ZHOU X C, WANG P, LI G Z, et al. Weibo-MEL, wikidata-MEL and richpedia-MEL: multimodal entity linking benchmark datasets[C]//Knowledge Graph and Semantic Computing: Knowledge Graph Empowers New Infrastructure Construction. Singapore: Springer, 2021: 315 – 320. DOI:10.1007/978-981-16-6471-7_27.
- [6] WANG X W, TIAN J F, GUI M, et al. WikiDiverse: a multimodal entity linking dataset with diversified contextual topics and entity types[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland. Stroudsburg, PA, USA: ACL, 2022: 4785-4797.. DOI:10.18653/v1/2022.acl-long.328.
- [7] SUN W X, FAN Y X, GUO J F, et al. Visual named entity linking: a new dataset and a baseline[EB/OL]. 2022: arXiv: 2211.04872. <https://arxiv.org/abs/2211.04872>.
- [8] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]//International Conference on Machine Learning., 2021.
- [9] JIA C, YANG Y F, XIA Y, et al. Scaling up visual and vision-language representation learning with noisy text supervision[C]//International Conference on Machine Learning., 2021.
- [10] LI J N, LI D X, SAVARESE S, et al. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models[C]//Proceedings of the 40th International Conference on Machine Learning. Honolulu, Hawaii, USA. ACM, 2023: 19730 – 19742. DOI:10.5555/3618408.3619222.
- [11] SINGH A, HU R H, GOSWAMI V, et al. FLAVA: a foundational language and vision alignment model[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, LA, USA. IEEE, 2022: 15617 – 15629. DOI:10.1109/CVPR52688.2022.01519.
- [12] MINDERER M, GRITSENKO A, STONE A, et al. Simple open-vocabulary object detection with vision transformers[EB/OL]. 2022: arXiv: 2205.06230. <https://arxiv.org/abs/2205.06230>.
- [13] ZHAI X H, MUSTAFA B, KOLESNIKOV A, et al. Sigmoid loss for language image pre-training[C]//2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France. IEEE, 2024: 11941 – 11952. DOI:10.1109/ICCV51070.2023.01100.
- [14] BAI J Z, BAI S, YANG S S, et al. Qwen-VL: a versatile vision-language model for understanding, localization, text reading, and beyond[EB/OL]. 2023: arXiv: 2308.12966. <https://arxiv.org/abs/2308.12966>.
- [15] YAO Y, YU T Y, ZHANG A, et al. MiniCPM-V: a GPT-4V level MLLM on your phone[EB/OL]. 2024: arXiv: 2408.01800. <https://arxiv.org/abs/2408.01800>.

论文引用格式

王芳, 王少煌, 胡明昊, 等. 基于科技信息的多模态实体链接数据集 SI-MEL[J/OL]. 中国科学数据, 2026, 11(1). (2026-02-05). DOI: 10.11922/11-6035.csd.2025.0024.zh.

数据引用格式

王芳, 胡明昊, 白晓颖. 基于科技信息的多模态实体链接数据集 SI-MEL[DS/OL]. V1. Science Data Bank, 2025. (2025-03-24). DOI: 10.57760/sciencedb.j00001.01301.

SI-MEL: A dataset of multi-modal entity linking based on scientific information

WANG Fang¹, WANG Shaohuang², HU Minghao^{3*}, GENG Guotong³, BAI Xiaoying^{3*}

1. Peking University, Beijing 100871, P.R. China

2. Xinjiang University, Urumqi 100195, P.R. China

3. Center of Information Research, Academy of Military Science, Beijing 100142, P.R. China

*Email: huminghao16@gmail.com (HU Minghao); baixy@aibd.ac.cn (BAI Xiaoying)

Abstract: Entity linking is one of the core tasks in natural language processing, and plays a crucial role in a wide range of downstream applications, such as information extraction, knowledge graph construction, and question answering. In the field of scientific information, however, due to the lack of large-scale, high-quality public datasets have led to notable performance limitations. Especially when dealing with phrases such as technical terms and abbreviations, these models often fail to effectively capture semantic relationships, resulting in suboptimal linking accuracy. To address this issue, this study designs a data processing framework based on Wikipedia articles related to scientific information. The framework includes raw data acquisition (article crawling and data cleaning), candidate entity generation (mention extraction and retrieval, image crawling, similarity computation), and final dataset construction (manual verification). Using this framework, we constructed the Science Information - Multimodal Entity Linking (SI-MEL), a multimodal entity linking dataset covering three main themes: organizations (1,415 entries), people (4,932 entries), and specialties (8,693 entries). Experimental results demonstrate that, through various input expansion strategies, the enhanced structure and diversity of the SI-MEL dataset significantly improve entity linking performance. The release of SI-MEL provides an important research resource for multimodal entity linking in the field of scientific information and holds substantial significance for advancing research in this area.

Keywords: scientific information; multi-modal; entity linking; dataset

Dataset Profile

Title	SI-MEL: A dataset of multi-modal entity linking based on scientific information
Data corresponding author	HU Minghao (huminghao16@gmail.com); BAI Xiaoying (baixy@aibd.ac.cn)
Data authors	WANG Fang, HU Minghao, BAI Xiaoying
Geographical scope	Global
Data volume	2.49 GB
Data format	.json
Data service system	https://doi.org/10.57760/sciencedb.j00001.01301
Sources of funding	National Natural Science Foundation of China (62476283, 62376284)
Dataset composition	The dataset comprises 15,040 entries of scientific information, covering various themes including organization, people and specialty. The data items include the entity name, Qid, Chinese & English attributes, images, and candidate entity sets.