



开放科学
(资源服务)
标识码
(OSID)

实体链接技术研究述评

张丹阳 李楠 陈翀

北京师范大学政府管理学院 北京 100875

摘要: 实体链接技术在知识组织、信息检索、语义出版等诸多领域有着重要研究意义和实践价值,可广泛应用于知识库扩容、机器翻译、自动问答等。本文论述了实体链接的主要研究问题,并对近年来实体链接的关键技术进行总结。此外,本文对基于短文本的实体链接技术也进行了介绍。最后,本文指出了实体链接研究目前的局限之处以及未来研究发展趋势。

关键词: 实体链接; 实体识别; 实体消歧; 短文本实体链接

中图分类号: G350, TP391.1

Review on Entity Linking Technology Research

ZHANG Danyang LI Nan CHEN Chong

School of Government Management, Beijing Normal University, Beijing 100875, China

Abstract: Entity linking technology is of great theoretical significance and practical value in many fields such as knowledge organization, information retrieval, semantic publishing, etc. It has been playing important roles in knowledge base population, machine translation, automatic question answering and so on. This article introduces the main research problems on entity linking. Key technologies of recent years are also summarized. In addition, this article also investigates entity linking technologies based on short-text. Finally, the article points out the current limitations and future research trends of entity linking research.

Keywords: Entity linking; entity recognition; entity disambiguation; entity linking for short-text

作者简介: 张丹阳 (1993-), 硕士研究生, 研究方向为文本挖掘; 李楠 (1991-), 硕士研究生, 研究方向为文本挖掘; 陈翀 (1975-), 通讯作者, 博士, 教授, 研究方向为知识组织、文本挖掘, E-mail: chenchong@bnu.edu.cn。

引言

由于计算机很难像人一样理解自然语言表达的复杂性,常常不能识别文本中的实体名称所指的对象或概念。这往往会影响到问题理解、知识库构建、知识点解释等一系列任务的正确性。因此,实体链接一直是自然语言处理、知识组织、信息检索、文本挖掘、语义出版等领域共同关心的问题。实体链接是将文本中代表实体的指称与特定知识库中的条目相链接,也被称为命名实体链接^[1]、实体消歧、实体共指消解等。其中,建立链接的重要依据是指称所在的文本上下文与特定知识库中实体的匹配程度,特定知识库包括 DBpedia^[2]、YAGO^[3]、TAP^[4]、Freebase^[5]、维基百科^[6]以及自建的知识库等。随着各领域迅速发展,很多研究会根据具体的问题域,在维基百科等大型开放知识库基础上补充领域词条自行构造知识库,如 Zhishi.me^①、CN-Dbpedia^②、Xlore^③等。它们主要是从百科网站等文本页面中提取信息,经过滤、融合、推断形成高质量的结构化数据,并在更新维护中逐渐完善和扩充。这些知识库在相关研究和实践中具有重要作用,可以支撑实体搜索、个性化推荐等智能服务。在互联网环境下,利用实体链接技术将网络文本与知识库桥接,从海量信息识别有用对象,还有助于实现语义互联。

本文主要贡献在于:总结实体链接的主要研究问题,对短文本实体链接研究进行了较为详细地阐述;对近年实体链接技术使用的新方

法进行了归纳和梳理;论述实体链接研究中的不足之处以及发展趋势。

1 实体链接研究问题

实体链接的任务是将给定文本中出现的实体指称与目标知识库中若干候选实体关联起来。研究中应用较多的工具有 TAGME^[7]、Wikifier^[8]。实体链接主要包括实体识别和实体词义消歧两个阶段。实体识别^[9]也叫做指称识别,指称包括具有特定意义的名词,例如人名、地名、专有名词术语等。实体识别的难点在于:从自然语言文本中确定实体的边界和类别;一些开放性任务在识别阶段不限定实体类型,在链接知识库阶段再根据实体所在上下文确定其类别信息。另外,由于自然语言中普遍存在一词多义和别名现象,识别出的实体指称在大多数情况下并不能唯一确定其在知识库中所链接的目标实体,因此需要利用给定实体指称所在的上下文语义来消除其在指代概念上的歧义。总而言之,实体识别是实体消歧的输入和前提;实体消歧是实体链接映射的过程和结果。

1.1 实体识别

实体识别在1995年第六届消息理解会议(MUC-6)中被设立为一种独立的信息提取任务。此后,实体类别由人名、地名和组织机构名扩充到产品名、书名、项目名以及领域术语等;已有的类别也被进一步细分,例如地名被细分为城市、州和国家等。

① <http://zhishi.me/>

② <http://kw.fudan.edu.cn/cndbpedia/search/>

③ <https://xlore.org/>

实体识别主要可以借助规则、字典以及结合机器学习的方法实现。在对文本进行分词和词性标注后,基于规则的方法通常依据设定的拼写规则、构词规则、指示词、前后缀字串定义模板来识别实体指称边界;基于词典的方法则是直接采用已有词典中的词条来识别实体指称。由于规则难以涵盖所有的语言现象、不同规则之间易产生冲突,而词典则受限于词条规模、前期建设成本很高,所以,这两种方法虽然准确率较高,但是可移植性较差。基于机器学习的方法一般通过预先标注的语料对模型进行训练,通过概率模型学习实体类别参数,并以此计算和预测测试语料中的实体所属类别信息,从而对实体进行识别。这种方法更加灵活、鲁棒性好,但需要依赖标注语料,且当训练语料较少时,数据稀疏问题会影响模型的学习效果。

1.2 实体消歧

歧义性和多样性是自然语言的固有属性,多个实体可能存在相同的实体指称,一个实体也可能存在多个实体指称。实体链接研究目前所面临最主要的困难就是实体歧义问题,它影响着整个实体链接系统的性能。实体消歧的主要目标是根据实体指称及其所在上下文,对候选实体集中的实体进行排序,选出最恰当的实体。生成候选实体集的方法主要包括借助外部词典^[10-11]、字面形式扩展^[12]以及借助搜索引擎返回结果^[13]等。外部词典虽然并不总是可得,但词典中往往包括实体的多种表达方式,包括变体、缩写、混淆名称等,例如维基百科实体的消歧页面、重定向页面可以提供多种匹配依

据。利用指称词汇直接与词典中的实体进行匹配准确率高,但是适用范围受限。基于字面形式扩展的方法是识别实体指称所在的相关文档中的其他可能的扩展变体,并利用这些扩展形式来生成候选实体集^[14]。还有方法^[13,15]将实体指称输入搜索引擎,检索出相关度高的词条类页面作为确定候选实体的依据。

在实体消歧中,可以利用上下文独立的特征和上下文相关的特征。基于上下文独立特征的方法只考虑实体指称与候选实体集的匹配关系,如根据候选实体使用率^[16,17]、实体类型^[13]、实体间名称相似度等。在维基百科中进行实体消歧时,实体使用率是指某个候选实体可以链接到维基百科页面词条的概率。实体类型主要依据实体指称与候选实体集中实体的类型,且与实体指称类型相同的实体成为最终目标实体的概率较大。实体间名称相似度利用编辑距离^[18]、Dice 系数^[15]、左右 Hamming 距离^[13]等比较字符串。基于上下文相关特征的方法在考虑实体本身特征的同时,结合实体所在上下文背景知识,主要包括文本相似度、实体映射的一致性^[14]等。将实体指称所在的文本上下文以及与候选实体相关的文档中包含的语义特征或统计特征用向量的形式表示,通过计算两种向量之间的相似度进行排序,从而确定其在特定知识库中对应的实体。其中,常用于衡量相似度的方法是余弦相似度^[6,11,13]等。

2 实体链接技术方法

目前,相关研究对实体链接技术方法有多种划分方式:根据模型的不同可以划分为基于

统计模型、基于规则模型；根据是否依赖人工知识分为无监督、半监督、有监督方法。本文从技术发展和演变的角度对实体链接主流方法分为基于统计模型的方法、基于深度学习的方法和基于多源知识的方法。本章主要阐述各类技术方法的本质和思路，以及不同方法所使用的核心特征。

2.1 基于统计模型的方法

基于统计模型的方法是早期实体链接研究中常用的方法。该类方法往往利用一些统计特征将实体指称和候选实体表示为特征向量，通过余弦相似度等计算方法进行实体排序和选择。基于统计模型方法主要是源于传统信息检索的文本建模方式，将文本看作词袋，统计词汇的特征。但这样容易割裂文本的上下文语境，缺乏对实体语义层面的考量。

在实体识别方面，最常用到的机器学习方法包括隐马尔可夫模型（HMM）、相对熵（KL）等，如利用最大熵马尔可夫模型（MEMM）将实体识别转换为字符串的分类任务，利用支持向量机（SVM）将实体识别转化为二分类问题，利用条件随机场（CRF）将问题转化为序列标注任务等。在这些方法中，通常将可能包含实体指称的文本作为输入，期望预测并输出候选实体及其类别信息。其中，统计模型对输入和输出之间的关联进行建模，机器学习算法对所建模型的参数进行学习^[19]。例如，基于HMM的实体识别中，HMM是一个生成模型，遵循特征分布独立假设，根据输入的已标注训练文本，统计其中关于文本序列、文本标注序列、状态转移概率、初始概率等信息，并根据这些

统计值计算测试语料中文本序列背后的标注序列，从而得到文本中的实体指称^[20]。而基于CRF的实体识别中，CRF是一个判别式模型，无特征分布独立假设，引入了实体本身特征和实体所在上下文的特征。CRF对输入的文本序列及其标注序列的条件概率进行建模，并利用这个模型对测试语料中文本序列的标注结果进行预测，从而识别其中的实体指称^[21]。

在实体消歧中，一般采用实体相关的统计信息、实体分布信息、实体相似度、文本主题信息等特征改善候选实体排序。Gottipati等^[22]利用上下文信息和全局信息扩展实体指称和候选实体的语言模型，并通过计算两者之间的KL距离对候选实体进行排序和筛选。Zhang等^[23]利用词法特征、词-类别对、实体类型等特征，采用支持向量机分类器进行实体消歧，并构造基于词袋和实体共现的特征向量，计算实体指称与候选实体间的余弦相似度，选择得分最高的候选实体。武川等^[24]利用候选实体上下文指称相似度特征进行实体排序和消歧提升了实体链接效果。谭咏梅等^[25]基于上下文信息，对实体指称进行扩充，在维基百科中检索候选实体列表，利用ListNet排序学习算法对候选实体进行排序。主流的实体排序方法包括二分类方法、Learning to Rank和基于图的方法等。一些系统将候选实体排序问题表述为二分类问题。这是一种监督学习的方法，使用二元分类器来确定给定的一对实体指称和候选实体是否存在指向关系。该方法用人工标注作为训练集进行分类，如果实体指称可以表示实体，则为正例，反之为负例。典型的二元分类器包括支持向量机^[26-27]、向量空间模型（VSM）和K近邻分类器等。但

是当实体指称的多个候选实体都被二分类方法判别为正例时,必须利用其他技术继续进行选择,从而确定最佳候选实体。同时,为避免二元分类器中忽略候选实体集和实体指称的候选实体之间的关系,许多实体链接系统^[10,13,22]利用 Learning to Rank^[28]对候选实体集排序,这种方法考虑同一实体指称的候选实体之间的关系。Learning to Rank也是一种监督学习方法,可以根据训练数据自动构建排序模型。这种方法中用于学习排名模型的训练数据包括了实体指称对应的候选实体集列表且其中部分实体已指定顺序,另外也对正确映射的实体位置进行了约束。一般研究利用 Ranking SVM 框架来学习排序模型。例如, Ratnov 等^[29]构建了局部特征和全局特征,通过 Ranking SVM 训练得到排序模型;Manahan^[15]也采用了 Ranking SVM 解决了实体链接中候选实体排序问题。基于图的方法将文本中的所有实体构建为图,并且图中包含实体间的关系、候选实体间的关系、实体指称与候选实体间的关系。这种方法全局模拟了文本中不同实体之间主题相互依赖性,与 PageRank 算法相类似,能够更好地利用知识库中实体间的语义关联关系,但在上下文信息不充分的情况下,较难区分语义相近的实体。例如,张涛等^[30]利用维基百科中概念之间的类别关系和链接关系构建了维基概念图模型,用随机游走算法来衡量实体指称与候选实体间的相似度,并根据相似度进行排序实现了实体链接。另外,还有一些研究结合实体链接与实体聚类的方法^[31]来提升实体消歧效果。

2.2 基于深度学习的方法

深度学习是基于多层人工神经网络的复杂

机器学习,主要学习样本数据的内在规律和不同层次的特征,并且广泛应用于计算机视觉、语音识别、自然语言处理等诸多领域。很多研究逐渐采用基于深度学习的方法来提升实体链接任务的效果^[32]。相比于统计模型方法,基于深度学习方法无需人为构造特征,通过神经网络算法对文本标注序列进行训练,学习实体、实体指称、上下文及其相互之间的关联关系,在低维稠密空间为不同实体及实体之间的语义关系构建统一表示,通过语义向量相似度排序得到目标实体。该方法主要优势在于表示了实体以及实体间的语义特征。

目前,利用卷积神经网络(CNN)、长短时记忆网络(LSTM)、双向长短时记忆网络(BiLSTM)等神经网络,与条件随机场(CRF)相结合的架构^[33]进行实体识别是一种有效的手段。这类方法将该任务转化为序列标注问题,利用有监督的方法标注输入文本,用神经网络进行特征训练,并将输出的特征向量通过 CRF 计算出最优联合概率,得到全局最优的标注序列,从而识别其中包含的实体指称及其向量表示。很多研究都单独或结合使用这些神经网络模型以提高识别效率,如 LSTM 与 CRF 相结合的模型,将实体的词向量表示和实体字符的向量表示相结合输入 LSTM 神经网络,并通过 CRF 层对 LSTM 的输出结果进行实体识别^[34];由于单向 LSTM 不能融合前后向信息,而 BiLSTM 神经网络与 CRF 相结合的模型在进行实体标注时能提高准确率并减少对词嵌入的依赖^[35];BiLSTM-CNNs-CRF 模型既结合了文本长序列的前后向信息,又具有一定的并行能力,不需要特征工程或数据预处理就可以得到

序列标注^[36]。张聪品等^[37]构建内嵌条件随机场的长短时记忆神经网络模型,使用长短时神经网络隐含层的上下文向量作为输出层实体的特征,用内嵌的条件随机场模型表示实体间的约束关系,从而能够有效识别出中文命名实体,为自动构建医疗知识图谱奠定了基础。针对深度学习对少量数据学习能力不足的问题,武惠等^[38]基于迁移学习用已有的知识源域数据构建出新的目标域数据集,并通过词向量、BiLSTM、CRF等融合上下文语义信息,克服了对人工特征和专家知识的依赖。针对实体消歧任务,有研究利用CNN捕获实体指称上下文和候选目标实体之间的语义对应关系的模型,同时利用多种粒度的主题信息衡量实体指称与候选实体在表征同一主题的能力以及它们之间的相似度特征,从而有效消除实体链接中存在的消歧问题^[39]。Huang等^[40]提出了一种基于深度神经网络(DNN)和语义知识图(KGs)的新型深度语义关联模型(DSRM),用于测量实体的语义相关性,并将深度学习模型训练得到的实体间语义相关度作为图中节点连边的权重。

除此之外,面对实际应用中各种任务和场景的切换,学者们致力于研究越来越多可移植性高、更加快速简便、适用于多种任务的模型。如注意力机制(attention)^[41]是神经网络的变形,能够依据应用场景将注意力集中到文本序列更重要的部分上,更好的提取上下文信息,常应用于基于词向量的文本处理任务中;基于迁移学习的思想^[37]可以与神经网络结合,实现跨语言、跨任务、跨领域的序列标注任务。一些新兴的基于深度学习的方法在实体表示、相似度计算等方面也进行了改进,如利用多源

异构文本信息进行实体的向量表示学习^[42];通过实体类型、实体在知识库中的描述及其所在上下文的联合编码进行实体链接^[43];通过神经网络计算文本细粒度相似性实现跨语言实体链接^[44]等。另有学者从三个不同方面利用DeepMatch、DeepType、DeepCosine三种模型来进行完成实体链接任务^[45]。其中,DeepMatch方法将实体链接问题转化为文本语义匹配问题,把实体之间的关系也纳入到向量嵌入中,对实体、实体指称、关系构成的元组进行嵌入学习^[46];DeepType^[47]方法在上下文相关字向量基础上,将实体候选序列中首尾两个字的向量连接在一起,最后进行全连接输出到各个候选实体类型进行多分类;DeepCosine方法^[48]将知识库中的三元组作为句子,获取实体的嵌入表示并进行实体预测。

2.3 基于多源知识的方法

在实体链接过程中往往依赖于背景知识库,而知识库的优劣在一定程度上也影响着实体识别和实体消歧的效果和性能。在实体链接研究中,特定知识库提供的背景信息对实体识别和消歧有着极其重要的作用。以维基百科为例,结构化的标题、锚文本等等页面信息对实体识别有很大的帮助,而详细文本描述信息则有助于实现上下文相似度对比,有利于进行实体消歧。目前,很多研究利用维基百科页面等信息提取语义知识来进行实体消歧^[49-50]。然而,在实体链接研究中,仅仅采用一种知识库会存在实体数量有限、实体更新不及时、实体获取来源单一等局限性。主要表现在:维基百科等知识库虽然被广泛认可和使用,但是也存在实体

缺失,信息并非全面完善;一些新实体往往存在于时效性很强的网络中,而一些比较规范的结构化知识库无法及时更新和补充。

因此,一些学者已经将基于多源知识的方法应用在实体链接相关的研究工作中。基于多源知识的方法本质是通过融合多种来源的知识扩展候选实体集,从而提升实体识别和消歧的准确性和全面性。一些研究在注重加强实体消歧效果的时候,往往会忽略利用多源知识来识别和构建候选实体集^[51]。一方面,可以利用多源知识构造词典。例如,利用 WordNet、SemCor 在线词汇知识库,构造出基于图的词义表示方法进行实体消歧^[52];利用维基百科和 Freebase 进行词典构建扩展实体识别结果^[53]。另一方面,可以基于多知识库对实体查询扩展。例如,管红英等^[54]把同义词词典、百科资源等知识与词袋模型相结合实现了基于多源知识的实体链接;周鹏程等^[55]以 Wikipedia 和 Freebase 为例融合多个知识库进行实体链接研究,表明了基于多知识库的实体链接方法的有效性。另外,Dredze 等^[13]还利用谷歌搜索结果补充了相关度较高的维基页面作为候选实体集合,扩展了外部背景知识。

目前,把来自不同数据源的知识融合,并将其与现有大规模知识库集成,使半结构化、非结构化数据都能有效获取的方法是基于多知识库算法的研究热点。虽然基于多源知识的方法能够较好的提升实体识别的质量以及实体消歧的效果,但在实体链接系统的整体性能方面还有较大的提升空间。融合更多的知识源对实体识别和实体消歧进行加强和补充,将进一步提高性能。

3 短文本实体链接

短文本包括微博、搜索 Query、用户对话内容以及文章标题等。由于缺乏充足的上下文信息、口语化程度较大以及中文语言自身特点等问题,基于短文本的实体链接在实体识别、歧义消解中存在更大的困难。因此,基于短文本的实体链接相关研究也很受重视。

以微博短文本为例,Meij 等^[56]采用多种特征识别微博信息中的实体,并将其链接到对应的维基百科页面。向宇等^[57]利用百度百科、实体专家库对实体进行规范化处理,并基于百度百科信息、微博数据、网络词语构建了消歧文本数据库,最后结合 TF-IDF 算法和 FastNewman 聚类算法进行实体链接。李禹恒等^[58]采用迭代消歧策略,将用户微博中已经识别出的实体作为先验知识来辅助后继的消歧任务,解决了微博上下文信息不足导致的准确率不高的问题。很多学者基于 2013 年和 2014 年自然语言处理与中文计算会议(NLPCC)发布的中文微博实体链接任务进行了短文本实体链接方法研究。例如,朱敏等^[59]改进拼音编辑距离算法和后缀词表匹配法,提出实体聚类消歧与基于百度百科词频的同类实体消歧相结合的消歧方法;陈万礼等^[60]利用 Ranking SVM 对候选实体集合进行排序,从而得到目标实体;冯冲等^[61]假设微博中的实体位于相近的语义空间,在词向量层面进行建模,并设计出一种基于词向量语义分类的实体链接方法。

此外,一些实体链接的相关会议,如 2019 年全国知识图谱与语义计算大会(CCKS 2019),也发布了面向中文短文本的实体链接

测评任务^[62]。在此任务中,最佳解决方案通过挖掘知识库中实体的描述文本获取实体的向量表示,利用字典匹配得到短文本中的实体,最后使用 BERT-EntityNameEmbedding (BERT-ENE) 模型对结果进行筛选,从而实现实体识别。在实体消歧部分使用基于 Transformer 的双向编码器 BERT^[63] 的二分类模型对候选实体进行预测并排序。

4 研究不足与发展

尽管在实体链接技术方面已有较多的研究,但是仍然存在以下一些局限和不足:

1) 由于中文语言自身的歧义性和多样性,对实体的表达不够规律,使得实体消歧存在一定困难。

2) 在上下文信息缺乏、短文本、跨语言、跨领域等情况中,相对精准地进行实体识别的方法不够完善。

3) 不同领域、不同形式的训练和标注语料的缺乏使得大部分依赖有监督模型的研究未能取得足够好的效果。

4) 语言的开放性和社会的发展将引起实体规模的扩张和含义的演变,而现有的技术方法并不能及时并全面地识别和抽取这些信息,且很少有模型可以适应这些变化^[19]。

5) 目前,缺少广泛认可的评价指标和评测工具对不同实体链接研究进行有效比较。

因此,未来实体链接研究将会呈现以下的发展趋势:

1) 更多的国际会议在推动实体链接研究发展的同时,逐步形成较为规范的实体链接技术

评测标准或框架。

2) 更多的学者和研究将会关注短文本实体链接研究,以及其他非文本类型的数据与知识库的关联研究。

3) 更多的研究将在实体消歧证据的挖掘、消歧算法性能的提升、高覆盖度的实体引用表的构建、信息自动抽取、迁移学习等方面有所深入^[19]。

4) 实体链接技术将应用于海量信息组织与发现,改善和提升文本信息以及资源的组织方式,从而提升自然语言处理任务性能和效果。

5) 不同领域实体链接问题的解决使得多源异构的知识相融合,与现有大规模知识库的集成,推动更有效的实体链接系统的出现以及语义网、知识图谱的改进。

5 结语

作为一项自然语言处理的热点研究内容,实体链接与多种自然语言处理任务息息相关。在很多实际应用中发挥重要作用。本文深入介绍和分析了实体链接的主要研究问题、关键技术核心及其特点,并对基于短文本的实体链接相关研究进行了阐述。最后,本文总结了实体链接研究面临的局限之处以及未来研究发展趋势。

参考文献

- [1] Zhang W, Sim Y C, Su J, et al. Entity Linking with Effective Acronym Expansion, Instance Selection and Topic Modeling[C]. In: Proceedings of the 22nd International Joint Conference on Artificial

- Intelligence, Barcelona, Catalonia, Spain. 2011: 1909-1914.
- [2] Auer S, Bizer C, Kobilarov G, et al. DBpedia: A Nucleus for a Web of Open Data[J]. Semantic Web, 2007, 4825:11-15.
- [3] Suchanek F M, Kasneci G, Weikum G. Yago:a core of semantic knowledge[C]. International Conference on World Wide Web. 2007.
- [4] Dill S, Eiron N, Gibson D, et al. SemTag and Seeker: Bootstrapping the Semantic Web via Automated Semantic Annotation [C]. Proceedings of the 12th International Conference on World Wide Web, Budapest, Hungary, 2003:178-186.
- [5] 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, Vancouver, BC, Canada, 2008: 1247-1249.
- [6] 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, Trento, Italy. 2006:9-16.
- [7] Ferragina P, Scaiella U. TAGME: on-the-fly annotation of short text fragments (by wikipedia entities)[C]. Conference on information and knowledge management. ACM, 2010:1625-1628.
- [8] Tsai C, Roth D. Illinois cross-lingual wikifier: grounding entities in many languages to the English wikipedia[C]. International conference on computational linguistics. ICCL, 2016:146-150.
- [9] Hachey B, Radford W, Nothman J, et al. Evaluating Entity Linking with Wikipedia[J]. Artificial Intelligence, 2013(194):130-150.
- [10] Shen W, Wang J, Luo P, et al. LINDEN: linking named entities with knowledge base via semantic knowledge[C]. The Web Conference, 2012: 449-458.
- [11] Shen W, Wang J, Luo P, et al. Linking named entities in Tweets with knowledge base via user interest modeling[C]. Knowledge Discovery And Data Mining, 2013: 68-76.
- [12] Varma V, Bysani P, B K R, et al. IIIT Hyderabad in Guided Summarization and Knowledge Base Population.[J]. Theory and Applications of Categories, 2010.
- [13] Dredze M, McNamee P, Rao D, et al. Entity Disambiguation for Knowledge Base Population[C]. International Conference on Computational Linguistics. Association for Computational Linguistics, 2010.
- [14] Shen W, Wang J, Han J. Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions[J]. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(2):443-460.
- [15] Monahan S, Lehmann J, Nyberg T, et al. Cross-Lingual Cross-Document Coreference with Entity Linking[J]. Theory and Applications of Categories, 2011.
- [16] Gattani A, Doan A H, Lamba D S, et al. Entity extraction, linking, classification, and tagging for social media[J]. Proceedings of the VLDB Endowment, 2013, 6(11):1126-1137.
- [17] Guo S, Chang M, Kiciman E, et al. To Link or Not to Link? A Study on End-to-End Tweet Entity Linking[C]. North American Chapter Of The Association For Computational Linguistics, 2013: 1020-1030.
- [18] Liu X, Li Y, Wu H, et al. Entity Linking for Tweets[C]. Meeting of The Association For Computational Linguistics, 2013: 1304-1311.
- [19] 中国中文信息协会. 知识图谱发展报告(2018)[EB/OL]. [2019-8-2]. <http://www.cipsc.org.cn/index.php>
- [20] 俞鸿魁, 张华平, 刘群, 等. 基于层叠隐马尔可夫模型的中文命名实体识别 [J]. 通信学报, 2006, 27(2):87-94.
- [21] 隋明爽, 崔雷. 结合多种特征的 CRF 模型用于化学物质-疾病命名实体识别 [J]. 现代图书情报技术, 2016(10):91-97.
- [22] Gottipati S, Jiang J. Linking Entities to a Knowledge Base with Query Expansion[C]. Empirical Methods In Natural Language Processing, 2011: 804-813.
- [23] Zhang W, Su J, Tan C L, et al. Entity Linking Leveraging Automatically Generated Annotation[C]. International Conference on Computational

- Linguistics, 2010: 1290-1298.
- [24] 武川, 陆伟. 基于上下文特征的短文本实体链接研究[J]. 情报科学, 2016(2):144-147.
- [25] 谭咏梅, 王睿, 李茂林. 基于上下文信息和排序学习的实体链接方法[J]. 北京邮电大学学报, 2015(5):33-36.
- [26] Chen Z, Ji H. Collaborative Ranking: A Case Study on Entity Linking[C]. Empirical Methods In Natural Language Processing, 2011: 771-781.
- [27] Pilz A, Paas G. From names to entities using thematic context distance[C]. Conference On Information And Knowledge Management, 2011: 857-866.
- [28] Liu T Y. Learning to Rank for Information Retrieval[J]. Acm Sigir Forum, 2010, 41(2):904-904.
- [29] Ratnoff L, Roth D, Downey D, et al. Local and Global Algorithms for Disambiguation to Wikipedia[C]. Meeting of the Association for Computational Linguistics, 2011: 1375-1384.
- [30] 张涛, 刘康, 赵军. 一种基于图模型的维基概念相似度计算方法及其在实体链接系统中的应用[J]. 中文信息学报, 2015, 29(2):58-67.
- [31] 谭咏梅, 杨雪. 结合实体链接与实体聚类的命名实体消歧[J]. 北京邮电大学学报, 2014(5):58-64.
- [32] Pan X, Zhang B, May J, et al. Cross-lingual name tagging and linking for 282 languages[C]. Meeting Of The Association For Computational Linguistics, 2017: 1946-1958.
- [33] Lample G, Ballesteros M, Subramanian S, et al. Neural architectures for named entity recognition[C]. North American Chapter Of The Association For Computational Linguistics, 2016: 260-270.
- [34] Zeng D, Sun C, Lin L, et al. LSTM-CRF for Drug-Named Entity Recognition[J]. Entropy, 2017, 19(6).
- [35] Huang Z, Xu W, Yu K, et al. Bidirectional LSTM-CRF Models for Sequence Tagging[J]. arXiv: Computation and Language, 2015.
- [36] Ma X, Hovy E. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF[J]. arXiv: Learning, 2016.
- [37] 张聪品, 方滔, 刘昱良. 基于 LSTM-CRF 命名实体识别技术的研究与应用[J]. 计算机技术与发展, 2019, 29(2):106-108+142.
- [38] 武惠, 吕立, 于碧辉. 基于迁移学习和 BiLSTM-CRF 的中文命名实体识别[J]. 小型微型计算机系统, 2019, 40(6):1142-1147.
- [39] Francis-Landau M, Durrett G, Klein D. Capturing semantic similarity for entity linking with convolutional neural networks[J]. arXiv preprint arXiv, 2016: 1604.00734.
- [40] Huang H, Heck L, Ji H. Leveraging Deep Neural Networks and Knowledge Graphs for Entity Disambiguation[J]. Computer Science, 2015:1275-1284.
- [41] Shaw P, Uszkoreit J, Vaswani A, et al. Self-Attention With Relative Position Representations[C]. North American Chapter Of The Association For Computational Linguistics. 2018: 464-468.
- [42] Ganea O, Hofmann T. Deep Joint Entity Disambiguation with Local Neural Attention[C]. Empirical Methods In Natural Language Processing. 2017: 2619-2629.
- [43] Gupta N, Singh S, Roth D, et al. Entity Linking via Joint Encoding of Types, Descriptions, and Context[C]. Empirical Methods In Natural Language Processing. 2017: 2681-2690.
- [44] Sil A, Kundu G, Florian R, et al. Neural Cross-Lingual Entity Linking[J]. arXiv: Computation and Language, 2017.
- [45] 多因子融合的实体识别与链指消歧 [2019-8-21]. <https://cloud.tencent.com/developer/article/1491361>
- [46] Le P, Titov I. Improving Entity Linking by Modeling Latent Relations between Mentions[C]. Meeting Of The Association For Computational Linguistics. 2018: 1595-1604.
- [47] Raiman J, Raiman O. DeepType: Multilingual Entity Linking by Neural Type System Evolution[J]. arXiv: Computation and Language, 2018.
- [48] Han X, Cao S, Lv X, et al. OpenKE: An Open Toolkit for Knowledge Embedding[C]. Empirical Methods In Natural Language Processing. 2018: 139-144.
- [49] 杜婧君, 陆蓓, 湛志群. 基于中文维基百科的命名实体消歧方法[J]. 杭州电子科技大学学报, 2013, 32(6):57-60.
- [50] 刘娇, 钟云, 刘瑶, 等. 基于语义一致性的集成实

- 体链接算法[J]. 计算机研究与发展, 2016(8):1696-1708.
- [51] 杨紫怡. 基于多知识库的实体链接研究[D]. 苏州: 苏州大学, 2018.
- [52] Navigli R, Velardi P. Structural semantic inter-connections: a knowledge-based approach to word sense disambiguation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(7):1075-1086
- [53] Sil A, Yates A. Re-ranking for joint named-entity recognition and linking[C]. Conference On Information And Knowledge Management. 2013: 2369-2374.
- [54] 管红英, 吴泳钢, 贾玉祥, 等. 基于多源知识的中文微博命名实体链接[J]. 山东大学学报(理学版), 2015(7):9-16.
- [55] 周鹏程, 武川, 陆伟. 基于多知识库的短文本实体链接方法研究——以 Wikipedia 和 Freebase 为例[J]. 现代图书情报技术, 2016(6):1-11.
- [56] Meij E, Weerkamp W, De Rijke M, et al. Adding semantics to microblog posts[C]. Web Search And Data Mining. 2012: 563-572.
- [57] 向宇, 郭云龙, 徐潇, 等. 多策略中文微博实体词消歧及实体链接[J]. 计算机应用与软件, 2016(8):12-17.
- [58] 李禹恒, 宋俊, 黄宇, 等. 基于微博文本的层次化实体链接方法[J]. 吉林大学学报(工学版), 2016(3):904-910.
- [59] 朱敏, 贾真, 左玲, 等. 中文微博实体链接研究[J]. 北京大学学报(自然科学版), 2014, 50(1):73-78.
- [60] 陈万礼, 管红英, 吴泳钢. 基于多源知识和 Ranking SVM 的中文微博命名实体链接[J]. 中文信息学报, 2015(5):117-124.
- [61] 冯冲, 石戈, 郭宇航, 等. 基于词向量语义分类的微博实体链接方法[J]. 自动化学报, 2016, 42(6):915-922.
- [62] CCKS& 百度 2019 中文短文本的实体链指[EB/OL].[2019-9-12]. https://biendata.com/competition/ccks_2019_el/
- [63] Devlin J, Chang M, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. arXiv: Computation and Language, 2018.