北京大学学报(自然科学版)
Acta Scientiarum Naturalium Universitatis Pekinensis
ISSN 0479-8023,CN 11-2442/N

《北京大学学报(自然科学版)》网络首发论文

题目：基于深度学习的实体链接方法研究
作者：李天然，刘明童，张玉洁，徐金安，陈钰枫
DOI：10.13209/j.0479-8023.2020.077
收稿日期：2020-06-06
网络首发日期：2020-10-13
引用格式：李天然，刘明童，张玉洁，徐金安，陈钰枫．基于深度学习的实体链接方法研究．北京大学学报(自然科学版).
https://doi.org/10.13209/j.0479-8023.2020.077

网络首发：在编辑部工作流程中，稿件从录用到出版要经历录用定稿、排版定稿、整期汇编定稿等阶段。录用定稿指内容已经确定，且通过同行评议、主编终审同意刊用的稿件。排版定稿指录用定稿按照期刊特定版式（包括网络呈现版式）排版后的稿件，可暂不确定出版年、卷、期和页码。整期汇编定稿指出版年、卷、期、页码均已确定的印刷或数字出版的整期汇编稿件。录用定稿网络首发稿件内容必须符合《出版管理条例》和《期刊出版管理规定》的有关规定；学术研究成果具有创新性、科学性和先进性，符合编辑部对刊文的录用要求，不存在学术不端行为及其他侵权行为；稿件内容应基本符合国家有关书刊编辑、出版的技术标准，正确使用和统一规范语言文字、符号、数字、外文字母、法定计量单位及地图标注等。为确保录用定稿网络首发的严肃性，录用定稿一经发布，不得修改论文题目、作者、机构名称和学术内容，只可基于编辑规范进行少量文字的修改。

出版确认：纸质期刊编辑部通过与《中国学术期刊（光盘版）》电子杂志社有限公司签约，在《中国学术期刊（网络版）》出版传播平台上创办与纸质期刊内容一致的网络版，以单篇或整期出版形式，在印刷出版之前刊发论文的录用定稿、排版定稿、整期汇编定稿。因为《中国学术期刊（网络版）》是国家新闻出版广电总局批准的网络连续型出版物（ISSN 2096-4188，CN 11-6037/Z），所以签约期刊的网络版上网络首发论文视为正式出版。

网络首发时间：2020-10-13 14:06:21
网络首发地址：https://kns.cnki.net/kcms/detail/11.2442.N.20201013.1203.002.html

北京大学学报(自然科学版)
Acta Scientiarμm Naturaliμm Universitatis Pekinensis
doi: 10.13209/j.0479-8023.2020.077

基于深度学习的实体链接方法研究

李天然　刘明童　张玉洁†　徐金安　陈钰枫

北京交通大学计算机与信息技术学院，北京　100044，† 通信作者，E-mail: yjzhang@bjtu.edu.cn

摘要　实体链接旨在将文本中的实体正确链接到知识库中的实体，其在本文语义理解中有十分重要的作用，广泛应用于问答系统、信息检索等自然语言处理任务中。实体链接的关键问题是语言的歧义性，包括同义词和一词多义的实体消歧问题。近年来，得益于深度学习的进步，实体链接性能得到了较大的改进，对其梳理总结，并对目前研究进展进行系统性的研究是十分必要的。本文详细介绍了实体链接的概念和步骤，对最近几年基于深度学习的命名实体链接相关问题和研究现状进行系统性的介绍，重点对实体链接存在的问题及相应解决模型进行详细分析,并介绍相关数据集和评测方法。最后，总结了国际评测会议中实体链接的现状，并对未来的研究方向进行了分析。

关键词　实体链接；实体消歧；实体识别；知识库；深度学习

Research on Entity Linking Method Based on Deep Learning

LI Tianran，LIU Mingtong，ZHANG Yujie†，XU Jin’an, CHEN Yufeng

School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044;
† Corresponding author, E-mail: yjzhang@bjtu.edu.cn

Abstract　Entity linking aims to correctly link the entities in the text to the entities in the knowledge base. It plays a very important role in the semantic understanding and is widely used in natural language processing tasks such as question answering and information retrieval. The key issue of entity linking is language ambiguity, including entity disambiguation of synonyms and polysemy. In recent years, thanks to the advancement of deep learning, entity linking performance has been greatly improved. It is necessary to sort out and summarize it and systematically study the current research progress. This article introduces the concept and steps of entity linking in detail, systematically introduces the problems and research status of named entity linking based on deep learning in recent years, and focuses on the detailed analysis of the problems and corresponding solution models of entity linking and introduces related data sets and evaluation methods. Finally, this article summarizes the current status of entity linking in international evaluation conferences and analyzes the future research directions.

Key words　entity linking; entity disambiguation; entity recognition; knowledge base; deep learning

随着大数据技术的提出与发展，海量信息的爆炸式增长导致信息过载，也对用户获取准确信息带来了挑战。为了准确获取目标信息，我们需要处理大量的虚假信息、冗余信息和噪声信息。出现这一问题的原因是自然语言表达的多样性，即一词多义、多词同义的现象。表述或指称(Mention)是指自然语言文本中表达实体的语言片段。实体链接(Entity Linking, EL)是指将文本中的表述链接到知识库 (Knowledge Base, KB)中相应实体(Entity)来进行实体消歧(Entity Disambiguation)，帮助人和计算机理解文本的具体含义的任务。

国家自然科学基金(61876198, 61976015, 61976016)资助
收稿日期: 2020–06–06; 修回日期: 2020–08–13

例如，在文本“苹果发布了最新产品”中，表述“苹果”在知识库中对应的实体有“苹果（科技产品）”、“苹果（苹果产品公司）”、“苹果（蔷薇科苹果属果实）”等，实体链接就是将表述“苹果”链接到知识库中的“苹果（苹果产品公司）”，消除其他义项所导致的歧义的过程。实体链接能够利用知识库丰富的语义信息，在许多领域起到了非常重要的基础作用，例如问答系统(Question Answering)、语义搜索(Semantic Search)和信息抽取(Information Extraction)等。实体链接也有着扩充知识库的重要应用，可以用于更新实体和关系，是知识图谱构建中的一个重要环节。实体链接研究中所使用的知识库包括 TAP、维基百科、Freebase、YAGO 等。

近年来，深度学习技术作为人工智能的关键技术，在计算机视觉和自然语言处理等领域取得的突破性进展,使得人工智能迎来新一轮爆发式发展。深度学习方法给实体链接任务提供了强有力的工具，得益于神经网络强大的特征抽象和泛化能力，基于深度学习的实体链接方法逐渐成为研究实体链接的主流方法。与传统的统计方法相比，深度学习方法主要有以下两点优势：一是深度学习方法的训练是端到端的过程，不需要手工定义相关特征；二是深度学习可以学习特定于任务的表示，建立不同模式、不同类型和不同语言之间的信息关联，从而实现更好的实体分析性能。

为了明确实体链接任务未来的研究方向，以便更好的应用到其他领域，本文总结了实体链接任务现有的研究。本文首先介绍实体链接的任务定义、关键技术、相关研究和发展历程，重点介绍任务目前存在的问题，接着对最近几年基于深度学习的实体链接相关问题和研究现状进行系统性的介绍,最后总结国际评测会议中实体链接的不足，并对未来的研究方向进行分析。

1 实体链接任务

1.1 实体链接任务

实体链接任务研究的对象为包含人名、地名、机构名在内的命名实体，研究任务是将非结构化文本中的表述指向其代表的真实世界实体, 关联到对应的知识库具体实体中，主要解决了实体名的歧义性和多样性问题。实体链接任务的关键技术包括两个步骤：候选实体生成和实体消歧。

候选实体生成是指为文本中的指称生成一个知识库中的相关实体集合。第一项工作是识别出文档中的实体指称，即需要链接到知识库进行消歧的词或短语。这一过程与自然语言处理中的命名实体识别任务较为类似。候选实体生成要求较高的召回率，目的是召回尽可能多的指称的可能的链接对象，以提高实体链接的准确性，同时尽可能排除不相关的实体，从而减少计算量。候选实体消歧是指通过计算实体集合中实体的相关性进行排序，选出最佳的对应候选实体的过程。本文先对上述相关研究基础进行简要介绍。

1.2 命名实体识别

候选实体生成的关键技术是命名实体识别，命名实体是指人名、机构名、地名以及其他所有以名称为标识的实体，命名实体识别包含识别出实体的边界和类型(人名、地名、机构名或其他)两个过程。实体链接中的候选实体生成和命名实体识别都需要识别出实体片段，但命名实体识别意在识别出所有实体，而候选实体生成则侧重找出存在相应链接到的知识库中条目的实体指称。实体链接中识别指称的方法主要使用了早期命名实体识别中基于规则和字典的方法，该方法也被广泛应用于各种实体链接系统，通过充分利用维基百科提供的各类信息，包括重定向页面、排歧页面、锚文本等，构建实体名称与所有可能链接的实体的映射关系字典，然后利用字典中的信息生成候选实体集合。此外，研究者也提出了基于上下文的实体名称字典的扩充法，文本中的实体通常用缩写词或实体全名的一部分来表示，而它们相应的全名通常出现在上下文中。因此，使用上下文信息来查找与这些缩写或局部词相对应的实体全名，从而扩展实体名称词典。

1.3 实体消歧

命名实体消歧是确定一个实体指称项所指向的真实世界实体的过程。实体链接中的命名实体消歧通过计算实体指称与候选实体之间的相似度并进行相似度排序来选择可能的候选实体。进行实体消歧的主流方法有基于概率生成模型的方法、基于主题模型的方法、基于图的方法、基于深度学习的方法、无监督方法等等。根据模型的差异，大致可以划分为基于统计学习的实体消歧方法和基于深度学习的实体消歧的方法。基于统计学习的方法侧重于计算实体之间的相似度，但需要借助有标注的实体链接语料库来进行。为了解决标注的语料库缺乏的问题，半监督、弱监督的方法也相继涌现。基于深度学习的方法的核心是构建多类型，多模态的上下文和知识的统一表示，需要借助性能较好的消歧模型来进行。近年来，得益于深度学习

的进步，基于深度学习的实体链接方法展示出明显优势，本文重点关注基于深度学习的实体链接方法。

2 实体链接技术的发展

早期的研究大多侧重于单独的为每个实体指称进行消岐，利用实体指称的上下文信息为每个指称生成每个候选实体与上下文的相关性得分。随后研究人员提出，同一篇文章内的被链接的实体应该存在制约关系,会影响最终的链接结果，应综合考虑多个实体间的语义关联，从而进行协同的实体链接。因此，根据可利用信息的不同和链接决策之间是否独立，现有的实体消歧模型可以分为两种：局部模型和全局模型。局部模型利用实体指称周围的局部文本上下文信息独立的解决每个实体指称的歧义问题 (Chen 和 Ji, 2011[1]; Chisholm 和 Hachey, 2015[2]; Lazic 等, 2015[3]; Yamada 等, 2016[4])。局部模型仅关注如何将文本中抽取到的实体链接到知识库中，忽视了位于同一文档的不同实体间存在的语义联系。而全局模型鼓励文档中所有指称的目标实体在主题上保持一致性，通过计算不同目标实体之间的主题一致性、实体关联度、转移概率、实体流行度特征等等来进行消歧(Han 等人, 2011[5]; Cassidy 等人,2012[6]; He 等人, 2013[7]; Cheng 和 Roth, 2013[8]; Durrett 和 Klein, 2014[9]; Huang 等人, 2014[10])。全局模型通常基于 KB 建立实体图，以捕获文档中所有已标识的指称的连贯的实体，具体而言，将文档中的实体指称及其候选实体构建为图结构,其中节点为实体，边表示其关系, 利用实体指称间、候选实体间、实体指称与候选实体间的关联关系进行协同推理。该图提供了局部模型无法使用的高度区分性语义信号（例如，实体相关性）（Eshel 等人，2017[11]）。

由于实体链接是一个相对下游的任务，其性能受限于命名实体识别任务的准确性，对于中文的实体链接任务而言，还会受到中文分词任务的影响，上游任务的错误会为实体链接任务带来不可避免的噪音。同时，实体间的歧义性问题较为严重，主要表现为多样性和歧义性。多样性指的是同一个实体对应了多个名称，而歧义性指的是同一个名称有多个含义。这给实体连接任务带来了很大的困难，目前实体链接任务存在了以下难以解决的问题：

（1）传统的基于机器学习的实体链接方法需要完整而标注准确的数据集，而人工标注的数据集较为缺乏，尤其是中文和其他语言的权威数据集。词向量模型的出现一定程度上有效的解决了这一问题，其采用无标注的文本为输入数据，将词表征为低维向量。然而传统的词向量模型不能有效表示出上下文语序信息，语义表示能力不够强，还需要进一步的改进。同时，有些带标签的数据集仅在狭窄领域使用，这可能导致过拟合问题或域偏差。

（2）与英语 Wikipedia，YAGO，Freebase 等知识库相比，中文百科全书和其他语言的知识库开起步相对较晚并且不成熟。

（3）尽管全局模型已经取得重大进步，但其仍有一定局限性：全局模型同样遭受着数据稀缺的问题，并且无法引入潜在的区分特征；全局模型中的联合推理机制导致计算量极大，尤其是在文档较长的情况下，实体图可能包含数百个节点。

3 基于深度学习的实体链接方法

为了避免过度依赖于人工设计的手工特征，减少对语言学知识的依赖,实体链接任务逐渐转向借助深度学习中神经网络强大的特征抽象和泛化能力,学习文本中潜在的语义信息等基本特征。基于深度学习的方法的核心是构建多类型，多模态的上下文和知识的统一表示，将不同含义不同类型的信息映射到同一特征空间，并对多源信息和多源文本之间的关系进行建模。目前，如何利用深度学习解决资源缺乏问题、如何在深度学习方法中融入知识指导以及考虑多任务之间的约束是当前的研究的热点。接来下分别详细介绍已有的研究中，实体指称识别的方法、候选实体集合生成方法以及基于深度学习的局部和全局实体消歧模型，最后介绍近 3 年性能较好的模型。

3.1 实体指称识别和候选实体集生成的方法

在实体指称的识别方面，传统的方法大多通过利用维基百科中的重定向页面、消歧页面、类别信息和超链接信息来构建实体别名词典，充分反映了实体指称与其候选实体之间的映射关系。利用这些信息，Bunescu 和 Pasca [12]使用实体的标题、重定向名称、消歧名称作为实体的名称集合，并将从名称到实体的一对多映射关系集成到一个字典里以进行指称识别。在候选实体集生成方面,已有研究通过统计维基百科以及其他公开知识库中实体表述和实体的共现情况来解决。但此方法存在了明显的问题，这种统计方法不分

领域，也不设上限，从而导致候选集中包含大量的噪声。以往的实体链接任务使用的知识库是 2014 年的维基百科(Wiki_2014),随着维基百科 2018 版本的发布,实体链接模型切换到了规模更大、内容更丰富的 Wiki_2018 上，本文将在 4.1 节里详细介绍两版维基百科的信息。

3.2 基于深度学习的实体消歧局部模型

对于局部模型，早期的研究大多侧重于设计有效的人为特征和复杂的相似性度量，以获得更好的消歧性能。相反，He 等人，2013[13]学习实体的分布式表示来测量相似性，而不需要人为的特征，这样单词和实体保留在联合语义空间中，候选实体可以直接基于向量相似性进行排名。他们使用了自编码器模型，实体表示由上下文文档表示和类别表示组成。基于深度神经网络(DNN)学习实体的文档表示，使用卷积神经网络获取类别表示。由使用简单的启发式规则过渡到将单词和实体用连续空间中的低维向量表示,实体的表述和实体的特征自动从数据中学习,最后对候选实体综合排名，链接到对应的实体。随后，Sun 等[14]在 2015 年提出将表述和实体以及上下文进行嵌入式表示,通过卷积神经网络提取特征,最后计算表述和实体的相似度进行链接。2016 年, Francis-Landau 等人[15]以 Sun 等[14]为基础，加入了堆叠去除噪声的自动编码器，来分别学习文本的上下文和实体的规范描述页面,一定程度上提升了链接性能。

Ganea 和 Hofmann，2017[16]构造了基于广泛的实体词共现数据的目标函数来调整传统的 Word2Vec 模型，从而提高了词向量模型的语义表示能力。同时，提出了用局部和全局模型结合的方式进行链接, 奠定了后续研究中局部模型与全局模型联合训练研究方法的基础。他们在局部模型中提出了使用软注意力(soft attention)和硬注意力(hard attention)来筛选上下文中的单词,从而进一步提升了链接性能。随后，通过分析 Ganea 和 Hofmann, 2017[16]的链接错误的案例，Chen 等，2020[17]发现模型经常将提及链接到类型错误的不正确实体，为了解决这一问题，他们将基于 BERT[18]的实体相似性评分集成到最新模型的局部模型中，以更好地捕获潜在实体类型信息，最终纠正了 Ganea 和 Hofmann, 2017[16]大部分的链接错误案例。

3.3 基于深度学习的实体消歧全局模型

在早期的全局实体消岐研究中，Han 等[5]构建了基于知识库的实体图，其以实体指称和候选实体为节点，包含了指称-实体、实体-实体的关系，同时提出了 PageRank\Random Walk 协同推理算法，得到实体指称所指向的实体。其中，基于图的随机游走算法描述如图一所示。之后 Hoffart 等[19]在 Han 等[5]的基础上，采用实体流行度、文本上下文相似度等对实体图中的实体指称-实体边进行加权，用映射实体一致性对实体-实体边加权，然后计算对每个指称只包含一条指称-实体边的稠密子图，得到指称-实体映射结果。然而，这些方法是不可微的，因此很难集成进入神经网络模型。

图的随机游走算法：
输入：初始化分布矩阵v_0，图模型的转移概率矩阵p
输出：图的稳定状态v_*
1：初始化$v=v_0$
2：循环
3：$v=v_{new}$
4：计算$v_{new}=\alpha * p^T * v+(1-\alpha) * v_0$
5：直到v稳定或者迭代次数超过某阈值

图 1　基于图的随机游走算法描述

Fig. 1　Graph based random walk algorithm description

对于全局模型的集体推理机制，其计算量极大的缺点通过近似优化技术得到了缓解。Globerson (Globerson 等，2016)[20]介绍了使用循环信念传播(LBP)(Murphy 等，1999)[21]用于集体推理。Ganea (Ganea 和 Hofmann, 2017)[16]通过截断拟合 LBP, 不滚动的可区分消息传递解决了全局训练问题。为了克服训练数据不足的第问题，(Gupta 等，2017)[22]探索了大量的维基百科超链接，使用多种信息源（例如其描述，提及的上下文及细粒度类型）为每个实体学习统一的密集表示，而无需任何特定于领域的训练数据或手工设计的功能。但这些潜在的注释包含很多噪音，这可能会给简单的消歧模型带来错误。

为了解决相同子句中相同的两个表述链接到知识库中不同实体的情况, Le 和 Titov，2018[23]在 Ganea 和 Hofmann, 2017[16]的基础上对表述进行关系建模并以特征的形式加入全局模型中,取得了较好的性能。

Guo 和 Barbosa, 2018[24]提出了一种贪婪的全局命名实体消歧算法，该算法利用了在知识库产生的子图上进行随机游走传播产生的概率分布之间的互信息,链接性能得以提高。同年，为了解决现有方法依赖于局部上下文独立的解析实体，可能会由于局部信息的数据稀疏而失败的问题，Cao 等人，2018[25]将图卷积网络应用到子图上，将实体链接的局部上下文特征和全局相关信息集成起来高效学习，提高了链接性能。尽管在这些模型中各个实体之间的语义依赖性能够通过构建神经网络自动建模，外部知识库的指导始终被忽略。为了解决上述问题，Xue 等人，2019[26]等采用具有随机游走层的神经网络利用外部知识来实现集体实体链接，进一步提高了链接性能，证实了探索外部知识库以建模不同的实体之间的全局语义相互依赖性是有效的。

为了解决全局模型尝试优化所有提及的整个链接配置，从而造成了很高的时间复杂度、内存消耗和计算量的问题，Yang 等，2019[27]提出了从先前链接的实体中积累知识作为动态上下文，以增强以后的链接决策的方法。积累的知识包括链接实体的固有属性和紧密相关的实体。与其他全局模型相比，此模型只需要遍历一遍所有实体指称，就可以在训练和推理上产生更高的效率。同时，在 5 个公开数据集上按不同链接顺序、注意力机制的大量实验表明，此模型具有良好的性能，使处理长文档的大规模数据成为了可能。

为了解决人工标注的数据集昂贵且缺乏的问题，Le 和 Titov，2019a[28]为未标记文档中的每个实体指称构建高召回率的候选实体列表，使用候选列表作为弱监督，以约束文档级实体链接模型。他们展示了如何使用 Wikipedia 和未标记的数据来构建一个精确的链接器，该链接器的性能可以与使用昂贵的人工监督而构建的链接器相媲美。同年，Le 和 Titov，2019b[29]针对对于不存在标记数据或标记数据非常有限的设置（例如，法律领域或大多数科学领域），实体链接工作进展甚微的现象，提出将实体链接任务定义为一个多实例学习问题，并依赖于表面匹配来创建初始的嘈杂标签作为弱/远程监督的方法，展示了如何将实体链接问题构造为一个远程学习问题。

上述不同模型在 5 个跨领域的实体链接数据集上的测试结果如表 1 所示，每个数据集上的最高得分将加粗展示，Le and Titov , 2019a [27]采用的是弱监督的方法，其他研究者均使用了人工标注的数据集来训练，5 个数据集将在 4.1 节详细介绍。可以看到，Xue 等人，2019[26]在多个数据集上均取得了最好的性能，证明了外部知识库的指导的有效性；Yang 等，2019[27]在 MSNBC 数据集上取得了最佳的效果，一定程度上解决了全局模型的联合推理机制造成高内存消耗和计算量的问题；Le 和 Titov，2019a[28]已经取得了与其他使用人工监督构建的模型相媲美的链接性能，证明了采用弱监督用来约束模型的有效性，一定程度上解决了人工标注的数据集缺乏的问题。在实验过程中，随着实体指称数量 k 的增加，大部分全局模型的运行时间将会显著增加，而 Yang 等，2019[27]的模型保持了线性的增长，同时一直保持了较低的内存占用，并且

表 1　不同模型实验结果对比

Table 1　Comparison of experimental results of different models

Model	MSNBC	AQUINT	ACE2004	CWEB	WIKI	Avg
Ganea and Hofmann, 2017	93.7	88.5	88.5	77.9	77.5	85.22
Le and Titov, 2018	93.9	88.3	89.9	77.5	78.9	85.51
Guo and Barbosa , 2018	92	87	88	77	84.5	85.7
Cao 等, 2018	—	88	90	—	—	—
Xue 等, 2019	94.43	91.94	90.64	79.65	85.47	88.43
Yang 等, 2019	94.57	88.53	90.14	75.59	78.84	—
Le and Titov , 2019a	92.2	90.7	88.1	78.2	81.7	86.18
Chen 等, 2020	93.4	89.8	88.9	77.9	80.1	86.02

节约了相比于 Le and Titov , 2018[23]约 80%的能耗，再次说明了他们的模型在处理大规模数据上的优势。由于需要在整个实体图上做推理，LBP 和基于 PageRank/random walk 的方法的时间复杂度为 $O(k^2n^2)$，而得益于只考虑相邻的指称，Cao 等, 2018[25]取得了较低的时间复杂度 $O(kn^2)$。关于实体的表示，Chen 等, 2020[17]

根据实体嵌入执行了实体类型预测任务，结果表明，他们使用 BERT[18]生成的实体嵌入的性能显著超过了 Ganea and Hofmann, 2017[16]的实体嵌入，表明了他们的模型可以更好的捕捉实体的类型信息。

在非端到端的模型中，分别处理实体指称检测和实体消歧两个步骤，他们之间的重要依存关系被忽略，由实体指称检测引起的错误将传播到实体消歧，而无可能恢复。Kolitsas 和 Ganea，2018[30]提出了第一个神经端到端的实体链接模型，将所有可能的区域视为潜在的指称，并学习其实体候选者的上下文相似性得分，这对实体检测和实体消歧的决策均有用。利用关键组件，即单词、实体和提及嵌入，他们证明，工程化特征几乎可以被现代神经网络完全取代。

4 实体链接的评测方法

4.1 实体链接常用数据集

AIDA-CoNLL[31]是最大的人工标注的实体消歧的数据集之一，它是在 CoNLL 2013 实体识别数据集标注的，题材是路透社新闻。实体链接模型通常使用 AIDA-CoNLL[31]数据集中的 AIDA-train 作为训练集, AIDA-A 作为验证集, AIDA-B 作为测试集。测试集还包含 Guo 和 Barbosa[32]发布的 MSNBC、AQUAINT、ACE2004 和 WNED-WIKI (WW),以及由 Gabrilovich [33]等人发布的 WNED-CWEB (CWEB)。在以上 6 个测试集中，只有 AIDA-B 与 AIDA-train 属于相同的领域,其他 5 个测试集都是来自不同的领域,这又增加了实体链接的难度，容易造成过拟合问题或地域偏差问题。表 2 详细地给出了所有数据集的详细信息,从每篇文档中拥有的实体指称的个数可以看出，数据集存在了一定的稀疏性问题。

表 2　实体链接数据集

Table 2　Datasets of entitiy linking

数据集	指称数	文档数	平均每个文档的指称数
AIDA-train	18448	946	19.5
AIDA-A	4791	216	22.1
AIDA-B	4485	231	19.4
MSNBC	656	20	32.8
AQUAINT	727	50	14.5
ACE2004	257	36	7.1
WW	6821	320	21.3
CWEB	11154	320	34.8

实体链接通常使用的知识库为维基百科，是基于网络的免费的百科全书，它包含有关传统百科全书的主题以及年历，地名词典和时事主题的条目。维基百科是互联网上最受欢迎的参考网站，每天收到数千万点击，分为 2018 版和 2014 版。其中，Wiki _2018 在规模上约是 Wiki_2014 知识库的 1. 5 倍,蕴含更丰富的信息，两者的详细数据如表 3 所示。其他常用的知识库还包括 Freebase、YAGO、DBpedia 等。

表 3 知识库的详细信息

Table 3　Detail information of KB

知识库	文档数	锚的数量	大小/GB
Wiki_2014	4459082	18611834	11.16
Wiki_2018	9618296	26916035	16.78

4.2 评测方法

随着实体链接研究的发展，如何对比不同的实体链接方法也成为研究者关注的重点。实体链接模型通常采用准确率 (Precision) 、召回率 (Recall) 以及 F1 值对实验结果进行评估，定义式如下所示。准确率

重点关注所有待消歧的实体指称中有多少能够被正确消歧，召回率则关注待消歧的实体候选集里含有正确实体的概率，F1 则可以综合评价模型的性能。

$$\text{Precision} = \frac{\text{正确消歧的实体指称}}{\text{待消歧的实体指称总数}} \times 100\%$$

$$\text{Recall} = \frac{\text{含有正确实体的候选集数}}{\text{待消歧的实体候选集数}} \times 100\%$$

$$\text{F1} = \frac{2 \times \text{准确率} \times \text{召回率}}{\text{准确率} + \text{召回率}} \times 100\%$$

4.3 实体链接工具

Dexter[34]是当前比较常用的开源实体链接框架之一，其利用维基百科中的词条来实现实体链接，提供了开发实体链接技术所需的工具。Dexter 是一个标准程序，无需高性能硬件或安装其他软件(例如，数据库等)，用户能够轻松使用。流行的开源实体链接服务还包括 TAGME(Ferragina 和 Scaiella，2010)[35]、AGDISTIS[36]等，TAGME 是第一个对短文本片段(搜索引擎结果的片段，推文，新闻等)进行准确且即时注释的系统，并将它们链接到相关的 Wikipedia 页面；AGDISTIS 能够可以有效地检测输入文本中给定的一组命名实体的正确地址，并将实体链接到对应的 DBpedia 界面。

5 结语

本文首先介绍了实体链接任务的定义、核心技术、相关研究和目前存在的问题，接着对最近几年基于深度学习的命名实体链接相关问题和研究现状进行系统性的介绍。评测会议上的研究现状表明，尽管实体链接领域已有多年的研究，但依然存在一些问题：目前已有的研究大多专注于英文实体链接，对非英语语言的实体链接关注较少；缺少受到广泛认可的实体链接的评测框架，不同实体链接研究在针对的问题、链接的步骤、选用的评测数据集等方面存在较大的差异，难以进行统一的有效的比较。展望未来，实体链接可能的研究方向如下：

（1）跨语言的实体链接。现有的研究大多针对某一种语言的实体链接，可以使用双语言或多语言的知识库进行联合学习,利用不同语言之间的互补性进一步提升实体链接的性能。同时，也可以利用高资源语言的丰富知识来帮助低资源语言中的实体链接。Upadhyay 等，2019[37]将多种语言的监督结合起来解决可用于监督的资源有限的问题，是首个训练一个模型用于多语言的方法，使用于监督的资源得以高效利用。

（2）实体链接的评测框架。Henry Rosales-Méndez 等，2019[38]提出了一种模糊召回指标来解决缺乏共识的问题，并以比较在线 EL 系统选择的细粒度评估结果作为结论，取得了不错的进展。

（3）端到端的实体链接模型。Kolitsas 和 Ganea，2018[30]提出了第一个神经网络端到端实体链接模型，并展示了共同优化实体识别和链接的好处，实体链接的应用性得以提升。

（4）弱监督/无监督的实体链接。Le 和 Titov，2019a[27]的模型在很大程度上优于以前的方法，因为以前的方法使用了相同形式的监督，并且他们的模型的性能与专门为实体链接问题而训练的全监督模型相抗衡。这个结果可以暗示人类注释的数据对实体链接不是必须的，还可以利用维基百科和网络链接，这两个信息来源可能是相辅相成的。

参考文献

[1] Zheng Chen and Heng Ji. 2011. Collaborative ranking: A case study on entity linking. In EMNLP.

[2] Andrew Chisholm and Ben Hachey. 2015. Entity disambiguation with web links. TACL.

[3] Nevena Lazic, Amarnag Subramanya, Michael Ringgaard, and Fernando Pereira. 2015. Plato: A selective context model for entity resolution. TACL.

[4] Ikuya Yamada, Hiroyuki Shindo, Hideaki Takeda, and Yoshiyasu Takefuji. 2016. Joint learning of the embedding of words and entities for named entity disambiguation. In CoNLL.

[5] Han X，Sun L，Zhao J．Collective Entity Linking in Web Text：a Graph-Based Method[C]//Proceedings of the 34th Internation- al ACM SIGIR Conference on Research and Development in Information Retrieval，

Beijing，China，2011：765—774．
[6] Taylor Cassidy, Heng Ji, Lev-Arie Ratinov, Arkaitz Zubiaga, and Hongzhao Huang. 2012. Analysis and enhancement of wikification for microblogs with context expansion. In COLING.
[7] Zhengyan He, Shujie Liu, Yang Song, Mu Li, Ming Zhou, and Houfeng Wang. 2013. Efficient collective entity linking with stacking. In EMNLP.
[8] Xiao Cheng and Dan Roth. 2013. Relational inference for wikification. In EMNLP.
[9] Greg Durrett and Dan Klein. 2014. A joint model for entity analysis: Coreference, typing, and linking. TACL.
[10] Hongzhao Huang, Yunbo Cao, Xiaojiang Huang, Heng Ji, and Chin-Yew Lin. 2014. Collective tweet wikification based on semi-supervised graph regularization. In ACL.
[11] Yotam Eshel, Noam Cohen, Kira Radinsky, Shaul Markovitch, Ikuya Yamada, and Omer Levy. 2017. Named entity disambiguation for noisy text. In CoNLL.
[12] Bunescu R C，Pasca M.Using Encyclopedic Knowledge for Named entity Disambiguation[C]//Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics,Tren to，Italy，2006：9-16 ．
[13] Zhengyan He, Shujie Liu, Mu Li, Ming Zhou, Longkai Zhang, and Houfeng Wang. 2013. Learning entity representation for entity disambiguation. In ACL.
[14] SunY, LinL, TangD, et al. Modeling mention, context and entity with neural networks for entity disambiguation [C]// Proceedings of the 24th International Conferenceon Artificial Intelligence. 2015 : 1333-1339.
[15] Francis-Landau M, Durrett G, Klein D.Capturing semantic similarity for entity linking with convolutional neural networks [C]// Proceedings of NAACL-HLT.2016 : 1256-1261.
[16] Ganea O E, Hofmann T. Deep joint entity disambiguation with local neural attention [C]//Proceedings of the 2017 Conferenceon Empirical Methods in Natural Language Processing.2017 : 2619-2629.
[17] Shuang Chen, Jinpeng Wang, Feng Jiang, Chin-Yew Lin. 2020. Improving Entity Linking by Modeling Latent Entity Type Information. In AAAI.
[18] Devlin J , Chang M W , Lee K , et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. 2018.
[19] Hoffart J，Yosef M A，BordinoI，et a1．Robust Disambiguation of Named Entities in Text [C]//Proceedings of the Conference on Empirical Methods in Natural Languag Processing，Edinburgh，Scotland，U K， 2011：782-792
[20] Amir Globerson, Nevena Lazic, Soumen Chakrabarti, Amarnag Subramanya, Michael Ringgaard, and Fernando Pereira. 2016. Collective entity resolution with multi-focal attention. In ACL.
[21] Kevin P Murphy, Yair Weiss, and Michael I Jordan.1999. Loopy belief propagation for approximate inference: An empirical study. In Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence, pages 467–475. Morgan Kaufmann Publishers Inc.
[22] Nitish Gupta, Sameer Singh, and Dan Roth. 2017. Entity linking via joint encoding of types, descriptions, and context. In EMNLP
[23] Le P , Titov I. Improving entity linking by modeling latent relations between mentions [C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics.2018 : 1595-1604.
[24] Zhaochen Guo and Denilson Barbosa. Robust named entity disambiguation with random walks. Semantic Web, pages 459–479, 2018.
[25] Yixin Cao, Lei Hou, Juanzi Li, and Zhiyuan Liu. Neural collective entity linking. In COLING, pages 675–686, 2018.
[26] Xue M , Cai W , Su J, et al. 2019.Neural Collective Entity Linking Based on Recurrent Random Walk Network Learning.In IJCAI.

[27] Yang X , Gu X , Lin S , et al. 2019.Learning Dynamic Context Augmentation for Global Entity Linking. In ACL.
[28] Le P , Titov I .2019a. Boosting Entity Linking Performance by Leveraging Unlabeled Documents. In ACL.
[29] Le P , Titov I .2019b. Distant Learning for Entity Linking with Automatic Noise Detection. In ACL.
[30] Kolitsas N , Ganea O E , Hofmann T. 2018.End-to-End Neural Entity Linking. In CoNLL.
[31] Hoffart J ,Yosef M A, Bordino I ,et al.Robust disambiguation of named entities in text [C]// Proceedings of the Conferenceon Empirical Methods in Natural Language Processing. Association for Computational Linguistics,2011 : 782-792.
[32] GuoZ , BarbosaD.Robust named entity disambiguation with random walks [J] .Semantic Web , 2018, 9(4): 459-479.
[33] Gabrilovich E , Ringgaard M , Subramanya A.FACC1 : Freebase annotation of ClueWeb corpora ,Version1 2013.
[34] Diego Ceccarelli ,Claudio Lucchese ,Raffaele Perego ,et al.2013.Dexter: an Open Source Framework for Entity Linking.In ESAIR.
[35] Paolo Ferragina and Ugo Scaiella. 2010. TAGME: On-the-Fly annotation of short text fragments (by wikipedia entities). In Proceedings of the 19th ACM International Conference on Information and Knowledge Management,CIKM 10, page 16251628, New York, NY, USA. ACM.
[36] Usbeck R , Ngomo A C N , Auer S , et al. AGDISTIS - Graph-Based Disambiguation of Named Entities Using Linked Data[C]// International Semantic Web Conference. Springer International Publishing, 2014.
[37] Upadhyay S , Gupta N , Roth D .2018. Joint Multilingual Supervision for Cross-lingual Entity Linking.In EMNLP.
[38] Henry Rosales-Méndez, Aidan Hogan, Barbara Poblete.2019. Fine-Grained Evaluation for Entity Linking.In ACL.