控制与决策
Control and Decision

实体消歧综述

段宗涛， 李菲， 陈柘

引用本文:
段宗涛, 李菲, 陈柘. 实体消歧综述[J]. 控制与决策, 2021, 36(5): 1025−1039.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2020.0388

实 体 消 歧 综 述

段宗涛, 李 菲, 陈 柘†

(长安大学 信息工程学院，西安 710064)

摘 要: 实体消歧是将文本中出现的命名实体映射到一个已知的无歧义的结构化知识库中的技术. 实体消歧是自然语言处理中的关键问题,对自然语言的发展起到重要作用.实体消歧对知识图谱构建、语义搜索、知识问答、推荐系统等应用有着重要的意义. 对此,从实体消歧的定义、分类和相关研究基础出发,对实体消歧技术进行全面的解析. 首先,对实体消歧的五元组定义进行说明,并给出实体消歧的常用分类以及相关研究基础;然后,分别对基于聚类的实体消歧、基于实体链接的实体消歧的研究内容以及研究现状进行详细综述;最后,对实体消歧的应用以及评测进行总结,并对未来研究方向进行了展望.

关键词: 知识库；知识图谱；实体消歧；自然语言处理；实体聚类；实体链接

中图分类号: TP391 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2020.0388 **开放科学(资源服务)标识码(OSID):**

引用格式: 段宗涛,李菲,陈柘. 实体消歧综述[J]. 控制与决策, 2021, 36(5): 1025-1039.

Entity disambiguation: A review

DUAN Zong-tao, LI Fei, CHEN Zhe†

(School of Information Engineering，Chang'an University，Xi'an 710064，China)

Abstract: Entity disambiguation is a technology that maps named entities that appear in text to a known unambiguous structured knowledge base. Entity disambiguation is a key issue in natural language processing and plays an important role in the development of natural language. Entity disambiguation is of great significance to the application of knowledge graph construction, semantic search, knowledge question answering, recommendation system and so on. Based on the definition, classification and related research basis of entity disambiguation, a comprehensive analysis of entity disambiguation technology is carried out. Firstly, the five-tuple definition of entity disambiguation is explained, and the common classification and related research foundation of entity disambiguation are given. Then, the research content of entity disambiguation based on clustering and entity disambiguation based on entity link, and the research status is reviewed in detail. Finally, the application and evaluation of entity disambiguation are summarized, and the future research directions are summarized.

Keywords: knowledge base；knowledge graph；entity disambiguation；natural language processing；entity clustering；entity linking

0 引 言

随着互联网的快速发展以及信息时代的到来,信息检索已成为人们获取信息的一条主要途径. 如何向检索者提供所需要的信息是信息检索技术研究关注的核心问题. 2012年谷歌提出了知识图谱[1]的概念,利用知识图谱增强搜索引擎的性能. 目前,在搜索引擎上检索常会得到多个同名但并非相关的实体内容,这一问题源于不同实体可能有多个文本表达. 通过实体消歧技术可以解决这一问题.

实体消歧是指解决同名实体存在的一词多义歧义问题. 实体消歧研究中常用的方法是基于实体链接的实体消歧,通常链接的目标知识库为Wikipedia(维基百科)[2]. 随着知识图谱的发展,基于知识图谱的实体消歧研究逐渐增多. 例如,YAGO[3]、DBpedia[4]、Freebase[5]等也可作为实体消歧的目标知识图谱. 实体消歧技术对于知识图谱的构建[6]以及语义检索、推荐系统、问答系统[7]有着重要的作用,也是建立语言表达和知识图谱联系的关键环节.

本文首先对实体消歧进行简介,阐述了实体消歧的定义、分类以及相关研究;然后,对实体消歧技术所涉及的研究内容以及研究方法进行详细说明,并介绍实体消歧的相关应用;随后,介绍实体消歧评测;最后,

收稿日期: 2019-04-08；**修回日期:** 2019-07-09.

基金项目: 陕西省重点研发计划项目(2019ZDLGY17-08, 2019ZDLGY03-09-01, 2020ZDLGY09-02).

†通讯作者. E-mail: zchen@chd.edu.cn.

指出实体消歧技术存在的问题与面临的挑战.

1 实体消歧简介

1.1 问题定义

命名实体的歧义指的是,一个实体指称项可对应于多个真实世界实体. 确定一个实体指称项所指向的真实世界实体就是命名实体消歧.

实体消歧系统通过以下一个五元组进行定义:

$$M=\{N,E,D,O,K\}.$$

其中: N 是待消歧的实体名集合; E 是待消歧实体名的目标列表,通常为知识库或者知识图谱的实体; D 是一个包含待消歧实体名的文本集,例如包含“陈光诚”的网页搜索集合; O 是 D 中的实体指称项集合,一个实体的指称项是在具体上下文中出现的待消歧实体名; K 是实体消歧任务所使用的背景知识,关于目标实体的描述.

1.2 实体消歧分类

目前,按照不同的分类标准,实体消歧技术可以有多种分类方法.

1) 按照实体任务领域划分,实体消歧分为基于结构化文本的实体消歧和基于非结构化文本的实体消歧.

基于结构化文本的实体消歧的实体指称项通常被存储在数据库中,表示为一个结构化的文本记录. 这种指称项缺少上下文信息,主要依赖字面意思和实体关系信息进行消歧.

基于非结构文本的实体消歧的实体指称项表示为一段非结构化的文本,含有大量的上下文信息,主要利用指称项上下文信息进行消歧.

2) 按照有无目标知识库划分,实体消歧包括基于无监督聚类的实体消歧(无目标知识库或知识图谱)和基于实体链接的实体消歧(有目标知识库或知识图谱).

基于聚类的实体消歧方法把所有实体指称项按其指向的目标实体进行聚类. 如图1所示,7个关于“迈克尔乔丹”的指称项经过聚类后得到3个类,每个类代表一个实体.

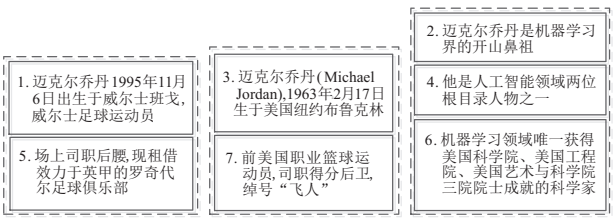

图 1 基于聚类的实体消歧实例

基于实体链接的实体消歧将实体指称项链接到目标候选实体列表中所对应的实体上实现实体消歧. 如图2所示,任务是将实体“迈克尔乔丹”链接到篮球运动员“迈克尔乔丹”,而不是其他“迈克尔乔丹”的实体.

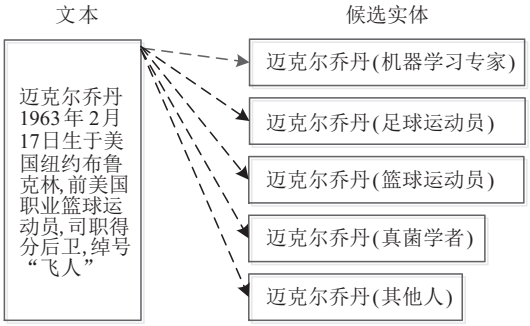

图 2 基于实体链接的实体消歧实例

3) 按照链接知识库类型划分,将基于实体链接的实体消歧分为基于知识库的实体链接和基于知识图谱的实体链接.

基于知识库的实体消歧重点是如何在大型文本知识库中提取上下文特征以及如何获取待消歧实体指称项的上下文信息.

基于知识图谱的实体链接主要利用知识图谱(KG)的结构来表示实体之间的关系以及候选实体的上下文特征.

1.3 实体消歧分类

1.3.1 词义消歧

词义消歧(WSD)[8-9]是一个开放性的自然语言处理问题,通过计算机分析并识别特定对象的词义信息. 传统的词义消歧采用的方法主要包括两种:基于知识库的消歧方法和基于语料库的消歧方法.

基于知识库的消歧方法也称为基于词典的消歧方法. 通常消歧依赖于词典对语义的区分. 消歧知识库有 Wordnet 和 Hownet 等. Patwardhan 等[10]将自适应 Lesk 算法推广到基于语义关联的词义消歧方法; Niu 等[11]提出了3种符号编码模型进行消歧.

基于语料库的消歧方法通常借助机器强大的计算能力实现词义消歧,主要包括无监督的消歧方法和有监督的消歧方法. 无监督消歧方法又称为聚类词义消歧[12-13],有监督词义消歧需要标注数据进行消歧[14-15].

词义消歧与实体消歧具有相似性,二者都解决了语言中词汇歧义的问题. 词义消歧与实体消歧的区别在于: 1) 词义消歧中的词义通常是固定的,可以通过词典进行列举;而在实体消歧中,实体词义无法列举. 2) 实体词的词义数目大于普通词;实体词消歧场

景比普通词消歧场景丰富;实体词消歧可利用特征比普通词更加丰富.

1.3.2 命名实体识别

命名实体识别,也称为实体抽取. 命名实体识别的任务是识别文本中人名、地名、机构名、时间、日期等指定类型的实体. 命名实体识别系统通常包括实体边界识别和实体类别标注两部分:实体边界识别确定一个字符串是否构成一个实体;实体归类将识别出的实体事先划分为指定的不同类别. 命名实体识别方法主要分为基于规则的方法、基于统计的方法和基于深度学习的方法.

基于规则的方法不需要标注训练语料,能直接根据词典和规则进行分词[16-17]. 然而,基于规则的方法有很大的局限性,可扩展性较差,难以适应各种数据的变化.

基于统计模型的方法通常使用统计模型来建模输入与输出之间的关联,并使用机器学习方法来学习模型的参数[18-19]. 隐马尔科夫模型、最大熵、支持向量机、条件随机场等都是常用的机器学习模型. 也有研究者采用统计学习与规则相结合的方法[20],取得了一些积极的研究成果.

近年来,随着深度学习的流行,研究者利用神经网络进行命名实体识别工作. 利用神经网络学习实体的低维表示,利用表示找出实体类别[21-23].

实体消歧任务的前提是识别出待消歧文本中的实体指称项,它与命名实体识别非常类似. 两者的相同之处是都要识别出文本片段中的实体,但它们之间也有不同之处. 命名实体识别的目标是识别出文本中所出现的所有实体,而实体指称项识别的目标是尽可能识别出目标库中已存在的实体所对应的实体指称项.

2 实体消歧方法概述

实体消歧方法主要按照目标列表是否给定分为基于聚类的消歧系统和基于实体链接的消歧系统. 本节分别对上述实体消歧方法的研究现状进行分析,并对特殊的实体消歧进行罗列.

2.1 基于无监督聚类的实体消歧

基于无监督聚类的实体消歧方法没有给定目标库,通过比较各个实体的相似程度将相似度高的聚集到一起,其核心问题是选取何种特征对指称项进行表示. 根据如何定义实体对象与指称项之间的相似度,聚类法可分为以下5种.

1) 基于词袋模型的聚类方法.

基于词袋模型的聚类方法也称为基于空间向量模型的聚类方法. 典型的方法是将当前语料库中实体指称项周围的词组成特征向量,然后利用向量的相似度对指称项进行比较,并将指称项划分到最接近的实体引用项集合中. 例如,Bagga等[24]利用向量空间模型(VSM)计算实体指称项词向量之间的相似度进行聚类;Liu等[25]利用标准空间向量模型以及HAC聚类算法进行消歧.

基于词袋模型的聚类方法采用的特征向量往往不能很好地代表实体本身,而且实体之间的向量区分不明确,从而影响聚类效果.

2) 基于语义特征的聚类方法.

基于语义特征的聚类方法与基于词袋模型的聚类方法类似,但两者的构造方法不同. 语义模型的特征向量不仅包括词袋向量,还包含语义特征. 例如,Pederson等[26]通过对文本进行分解得到实体的语义向量,并结合词袋向量得到更精确的聚类结果;Bollegala等[27]先从一组文档中的名称获取语境表征和词袋向量,再利用向量对这组文档进行聚类. 但是,基于语义特征的聚类方法很难达到最优.

3) 基于社会化网络的聚类方法.

基于社会化网络的聚类方法遵循“物以类聚,人以群分”的原则. 该类方法先构造社会化网络,再利用网络中的社会关系计算实体指称项之间的相似度[28-29]. Emami[30]提出了一个基于聚类的人名消歧系统,将从文本中提取实体之间的个人属性和社会关系映射到一个无向加权图(属性-关系图),使用聚类算法对图进行聚类,其中每个聚类包含指向一个人的所有web页面.

基于社会化网络的聚类方法较为注重实体之间的关系而忽略实体本身的特征以及实体的上下文特征,并且网络构造难度大、复杂度高.

4) 基于百科知识的聚类方法.

百科类网站通常会为每个实体(指称项)分配一个单独页面,其中包括指向其他实体页面的超链接,百科知识模型正是利用这种链接关系来计算实体指称项之间的相似度. 例如,Han等[31]从维基百科中构建了一个大规模的语义网络,根据语义网络中的百科语义知识进行消歧;Sen[32]提出了主题模型,利用群体学习主题模型进行集体消歧. 然而,百科知识覆盖性有限且实体种类较少,因此此类方法使用率较低.

5) 基于多源异构语义知识融合的聚类方法.

传统的聚类实体消歧方法所使用的目标知识库通常只有一种,覆盖度有限. 采用多源异构知识可以克服这一缺点. 多源异构知识是指知识源中存在大

量的多源异构知识,挖掘和集成不同知识源中的结构化语义知识表示模型来统一表示这些语义知识可以提高实体消歧效率. 这种方法[33-34]的多源异构知识表示框架为结构化语义关联图. 语义关联图中每个节点代表一个独立的概念,节点之间的边代表概念之间的语义关系,边的权重代表语义关系的权重. 但是,该方法使用多个知识库进行聚类,多种数据源之间表达方式略有差异且组合难度大,从而导致实体聚类效果差.

2.2 基于实体链接的实体消歧

基于实体链接的实体消歧的任务是将给定实体指称项链接到目标知识库中的相应实体上. 主要分为两个步骤: 候选实体的生成和候选实体的链接. 实体链接又分为基于知识库的实体链接以及基于知识图谱的实体链接.

2.2.1 候选实体的生成

候选实体的生成首先需要给定一个实体指称项,然后根据知识、规则等信息找到实体指称项所对应的候选实体列表. 候选实体集合的质量主要由两个因素决定: 1)是否包含目标实体; 2)候选实体的数目. 候选实体生成的方法主要有3种: 基于词典构建的方法、基于表面形式扩展的候选生成方法以及基于目标库的候选生成方法.

1) 基于词典构建的方法.

这种方法主要针对目标库为维基百科知识库. 利用维基百科的页面信息可构建实体指称与实体之间的映射关系,生成指称-实体映射词典. 常用方法为构建同义词词典及歧义词典. 首先通过同义词词典将实体指称映射为规范形式,然后通过歧义词典获得实体指称的初始候选实体集合. 一般通过字典生成的候选集合往往比较大,为了有效减小候选实体集合大小,需要对初始候选集合中候选实体进行排序和过滤. 排序指标主要有字表面相似性、上下文相似性以及实体流行度. 例如, Ratinov等[35]使用实体流行度对候选实体进行筛选.

基于词典构建的方法其候选生成效果并不理想,一方面会产生过多的候选实体,另一方面对目标实体的覆盖度还不够高.

2) 基于表面形式扩展的候选生成方法.

命名实体指称通常情况为全名,但有时会碰到缩写的形式,通过扩展技术识别实体指称可能会出现的相关扩展变化. 基于表面形式扩展的候选生成方法包括基于启发式的方法和基于监督学习的方法.

①基于启发式的方法.

对于实体指称的缩写形式,通过启发式模式匹配搜索实体指称周围的文本来扩展缩写. 最常见的模式是利用规则. Varma等[36]以及Gotipati等[37]将已经被识别的实体看成一个子串,如果实体指称包含一个子串,则该实体为实体指称的扩展形式. Cucerzan[38]采用一个缩写检测器,主要利用网页数据识别缩写的扩展. 然而,基于启发式方法的表面形式扩展无法识别一些复杂的缩写的扩展形式.

②基于监督学习的方法.

基于监督学习的方法需要标记数据,利用标记数据找到候选实体. Zhang等[39]提出了一种基于监督学习的缩略语展开算法,利用SVM分类器对每个候选缩写扩展输出一个置信得分,将得分最高的扩展实体作为候选实体.

3) 基于目标库的候选生成方法.

由于目标知识库(例如维基百科、DBpedia等)包含多种页面数据,可以利用这些页面数据找到候选实体. 主要利用消歧页面以及重定向页面的信息生成候选实体. 对于有歧义的实体,消歧页面进行了总结,重定向页面中汇总了提及以及其对应的别名. 例如,杨光等[40]利用DBpedia知识图谱数据中提供的数据集进行候选实体生成. 从消歧数据集中添加候选实体并利用提供的数据集,结合实体先验概率生成候选实体列表.

2.2.2 基于知识库的实体链接系统

基于知识库的实体链接系统的目标知识库通常为维基百科知识库. 最常用的两种候选实体链接方法是局部实体链接和协同实体链接.

1) 局部实体链接.

局部实体链接通常得到实体指称以及实体的上下文信息的特征表示,然后计算实体指称以及实体表示的相似度以选出目标实体. 局部实体链接方法主要包括传统特征方法和表示学习方法两种.

①传统特征方法.

传统特征方法的核心是如何手工设计有效的特征,其中实体的表示很简单. 例如, Honnibal等[41]利用Bow模型得到实体指称项和候选实体的向量,将余弦相似度得分最高的作为候选实体.

由于候选实体的背景知识、先验知识和实体类别信息对于实体消歧也很重要,许多研究者将这些信息考虑进来从而提高消歧的准确性[42-43]. 候选实体的背景知识和先验知识包括实体流行度(实体在知识库中的概率)、实体指称项与实体的关系(指称项指向实体的概率).

传统特征方法对目标实体和实体指称项表示都是启发式的,如词袋模型、TF-IFD等. 这些启发式算法很难调整,而且很难捕获更细粒度的语义信息和结构信息,所以传统特征方法不是主流的方法.

②表示学习方法.

表示学习方法的核心是如何获得实体和实体指称项上下文的分布式表示. 一般实体的表示比较复杂,可能从不同粒度来表示实体,可能会用到实体的类别(entity type)信息. 通常采用神经网络的方法自动学习实体以及实体指称项的分布式表示. 神经网络常用的有LSTM、CNN、RNN等.

采用神经网络进行实体链接有两种方法:排序方法和二值分类方法. 排序方法[44-46]训练一个排序模型,对所有候选实体进行排序,取排序最高的作为目标实体;二值分类方法[47-49]训练一个分类器来决定实体指称项与候选实体是否相同.

通常,实体指称项以及候选实体的上下文信息较多,然而,上下文中的有些词与实体指称或者候选实体的关联性并不大,这样训练的上下文连续表示含有噪声,影响实体消歧的准确率. 研究者们提出将注意力机制与深度神经网络相结合训练上下文的语义特征向量以改进实体消歧模型[50]. Sun等[51]通过注意机制自动从周围的上下文中发现实体指称以及候选实体的重要线索,并利用这些线索促进实体消歧. Zeng等[52]将长短时记忆网络(LSTM)与双重注意力相结合进行实体消歧. 第1个注意力机制将实体嵌入作为注意向量来突出实体描述中的信息部分;第2个注意力机制将实体上下文作为注意向量来突出实体指称上下文中的信息部分;最后结合相似度以及先验概率得到正确实体.

近年来，研究者还提出了将符号知识集成到神经网络中进行实体消歧,这样可以降低实体消歧的时间复杂度. 例如,Raiman等[53]提出了一个将符号信息显式集成到神经网络推理过程的策略,并将这种策略运用到实体消歧任务中,通过训练一个类型系统进行实体链接.

还有一些研究者利用远程监督的方法进行实体消歧. 远程监督的方法不需要标记数据训练网络,因而减少了工作量. 例如,Le等[54]将实体链接问题构造成一个远程学习问题,设置只包含正确实体的集合和只包含不正确实体的集合,利用训练监督模型决定正确实体集合中的哪一个实体最有可能成为正确实体.

局部实体链接只处理单个实体指称项的链接问题,忽略了单篇文档内所有实体指称项的目标实体之间的关系.

2) 协同实体链接.

协同实体链接认为,一个文档中的实体具有一定的关联性,因而在局部链接之上增加了一个全局项(协同策略)来综合考虑目标实体之间的一致性. 对文档内所有实体指称项进行协同链接可以提升实体链接的性能. 全局链接方法有基于图的方法、基于条件随机场的方法、基于Pair-Linking的方法和基于深度学习的方法.

①基于图的方法.

基于图的方法通常将所有实体指称的候选实体作为图的节点,指称之间的联系作为边的权重构成图模型,在此基础上采用消歧算法为实体指称选出一组最有可能的实体组合[55]. 采用图方法主要有3个步骤:候选实体生成、实体相关图构造和集成实体链接. Han等[56]提出的集成实体链接算法以维基百科作为本地知识库,对给定的文本首先提取出所有实体指称项,并通过查询确定每个实体指称项在知识库中的候选链接对象;然后采用随机游走方法对实体-候选构成图(如图3所示)中的候选实体进行排序,得到实体链接的推荐结果. Alhelbawy等[57]对图中每个节点设置一个初始置信度评分,使用页面排序算法对节点进行排序,将最后的排序与候选实体的初始置信度相结合选出正确的实体. Ma等[58]构建了实体相关图,图的节点为所有提及的候选实体,边为实体之间的转换概率, 在实体相关图上采用动态PageRank算法选出所有提及的正确实体.

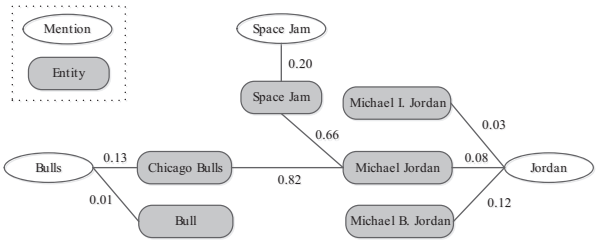

图3　实体-侯选构成图[56]

基于图的方法对于全局消歧有很好的准确性,但很难与局部方法联合对消歧进行优化.

②基于条件随机场的方法.

基于条件随机场(CRF)模型全局方法可以很好地与局部方法联合起来. 例如,Durrett等[59]研究的模型在形式上是一个结构化的传统随机场. 一元因子从每个任务的强基线编码本地特性,添加二进制和三元因子来捕获跨任务交互,将实体识别与实体消歧联合实现. Ganea等[50]利用最大化似然函数对条件随机场模型进行训练,采用环路信念传播(loopy belief

propagation)对条件随机场模型进行解码. 在局部注意力机制基础上利用条件随机场来建模全局项以进行消歧.

③基于Pair-Linking的方法.

现有的协同链接方法假设每一个链接到的候选实体都要与其他所有的实体相关,这一假设在多主题的长文档中并不一定成立;而且要考虑所有链接的实体之间的一致性,现有方法计算复杂度高. 通过对实体做Pair-Linking[60]可以克服这一弊端. 例如, Phan等[61]利用Pair-Linking算法通过模拟Kruskal算法来近似MINTREE(基于树的实体消歧目标)的解,从而得到正确实体集合.

④基于深度学习的方法.

现有的协同链接方法实现较为复杂,需要时间较多. 近年来,一些研究者通过深度学习的方法[62]对局部信息以及全局信息进行编码,从而大大提高了实体消歧效率. 例如, Xue等[63]提出了RRWEL模型,模型使用CNN学习局部上下文、提及、实体、类型信息的语义表征,使用随机漫步网络对文档信息进行学习,结合局部信息和全局信息得到文档中每个提及所对应的正确实体.

虽然基于深度学习的方法消歧效率较高,但由于文档较多,训练起来工程很大,一些研究者将深度学习的方法与基于图的方法结合起来进行集体消歧[64]. 文献[65]和文献[66]将所构建的实体图输入到图神经网络中进行学习. Deng等[67]构造了作者-文档的图网络,并提出了一种新的模型HRFAENE(异构关系融合和属性增强网络嵌入模型)进行集体消歧.

基于知识库的实体链接系统的目标知识库通常提供实体的上下文信息,而忽略了实体与实体之间的信息,这部分信息对实体消歧也很重要.

2.2.3 基于知识图谱的实体链接系统

知识图谱虽然从概念上来说是一个新的研究领域,但它其实是一个结构化的语义知识库,数据内容通常采用三元组形式表示. 基于知识图谱的实体消歧所使用的候选实体多侧重于从图结构中获取上下文信息,涉及图拓扑结构. 目前,研究者对基于知识图谱的实体消歧进行了研究,其中包括局部实体链接和协同实体链接.

1) 局部实体链接.

局部实体链接主要利用实体指称以及候选实体的上下文信息选出目标实体候选实体[68-70]. Shao等[71]在论文知识图XLore上提出了一个论文实体消歧框架,并设计了一个实体链接的概率公式以计算每个候选实体的概率,最后选出概率最高的实体作为正确实体.

近年来,一些研究者利用深度学习的方法对知识图谱的实体链接系统进行改进,从而提高了消歧系统的性能. 例如, Luo等[72]提出了一个深层语义匹配模型,模型使用字-LSTM和词-LSTM学习得到字以及上下文的匹配分数,并进行加权求和后对所有候选实体排序. Kartsaklis等[73]利用随机游走技术将图数据映射到多维实体空间中,利用Multi-Sense LSTM模型实现实体链接.

知识图谱利用图神经网络进行学习,能更好地学习到图结构数据的特征表示. 一些研究者利用图神经网络(GCN[74]、GAT[75])学习知识图的连续性表示,使得链接准确率得到提高.

2) 协同实体链接.

基于知识图谱的协同实体链接假设文档中所有实体指称在知识图谱中所对应的目标实体是相关的,所以对一个文档中的多个指称项一起连接到目标知识图谱中[76-77]. Wang等[78]构建了地理知识图,并在图上进行集体实体消歧. 在图数据中找出每个文本所对应的候选实体构成提及-实体图,利用局部相似度为图模型提及、实体的每条边赋予权重,为图中每个节点进行打分,并与节点的嵌入分数相结合得到最终实体得分. 这种方法通过考虑文档之间实体的关联性进行消歧比局部消歧更加高效,但由于目标知识图谱的数据关系结构较复杂,会降低消歧率.

基于知识图谱的实体链接系统的目标知识图谱是结构化的数据方式,实体的邻居节点可作为上下文信息,实体与实体之间的关系也可对链接提供帮助. 基于知识图谱的链接系统会成为未来实体消歧研究热点.

2.3 其他实体消歧

1) 跨语言实体消歧.

跨语言实体消歧是将一种语言表述的实体指称项链接到另一种语言的知识库[79]中. 例如, Wang等[80]提出一个链接因子图模型,将英文的维基百科以及中文维基百科的文章进行了链接. Zhang等[81]利用双语隐含主题模型,通过将实体指称项与候选实体映射到同一个主题,它们在语义空间的余弦相似度为匹配得分. Tsai等[82]利用Skip-gram模型分别训练两个单词的实体向量和词向量,并将超链接信息替换为对应的实体.

跨语言实体消歧的难点有以下几个方面:①很

多语言Wikipedia不完备,造成实体信息缺乏;②跨语言候选实体生成很难;③神经网络跨语言实体链接需要解决实体指称项所描述的语言词向量和英文词向量位于不同语义空间的问题.

2) 社交数据中的实体消歧.

社交数据实体消歧[83]将社交数据作为消歧数据集进行消歧. 社交数据具有字数受限、用户多、数目大等特点,所以在社交数据中只利用上下文信息是不充分的,还要利用用户发布的其他推文来辅助链接. 社交媒体数据中一般会有时间戳,有些文本还有地点信息以及候选实体的先验信息会随着时空信息发生变化. Fang等[84]考虑了社交媒体的时空特性,在模型中引入时间地点信息,并选用隐变量充当标注数据中的时间地点的监督信息,使得实体消歧更加准确.

社交数据中实体消歧面临的挑战主要有:①篇章级实体链接中实体指称项上下文与知识库中实体的描述之间相似度是重要特征,但由于社交文本的特性造成很难计算这一相似度而导致社交数据很难进行消歧;②篇章级实体链接中协同链接对提升链接性能作用很大,但在社交数据中可能作用不大.

3) 受限知识库的实体消歧.

现有的基于实体链接的实体消歧方法要借助于知识库中实体的丰富的信息,例如实体的描述、实体的不同属性、实体之间的超链接等. 有些知识库的存储内容以及种类较少,对消歧带来一定的影响. 例如Yelp是一个类似于大众点评类的平台,Yelp中很多实体可能并不出现在Wikipedia中,同时,很多普通用户也不会出现在Wikipedia中,但是他们都在Yelp这类平台上有账号. Xie等[85]利用Yelp中独特的社交信息资源,提取了传统实体链接特征、社交特征和地点特征,实现了受限知识库的实体消歧.

3 实体消歧应用

实体消歧旨在解决文本中广泛存在的名称歧义问题,在知识图谱构建、语义化搜索、问答系统、推荐系统等领域有着广泛的应用.

知识图谱构建: 知识图谱构建技术离不开实体消歧的支撑. 对于一段自然语言文本,例如“迈克尔·乔丹教授昨天访问了CMU”,需要从自然语言文本中抽取信息以构成知识图谱. 处理流程如下: 首先进行命名实体识别(“[迈克尔·乔丹]/PER教授昨天访问了[CMU]/ORG”);然后进行关系抽取(迈克尔·乔丹, visit, CMU). 抽取出三元组并不能直接构造知识图谱,因为不知道迈克尔·乔丹到底是哪个迈克尔·乔丹,CMU到底指的是哪个机构. 实体消歧技术将实体的歧义进行消除,经过实体抽取的实体都能够得到正确的链接. 实体消歧是知识图谱构建中必不可少的一步,对知识图谱的构建有着重要的作用.

语义搜索: 语义检索需要利用关键词检索用户所需的信息. 知识图谱的出现为语义带来新的发展前景. 基于知识图谱的语义检索搜索更加精准化. 在知识图谱知识的支持下,利用实体链接技术对关键词与知识图谱中的实体进行链接从而获取信息. 借助实体消歧技术将查找内容链接到正确的实体上,通过知识图谱中实体之间的关联可直接给出满足用户搜索意图的答案并扩展用户的搜索范围,联系更多的相关知识以反馈给用户. 例如,Zhu等[86]提出了基于自然语言查询处理、实体链接、实体类型链接和基于语义相似性的查询扩展的知识图实体搜索框架.

问答系统: 问答系统是指让计算机自动回答用户所提出的问题,是信息服务的一种高级形式. 问答系统依赖于它们背后支持的知识库来回答用户的问题. 问答系统包括检索式问答系统、社区问答系统以及面向知识图谱的问答系统. 每一种问答系统都需要将问答信息与知识库中所对应的信息链接,然后才能反馈答案[87]. Wu等[88]基于问答系统任务提出了基于序列标注的主题实体提及识别算法以及一种基于扩展信息相似度的实体消歧算法.

推荐系统: 如何为用户提供个性化推荐并提高推荐的准确度和用户满意度,是当前推荐系统研究所面临的主要问题. 知识图谱的出现为推荐系统的改进提供了新的途径. 传统的推荐系统需要将物品先链接到知识图谱中,然后为用户生成推荐列表. 实体消歧技术为推荐系统提供关键词到知识图谱的定位,通过定位才能完成个性化推荐任务[89].

4 实体消歧评测

4.1 实体消歧评测资源

随着实体消歧技术的发展,实体消歧方法的评价技术也得到了重视. 主要包括实体消歧评测会议、实体消歧评测框架、实体消歧宏观评测指标.

4.1.1 实体消歧评测会议

1) 实体消歧是信息抽取的一部分,评测会议主要采用信息抽取的评测会议. 信息抽取的评测会议主要有MUC、ACE、TAC-KBP.

①MUC: 由美国国防高级研究计划委员会DARPA资助,评测任务包括命名实体识别、共指消解、模板关系抽取等. 语料范围主要是英文.

② ACE: 由美国国家标准与技术研究所NIST主

办,评测任务包括命名实体识别、关系抽取、事件抽取等. 语料范围主要是英文、中文、阿拉伯文.

③ TAC-KBP: 由美国国防高级研究(DARPA)资助,评测任务包括实体链接、属性抽取、事件、事件关系抽取. KBP任务中实体链接任务主要对基于知识库的链接系统进行评测. KBP任务中评测任务将文本中的实体指称项链接到Wikipedia中的真实概念,达到消歧的目的. KBP是主流的基于实体链接的实体消歧评测会议.

2) 实体消歧包括一些国际评测会议,主要有SemEval、WWW、TREC、INEX、CLP.

① WePS是由西班牙国家远程教育大学举办的网络人名搜索评测会议. WePS评测任务主要集中在网络搜索中的人名实体消歧上. WePS是主流的基于聚类的实体消歧评测方法. WePS1是SemEval的任务,WePS2和WePS3是WWW的子任务.

② TREC的KBA任务要求识别出文档集中与特定实体相关的文档,并标注文档与实体之间的相关程度. 主要针对基于实体链接的实体消歧进行评测.

③ INEX的“Link the Wiki”任务主要是探索如何在Wikipedia文章中自动发现应当被创建连接的文本. 主要针对基于实体链接的实体消歧进行评测.

④ CLP会议主要针对人名消歧任务进行评测,以及对基于聚类的实体消歧系统进行评测.

3) 主要评测会议对比分析.

目前,国际上受研究者认可的主流实体消歧方面的评测有两个: TAC-KBP会议和WePS会议.

TAC-KBP评测与WePS评测都是对实体消歧任务进行评测. 不同点在于: ① TAC-KBP评测主要针对基于实体链接的实体消歧任务,WePS评测主要针对基于聚类的实体消歧任务; ② TAC-KBP评测的评价指标使用宏观准确率以及微观准确率,而WePS评测使用纯净度、倒纯净度和F值; ③ TAC-KBP评测主要针对人名、地名、组织名等实体的研究,而WePS评测主要针对人名的研究.

4) 团队机构评测对比分析.

①国内团队机构评测对比分析.

通过知网文献数据检索分析,目前,国内对于实体消歧研究较多的机构如表1所示. 表1中对各个机构的评价指标、评测数据集进行了说明.

②国外团队机构对比分析.

通过Web of science文献数据检索分析,目前,国外对于实体消歧研究较多的机构以及各个机构的评价指标、评测数据集如表2所示.

表 1 国内机构评测指标

机构	评测指标	评测数据集
哈尔滨工业大学	TAC-KBP、WePS评价指标	TAC-KBP、WePS评测数据、AIDA数据集
北京邮电大学	TAC-KBP、WePS、Entity Linking评价指标	TAC-KBP、WePS评测数据、Entity Linking数据集
昆明理工大学	CLP-2012评价指标	CLP-2012提供的数据集
国防科学技术大学	宏观标准评价指标	ACE2005、AIDA数据集
解放军信息工程大学	WePS的标准评价指标	CLP-2010、CLP-2012提供的数据集
中国科学院大学	TAC-KBP的标准评价指标	SimpleQuestions、Web-Questions数据集
东南大学	宏观标准评价指标	ACE2004、MSNBC、Web-Questions数据集

表 2 国外机构评测指标

机构	评测指标	评测数据集
Microsoft	TAC-KBP评价指标、宏观标准评价指标	TAC-KBP2009、TAC-KBP2010数据集
Centre National De La Recherche	TAC-KBP评价指标	AIDA、TAC-KBP2009数据集
Google公司	宏观标准评价指标	查询日志
Pennsylvania Commonwealth	宏观标准评价指标	ACM数字图书馆获取数据集
Unversity of California System	WePS评价指标、宏观标准评价指标	列表数据集、WePS数据集
University of Tokyo	TAC-KBP的标准评价指标	AIDA、ACE2004、MSNBC数据集
Chinese Academy of Sciences	宏观标准评价指标	AIDA、ACE2004、MSNBC、WW数据集

从表1可以看出,每个机构使用的评测指标都是标准的评测指标,评测数据集也是标准的评测数据集. 其中: 哈尔滨工业大学在基于实体链接的评测中使用的评测指标为主流的评测指标,评测数据集也比较广泛,系统消歧更好; 北京邮电大学在基于聚类的实体消歧研究中使用评测指标较广泛,聚类效果好; 昆明理工大学和解放军信息工程大学主要在中文上进行评测,为中文实体消歧发展做出一定的贡献; 国防科学技术大学和东南大学的研究中使用宏观的评测指标更利于系统之间的对比分析; 中国科学技术大学使用问答系统的评测数据集,使得消歧系统可以更好地应用于问答系统中.

从表2中可以看出,国外的一些团队大都使用国际认可的主流的TAC-KBP评测会议中的标准评测指标,数据集也大多使用TAC-KBP评测会议中标准的评测数据集以及公开的AIDA评测数据集. 其

中: Microsoft 和 Centre National De La Recherche 对实体消歧研究的贡献较大, 评测指标都使用主流的评测指标, 有利于系统的对比评价; Google 公司在 Web 查询任务中使用宏观评测指标, 有助于查询消歧的发展; Pennsylvania Commonwealth 研究中的人名消歧评测数据集来自 ACM 数字图书馆, 使得人名消歧可以更好地利用图书馆资源; Unversity of California System 研究得比较全面, 评价指标也较为全面, 系统认可度高; University of Tokyo 和 Chinese Academy of Sciences 的研究评测数据集较多, 都是一些公开的评测数据集, 评测的系统复用性高.

4.1.2 实体消歧评测框架

1) Chen 等[90] 提出了一个被称为 EUEF 的基准 ERD 系统的评估框架. EUEF 旨在促进评估过程, 并对各种实体识别以及实体消歧 (ERD) 系统进行公平比较和详细分析. EUEF 是灵活的、易于使用的, 可以方便地结合新的 ERD 系统、数据集和评估指标进行扩展. EUEF 定义了几个新的模糊匹配指标, 并提出了一种新的评价 NILs 的方法. 通过基于 EUEF 的公平详尽的比较, 更容易发现各种 ERD 系统的优缺点.

2) GERBIL 平台是近年来被广泛使用的标准评测平台, 该平台提供了一个 Web 服务 API (http://aksw.org/Projects/GERBIL.html). 它里边提供工具以便对用户指定的数据集自动进行评测.

4.1.3 实体消歧宏观评测指标

从宏观角度, 命名实体消歧评估指标通常包括准确率、召回率以及 F 值.

准确率的定义为

$$\text{precision} = \frac{\{\text{正确消歧的待消歧实体}\}}{\{\text{系统生成的待消歧实体}\}};$$

召回率的定义为

$$\text{recall} = \frac{\{\text{正确消歧的待消歧实体}\}}{\{\text{应被消歧的待消歧实体}\}};$$

F 值是将准确率与召回率放在一起进行计算, F 值的定义为

$$F = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}.$$

4.2 实体消歧评测数据集

实体消歧中的评测数据集包括实体提及文本集、目标实体知识库、目标实体知识图谱.

1) 实体提及文本集如表 3 所示. 其中比较常用的有 AIDA、WNED、MSNBC、AQUAINT、ACE2004.

AIDA: 由马普研究所公开的数据集, 是目前最大的手工标注实体链接数据集. 它是基于 CONLL-2013 实体识别数据集上标注的, 题材是路透新闻.

表 3 实体指称项文本集

数据集	类型	实体指称项个数	文档的数目
AIDA-train	news	18 448	946
AIDA-A(valid)	news	4 791	216
AIDA-B(test)	news	4 485	231
WNED-CWEB	news	11 154	320
WNED-WIKI	news	6 821	320
MSNBC	news	956	20
AQUAINT	news	727	50
ACE2004	news	257	36
RSS500	RSS-feeds	518	343
KORE50	short sentences	144	50
N3-Reuters	news	650	128
Reuters128	news	637	111
ITTB	web-texts	11 245	103
Meij	tweets	812	502

WNED: 包括 WNED-CWEB 和 WNED-WIKI 两种. WNED-CWEB 是从 ClueWeb 中自动构建的, 而 WNED-WIKI 是从 Wikipedia 中自动构建的.

MSNBC: 数据题材是新闻, 包含 10 个不同主题的 20 篇新闻文章 (每个主题 2 篇).

AQUAINT: 数据来自 3 家不同新闻机构的 50 篇新闻报道.

ACE2004: 由众包注释的数据集, 是 ACE2004 的参考文献的一个子集.

实体消歧主要评测会议为 TAC-KBP 和 WePS, 这两个会议提供了评测数据集.

TAC-KBP: 数据来源是新闻和论坛, 是手工标注的数据集, 如表 4 中以较为流行的 TAC-KBP2015 文本集为例进行说明.

表 4 TAC-KBP2015 文本集

	训练	测试
总量	30 838	32 533
中文	13 116	11 066
西班牙文	4 177	5 822
英文	13 545	15 645

WePS: 数据来源于网络. 对于每一个歧义人名, WePS 数据集提供其在搜索引擎中的前 N 个结果, 每个结果包括以下信息: 在原来搜索引擎中的排序、URL、Snippet 和标题等. 主要包括 WePS1 和 WePS2 两种数据集, 如表 5 所示.

表 5 WePS 数据集

数据集名称	数据集规模
WePS1	3 489
WePS2	3 432

2) 目前常用的目标实体知识库是 Wikipedia, 在一些任务中使用特定领域知识库, 例如 Yelp、IMDb.

Wikipedia: 维基百科是一个网络百科全书项目, 包含多种语言. 其中已收录了超过 3 000 万篇条目, 其

中英语维基百科以超过450万篇条目在数量上位居首位.

Yelp: Yelp是美国最大点评网站,用户可以在Yelp网站中给商户打分、提交评论、交流购物体验等.

IMDb:互联网电影资料库是一个关于电影演员、电影、电视节目、电视明星和电影制作的在线数据库.

3) 常用目标实体知识图谱有DBpedia、Freebase、YAGO2、CN-DBpedia等,特定领域知识图有XLore等.

DBpedia:维基百科的结构化数据版本. DBpedia的数据集包含458万个实体和超过30亿关系数量. DBpedia官网中提供了多个版本,目前实体消歧中常用的版本是DBpedia2016-10.

Freebase: Freebase中的条目都是结构化数据的形式,拥有约2 000万个主题或实体的信息. Freebase有许多子集,如FB15K、FB5M. 目前,实体消歧中常将FB5M子集作为目标知识图. FB5M在简单问答数据集中发布,它包含4 904 397个实体、7 523个关系和22 441 880个事实.

YAGO2:一个大型的语义知识库,拥有超过1 000万个实体的知识,并包含超过1. 2亿个关于这些实体的事实.

CN-DBpedia:由复旦大学知识工厂实验室研究,数据源来自中文类百科网站. 包含1 686万实体数量和2 228.6万关系数量.

XLore:针对文献领域的知识图谱. 包含实例16 284 901个、概念2466956个以及属性446 236个.

5 实体消歧总结与展望

近年来,实体消歧任务在自然语言处理领域受到广泛的关注,得到了很好的发展. 然而,实体消歧距离真正实用还有很远的距离,本节对实体消歧进行总结并对未来发展方向进行展望.

5.1 实体消歧总结

实体消歧按照有无目标知识库可以划分为基于无监督聚类的实体消歧和基于实体链接的实体消歧. 基于无监督聚类的实体消歧包括基于词袋模型的聚类方法、基于语义特征的聚类方法、基于社会化网络的聚类方法、基于百科知识的聚类方法以及基于多源异构语义知识融合的聚类方法. 基于实体链接的实体消歧方法包括基于知识库的实体链接系统和基于知识图谱的实体链接系统. 各类方法的优缺点汇总如表6～表8所示.

表6 实体消歧优缺点

方 法	优 点	缺 点
基于无监督聚类的实体消歧	不需要候选实体集合以及标记训练数据	实体之间特征区分不明确
基于实体链接的实体消歧	有目标库,消歧更加准确	需要大量有标签数据,耗费人力

表7 基于无监督聚类的实体消歧优缺点

方 法	优 点	缺 点
基于词袋模型的聚类方法	思路简单,易于实现	实体向量之间难以区分
基于语义特征的聚类方法	向量特征表示准确,聚类效果好	算法匹配程度很难最优
基于社会化网络的聚类方法	能够利用社会关系进行聚类	忽略实体本身特征,网络构造难度大
基于百科知识的聚类方法	百科网站知识特征表示全面	百科知识覆盖性有限且实体种类较少
基于多源异构语义知识融合的聚类方法	利用多种数据源可提供多种特征	知识库表达方式有差异组合难度大

表8 基于实体链接的实体消歧优缺点

方 法	优 点	缺 点
基于知识库的局部实体链接	词条内容丰富	上下文信息对实体表示不够充分
基于知识库的协同实体链接	增加实体之间相关性,消歧准确率高	文档信息量大,链接复杂性高
基于知识图谱的局部实体链接	图数据实体的上下文信息丰富	图谱数据标记样本较为复杂
基于知识图谱的协同实体链接	图数据协同实体链接准确率高	图谱数据关系较多,检索较为麻烦

5.2 实体消歧展望

虽然实体消歧技术已经有不少研究工作,但在实体消歧的各个环节还存在不少的问题与挑战. 本节对实体消歧未来方向进行展望.

5.2.1 空实体链接

在实体链接阶段,由于链接到的知识库不完备性,并不是每一个实体指称项在知识库中都能找到对应的实体. 对于这类实体指称项,实体链接系统通常将其链接到一个特殊的空实体(NIL)上,并将空实体聚类. 无链接实体指称项判别标准有3种: 1) 如果一个实体指称项没有对应的候选实体集合,则该实体指称项的链接结果为NIL. 2) 如果一个实体指称项所对应排名最高的候选实体得分低于一个预先设定的阈值,则该实体指称项的链接结果为NIL. 3) 给定一个

实体指称项及其对应排名最高的候选实体,使用二分类器对齐进行分类. 如果分类结果为1,则返回候选实体作为实体链接结果;否则,该实体指称项的链接结果是NIL.

本文认为NIL问题有以下两方面需要研究:

1) 无链接实体指称项预测. 目前由于知识库的实体类别有限,有很多实体不能链接到知识库中. 另外,文本指称项较多,通常不能很准确地将这类提及找出,如何识别出无对应实体的指称项是NIL问题需要解决的一点.

2) 无链接实体指称项聚类. 识别出文档中的NIL实体后,还需要将这类提及链接到NIL实体上并进行聚类. 但空实体的聚类存在难度,未来可进行研究.

5.2.2 实体消歧与实体识别的联合学习

在知识库构建中,实体识别是实体消歧的前提,实体识别可以为实体消歧提供更多的有效信息. 实体消歧与实体识别任务联合学习可以减少工作量[91]. Kolitsas等[92]利用先验知识识别出所有可能的实体指称项,再将所有可能的实体指称项与实体表示进行相似度打分,自动筛选出可能性最大的实体指称项和最可能的连接结果. Martins等[93]利用NER与EL之间的相关性,得到一个更健壮、更通用的系统,实现对NER和EL的多任务学习. 与单独目标训练的模型相比,联合学习提高了这两个任务的性能. 由此看出,实体识别与实体消歧任务的联合解决既能提高命名实体识别的性能,也能提高实体消歧的性能,是未来的研究重点.

5.2.3 基于多种语言的实体消歧

目前实体消歧系统主要针对的是英文语料,中文或者其他语言的消歧系统非常缺乏.

本文认为在多种语言的实体消歧中有以下3个方面需要开展.

1) 多种语言实体消歧数据集构建.

目前,实体消歧在英文领域已经构建了一些标准的公开的实体指称项文本集. 然而,在中文、法文、俄文等语言领域还没有公开的标准实体指称项文本集. 因此,构建标准的其他语言的消歧文本集是当前实体消歧的一个研究方向,尤其是中文文本集.

对于实体链接任务,需要将提及链接到目标知识库上. 维基百科知识库中涵盖了英语、俄语等语言的维基百科知识库,但除英文外的其他语言的知识库实体种类包含较少,目前也没有高质量的中文知识库可以支持实体消歧任务. 因此,构建中文等目标知识库也成为未来的一个研究方向.

2) 基于多种语言的聚类实体消歧.

在英文文本集上的聚类消歧方法已经很成熟,但在其他语言文本集上聚类消歧方法短缺. 例如,对于人口较多的中国而言,使用基于社会化网络的聚类方法可很好地将人进行分类统计. 因此,基于多种语言的聚类实体消歧是未来需要解决的问题.

3) 基于多种语言的实体链接.

尽管实体链接任务技术发展很好,但主要是针对英文语料的实体链接. 而由于语言特征的多样性和目标知识库的质量参差不齐这两个突出的因素,使得英语语料库不能直接应用于其他语言. 一些研究者利用深度学习的方法在其他语言上实现了实体链接[94].

由于汉语以及其他语言与类似英语的语言不同,这使得实体链接难度增加,可见,对于中文和其他语言的实体链接系统也需要重点研究.

5.2.4 其他研究方向

1) 多领域数据集构建问题. 现有的实体消歧数据集大部分都是基于新闻语料构建的,其他领域内语料规模和数量都远远不够. 在多个领域数据集上进行实体消歧是未来的发展方向.

2) 别名实体候选生成问题. 在候选实体生成阶段,人名通常会有别名隐藏在Wikipedia页面中,很难通过超链接、重定向等手段发现. 在判断别名实体指称项时很难正确找到对应人名的候选实体列表,导致实体消歧的准确率下降,因此,解决别名实体候选生成是未来的研究重点.

3) 实体消歧中,实体、实体的类别信息、关系信息以及实体上下文信息对实体消歧非常重要,但经常会出现实体数据集不完整的情况,使得实体消歧效果不是很好. Idrissou等[95]在实体数据不完整的情况下,提出了一种方法来支持实体消除歧义. 在实体指称项以及实体数据集信息不完整时,通过提供的仅有的实体相关信息进行消歧是该领域面临的又一大挑战.

6 结 语

知识图谱是一个新兴的研究领域,实体消歧作为知识图谱构建技术中的一个环节有着重要的研究意义. 实体消歧对知识问答、语义检索、推荐系统等领域有着重要的潜在价值. 本文对实体消歧的定义与分类进行了描述,对实体消歧的关键方法进行了全面的分析,对实体消歧相关应用以及实体消歧的评测进行了综述,并对实体消歧面临的问题与挑战进行了总结.

实体消歧的重要性不仅在于它是知识图谱构建技术中必不可少的一步,而且它是将自然语言与知识连接起来的桥梁. 实体消歧为许多相关科学领域带来了先机. 因此,本文旨在让更多人理解实体消歧技术,并积极地投入这项研究工作中.

参考文献(References)

[1] 李涓子, 侯磊. 知识图谱研究综述[J]. 山西大学学报: 自然科学版, 2017, 40(3): 454-459.
(Li J Z, Hou L. Rerviews on knowledge graph research[J]. Journal of Shanxi University: Natural Science Edition, 2017, 40(3): 454-459.)

[2] Dredze M, McNamee P, Rao D, et al. Entity disambiguation for knowledge base population[C]. Proceedings of the 23rd International Conference on Computational Linguistics. New York: ACM, 2010: 277-285.

[3] Suchanek F M, Kasneci G, Weikum G. Yago: A large on-tology from wikipedia and wordnet[J]. Web Semantics: Science, Services and Agents on the World Wide Web, 2008, 6(3): 203-217.

[4] Bizer C, Lehmann J, Kobilarov G, et al. DBpedia — A crystallization point for the web of data[J]. Journal of Web Semantics, 2009, 7(3): 154-165.

[5] Bollacker K, Cook R, Tufts P. Freebase: A shareddatabase of structured general human knowledge[C]. Proceedings of the 22nd National Conference on Artificial Intelligence. Menlo Park: AAAI, 2007: 1962-1963.

[6] 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.
(Liu Q, Li Y, Duan H, et al. Knowledge graph construction techniques[J]. Journal of Computer Research and Development, 2016, 53(3): 582-600.)

[7] Wang F, Wu W, Li Z J, et al. Named entity disambiguation for questions in community question answering[J]. Knowledge-Based Systems, 2017, 126: 68-77.

[8] Ide N, Véronis J. Introduction to the special issue on word sense disambiguation: The state of the art[J]. Computational Linguistics, 1998, 24(1): 2-40.

[9] 卢志茂, 刘挺, 李生. 统计词义消歧的研究进展[J]. 电子学报, 2006, 34(2): 333-343.
(Lu Z M, Liu T, Li S. The research progress of statistical word sense disambiguation[J]. Acta Electronica Sinica, 2006, 34(2): 333-343.)

[10] Patwardhan S, Banerjee S, Pedersen T. Using measures of semantic relatedness for word sense disambiguation[C]. Proceedings of the 4th International Conference on Intelligent Text Processing and Computational Linguistics. Berlin: Springer Verlag, 2003: 241-257.

[11] Niu Y L, Xie R B, Liu Z Y, et al. Improved word representation learning with sememes[C]. Procedings of the 55th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2017: 2049-2058.

[12] Agirre E, Rigau G. A proposal for word sense disambiguation using conceptual distance[C]. Proceedings of the International Conference on Recent Advances in Natural Language Processing. Bu1garia: RANLP, 1995: 258-264.

[13] Bordag S. Word sense induction: Triplet-based clustering and automatic evaluation[C]. Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics. Trento: EACL, 2006: 137-144.

[14] Gliozzo A, Giuliano C, Strapparava C. Domain kernels for word sense disambiguation[C]. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Morristown: ACL, 2005: 403-410.

[15] Raganato A, Delli Bovi C, Navigli R. Neural sequence learning models for word sense disambiguation[C]. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen: EMNLP, 2017: 1156-1167.

[16] Rau L F. Extracting company names from text[C]. Proceedings of the 7th IEEE Conf on Artificial Intelligence Applications. Alomitos: IEEE, 1991: 29-32.

[17] Chua T, Liu J M. Learning pattern rules for Chinese named entity extraction[C]. Proceedings of 18th National Conference on Artifical Intelligence. Cambridge: MIT Press, 2002: 411-418.

[18] Liu X H, Zhang S D, Wei F R, et al. Recognizing named entities in tweets[C]. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: ACL, 2011: 359-367.

[19] Chen Y K, Lask T A, Mei Q Z, et al. An active learning-enabled annotation system for clinical named entity recognition[J]. BMC Medical Informatics and Decision Making, 2017, 17(2): 82.

[20] Lin Y F, Tsai T H, Chou W C, et al. A maximum entropy approach to biomedical named entity recognition[C]. Proceedings of the 4th International Conference on Data Mining in Bioinformatics. New York: ACM, 2004: 56-61.

[21] Lample G, Ballesteros M, Subramanian S, et al. Neural architectures for named entity recognition[C]. Proceedings of North American Chapter of the Association for Computational Linguistics. Stroudsburg: ACL, 2016: 260-270.

[22] Peng N Y, Dredze M. Learning word segmentation representations to improve named entity recognition for chinese social media[C]. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2016: 149-155.

[23] Yadav V, Bethard S. A survey on recent advances in named entity recognition from deep learning models[C]. Proceedings of the 27th International Conference on Computational Linguistics. Stroudsburg: ACL, 2018:

2145-2158.

[24] Bagga A, Baldwin B. Entity-based cross-document coreferencing using the vector space model[C]. Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 1998: 79-85.

[25] Liu Z Z, Lu Q, Xu J. High performance clustering for web person name disambiguation using topic capturing[C]. Proceedings of the 1st International Workshop on Entity-Oriented Search. New York: ACM, 2012: 1-6.

[26] Pedersen T, Purandare A, Kulkarni A. Name discrimination by clustering similar contexts[C]. Proceedings of the 6th International Conference on Intelligent Text Processing and Computational Linguistics. Berlin: Springer, 2005: 226-237.

[27] Bollegala D, Matsuo Y, Ishizuka M. Disambiguating personal names on the web using automatically extracted key phrases[J]. Frontiers in Artificial Intelligence and Applications, 2006, 141: 553-557.

[28] Bekkerman R, McCallum A. Disambiguating web appearances of people in a social network[C]. Proceedings of the 14th International Conference on World Wide Web. New York: ACM, 2005: 463-470.

[29] Minkov E, Cohen W W, Ng A Y. Contextual search and name disambiguation in email using graphs[C]. Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2006: 27-34.

[30] Emami H. A graph-based approach to person name disambiguation in web[J]. ACM Transactions on Management Information Systems, 2019, 10(2): 1-25.

[31] Han X P, Zhao J. Named entity disambiguation by leveraging wikipedia semantic knowledge[C]. Proceedings of the 18th ACM Conference on Information and Knowledge Management. New Nork: ACM, 2009: 215-224.

[32] Sen P. Collective context-aware topic models for entity disambiguation[J]. Proc of the 21st Annual Conference on World Wide Web. New Nork: ACM, 2012: 729-738.

[33] Han X P, Zhao J. Structural semantic relatedness: A knowledge-based method to named entity disambiguation[C]. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Somerset: ACL, 2010: 50-59.

[34] 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.
(Zhou P C, Wu C, Lu W. Research on entity link method of short text based on multi-knowledge base — A case study of Wikipedia and Freebase[J]. New Technology of Library and Information Service, 2016(6): 1-11.)

[35] Ratinov L, Roth D, Downey D, et al. Local and global algorithms for disambiguation to wikipedia[C]. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2011: 1375-1384.

[36] Varma V, Bysani P, Reddy K, et al. IIIT hyderabad in guided summarization and knowledge base population[C]. Proceedings of the Text Analysis Conference 2009. Washington DC: TAC, 2009: 213-222.

[37] Gottipati S, Jiang J. Linking entities to a knowledge base with query expansion[C]. Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2011: 804-813.

[38] Cucerzan S. Entity linking by performing full-document entity extraction and disambiguation[C]. Proceedings of the Text Analysis Conference. Washington DC: TAC, 2011: 430-441.

[39] Zhang W, Sim Y C, Su J, et al. Entity linking with effective acronym expansion, instance selection and topic modeling[C]. Proceedings of the 22th International Joint Conference on Artificial Intelligence. Barcelona: IJCAI, 2011: 1909-1914.

[40] 杨光, 刘秉权, 刘铭. 基于图方法的命名实体消歧 [J]. 智能计算机与应用, 2015, 5(5): 52-55.
(Yang G, Liu B Q, Liu M. Graph-based method for named entity disambiguation[J]. Intelligent Computer AND Applications, 2015, 5(5): 52-55.)

[41] Honnibal M, Dale R. DAMSEL: The DSTO/Macquarie system for entity-linking[C]. Proceedings of the Theory and Applications of Categories. Washington DC: TAC, 2009: 1-4.

[42] Bunescu R, Paşca M. Using encyclopedic knowledge for named entity disambiguation[C]. Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics. Stroudsburg: ACL, 2006: 9-16.

[43] Han X P, Sun L. A generative entity-mention model for linking entities with knowledge base[C]. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2011: 945-954.

[44] Sun Y, Lin L, Tang D, et al. Modeling mention, context and entity with neural networks for entity disambiguation[C]. Proceedings of the 24th International Conference on Artificial Intelligence. Freiburg: IJCAI, 2015: 1333-1339.

[45] Francis-Landau M, Durrett G, Klein D. Capturing semantic similarity for entity linking with convolutional neural networks[C]. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologie. Stroudsburg: ACL, 2016: 1256-1261.

[46] Shahbazi H, Fern X Z, Ghaeini R, et al. Entity-aware ELMo: Learning contextual entity representation for entity disambiguation[J]. arXiv: Computation and Language, 2019, 2(3): 6-11.

[47] Wei C H, Lee K, Leaman R, et al. Biomedical mention disambiguation using a deep learning approach[C]. Proceedings of the 10th ACM International Conference

on Bioinformatics, Computational Biology and Health Informatics. New Nork: ACM, 2019: 307-313.

[48] Zuheros C, Tabik S, Valdivia A, et al. Deep recurrent neural network for geographical entities disambiguation on social media data[J]. Knowledge Based Systems, 2019, 173(1): 117-127.

[49] Alokaili A, Menai M E B. SVM ensembles for named entity disambiguation[J]. Computing, 2020, 102(4): 1051-1076.

[50] Ganea O E, Hofmann T. Deep joint entity disambiguation with local neural attention[C]. Proceedings of the Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2017: 2619-2629.

[51] Sun Y M, Ji Z Z, Lin L, et al. Entity disambiguation with memory network[J]. Neurocomputing, 2018, 275: 2367-2373.

[52] Zeng W X, Tang J Y, Zhao X, et al. Entity linking on Chinese microblogs via deep neural network[J]. IEEE Access, 2018, 6(3): 25908-25920.

[53] Raiman J R, Raiman O M. Deeptype: Multilingual entity linking by neural type system evolution[C]. Proceedings of the 32nd AAAI Conference on Artificial Intelligence. Palo Alto: AAAI, 2018: 5406-5413.

[54] Le P, Titov I. Distant learning for entity linking with automatic noise detection[C]. Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2019: 4081-4090.

[55] Hoffart J, Yosef M A, Bordino I, et al. Robust disambiguation of named entities in text[C]. Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2011: 782-792.

[56] Han X P, Sun L, Zhao J. Collective entity linking in web text: A graph-based method[C]. Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2011: 765-774.

[57] Alhelbawy A, Gaizauskas R. Graph ranking for collective named entity disambiguation[C]. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2014: 75-80.

[58] Ma N, Liu X, Gao Y, et al. Entity linking based on graph model and semantic representation[C]. Knowledge Science, Engineering and Management. Cham: Springer, 2019: 561-571.

[59] Durrett G, Klein D. A joint model for entity analysis: Coreference, typing, and linking[J]. Transactions of the Association for Computational Linguistics, 2014, 2(1): 477-490.

[60] Phan M C, Sun A X, Tay Y, et al. Attention-based semantic matching and pair-linking for entity disambiguation[C]. Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. New York: ACM, 2017: 1667-1676.

[61] Phan M C, Sun A X, Tay Y, et al. Pair-linking for collective entity disambiguation: Two could be better than all[J]. IEEE Transactions on Knowledge and Data Engineering, 2019, 31(7): 1383-1396.

[62] Fang Z, Cao Y N, Li Q, et al. Joint entity linking with deep reinforcement learning[C]. Proceedings of the World Wide Web Conference. New York: ACM, 2019: 438-447.

[63] Xue M G, Cai W M, Su J S, et al. Neural collective entity linking based on recurrent random walk network learning[C]. Proceedings of the 28th International Joint Conference on Artificial Intelligence. Freiburg: IJCAI, 2019: 5327-5333.

[64] Cao Y X, Hou L, Li J Z, et al. Neural collective entity linking[J]. International Conference on Computational Linguistics, 2018, 5(1): 675-686.

[65] Xin K, Hua W, Liu Y, et al. Entity disambiguation based on parse tree neighbours on graph attention network[C]. Web Information Systems Engineering. Cham: Springer, 2019: 523-537.

[66] Hu L M, Ding J Y, Shi C, et al. Graph neural entity disambiguation[J]. Knowledge Based Systems, 2020, 195(11): 716-723.

[67] Deng C H, Deng H F, Li C R, et al. A scholar disambiguation method based on heterogeneous relation-fusion and attribute enhancement[J]. IEEE Access, 2020, 8(22): 28375-28384.

[68] Huang H Z, Heck L P, Ji H. Leveraging deep neural networks and knowledge graphs for entity disambiguation[J]. Computation and Language, 2015, 6(8): 1275-1284.

[69] Radhakrishnan P, Talukdar P, Varma V. ELDEN: Improved entity linking using densified knowledge graphs[C]. Proceedings of the North American Chapter of the Association for Computational Linguistics. Stroudsburg: ACL, 2018: 1844-1853.

[70] Zhang K, Zhu Y W, Gao W J, et al. An approach for named entity disambiguation with knowledge graph[C]. Proceedings of the 2018 International Conference on Audio Language and Image Processing. Piscataway: IEEE, 2018: 138-143.

[71] Shao Z, Cao X Y, Yuan S, et al. ELAD: An entity linking based affiliation disambiguation framework[J]. IEEE Access, 2020, 8: 70519-70526.

[72] Luo A, Gao S, Xu Y, et al. Deep semantic match model for entity linking using knowledge graph and text[C]. Proceedings of the 6th International Conference on Identification, Information and Knowledge in the Internet of Things. Amsterdam: Elsevier Science, 2018: 110-114.

[73] Kartsaklis D, Pilehvar M T, Collier N, et al. Mapping text to knowledge graph entities using multi-sense LSTMs[C]. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2018: 1959-1970.

[74] Cetoli A, Akbari M, Bragaglia S, et al. Named entity disambiguation using deep learning on graphs[J]. Computation and Language, 2018, 7(2): 8-15.

[75] Zhang F J, Liu X, Tang J, et al. OAG: Toward linking large-scale heterogeneous entity graphs[C]. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM, 2019: 2585-2595.

[76] Parravicini A, Patra R, Bartolini D B, et al. Fast and accurate entity linking via graph embedding[C]. Proceedings of the 2nd Joint International Workshop on Graph Data Management Experiences Systems and Network Data Analytics. New York: ACM, 2019: 10-18.

[77] Sevgili Ö, Panchenko A, Biemann C, et al. Improving neural entity disambiguation with graph embeddings[C]. Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2019: 315-322.

[78] Wang Y T, Li Z X, Yang Q, et al. WebEL: Improving entity linking with extra web contexts[C]. Web Information Systems Engineering — WISE 2019. Cham: Springer, 2019: 507-522.

[79] Wentland W, Silberer C, Hartung M. Building a multilingual lexical resource for named entity disambiguation, translation and transliteration[C]. Proceedings of the International Conference on Language Resources and Evaluation. Marrakech: LREC, 2008: 3230-3237.

[80] Wang Z C, Li J Z, Wang Z G, et al. Cross-lingual knowledge linking across wiki knowledge bases[C]. Proceedings of the 21st International Conference on World Wide Web. New York: ACM, 2012: 459-468.

[81] Zhang T, Liu K, Zhao J. Cross lingual entity linking with bilingual topic model[C]. Proceedings of the 23th International Joint Conference on Artificial Intelligence. Menlo Park: AAAI, 2013: 2218-2224.

[82] Tsai C T, Roth D. Cross-lingual wikification using multilingual embeddings[C]. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: ACL, 2016: 589-598.

[83] Shen W, Wang J Y, Luo P, et al. Linking named entities in tweets with knowledge base via user interest modeling[C]. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2013: 68-76.

[84] Fang Y, Chang M W. Entity linking on microblogs with spatial and temporal signals[J]. Transactions of the Association for Computational Linguistics, 2014, 2(10): 259-272.

[85] Xie Q Z, Lai G K, Dai Z H, et al. Large-scale cloze test dataset designed by teachers[C]. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2020: 2344-2356.

[86] Zhu G G, Iglesias C A. Sematch: Semantic entity search from knowledge graph[J]. Results Evaluation of Pruning Methods with Varying Threshold, 2015, 1556(1): 1-6.

[87] Sorokin D, Gurevych I. Mixing context granularities for improved entity linking on question answering data across entity categories[C]. Proceedings of the Joint Conference on Lexical and Computational Semantics. Stroudsburg: ACL, 2018: 65-75.

[88] Wu G, Wu W F, Ji H, et al. Enhanced entity mention recognition and disambiguation technologies for chinese knowledge base Q&A[C]. Proceedings of the 9th Joint International Conference, Lecture Notes in Computer Science. Cham: Springer, 2020: 99-115.

[89] Wang H W, Zhang F Z, Xie X, et al. DKN: Deep knowledge-aware network for news recommendation[C]. Proceedings of the 27th World Wide Web Conference. New York: ACM, 2018: 1835-1844.

[90] Chen H, Wei B G, Li Y M, et al. An easy-to-use evaluation framework for benchmarking entity recognition and disambiguation systems[J]. Journal of Zhejiang University Science C, 2017, 18(2): 195-205.

[91] Sil A, Yates A. Re-ranking for joint named-entity recognition and linking[C]. Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. New York: ACM, 2013: 2369-2374.

[92] Kolitsas N, Ganea O E, Hofmann T, et al. End-to-end neural entity linking[C]. Proceedings of the Conference on Computational Natural Language Learning. Stroudsburg: ACL, 2018: 519-529.

[93] Martins P H, Marinho Z, Martins A F T. Joint learning of named entity recognition and entity linking[C]. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. Stroudsburg: ACL, 2019: 190-196.

[94] Sysoev A, Nikishina L. Deep JEDi: Deep joint entity disambiguation to wikipedia for russian[C]. Proceedings of the 8th International Conference the Lecture Notes in Computer Science. Cham: Springer, 2019: 230-241.

[95] Idrissou A K, Zamborlini V, Van H F, et al. Contextual entity disambiguation in domains with weak identity criteria: Disambiguating golden age amsterdamers[C]. Proceedings of the 10th International Conference on Knowledge Capture. New York: ACM, 2019: 259-262.

作者简介

段宗涛 (1977−), 男, 教授, 博士, 从事大数据、知识图谱等研究, E-mail: ztduan@chd.edu.cn;

李菲 (1996−), 女, 硕士生, 从事知识图谱的研究, E-mail: feili@chd.edu.cn;

陈柘 (1969−), 男, 副教授, 博士, 从事计算机视觉、知识图谱等研究, E-mail: zchen@chd.edu.cn.

(责任编辑：李君玲)