命名实体消歧研究综述

李欣宇,赵 震*
(渤海大学 信息科学与技术学院,辽宁 锦州 121013)

摘 要:实体消歧是指在一个具体的知识库中,把一个被标识的实体指称链向它对应条目的过程。实体消歧的任务是根据上下文信息解决一个命名实体指称项对应多个实体概念的一词多义问题,它在从海量数据准确提取信息的知识图谱构建过程中起到重要作用,是自然语言处理中的一项基本任务。该文主要对实体消歧技术的相关研究内容进行综述。首先,阐述了实体消歧的国内外研究背景,并对命名实体识别、候选实体生成、候选实体排序等实体消歧相关理论进行全面梳理。其次,对实体消歧的具体含义及其研究内容进行详细综述,并对实体消歧研究内容的特点进行了分析。再次,将实体消歧技术的实现方法划分为三类并对涉及到的数据集进行归纳,并从四个方面讨论了实体消歧领域存在的难点和提高实体消歧准确率的途径,对消歧方法的优缺点及评价指标进行了总结,意在为改善实体消歧效果提供新的解决思路。最后,对实体消歧技术的应用和发展前景进行总结。

关键词:实体消歧;命名实体识别;知识图谱;自然语言处理;综述
中图分类号:TP182 **文献标识码**:A **文章编号**:1673-629X(2024)02-0001-08
doi:10.3969/j.issn.1673-629X.2024.02.001

Review of Named Entity Disambiguation Studies

LI Xin-yu,ZHAO Zhen*
(School of Information Science and Technology,Bohai University,Jinzhou 121013,China)

Abstract:Entity disambiguation is the process of chaining an identified entity referent to its corresponding entry in a specific knowledge base. The task of entity disambiguation is to solve the word polysemy problem where a named entity referent term corresponds to multiple entity concepts based on contextual information,and it plays an important role in the construction of knowledge graphs for accurate extraction of information from massive data,which is a fundamental task in natural language processing. We mainly review the research content related to entity disambiguation techniques. Firstly,the background of the domestic and international research on entity disambiguation is described,and the theories related to entity disambiguation such as named entity identification,candidate entity generation,and candidate entity ranking are comprehensively reviewed. Secondly,a detailed overview of the specific meaning of entity disambiguation and its research content is presented,and the characteristics of the research content of entity disambiguation are analyzed. Thirdly,the implementation methods of entity disambiguation techniques are classified into three categories and the data sets involved are summarized,and the difficulties in the field of entity disambiguation and the ways to improve the accuracy of entity disambiguation are discussed from four aspects,and the advantages and disadvantages of disambiguation methods and evaluation indexes are summarized,with the intention of providing new solutions for improving the effectiveness of entity disambiguation. Finally,the application and development prospects of entity disambiguation techniques are summarized.

Key words:entity disambiguation;named entity identification;knowledge graph;natural language processing;review

0 引 言

在信息化发展迅速的今天,众多通用知识图谱和特定领域知识图谱应运而生。但随着网络上不断增加的数据量,针对一词多义和多词一义的语言现象,如果双方对于同一事物的理解不一致,就会造成非常多的误解和问题。如何确定某个实体指向的精确实体概念就变得更加重要,这也就是实体消歧的主要研究内容。实体消歧是自然语言处理中的一项基础环节,如何提高实体消歧准确率,解决实体消歧的难点问题,已经成为各领域当前的研究重点。

收稿日期:2023-04-10 **修回日期**:2023-08-15
基金项目:国家自然科学基金项目(61976027);辽宁省教育厅基本科研项目(LJKZ1028);渤海大学 2021 年研究生教育教学改革项目(YJG20210022)
作者简介:李欣宇(2000-),女,硕士研究生,CCF 会员(O7927G),研究方向为知识图谱构建等;通讯作者:赵 震(1977-),男(满族),博士,CCF 会员(D0145M),研究方向为人工智能与知识图谱。

该文主要的工作内容如下：

(1)对国内外研究现状进行分析，并整理了命名实体识别、候选实体生成等实体消歧相关研究理论。

(2)介绍了实体消歧的具体含义及其研究内容，同时以一个新颖的角度对实体消歧方法进行综述，详细阐述了基于全局和局部特征的实体消歧、基于上下文特征的实体消歧和基于字符串相似度的实体消歧方法。

(3)详细描述了实体消歧领域存在的难点，对实体消歧方法的优缺点及评价指标进行了总结，同时对如何提高实体消歧的准确率进行了讨论。

(4)对实体消歧领域的应用及未来发展进行了总结。

总体框架如图 1 所示。

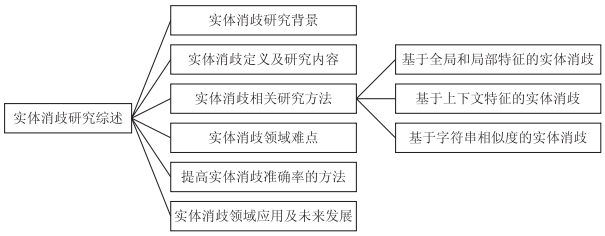

图 1　总体框架

1　实体消歧研究背景

在 Web of science 上对字段“Entity Disambiguation”进行检索，如图 2 所示，分析检索结果得到：第一，中国学者们对实体消歧相关研究发表的文章数多于外国，说明相比国外，中国对实体消歧的相关研究更加感兴趣。第二，国外对实体消歧相关研究的文章最早发表于 2013 年，近十年发表的文章总数呈上升趋势，说明实体消歧已经成为外国学者们越来越关注的研究内容。

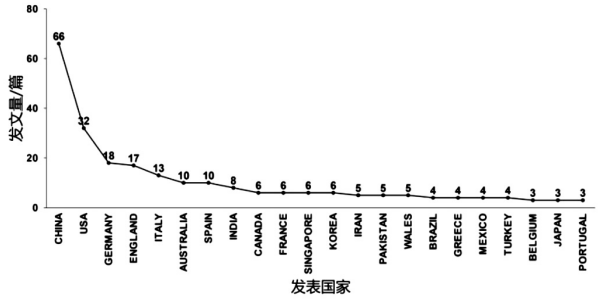

图 2　实体消歧国内外研究趋势

同时，在中国知网上对关键词“实体消歧”进行检索，如图 3 所示，分析检索结果得到：2008 年中国发表首篇实体消歧相关的文章，2020 年的发文量达至顶峰。2008 年至今，总体上看中国学者们在实体消歧研究领域的发文量呈上升趋势，但近三年文献数量显著减少，说明中国学者们也更加关注实体消歧领域的相关研究，但是近三年对实体消歧领域的研究热度有所减弱。

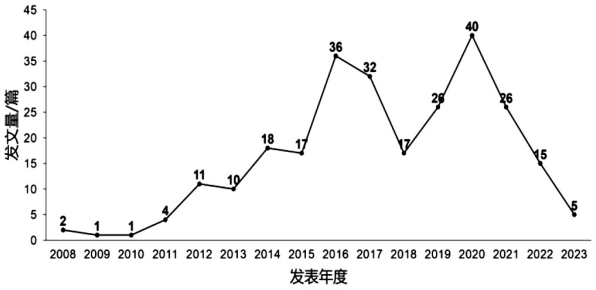

图 3　实体消歧中国研究趋势

通过对国内外实体消歧研究背景的分析可以看出，近年来，实体消歧技术取得了较大进展。同时，从温萍梅、段宗涛等一些研究人员在实体消歧方面的综述文章中可以看出，从文献全文等长文本到推特、微博、查询语句等短文本，再到专业领域语料，针对不同语料特征，学者们提出了对应的消歧策略[1-2]。然而，实体消歧技术还具有一定的提升空间，对于实体消歧技术的各个环节中仍存在着的一些问题和挑战，该文主要从命名实体识别、候选实体生成、候选实体排序、实体链接四个方面进行总结。

1.1　命名实体识别

命名实体识别是搭建知识库与自然语言之间的桥梁，它负责从给定文本中准确地识别出人名、地名、机构名、时间等所有类型的实体命名指称。

近年来，随着深度学习的流行，研究人员逐渐利用神经网络进行命名实体识别工作[2]。Lample 等[3]介绍了两种用于序列标记的神经结构，一种基于双向长短期网络(Bi-LSTM)和条件随机场(CRF)，另一种使用基于转移的方法，即使与使用外部资源的模型相比，它们也能在标准评估设置中提供最好的 NER(Named Entity Recognition)结果。Kuru 等[4]描述了一个采用深度 Bi-LSTM 架构的字符级标记器，句子不是用单词表示的，而是用字符序列表示的，评估结果展示了相同的深度字符级模型能够在多种语言上获得良好的 NER 性能。Rocktäschel 等[5]介绍了一个从自然语言文本中提取化学实体的混合系统 ChemSpot，该系统使用了将 CRF 与字典相结合的混合方法，通过结合这两种方法的优点，ChemSpot 实现了 NER 性能的大幅提高。

1.2　候选实体生成

简单来说，候选实体生成就是为每个实体指称在知识库中生成其可能的候选实体集合的过程。

Sui 等[6]提出一种分层多任务模型，将提取的超细类型信息引入到候选生成任务中，改进了高级零射实体链接候选生成任务，实验结果证明了该方法的有

效性。Fang 等[7]通过从多方面检索候选实体,提高了候选集的质量和候选实体生成方法的召回率。Hebert 等[8]提出一种密集检索方法进行候选生成,在 Twitter 领域该方法通过使用两个独立的语言模型分别对推文和实体的语义内容进行编码来实现,有效提高了候选生成的召回率。

1.3　候选实体排序

候选实体排序问题研究的内容是:给定一个查询 q 和一个由实体 $e \in E$ 填充的知识库(Knowledge Base,KB),找到满足该查询 q 的最佳匹配实体 e。

近年来,学者们对实体排序问题的研究有所增加。Hasibi 等[9]建立一个可以插入不同候选实体排名和消歧方法的框架,对于其中的每一个组件都用无监督和有监督的替代方案进行了实验,研究结果表明,在候选实体排序步骤中使用监督学习更有益。Cao 等[10]提出一种基于二部图的实体排名方法,该方法利用候选实体之间的 Co-List 关系来帮助提高实体排名,实验结果验证了该方法尤其在提升那些相关但不受欢迎实体的有效性。Mondal 等[11]提出一种基于候选知识库条目与疾病提及的相似度对候选知识库条目进行排序的方法,该方法使用三元网络进行候选人排名,结果表明其很大程度上优于现有的排序系统。

1.4　实体链接

实体链接是自然语言处理中的一项重要技术,它负责把给定文本中的实体指称链接到知识库中的一个无歧义实体,通常将维基百科作为知识库[12]。一个准确的实体链接系统对于许多与知识相关的任务,如智能问答和信息提取等是至关重要的。

为了严格解决 Twitter 上几乎没有上下文的实体链接问题,Guo 等[13]提出一种用于实体链接的结构化 SVM(Support Vector Machine)算法,通过同时考虑提及检测和实体消歧,构建了一个优于当前最先进系统的端到端实体链接系统。Le 等[14]使用 MIL(Multiple Instance Learning)方法,同时引入一个新的组件,即噪声检测分类器,与实体链接模型联合估计,从而产生更准确的实体链接模型。为了解决潜在实体类型常被忽略的问题,Chen 等[15]提出将潜在的实体类型信息注入到基于预训练 BERT 的实体嵌入中,并将基于 BERT 的实体相似度评分集成到最新模型的本地上下文模型中,来更好地捕获潜在的实体类型信息,实验结果表明该方法可以有效地改进实体链接。

2　实体消歧定义及研究内容

实体消歧是指将特定文档中的文本提及链接到 KB 中的正确命名实体的过程,它是自然语言处理的一个基本任务。

2.1　基于词典的语义消歧

基于词典的语义消歧是对词的处理,指根据一个多义词在特定文本中出现的上下文语义环境来确定其词义,通过使用词典或者类似词典的知识库进行消歧。基于词典的语义消歧是自然语言处理的核心和基础环节,有效解决基于词典的语义歧义问题,也会带动自然语言处理领域的新发展。整个基于词典的语义消歧过程可用公式描述如下:

$$S' = \operatorname{argmax} R(S_k \mid C)$$

其中,C 表示词语 W 所在的特定上下文语言环境,S_k 表示词语 W 在特定上下文语言环境 C 中的每个词义,$R(S_k \mid C)$ 表示词语 W 的每个词义 S_k 和特定上下文语言环境 C 存在的不同强弱的关系,S' 表示词语 W 在 N 个词义中的确定词义。

形式化地,基于词典的语义消歧过程就是通过分析和计算词语 W 所在的特定上下文语言环境 C 与每个词义 S_k 间的关系 R,找到关系最强的 S_k 即词语 W 的确定语义 S'。如“苹果”在百度百科中共有 24 个义项,常用的义项有“蔷薇科苹果属植物”“美国高科技苹果公司”“2007 年李玉执导电影”,基于这三个义项,“苹果”这个多义词的消歧示例如图 4 所示。

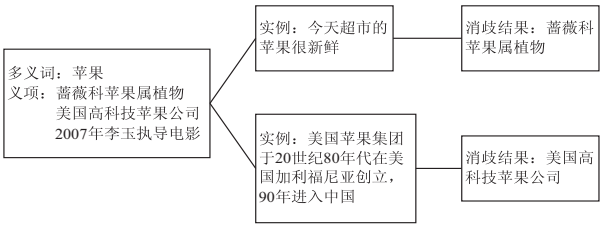

图 4　“苹果”消歧示例

2.2　基于实体的语义消歧

基于实体的语义消歧研究的主要内容是解决同一个实体指称项在不同的上下文中可以对应到不同真实世界实体的语言现象。例如,给定如下两个包含“Zhang Wei”的句子:

(1)Zhang Wei is a responsible entrepreneur.

(2)Zhang Wei is a famous piano player.

基于实体的语义消歧过程:在给定的特定文本“Zhang Wei won the piano competition.”中,“Zhang Wei”是待消歧实体指称项,通过将实体指称项在知识库中的两个定义“entrepreneur”,“piano player”和待消歧文本中“piano competition”一词分别计算并比较其语义相关度得到:特定文本中的“piano competition”与该实体定义中的“piano player”具有较高的语义相关度,所以该实体指称项“Zhang Wei”应指的是“一个著名的钢琴演奏家”。

基于实体的语义消歧研究的角度可以分为如下两种:

(1)实体指称多样性:一个命名实体可以有多种不同的方式表达。

(2)实体指称歧义性:一个实体指称在不同的上下文语言环境中可能表示不同的实体含义。

实体消歧研究内容的特点总结如表1所示。

表1 实体消歧研究内容的特点总结

研究内容	定义	词义	消歧特点	消歧方法	消歧难点	目标	未来研究重点
基于词典的语义消歧	根据上下文语境来确认某个多义词的确切语义	词义数目少且固定,可通过词典列举	消歧场景和可利用特征不够丰富	有监督、无监督和基于知识库的消歧方法	现有词典知识完备性不足,缺乏可扩展性和灵活性	确定文本中歧义词的正确含义	围绕词义消歧知识源的研究是最重要的,也是目前国内外研究热点
基于实体的语义消歧	解决某个实体指称项可对应到多个真实世界实体的问题	词义数目多且不固定,无法列举	消歧场景丰富,可利用特征多	命名实体聚类消歧、命名实体链接消歧	实体指称具有多样性、歧义性的特点	确定文本中歧义词的正确含义	未来可利用人工智能方法和领域资源来提升消歧效果

3 实体消歧相关方法

3.1 基于全局和局部特征的实体消歧

在局部消歧方面常用基于Bi-LSTM和注意力机制(Attention)相结合的方法,在全局消歧方面常用基于关联图和PageRank算法相结合的方法。目前,将局部和全局两种模型结合起来的消歧方法可以有效改善实体消歧的准确率。

例如,NCEL(Neural Collective Entity Linking)[16]方法就是应用图卷积网络集成局部上下文特征和全局图特征进行实体消歧。Yang等[17]第一个使用SGTB(Structured Gradient TreeBoosting)算法,并将全局特征与局部特征联合建模,来消除集体实体的歧义。Shahbazi等[18]提出一种新的实体消歧模型,该模型通过LDS(Limited Discrepancy Search)方法结合了局部上下文信息和全局证据,以从全局角度改进局部解决方案。Hu等[19]采用GNED(Graph Neural Entity Disambiguation)图神经网络模型,该模型充分利用了在同一文档中的候选实体之间的全局语义关系,解决实体消歧问题。Tang等[20]使用图注意网络捕获全局主题连贯性,图注意网络通过一种特殊的自注意机制,动态获取不同邻居节点的重要信息。

3.2 基于上下文特征的实体消歧

随着机器学习技术的发展,卷积神经网络(CNN)和循环神经网络(RNN)被用于实体消歧任务中,但针对其存在的上下文文本特征提取不充分、语义信息获取较少的问题,姜丽婷等[21]提出一个新的混合卷积网络(MCN)模型,该模型融合了CNN和图卷积网络(GCN)两种模型的优势来解决上下文文本特征提取不充分的实体消歧问题,获得了很好的结果。

Wang等[22]提出一个具有多视角注意力的神经网络,以丰富不同视角下的提及和实体表示,捕捉更多信息特征,提高消歧性能。Deng等[23]提出关联图和候选实体关联图,利用图神经网络(GNNs)获得同一文档的多主题相干特征进行消歧。Veloso等[24]提出EAND(Eager Associ-ative Name Disambiguation),LAND(Lazy Associative Name Disambiguation),SLAND(SelfTraining LAND)三种关联作者姓名消歧方法,特别是SLAND,它利用引文特征,扩展了LAND的自我训练能力,大大减少了构建有效消歧函数所需的示例数量,从而很好地实现了作者姓名消歧效果。

3.3 基于字符串相似度的实体消歧

机器学习用于教机器如何更有效地处理数据,赋予计算机在没有明确编程的情况下学习的能力,它依靠不同的算法来解决数据问题。孙笑明等[25]使用半监督学习算法,以特征向量(如姓名相似度、分类号相似度等)为信息提取源,构造基于决策树C4.5算法的分类模型,识别姓名歧义问题。

近年来,神经网络也得到了广泛的研究,并被证明可以有效地用于各种数据挖掘和分析任务。神经网络结合结构信息和语义特征的能力对自然语言处理任务中的实体消歧工作至关重要。例如,He等[26]次将深度神经网络引入实体链接框架,提出了一种基于深度神经网络(DNN)的实体消歧模型,通过直接优化给定相似性度量的文档和实体表示来消除实体歧义,进一步提高了消歧性能。Phan等[27]提出一种NeuPL(Neural network model combined with Pair-Linking)方法来计算实体之间的语义相似度,进而更好地实现实体消歧。

另外,DoSeR[28]设计了一种利用实体知识图上的个性化PageRank值的集体消歧方法,该方法利用语义嵌入来计算实体间的语义相似性从而进行实体消歧。Mingke等[29]提出了一种基于分类语义关联和结构语义关联的命名实体消歧方法,该方法综合考虑了实体之间的显式和隐式语义关联,通过计算结构语义相关度和分类语义相关度显著提高了实体消歧效果。Zhu

等[30]提出了一种基于词和类别嵌入联合学习的 Category2Vec 嵌入模型,该模型可以更好地计算词类别的相似性,有效地解决了上下文信息有限的短文本实体消歧问题,改善了实体消歧性能。

各类实体消歧方法涉及的主要技术和数据集如表 2 所示。

表 2　实体消歧方法总结

方法	主要技术	数据集
基于全局和局部特征的实体消歧	应用图卷积网络集成局部上下文特征和全局图特征[16]	使用收集的维基百科超链接
	定义全局特征,并将其与局部特征联合建模[17]	AIDA-CoNLL、AQUAINT、MSNBC、ACE、WIKI、CWEB
	用 LDS(Limited Discrepancy Search)方法结合局部上下文信息和全局证据[18]	CoNLL 2003 和 TAC 2010
	采用图神经网络模型,可以提高消歧准确率[19]	域内场景:AIDA-CoNLL;域外场景 AIDA-train,并在五个流行的测试集上进行测试:MSNBC、AQUAINT、ACE 2004、WNED-CWEB、WNED-WIKI
	使用图注意网络,动态获取不同邻居节点的重要信息[20]	(1)语言网络:构建了一个词的共现网络 (2)社交网络:Flickr 网络和 Youtube 网络 (3)引文网络:用 DBLP 数据集来构建
基于上下文特征的实体消歧	用混合卷积网络(MCN)模型,解决文本特征提取不充分、语义信息获取较少的问题[21]	数据集采用全国知识图谱与语义计算大会评测任务中的面向中文短文本的实体识别与链指任务数据
	利用多视角注意力神经网络,捕捉更多信息特征[22]	ACE2004、MSNBC 、AQUAINT、CWEB12、WW
	提出一个新的全局模型并结合图神经网络(GNNs)获得同一文档的多主题相干特征[23]	ArnetMiner 数据集
	从训练示例中提取将引用特征与特定作者相关联的规则来进行识别[24]	从 DBLP 和 BDBComp 中提取的引文集合
	使用半监督学习算法,构造基于决策树 C4.5 算法的分类模型[25]	从全国各大通讯企业的专利数据发明人中通过分层抽样选取了 400 个姓名对作为样本数据
	直接优化给定相似性度量的文档和实体表示[26]	非集体方法:TAC-KBP 集体方法:2010AIDA
基于字符串相似度的实体消歧	用 NeuPL(Neural network model combined with Pair-Linking)方法进行语义相似度计算[27]	NeuPL, Reuters128、ACE2004、MSNBC、DBpedia Spotlight 、RSS500、KORE50、Microposts2014
	利用一种个性化 PageRank 值的集体消歧方法[28]	ACE2004、AIDA/CO-NLL-TestB、AQUAINT、DBpedia Spotlight、MSNBC、N3-Reuters128、IITB、Microposts-2014 Test、N3 RSS-500
	结合显式和隐式语义关联,利用领域知识的主要属性进行消歧[29]	中国林业信息网(CFI)的中国木本植物资源数据库和中国植物学科数据库(SDCP)
	提出一种基于词和类别嵌入联合学习的 Category2Vec 嵌入模型,计算词类别的相似性[30]	Web Questions 数据集、Web Queries 数据集、Tweets 数据集

4　实体消歧领域难点

4.1　中文比英文消歧难度高

与中文的知识资源相比,英文的知识资源更加成熟和丰富。由于汉语语义知识资源的稀缺,知识获取瓶颈在汉语中更为严重,这也就导致了中文实体消歧的困难性。邵发等[31]针对开放文本中中文实体关系抽取的一词多义问题,利用贝叶斯分类器和模式合并法提高实体关系抽取性能。Lu 等[32]为了解决中文消歧知识瓶颈的问题,提出一种基于图的多知识集成中文 WSD(Word Sense Disambiguation)方法来消除歧义。

4.2　短文本比长文本消歧难度高

短文本的上下文通常是嘈杂和稀缺的,具有信息模糊和不完整的问题。无法提供给实体消歧任务所必需的丰富的上下文信息,这一局限性给实体消歧任务增加了难度。Jiang 等[33]提出一种基于神经网络的胶囊网络和 CNN 的实体消歧方法,充分利用了短文本数据的全部语义信息来执行实体消歧任务。Feng 等[34]针对短文本信息模糊和不完整的问题,提出一种知识

增强的短文本实体消歧方法,可以显著提高短文本实体消歧任务的性能。

4.3 跨语言比单语言消歧难度高

跨语言实体消歧在过去几年得到了学者们的关注,不同语言之间的翻译是跨语言信息抽取的难点,而在翻译过程中自然存在着实体歧义性问题。Barrena等[35]提出一个0-shot XNED(zero-shot cross-Lingual Named Entity Disambiguation)架构,为每个可能的提及字符串提供了一个模型,从而消除了本机先验概率的需要,而不是一个单一的消歧模型。Maeda等[36]针对基于查询翻译的CLIR(Cross Language Information Retrieval)方法所需要的自然语言资源不易获得的问题,提出一种基于词典的查询翻译的消歧方法,实现足够的检索效率。

4.4 低程度的相关性消歧难度高

传统的实体消歧方法通常是基于文档中提到的所有实体都紧密相关,但研究表明,在一些新闻、推文中也常常存在着一些低相关性的实体。Phan等[37]针对文档中提及到的实体存在低程度的一致性问题,提出一种新的基于树的集合链接模型MINTREE,该模型利用最小生成树的权值来度量实体图中的相干性。Zhang等[38]在同一份文件中的提及通常对应不同的主题,提出一种多主题全局一致性特征提取的全局模型。

5 提高实体消歧准确率方法

5.1 改善相似度计算方法

因为实体间存在着较为复杂的关系,所以应用更优异的相似度计算方法能够更准确地描述出它们之间的关联度,进而可以提高实体消歧的准确率。

汪沛等[39]采用一种基于图的随机游走算法辅助计算相似度,可以高效地获取实体指称项与目标实体间的相似度,进一步提升了特定领域实体消歧的准确率。Fan等[40]提出一种基于图的GHOST(abbreviation for GrapHicalframewOrk for name diSambiguaTion)算法,结合AP(A-ffifinity Propagation)聚类算法进行相似度计算,在人名消歧方面取得了较好的实验结果。

5.2 提升词间依赖性

丰富的上下文依赖关系,可以增强实体间的关联程度,进而来帮助实现消歧的过程。

曾维新等[41]提到现有的实体消歧方法大多采用集体排序方法以更好地捕捉实体指称间的依赖性,进而提升消歧效果。Li等[42]提出一种结合双注意机制和分布强化的图卷积网络关系提取模型,该方法通过两个并行注意模块聚合全局特征语义信息,增强特征全局依赖性。

5.3 挖掘实体隐藏语义关系

大多数关系提取方法都需要足够的数据来实现良好的性能,挖掘实体的隐藏语义可以更准确地提取文本中的实体关系,改善实体消歧性能。

Guo等[43]提出一种以结构化数据库为领域知识的连体图神经网络——传记神经网络模型,提高了从生物医学文献中提取实体关系的准确性。它还可以在生物医学文献中发现一些潜在的、未被发现的关系。Zeng等[44]提出利用潜在文本特征,通过基于双注意的长短期记忆网络(LSTM)生成提及和实体的表示,挖掘表面形式下的语义关系,并进一步用于计算提及-实体相似度。

5.4 解决有效信息利用不足问题

充分考虑实体特征并综合提取文档属性特征,可以进一步提高实体消歧的精确度。

Deng等[23]提出一种新的HRFAENE(Heterogeneous Relation Fusion and Attribute Enhanced Network Embe-dding)模型,该模型通过对网络结构和属性的多次学习,有效地解决了有效信息利用不足的问题,提高消歧效果。贺紫涵[45]针对文档级实体消歧问题中一致性特征提取不精确的现象,在实体局部一致的前提下,提出一种GN-CED(Graph Neural Collaborative Entity Disambiguation)协同实体消歧模型,实验结果表明,相比于其它方法,应用该模型可以改善实体消歧的准确率。

前文所介绍的实体消歧方法具有良好的准确率,对它们的优缺点和评价指标进行汇总如表3所示。

表3 实体消歧方法优缺点及评价指标汇总

方法	优点	缺点	评价指标
基于全局和局部特征的实体消歧	充分利用全局语义,灵活地探索实体之间的相关性,有很好的泛化能力	复杂性高、成本高	in-KB、Micro-F1、精确度、召回率和F1值
基于上下文特征的实体消歧	捕获更多信息特征,具有更好的准确性和稳定性	存在参数自适应问题,未考虑时间和其他可变因素的影响	精确度,召回率,F1指标,Micro-F1和Macro-F1
基于字符串相似度的实体消歧	适用于有限上下文信息的短文本	存在一定的误差,尚未针对高性能消歧进行优化	Micro-F1、Macro-F1、Micro-averaged F1;精确度、召回率和F-Measure

6　结束语

实体消歧在自然语言处理工作中扮演重要角色,在智能推荐、智能问答、信息检索、知识库构建领域都有着广泛的应用价值。其中,武汉音乐学院构建了智慧型博物馆,当人们对某件乐器感兴趣,在该乐器前停下来时,VR、AR 设备就会展示出该乐器的知识图谱,参观者可以根据自身专业和兴趣提取相关知识。其知识图谱中存在的非结构化类型数据就需要实体消歧技术做进一步处理。该智慧型博物馆旨在实现网络与实体馆的多维度互动,进而探索音乐与科技融合的新模式。

实体消歧领域未来发展方向:

(1)充分考虑消歧特征,提高中文短文本的实体消歧效果。

(2)优化不同语言文本间的相似度计算方法,保证较好的通用性。

(3)深入研究结合文本和图形信息的实体消歧方法,提高消歧准确率。

(4)将图卷积网络应用于实体消歧。应该更多的将图卷积网络应用于实体消歧,可以获得更高的消歧性能。

未来对实体消歧领域的研究应该更多地结合卷积神经网络、图卷积网络等深度学习方法,更多地应用于文本语言环境不理想的实体歧义现象中。知识图谱与自然语言处理关系密切,实体消歧是知识图谱构建中的一个关键技术,期待研究人员更多地关注到实体消歧领域,探索提高实体消歧准确率的方法,从而进一步推动自然语言的发展。

参考文献:

[1] 温萍梅,叶志炜,丁文健,等. 命名实体消歧研究进展综述[J]. 数据分析与知识发现,2020,4(9):15-25.

[2] 段宗涛,李　菲,陈　柘. 实体消歧综述[J]. 控制与决策,2021,36(5):1025-1039.

[3] LAMPLE G, BALLESTEROS M, SUBRAMANIAN S, et al. Neural architectures for named entity recognition[J]. arXiv:1603.01360v3,2016.

[4] KURU O, CAN O A, YURET D. Charner: character-level-named entity recognition[C]//Proceedings of COLI-NG 2016, the 26th international conference on computational linguistics. Osaka: [s. n.],2016:911-921.

[5] ROCKTÄSCHEL T, WEIDLICH M, LESER U. ChemSpot: a hybrid system for chemical named entity recognition[J]. Bioinformatics,2012,28(12):1633-1640.

[6] SUI X, ZHANG Y, SONG K, et al. Improving zero-s-hot entity linking candidate generation with ultra fine entity type information[C]//Proceedings of the 29th international conference on computational linguistics. Gyeongju: ICCL,2022: 2429-2437.

[7] FANG Z, CAO Y, LI R, et al. High quality candidate generation and sequential graph attention network for entity linking[C]//Proceedings of the web-conference 2020. Taipei, China: ACM,2020:640-650.

[8] HEBERT L, MAKKI R, MISHRA S, et al. Robust candidate generation for entity linking on short social media texts[J]. arXiv:2210.07472,2022.

[9] HASIBI F, BALOG K, BRATSBERG S E. Entity linking in-queries: efficiency vs. effectiveness[C]//Advances in information retrieval: 39th European conference on IR research. Aberdeen: Springer International Publishing,2017:40-53.

[10] CAO L, GUO J, CHENG X. Bipartite graph based entity ranking for related entity finding[C]//2011 IEEE/WIC/ACM international conferences on web intelligence and intelligent agent technology. Lyon: IEEE,2011:130-137.

[11] MONDAL I, PURKAYASTHA S, SARKAR S, et al. Medical entity linking using triplet network[J]. arXiv:2012.11164, 2020.

[12] 李天然,刘明童,张玉洁,等. 基于深度学习的实体链接研究综述[J]. 北京大学学报,2021,57(1):91-98.

[13] GUO S, CHANG M W, KICIMAN E. To link or not to link? a study on end-to-end tweet entity linking[C]//Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics. Minneapolis: ACL,2013:1020-1030.

[14] LE P, TITOV I. Distant learning for entity linking with automatic noise detection[J]. arXiv:1905.07189,2019.

[15] CHEN S, WANG J, JIANG F, et al. Improving entity linking by modeling latent entity type information[C]//Proceedings of the AAAI conference on artificial intelligence. New York: AAAI,2020:7529-7537.

[16] CAO Y, HOU L, LI J, et al. Neural collective entity linking[J]. arXiv:1811.08603,2018.

[17] YANG Y, IRSOY O, RAHMAN K S. Collective entity disambiguation with structured gradient tree boosting[J]. arXiv:1802.10229,2018.

[18] SHAHBAZI H, FERN X Z, GHAEINI R, et al. Joint neural entity disambiguation with output space search[J]. arXiv: 1806.07495,2018.

[19] HU L, DING J, SHI C, et al. Graph neural entity disambiguation[J]. Knowledge-Based Systems,2020,195:105620.

[20] TANG J, QU M, WANG M, et al. Line: large-scale information network embedding[C]//Proceedings of the 24th international conference on world wide web. Florence: IW3C2, 2015:1067-1077.

[21] 姜丽婷,古丽拉·阿东别克,马雅静. 基于混合卷积网络的短文本实体消歧[J]. 中文信息学报,2021,35(11):101-108.

[22] WANG C, SUN X, YU H, et al. Entity disambiguation leveraging multi-perspective attention[J]. IEEE Access,2019,7:

113963-113974.
[23] DENG C,DENG H,LI C. A scholar disambiguation method based on heterogeneous relation - fusion and attribute enhancement[J]. IEEE Access,2020,8:28375-28384.
[24] VELOSO A,FERREIRA A A,GONCALVES M A,et al. Cost-effective on-demand associative author name disambiguation[J]. Information Processing & Management,2012,48(4):680-697.
[25] 孙笑明,余武憬,任若冰,等. 基于决策树算法的专利发明人姓名消歧研究[J]. 科学与管理,2023,8:1-20.
[26] HE Z,LIU S,LI M,et al. Learning entity representation for entity disambiguation[C]//Proceedings of the 51st annual meeting of the association for computational linguistics. Bulgaria Sofia:ACL,2013:30-34.
[27] PHAN M C,SUN A,TAY Y,et al. NeuPL:attention based semantic matching and pair-linking for entity disambiguation[C]//Proceedings of the 2017 ACM on conference on information and knowledge management. Turin:ACM,2017:1667-1676.
[28] ZWICKLBAUER S,SEIFERT C,GRANITZER M. Robust andcollective entity disambiguation through semantic embeddings[C]//Proceedings of the 39th international ACM SIGIR conference on research and development in information retrieval. Pisa:ACM,2016:425-434.
[29] MINGKE C,DONGMEI L,TINGTING Z,et al. Named entity disambiguation based on classified and structural semantic relatedness[J]. Chinese Journal of Electronics,2018,27(6):1176-1182.
[30] ZHU G,IGLESIAS A C. Exploiting semantic similarity for named entity disambiguation in knowledge graphs[J]. Expert Systems with Applications,2018,101:8-24.
[31] 邵 发,黄银阁,周兰江,等. 基于实体消歧的中文实体关系抽取[J]. 山东大学学报,2014,44(6):32-37.
[32] LU W,MENG F,WANG S,et al. Graph-based Chinese word sense disambiguation with multi-knowledge integration[J]. Comput. Mater. Continua,2019,61(1):197-212.
[33] JIANG L,ALTENBEK G,WU D,et al. Chinese short text entity disambiguation based on the dual-channel hybrid network[J]. IEEE Access,2020,8:206164-206173.
[34] FENG Z,WANG Q,JIANG W,et al. Knowledge-enhanced named entity disambiguation for short text[C]//Proceedings of the 1st conference of the Asia-Pacific Chapter of the association for computational linguistics and the 10th international joint conference on natural language processing. Qingdao:CCF-NLP,2020:735-744.
[35] BARRENA MADINABEITIA A, SOROA ECHAVE A, AGIRRE BENGOA E. Towards zero-shot cross-lingual named entity disambiguation[J]. Expert Systems with Applications,2021,184:115542.
[36] MAEDA A,SADAT F,YOSHIKAWA M,et al. Query term disambiguation for Web cross-language information retrieval using a search engine[C]//Proceedings of the fifth international workshop on information retrieval with Asian languages. Sanya:NT-CIR,2000:25-32.
[37] PHAN M C,SUN A,TAY Y,et al. Pair-linking for collective entity disambiguation:two could be better than all[J]. IEEE Transactions on Knowledge and Data Engineering,2018,31(7):1383-1396.
[38] ZHANG C,LI Z,WU S,et al. Multitopic coherence extraction for global entity linking[J]. Electronics,2022,11(21):3638.
[39] 汪 沛,线岩团,郭剑毅,等. 一种结合词向量和图模型的特定领域实体消歧方法[J]. 智能系统学报,2016,11(3):366-375.
[40] FAN X,WANG J,PU X,et al. On graph-based name disambiguation[J]. Journal of Data and Information Quality,2011,2(2):1-23.
[41] 曾维新,赵 翔,冯 滔,等. 面向领域的命名实体消歧方法改进研究[J]. 计算机工程与应用,2018,54(17):126-134.
[42] LI Z,SUN Y,ZHU J,et al. Improve relation extraction with dual attention-guided graph convolutional networks[J]. Neural Computing and Applications,2021,33:1773-1784.
[43] GUO S,HUANG L,YAO G,et al. Extracting biomedical entity relations using biological interaction knowledge[J]. Interdisciplinary Sciences:Computational Life Sciences,2021,13:312-320.
[44] ZENG W,TANG J,ZHAO X. Entity linking on Chinese microblogs via deep neural network[J]. IEEE Access,2018,6:25908-25920.
[45] 贺紫涵. 面向实体消歧的特征增强方法研究[D]. 重庆:重庆大学,2021.