计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2016.20150832
53(2):270-283,2016

基于图的中文集成实体链接算法

刘 峤 钟 云 李 杨 刘 瑶 秦志光
(电子科技大学信息与软件工程学院 成都 610054)
(qliu@uestc.edu.cn)

Graph-Based Collective Chinese Entity Linking Algorithm

Liu Qiao, Zhong Yun, Li Yang, Liu Yao, and Qin Zhiguang
(*School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu* 610054)

Abstract Entity Linking technology is a central concern of the knowledge base population research area. Traditional entity linking methods are usually limited by the immaturity of the local knowledge base, and deliberately ignore the semantic correlation between the mentions that co-occurr within a text corpus. In this work, we propose a novel graph-based collective entity linking algorithm for Chinese information processing, which not only can take full advantage of the structured relationship of the entities offered by the local knowledge base, but also can make use of the additional background information offered by external knowledge sources. Through an incremental evidence minning process, the algorithm achieves the goal of linking the mentions that are extraced from the text corpus, with their corresponding entities located in the local knowledge base in a batch manner. Experimental results on some open domain corpus demonstrate the validity of the proposed referent graph construction method, the incremental evidence minning process, and the coherence criterion between the mention-entity pairs. Experimental evidences show that the proposed entity linking algorithm consistently outperforms other state-of-the-art algorithms.

Key words collective entity linking; knowledge base population; knowledge graph; referent graph; Chinese information processing

摘 要 实体链接(entity linking)是知识库扩容的核心关键技术,传统的实体链接方法通常受制于本地知识库的知识水平,而且忽略共现实体间的语义相关性.提出了一种基于图的中文集成实体链接方法,不仅能够充分利用知识库中实体间的结构化关系,而且能够通过增量证据挖掘获取外部知识,从而实现对同一文本中出现的多个歧义实体的批量实体链接.在开放域公开测试语料上的实验结果表明,所提出的实体相关图构造方法、增量证据挖掘方法和实体语义一致性判据是有效的,算法整体性能一致且显著地优于当前的主流算法.

关键词 集成实体链接;知识库扩容;知识图谱;实体相关图;中文信息处理

中图法分类号 TP391

收稿日期:2015-09-24;**修回日期**:2015-11-16

基金项目:国家自然科学基金项目(61133016,61272527,61202445);教育部-中国移动科研基金项目(MCM20121041);中央高校基本科研业务费专项资金(ZYGX2014J066)

This work was supported by the National Natural Science Foundation of China (61133016,61272527,61202445), Chinese Ministry of Education-ChinaMoblie Communications Corporation Research Funds (MCM20121041), and the Fundamental Research Funds for the Central Universities (ZYGX2014J066).

实体链接(entity linking)是文本分析会议(text analysis conference, TAC)知识库构建领域设定的基本挑战,任务目标是将从文本中提取到的实体指称项正确地链接到知识库中对应的实体对象上[1].

实体链接是知识库扩容的核心关键技术.随着面向开放域的信息抽取技术的发展,人们有可能从海量开放数据中自动抽取出实体、关系和属性信息[2-4].然而,通过开放域抽取得到的知识元素间的关系是扁平化的,缺乏层次性和逻辑性,甚至可能包含大量冗余和错误信息.为建立结构化的知识库,首先必须解决知识融合的问题,实体链接技术就是用于解决知识库构建过程中遇到的实体映射问题的信息融合技术.通过实体链接,可以消除概念的歧义,剔除冗余和错误概念,从而确保知识的质量[5].

具体说来,通过实体链接可以解决实体指称项的歧义性和多样性问题[6].实体指称项的歧义性是指相同的实体指称项在不同的上下文环境中有可能指向不同的实体对象,例如实体指称项"张三"在不同的语境下可能指代不同的实体对象.实体指称项的多样性则是指某个特定的实体对象,可能与多个不同的实体指称项(如别名、缩写等)相对应,例如NBA球星"迈克尔·乔丹"在不同的语境中可以采用"乔丹"、"飞人"甚至姓名缩写"M.J"来指代.

实体链接技术不仅具有重要的理论研究价值,而且有着重要和迫切的实际应用价值.知识库扩容是自然语言处理、人工智能和专家系统等相关领域共同关注的热点研究领域,而实体链接问题是当前该领域面临的主要研究挑战[5].近年来,随着实体链接技术的发展,知识库自动构建和扩容技术也不断取得进展,一些商用和公益性知识库的规模得到了迅速扩张,例如,WolframAlpha知识库的实体总数已超过10万亿条,而谷歌知识图谱则拥有5亿个实体和350亿条实体间的关系.然而,现有的实体链接技术仍存在明显的局限性,如依赖百科知识作为实体链接的知识来源,导致处理开放域实体链接任务时的性能不稳定和计算效率低下.一旦面向开放域的实体链接技术取得突破,将对知识库的扩容产生极大的推动作用,进而对知识库应用产生深远影响[7].

现有的实体链接研究成果主要面向英文处理,相对而言,中文实体链接技术的发展稍微有些滞后,主要有如下3方面原因:1)英文的开源知识库建设起步较早,已建成一些较为成熟的知识库,如DBpedia①,Freebase②等,而中文开源知识库目前仍处于起步阶段,对实体链接研究工作形成一定的制约;2)中文实体抽取技术受制于分词技术,分词和词性标注是中文信息处理技术的难点,也是制约实体链接技术发展和应用的关键问题之一;3)中文实体的共指和消歧处理难度比英文更大,原因是中文的语法更为灵活,语义也更加丰富[8].中文是仅次于英语的世界性语言,对中文实体链接的研究可以促进中文知识库的扩容,进而提高对中文信息的智能处理水平,因此是极具前景的研究方向,近年来吸引了大量的研究努力,TAC 2015会议也将跨语言实体链接(中文、英文、西班牙语)定为主要挑战.本文的研究目标就是致力于解决中文实体链接研究中当前面临的主要挑战性问题.

当前主流的实体链接方法采用基于相似度比较的思路,即通过计算实体指称项与其相应的候选实体间的上下文相似度,选择相似度最高的候选实体作为链接目标[9-10].该方法的局限性在于每次仅处理文本中出现的一个待定实体指称项,计算效率低,且未考虑该文本中共现实体间的语义相关性,造成信息浪费和实体链接准确率降低.研究表明,利用词语间的共现关系能够有效提高消歧的准确性[11].本文提出一种基于图的中文集成实体链接算法(graph-based collective Chinese entity linking algorithm, GCCEL),通过将文本中出现的实体指称项以及其候选实体集合视为图的顶点集合,利用实体间的语义相关性构造实体相关图,以图中顶点的语义一致性为判据,实现对同一文本中出现的多个实体的批量实体链接.与相关工作相比,本文的主要贡献在于:

1) 所设计的实体相关图综合考虑了实体间的语义相关度、上下文相似度、实体的知名度(流行程度)以及实体在知识库和外部知识源中表现出的间接语义关联等要素,能够更准确地辅助实现候选实体的区分和判别,达成精准实体链接的目标;

2) 在实体相关图构造过程中引入了增量证据挖掘的思想,在充分利用本地知识库中既有知识的基础上,能够有效利用第三方知识源提供的实体背景知识,从而在降低对本地知识库的依赖的同时,显著提升实体链接算法的准确率和召回率;

① http://datahub.io/dataset/dbpedia

② https://www.freebase.com

3) 提出了一个完整的基于实体相关图的中文实体集成链接算法原型和一种全新的实体语义一致性计算方法,并基于实体相关图实现了对开放域文本语料的批量实体链接.实验结果表明,该算法的准确率和召回率优于当前主流的相关工作,且所需的训练样本规模较小,方法适应性和推广性较好.

1 相关工作

实体链接任务是知识库构建领域当前面临的关键问题和基本挑战之一,由于该技术对于知识库扩容具有重要的基础研究价值,近年来受到了学术界的广泛关注.早期的实体链接研究思路主要针对单一实体进行考虑,即逐一地将从外部语料中抽取得到的实体映射到知识库中.近年来,随着一系列集成实体链接方法的提出,该类方法逐渐成为研究热点.本节将首先简要介绍实体链接方法的研究进展概况,然后重点讨论与本文工作密切相关的集成实体链接方法.

1.1 单实体链接方法

单实体链接方法一次仅对文本中的一个实体进行链接,而不考虑文本中其他共现实体的影响.基本研究思路是通过计算从文本中抽取得到的实体指称项与从知识库查询得到的候选实体之间的上下文相似度,选择相似度最大的候选实体作为链接目标.代表性工作是 Bunescu 等人提出的基于上下文的相似度计算模型,该模型以维基百科为知识库,对于给定文本中抽取得到的实体指称项,在维基百科上查找相应的候选实体构成集合,然后利用词袋模型计算给定文本和候选实体所在的维基页面之间的余弦相似度,选择相似度最大的候选实体作为链接对象[9].

该项研究工作引发了学术界对基于相似度计算的实体链接方法的关注,产生了一些具备实用价值的成果.其中,Silviu 在计算实体间的余弦相似度时加入了对实体间类别相关性的考虑,在维基百科和新闻网页语料上分别取得了 88.3%和 91.4%的实体链接准确率[10].类似的方法扩展工作还包括 Nguyen 等人提出的相似度计算模型,该模型在计算相似度时加入了候选实体在维基百科页面的上下文特征(关键词)和页面结构特征(如页面重定向、实体类别、锚文本等),从而有效提高了算法的准确性[12].

针对多个候选实体可能具有相同的余弦相似度的问题,Zeng 等人提出采用外部知识扩展实体指称项特征向量的解决方案(在该论文中是以实体指称项上下文词的维基百科页面作为外部知识源,对输入文本的特征向量进行扩展,然后在迭代计算实体指称项与其候选实体的上下文相似度),该方法在新浪微博数据集上取得了 88.5%的实体链接准确率[13].

基于实体上下文相似度的实体链接算法通常受制于上下文信息的不足,为此 Zhang 等人提出了一种基于图模型的维基概念相似度计算方法,该方法借助维基百科提供的实体上下文,能够有效提高实体指称项与候选实体间的语义相似度计算准确性,在 TAC2011 会议的 KBP 数据集取得了 80.40%的准确率[14].本文提出的 GCCEL 算法与该方法的主要区别在于实体相关图的构造方法不同.首先,Zhang 等人提出的方法仅考虑实体间的直接上下文关联关系,而 GCCEL 算法则在此基础上进一步考虑到了实体间的间接语义关联关系,即 2 个目标实体均与第三方存在直接关系的情况.其次,Zhang 等人提出的算法是基于全局的维基概念图的(图中包含 260 万个节点、5 100 万条边),而本文提出的算法仅针对输入文本中出现的实体构造相关子图,因此计算效率更高.

除了基于相似度计算的方法外,一些学者还尝试将统计机器学习方法引入到实体链接工作中.例如,Zuo 等人提出了一个投票模型,思路是将奇数个实体链接方法作为分类器,在链接时分别对每个候选实体进行 0/1 判定,获得半数以上选票的候选实体将成为最终的链接对象.该方法在 KORE,CoNLL-YAGO,CUCERZAN 等基准数据集上测试的结果显示,F_1 值分别达到了 77.83%,87.98%,88.61%[15].

1.2 集成实体链接方法

单实体链接方法的主要缺点在于没有考虑同一篇文档中出现的实体间的语义相关性,而这种由共现关系导致的语义相关性对于区分有歧义的实体通常是有帮助的.为了解决这一问题,Han 等学者提出了基于实体共现关系的集成实体链接方法[16].

集成实体链接方法的基本思想是对给定文本中出现的所有实体指称项,在当前的上下文环境中根据语义关联关系同步进行消歧,即批量地将其链接到本地知识库中对应的实体对象上.与单实体链接方法相比,集成实体链接方法的另一个优点是采用语义相关图的方式表示文档中实体间的语义关系,从而避免了逐一扫描待定实体,分别进行消歧处理的单线程处理模式,因此有助于提高实体链接效率.

Han 等人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,将实体指称项和候选实体视为图的顶点,以实体间的谷歌距离(Google distance)作为语义相关性测度,建立与该文本对应的实体相关图;最后,采用随机游走方法对图中的候选实体进行排序,得到实体链接的推荐结果. 在维基百科和 IITB 等基准语料上的实验结果表明,基于图的集成实体链接算法性能优于当前主流的单实体链接方法[16].

该方法的提出在业界和学术界形成了广泛影响,近年来涌现出大量相关工作. 其中,Shen 等人基于维基百科和 Yago 知识库提出的 LINDEN 模型将实体间的语义关联进一步区分为语义相似性和全局相关性,在 TAC2009 会议数据集上实现了高达 84.32%的实体链接准确率[17]. Johannes 等人则进一步考虑了实体流行度和相似度等因素,并据此设计实现了一个面向实际应用的 AIDA 实体链接原型系统[18]. 在上述工作的基础上,Ayman 等人通过修改顶点初始(概率)值的方式,将实体上下文相似度和实体流行度等因素结合到 PageRank 算法原型中,在 AIDA 数据集上实现了 86.10%的实体链接准确率[19].

然而,这些方法存在的共同问题是依赖实体所在的百科页面作为知识源,对于非知名实体的消歧任务而言,方法的适用性较差[20]. 为解决该问题,Andrea 等人利用 BabelNet 语义网络,采用带重启的随机游走(random walk with restart, RWR)算法来计算实体间的语义相关性(称为语义签名),以此为基础构造实体相关图,并采用抽取密集子图的方式实现实体消歧. 所提出的 Babelfy 算法在 KORE 和和 CoNLL 等基准数据集上的准确率分别达到 71.5%和 82.1%,是目前性能表现最好算法之一[21]. Babelfy 算法虽然部分降低了对百科知识的依赖,但却增加了对本地知识库的依赖. 为充分利用海量公开数据中包含的实体区分性证据,Li 等人提出了增量证据挖掘的思想,基本思路是采用外部知识库作为对本地知识库中实体知识的补充,在 Twitter 数据集上的模拟实验表明,采用基于生成模型的增量证据挖掘方法,能够有效地提升实体链接的准确率[22].

与其他基于图的集成实体链接方法相比,本文提出的 GCCEL 算法的主要特点是:在实体相关图的构造阶段考虑到语义关系相近的实体之间的间接关联,并引入了增量证据挖掘机制,在实体链接阶段综合考虑了实体相关图的拓扑结构特征和实体间的语义相似性特征. 通过上述改进,GCCEL 算法能够有效降低现有基于图的集成实体链接方法对本地知识库和单一外部知识源的依赖性,显著提高实体链接算法的准确率和召回率. 此外,与 Li 等人提出的算法相比,本文提出的增量证据挖掘方法不是基于生成模型,而是基于实体间的上下文关联性,因此在计算上更为简捷高效,模型本身也更为直观,易于理解和扩展.

2　基于图的中文集成实体链接算法

本文提出的基于图的中文实体集成链接方法由 3 个模块构成:候选实体生成模块、实体相关图构造模块和集成实体链接模块,如图 1 所示:

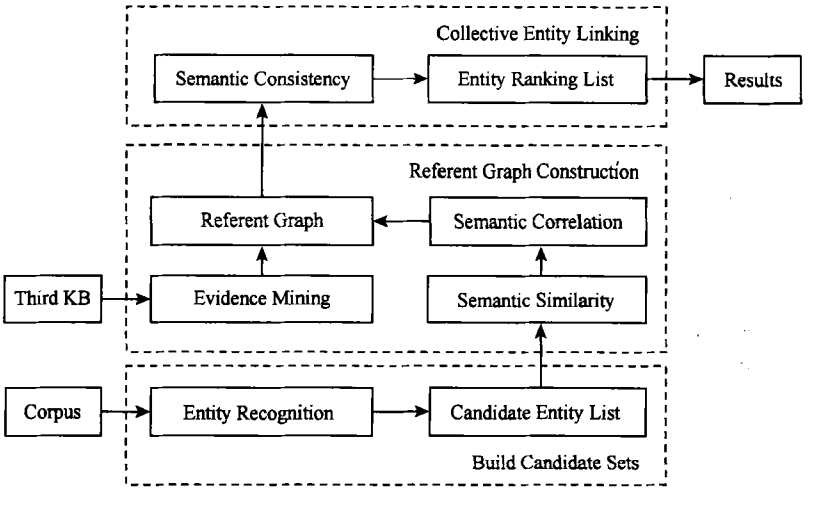


Fig. 1　The framework of the graph-based collective Chinese entity linking algorithm.

图 1　基于图的中文集成实体链接算法框架

候选实体生成模块的主要功能是对于给定的输入语料,识别出其中的所有实体指称项,据此分别查找本地知识库,得到与该实体指称项同名的候选实体集合,作为后续构造实体相关图的顶点集合.

实体相关图构造模块的主要功能是针对从同一文本中抽取得到的所有实体指称项和相应的候选实体集合,构造出一张该文本的实体参考关系图(referent graph),作为集成实体链接的依据.为简化描述,下文将实体参考关系图简称为实体相关图.为弥补本地知识库知识容量不足的问题,在实体相关图的构造阶段,引入了增量证据挖掘机制.

集成实体链接模块的主要功能是借助实体相关图实现对输入语料中歧义实体的消歧,将其正确链接到本地知识库中的正确的实体对象上.基本方法是基于实体指称项所在文本语料与其候选实体百科页面的余弦相似度和候选实体的出入度,计算每个候选实体与其对应的实体指称项的语义一致性,并选择语义一致性最高的候选实体作为最终的目标链接对象.

2.1 候选实体生成

本文使用的知识库(以下称本地知识库)以清华大学发布的中文知识库为基础构建[①],该知识库包含 19 542 个概念(类别)和 802 593 个实体.为提高实体语义相似度计算结果的准确性,本文采用百度百科作为外部知识源对该知识库中的实体属性知识进行了扩充.具体方法是针对知识库中的每个实体,抓取实体对应的百科页面,借助页面的 Infobox 抽取该实体的基本属性和关系属性信息,所抽取的实体基本属性包括别名、身高、性别等,所抽取的实体关系属性包括朋友关系、夫妻关系、父母关系等.经过扩充后的实体知识库采用 Neo4j 图数据库进行管理.

本地知识库中的同名歧义实体带有后缀标签,例如,对于实体指称项"李娜",在本地知识库中有三个实体对象与之对应,后缀标签分别为"北京大学教授"、"歌手"和"网球运动员".采用带括号的后缀标签来表达和区分这些歧义实体,如:李娜(歌手).

为提高检索效率,在构建本地知识库时,为每个实体节点建立索引,同名的候选实体具有相同的索引,指向同一实体指称项.这样在查询时,可以一次性查找到给定实体指称项所对应的所有候选实体对象.本文采用 Lucene[②] 对知识库中的实体建立索引.

候选实体的生成过程包括 2 个步骤:首先,从给定文本中识别出所有的实体指称项.分词工具采用中科院计算所发布的 NLPIR 汉语分词系统,根据输出的词性标注结果进行实体识别.例如,词性 nr 表示人名、ns 表示地名、nt 表示机构名、nz 表示其他专用名词等.为提高实体识别结果的精确率和召回率,采用自定义规则和添加用户字典的方式,对 NLPIR 系统进行了修正.通过实体识别,可得到文本中所有待定实体指称项集合 $M=\{m_1,m_2,\cdots,m_n\}$.

然后,针对从当前语料中识别出的每个实体指称项 m_i,在本地知识库中进行索引查找,若知识库中存在与之同名的索引项,则将该索引项对应的所有实体对象作为其初步的候选实体集合 $N_i=\{n_{i1},n_{i2},\cdots\}$,其中,$n_{ik}$ 表示实体指称 m_i 所对应的第 k 个候选实体对象.最终,得到实体指称项集合 M 的初步候选实体对象集合 $N'=\{N_1,N_2,\cdots,N_n\}$.

2.2 实体相关图构造

本文提出的基于图的集成实体链接算法对每篇输入的文本语料构造一张实体相关图 $G=(V,E)$,以实现对文中实体的集成链接目标.其中,V 表示顶点集合,E 表示边集合.实体相关图的构造是该算法的关键环节,由于实体链接的最终决策是依据图的拓扑结构得出的,因此图的质量对于实体链接算法的性能具有关键性影响.本节从图的顶点集构造方法、边的构造方法以及利用外部知识完善图的结构等 3 个层面,完整介绍实体相关图的构造方法.

2.2.1 构造顶点集合

本文所使用的实体相关图为有向图,图中的顶点为从文本中识别得到的实体指称项及相关的候选实体集合(以后缀区分).在得到初步的候选实体集合 N' 之后,首先需要对其进行筛选,以确保在顶点集合中,从文本中识别得到的待链接实体指称项与其候选实体对象之间尽可能具有明确的语义相关性.当实体指称项在知识库中只有一个候选实体且该候选实体不带后缀标签时,则认为该候选实体即是实体指称项所指的实体对象.为简化描述,以下将从待处理文本语料中抽取得到的待链接实体指称项简称为待定实体,将从本地知识库中查询得到的同名实体对象简称为候选实体.

① http://keg.cs.tsinghua.edu.cn/project/ChineseKB

② https://lucene.apache.org

为确定待定实体和候选实体的相关性,首先需要设法确定一个待定实体的候选实体与该实体是否具有某种语义上的相关性,本文所采用的方法是计算待定实体所在的文本与其候选实体所对应的互动百科页面的余弦相似度,作为判定其语义相关性的依据[23].计算方法如下:首先采用 NLPIR 汉语分词系统对输入语料与候选实体的互动百科页面进分词和去除停用词等预处理,选择其中的名词和命名实体作为特征,分别构建特征向量.设待定实体 m_i 所在文本的特征向量为 $\boldsymbol{X}_i=(x_1,x_2,\cdots,x_n)$,其候选实体对象 n_{ij} 对应的互动百科页面的特征向量为 $\boldsymbol{Y}_j=(y_1,y_2,\cdots,y_n)$,余弦相似度的计算公式如下:

$$sim(m_i,n_{ij})=\frac{\boldsymbol{X}_i\cdot\boldsymbol{Y}_j}{\|\boldsymbol{X}_i\|\times\|\boldsymbol{Y}_j\|},\qquad(1)$$

其中,符号 $\boldsymbol{X}\cdot\boldsymbol{Y}$ 表示向量 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的内积,$\|\boldsymbol{X}\|$ 表示向量 $\boldsymbol{X}$ 的长度.若候选实体没有对应的百科页面,则假设其与实体指称项的余弦相似度为零.

在进行顶点(实体)筛选时,若计算得到的余弦相似度小于预先设定的阈值 λ,则认为该候选实体与待链接实体在语义上不相关,从初始候选实体集合 N' 中删除该候选实体节点.λ 的取值采用交叉验证方法经验地得到,本文的实验取 $\lambda=0.2$.

此外,研究表明,在百科语料中出现频率在前 40%的候选实体,可以覆盖约 90%的正确实体链接的目标对象[15].因此为提高算法执行效率,本文根据候选实体所在互动百科页面的被浏览次数对实体进行排序,选择排在前 40%的实体作为候选实体对象,由此可以进一步缩小候选实体的选择范围.经过上述 2 步筛选过程,得到最终的候选实体对象集合 N.

实体相关图的顶点集合 V 定义为:待链接实体指称项集合 M 和候选实体对象集合 N 的并集.

2.2.2　构造边集合

实体相关图的基本假设是位于同一文本中的实体之间通常存在语义上的相关性,利用这种位置上的语义相关性,能够提高实体链接的准确率[24].

当前流行的实体相关图构造方法是采用谷歌距离作为实体间的语义相关性测度,据此建立顶点之间的关联关系[16,25-26].然而,采用谷歌距离作为相关性测度的主要缺点是计算量较大,例如以维基百科作为实体知识来源的情况下,为了计算 2 个实体间的谷歌距离,需要分别统计这 2 个实体在所有百科页面上单独出现和共同出现的频率,对于大规模语料而言,这样的计算开销是不可接受的.因此,本文采用一种更为直观和高效的方法来构造边集合.

基本思路是基于本地知识库的实体关系拓扑结构来构造边的集合.具体实现方法是将实体间的关系划分为直接关系和间接关系,分别进行处理.

1) 直接关系.是指 2 个实体在本地知识库中存在关系边.如果实体相关图的顶点集合 V 中的 2 个实体顶点 v_i 和 v_j 在知识库中存在直接关系,则在这 2 个顶点间添加一条有向边,边的方向与知识库中 2 个实体间的关系方向保持一致.在当前的基于图的实体链接方法研究领域,主要采用这种方式来确定实体间的相关关系[14,19].

然而研究表明,仅考虑知识库中存在的直接关系是不够的,因为相对于复杂的实体关系而言,知识库中已有的显式知识通常是不足的,仅依赖这些直接关系进行建模,得到的实体相关图很可能无法正确反映实体关系网络,从而导致错误的实体链接结果[27].为减轻该问题的影响,本文提出一种间接关系定义,用于帮助完善实体相关图的结构.

2) 间接关系.若 2 个实体在本地知识库中均与一个以上的第三方实体存在直接关系,则称二者间存在间接关系.如果实体相关图的顶点集合 V 中的 2 个候选实体顶点 v_i 和 v_j 在本地知识库中存在间接关系,则在这 2 个顶点间添加一对有向边.

综上,对于给定的顶点集合 V,所构造的实体相关图的邻接矩阵的元素(有向边)的取值如式(2)所示:

$$\begin{cases}e_{ij}=1,\text{若 } v_i \text{ 和 } v_j \text{ 有直接关系}(i\neq j),\\ e_{ij}=1,e_{ji}=1,\text{若 } v_i \text{ 和 } v_j \text{ 有间接关系}(i\neq j),\\ e_{ij}=0,\text{其他},\end{cases}\qquad(2)$$

其中,$e_{ij}=1$ 表示从顶点 v_i 到 v_j 存在一条有向边 $\langle v_i,v_j\rangle$,$e_{ij}=0$ 表示从顶点 v_i 到 v_j 不存在直接路径.注意,当 $i=j$ 时,$e_{ij}=0$ 表示在实体相关图中不存在自环.同时需要注意的是实体指称项与其对应的所有候选实体间均存在一条有向边,方向由前者指向后者,而同一实体指称项对应的候选实体之间不存在路径.式(2)表明,本文在考查实体间的语义相关性时优先考虑实体间的直接关系,若实体间不存在直接关系,则进一步考虑实体间的间接关系.

2.2.3　增量证据挖掘

当前主流的实体链接方法大多基于本地知识库中现有的结构化知识,其中隐含的假设为知识库中所包含的实体知识结构是完整的,能为实体链接提供足够的背景知识.然而事实上现有的知识库技术

并不满足这一要求,其中的实体关系数据可能存在错误、滞后和缺失等问题,而且知识库的主要知识来源(如百科类网站等)也在不断地动态更新,因此不应仅仅依靠本地知识库作为实体链接的唯一知识来源[22].

为充分利用第三方知识库和百科类网站知识更新迅速的优点,本文提出一种增量证据挖掘方法,能够有效地利用外部知识对实体相关图进行修正和完善,从而进一步提高实体链接的准确率.作为示例,本文采用互动百科(http://www.baike.com/)网站作为增量证据挖掘的外部知识来源,但在实际应用中,该方法可以很方便地推广到同时使用多个外部数据源的场景,且对外部知识源没有特殊的要求.

增量证据挖掘方法简述如下.若实体相关图的顶点集合 V 中的 2 个候选实体顶点 v_i 和 v_j 在当前知识库中并无直接或间接关系,则以这 2 个顶点对应的实体对象为查询条件,利用互动百科网站提供的查询接口查找其对应的主页,进而通过模式匹配和规则过滤,确定这 2 个给定实体之间是否存在语义上的相关性.

若实体顶点 v_i 的互动百科页面包含指向实体顶点 v_j 的链接,则认为顶点 v_i 和 v_j 具有语义上的直接相关性,相应地在实体相关图上增加一条从 v_i 指向 v_j 的有向边.若实体顶点 v_i 的互动百科页面不包含指向实体顶点 v_j 的链接,则进一步提取出 v_i 和 v_j 所在页面的实体对象,并求出这 2 个实体对象集合的交集.对所得到的交集应用预定义语法规则进行过滤,去除其中的高频词(如地名和建筑物名)和代词.若经过滤后得到的实体交集 U 不为空,则认为实体 v_i 和 v_j 之间存在间接语义关联,相应地在实体相关图的顶点 v_i 和 v_j 之间增加一对有向边;若 U 为空集,则采用式(1)计算实体 v_i 和 v_j 所在页面的余弦相似度,若相似度大于 50%,则认为实体 v_i 和 v_j 之间存在间接语义关联,相应地在实体相关图的顶点 v_i 和 v_j 之间增加一对有向边.经过增量证据挖掘过程,得到最终的实体相关图,随后算法转入集成实体链接阶段,基于实体相关图对文本中提取得到的实体进行批量实体链接.算法 1 给出了实体相关图构造过程的算法伪代码框架.

算法 1. 实体相关图构造算法.

输入:实体指称项及其候选实体构成的顶点集合 $V=M\cup N$;

输出:实体相关图 G.

```
foreach m_i∈M and n_ij∈N_i do
  set edge(m_i,n_ij)=1;
foreach v_i∈ N_i and v_j∈N_i(i≠j) do
  if v_i 与 v_j 在本地知识库中有直接关系
    set edge(v_i,v_j)=1;
  else if v_i与 v_j 在本地知识库中有间接关系
    set edge(v_i,v_j)=1,edge(v_j,v_i)=1;
  else
    evidenceMining(v_i,v_j)}; /*增量证据挖掘*/
return 实体相关图 G.
```

2.3 集成实体链接

集成实体链接算法的基本思想是针对一篇文本中出现的多个实体,利用其内在的语义关系辅助消歧,实现批量实体链接.本文采用实体相关图实现基于上下文语义关系的集成实体链接,具体思路是对于图中的每个待定实体求出该实体与所有候选实体之间的语义一致性,从而将实体链接问题转化为候选实体的语义一致性排序问题,从中选择与待定实体语义一致性最高的候选实体作为最终的链接对象.方法流程简述如下:

1) 对每篇待处理文本语料构造实体相关图,并求出图中所有顶点的出入度.

2) 利用 2.2.1 节计算得到的待定实体和候选实体之间的余弦相似度以及候选实体的出入度,按式(3)求得待定实体和候选实体之间的语义一致性:

$$coherence(m_i,v_i)=\frac{1}{2}\left(\frac{D(v_i)}{\sum\limits_{v_j\in N_i}D(v_j)}+sim(m_i,v_i)\right),\tag{3}$$

其中,$coherence(m,v)$ 表示待定实体 m 和与其候选实体 v 之间的语义一致性;$D(v)$ 表示实体相关图 G 中顶点 v 的度,由于 G 为有向图,因此 $D(v)$ 的取值为顶点 v 的出度与入度之和. $\sum\limits_{v_j\in N_i}D(v_j)$ 表示关于给定的待定实体 m_i,求其所有候选实体顶点的度之和.从式(3)可以看出,本文提出的语义一致性定义包含 2 部分内容:$\frac{D(v_i)}{\sum\limits_{v_j\in N_i}D(v_j)}$ 表示候选实体 v_i 与当前文本中的实体间的上下文关联程度;$sim(m_i,v_i)$ 则表示候选实体 v_i 所在的百科页面与当前文本的语义相似度.由于相似度的取值范围均为[0,1],因此加权之后函数 $coherence$ 的取值范围也是[0,1].加权因子 1/2 表示对等加权,也可以考虑不对等加

权的情况,但初步实验结果表明,少量的权值修正对实体链接结果的影响不大,因此本文采用 1/2 作为加权因子,对加权因子选择和加权项的系统研究留作下一步工作.

3) 在计算出待定实体和其所有候选实体间的语义一致性参数之后,对候选实体按语义一致性参数值从大到小进行排序,选择其中排名最高的候选实体对象作为待定实体的链接对象,形式化表示为:

$$Link(m_i,v_k)=\arg\max_{v_k\in N_i}(coherence(m_i,v_k)),\quad(4)$$

其中,$Link(m_i,v_k)$表示将待定实体 m_i 链接到本地知识库中的候选实体 v_k 之上,v_k 为等式右侧的函数返回值.等式右侧的内容表示对于给定的待定实体 m_i,在其候选实体集合 N_i 中选出与待定实体的语义一致性最高的候选实体,作为实体指称项 m_i 的链接目标.

由于待定实体指称项所对应的实体对象可能不在本地知识库中,可能导致候选实体集合为空集(NIL)的情况出现[17,20].对于待定实体 m,采用如下规则判定其候选实体集合为空集:

1) 若以待定实体 m 为查询条件,在本地知识库中查询结果为空,则判定候选实体集合为空集.

2) 若针对待定实体 m 的查询结果非空,但所返回的候选实体与待定实体的余弦相似度均小于阈值 λ,则判定其候选实体集合为空集.

3) 若针对待定实体 m 的查询结果非空,但所返回的候选实体在实体相关图中均为孤立节点(仅与 m 相关联),即该节点与图中其他节点间不存在语义上的关联关系,则判定其候选实体集合为空集.

2.4　基于图的实体链接算法示例

接下来以一个例子完整演示 GCCEL 算法的实现细节.设给定语料为:"李娜的妈妈李艳萍是全国劳动模范".通过人工查询,了解到其中的实体指称项李娜为我国著名的网球运动员,即知识库中的"李娜(网球运动员)"是我们希望正确链接的实体对象.采用 GCCEL 算法,实体链接的实现过程如下.

首先通过实体抽取过程识别出待定实体指称项李娜和李艳萍,然后分别对其进行知识库查询,得到李娜的候选实体集合为{李娜(歌手)、李娜(网球运动员)、李娜(北京大学教授)};李艳萍的候选实体集合为{李艳萍(山西妈妈)、李艳萍(全国优秀共青团干部)}.采用式(1)进行相似度计算,得到每个候选实体所在的互动百科页面与给定语料的余弦相似度,计算结果如表 1 所示:

Table 1　Cosine Similarity Between the Undetermined Entitiy and the Candidate Entities

表 1　候选实体与待定实体的余弦相似度

Mention	Candidate Entities	Cos Similarity
Li Na	Li Na (Tennis Player)	0.886
	Li Na (Singer)	0.665
	Li Na (Professor of PKU)	0
Li Yanping	Li Yanping (Shanxi Mom)	0.5
	Li Yanping (League Cadre)	0.5

由于候选实体李娜(北京大学教授)的互动百科页面与待消歧语料的余弦相似度小于预先设定的阈值 $\lambda=0.2$,所以从候选实体集合中删除该实体.以保留下来的实体集合(包括待定实体和候选实体)为顶点集,采用 2.2 节介绍的边集合构造方法和增量证据挖掘方法构造实体相关图,如图 2 所示.其中黑色顶点表示待定实体,白色顶点表示候选实体.

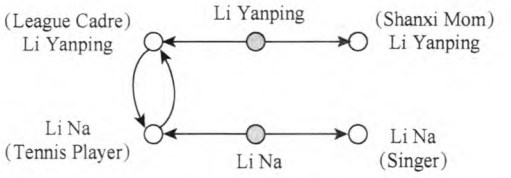


Fig. 2　Referent graph of the example corpus.

图 2　示例文本语料的实体相关图

根据图 2,可以计算出图中所有候选实体的出入度,然后依据式(3)求得待定实体和候选实体之间的语义一致性,结果如表 2 所示:

Table 2　Semantical Coherence between the Undetermined Entitiy and the Candidate Entities

表 2　待定实体和候选实体之间的语义一致性

Candidate Entities	Indegree	Outdegree	Consistency
Li Na(Tennis Player)	2	1	0.818
Li Na(Singer)	1	0	0.478
Li Yanping(Shanxi Mom)	1	0	0.375
Li Yanping(League Cadre)	2	1	0.625

最后,按照 2.3 节介绍的实体链接算法,根据语义一致性对候选实体进行排序,选择一致性最高的候选实体作为待定实体的链接对象.在本例中,将待定实体李娜链接到知识库中的李娜(网球运动员),将待定实体李艳萍链接到知识库中的李艳萍(全国优秀共青团干部).至此,对于给定文本语料,批量地完成了将给定文本中的(歧义)实体指称项正确链接到知识库中的对应实体对象上的实体链接任务.

3 实验结果与讨论

3.1 实验数据

为验证所提出的集成实体链接算法的有效性，本文采用3组可公开获得的语料进行测试.

第1组语料是哈尔滨工业大学智能技术与自然语言处理研究室公开发布的搜狗人名消歧语料①. 该语料根据国内常用的人名，选取其中相关新闻报道最多的12个人名，对11876篇文档进行了人工标注. 本文从中随机抽取120篇文档进行测试，在预处理阶段去除了文档中的标注信息，仅保留纯文本. 经人工统计，120篇测试文档中共包含1170个命名实体，其中在知识库中存在对应实体对象的实体973个(InKB类型)，NIL类型实体的个数为197个.

为验证GCCEL算法对短文本的实体链接效果，本文进一步采用2组公开数据进行了算法性能测试. 其中，第2组语料是从新浪娱乐新闻网和凤凰娱乐新闻网上随机地采集得到的180条短新闻文本(平均字符长度为107字). 随机选取其中的30条作为训练样本，剩余的150条作为测试样本. 经统计，该训练样本和测试样本所包含的实体总数分别为125个和723个，其中InKB类型实体个数分别为99个和650个，NIL类型实体的个数分别为26个和73个.

第3组语料是NLP&CC国际会议2013年公开发布的微博实体链接评测语料②. 该数据集包含779篇微博文档，每篇微博文档的长度不超过150字. 经人工统计，779篇微博中共包含实体总数1232个，其中InKB类型实体个数为843个，NIL类型实体个数为389个.

需要说明的是，为了模拟开放域环境下的实体链接任务场景(即训练样本对真实数据的覆盖率很低)，本文仅采用30条短新闻文本作为训练样本用于模型参数学习，因此在3组测试数据集上所使用的模型是完全相同的. 实验数据的统计情况如表3所示.

Table 3 Statistics of the Corpus

表3 训练样本与测试语料一览

Corpus	Entities	InKB Entities	NIL
News train corpus	125	99	26
Sougou NED corpus	1170	973	197
News test corpus	723	650	73
NLP&CC corpus	1232	843	389

3.2 实验方法

为了验证GCCEL算法的有效性，本文从近年的相关工作中分别选择了2类具有代表性和较高学术影响力的方法作为实验比较的对象，分别介绍如下.

WTCosSim算法是基于实体上下文相似度计算的经典实体链接算法[12]. 该算法基于维基百科上的实体知识进行实体消歧，方法是利用文本中的命名实体和上下文关键词构造特征向量，并对查询维基百科得到的若干候选实体所在页面进行向量化处理，据此计算从文本中抽取得到的实体指称项的上下文与候选实体的维基百科页面上下文之间的余弦相似度，选择相似度最高的候选实体作为目标链接对象.

Babelfy③是一个基于图的实体链接软件[21]. 该软件基于开源的百科字典BabelNet④构建，因此支持多语种(包括英语、汉语、俄语等)实体消歧任务. 其实体链接过程包含3个步骤：实体识别、候选实体选择和实体消歧. 基本思想是利用实体(节点)在BabelNet语义网络中的三角形关系计算每个实体节点的结构权值，据此构造转移矩阵，然后利用带重启的随机游走算法得到实体间的语义相关性(称为语义签名). 对于输入的待处理文本，Babelfy首先基于语义签名构造其实体语义关系图，然后通过抽取密集子图对歧义实体进行消歧. 该方法在基准数据集上的表现非常优异，在KORE50和CoNLL数据集上的准确率分别高达71.5%和82.1%. 该方法的主要局限性在于性能完全依赖于BabelNet知识库的知识规模.

除上述2类经典算法外，本文还选择在NLP&CC 2013实体链接竞赛中获奖的CASIA_EL算法进行比较. 该算法也属于基于实体上下文相似度计算的传统实体链接算法[13]，与WTCoSim的区别在于，针对因实体上下文信息不足而导致的余弦相似度区分度不足的问题，CASIA_EL采用文本中实体指称项上下文词的维基百科页面作为外部知识源，对输入文本的特征向量进行了扩展，然后再进行实体

① http://www.datatang.com/data/44022

② http://www.datatang.com/data/44052

③ http://babelfy.org/index

④ http://babelnet.org

相似度计算,由此提高了实体相似度的计算精度,但同时也导致了算法模型计算复杂度的大幅提高.

除与相关工作进行比较外,本文还针对所提出的GCCEL算法设计了一个对比实验,用于演示和讨论本文提出的增量证据挖掘方法对于系统性能的影响.为简化描述,将参与对比的算法称为Baseline模型,该模型与GCCEL算法模型几乎完全一致,差别仅在于Baseline中不包含增量证据挖掘过程,在构造候选实体关系图时,完全依赖于知识库现有的知识结构.

3.3 评估方法

为客观评价实验结果,对每组实验数据,分别记录所采用的实验方法在该数据集上的准确率(*Accuracy*)、精确率(*Precision*)、召回率(*Recall*)、并计算出相应的F_1值[16,20].

实验结果的取得方法如下.首先,对2组测试数据集中的歧义实体进行人工消歧,即人工地将其链接到本地知识库中正确的实体对象上,对于不在知识库中的实体,将其标记为NIL.由此得到评估实验结果所需的基本事实集合T,即该语料中出现的实体指称项集合.以T_1表示测试语料中由人工将其正确链接到本地知识库中对应实体对象上的实体指称项集合,T_2表示测试语料中相应实体对象不在本地知识库中的实体指称项集合(即标记为NIL的实体指称项集合),则有:$T=T_1\cup T_2$.

然后,对本文提出的GCCEL算法进行测试,记录算法GCCEL输出的实体链接结果集合,以符号S表示.以S_1表示输出结果中链接到本地知识库中实体对象上的实体指称项集合,S_2表示输出结果中的实体对象不在本地知识库中的实体指称项集合,则有:$S=S_1\cup S_2$.需要说明的是,S中包含的实体指称项个数与T相同,二者的区别在于:T中的实体链接结果是经人工验证过的,可以视为基本事实;S中可能包含错误的实体链接结果,即S_1中可能包含T_2中的实体指称项,S_2中可能包含T_1中的实体指称项,这2种情况均与基本事实T产生冲突.

通过将GCCEL算法的输出结果S与基本事实T进行对比,可以根据式(5)计算得到算法GCCEL在该数据集上的实体链接准确率:

$$Accuracy=\frac{|S_1\cap T_1|+|S_2\cap T_2|}{|T|}\times 100\%,\quad(5)$$

其中,$|T|$表示集合T中的元素个数,$|S_1\cap T_1|$表示算法输出结果中正确链接到本地知识库中对应的实体对象上的实体指称项个数,$|S_2\cap T_2|$则表示被算法正确判定为NIL的实体指称项个数.从式(5)可以看出,准确率指标综合考虑了算法对于在本地库中和不在本地库中的实体指称项的链接效果,是对实体链接算法综合性能的评价指标.

通过统计对应的实体对象在本地知识库中的实体指称项集合S_1和算法输出的链接到本地知识库中实体对象上的实体指称项集合T_1的数目,可以根据式(6)与式(7)进一步计算得到算法GCCEL在该数据集上的精确率和召回率:

$$Precision=\frac{|S_1\cap T_1|}{|S_1|}\times 100\%,\quad(6)$$

$$Recall=\frac{|S_1\cap T_1|}{|T_1|}\times 100\%.\quad(7)$$

精确率的含义是:GCCEL算法正确链接到知识库的实体数量,占GCCEL算法输出的实体链接总数的百分比.该指标反映的是GCCEL算法的精确性,精确率越高,表明算法对于已经存在于本地库中的实体执行消歧操作时正确结果的比率越高.召回率的含义是:GCCEL算法正确链接到知识库的实体数量,占测试集中能够准确链接到知识库中的实体总数的百分比.该指标反映的是GCCEL算法的查全率,召回率越高,表明算法输出结果中对于本地库中已有实体而言,遗漏正确结果的可能性越低.

精确率和召回率是一对具有内在矛盾的指标,通常情况下,精确率的提高意味着召回率的降低,在实际应用中人们通常会在这2个指标间进行折衷,根据F_1值来客观地评估算法的实际性能,公式如下:

$$F_1=\frac{2\times Precision\times Recall}{Precision+Recall}.\quad(8)$$

从式(8)可以看出,F_1值受算法精确度和召回率的共同影响,当二者均趋近于1时,F_1值也趋近于最大值1.显然,F_1值越大,说明算法对于本地知识库中已有实体的消歧性能越好.由于实体链接任务的目标本身是针对本地知识库进行实体消歧与匹配,因此F_1指标是衡量实体链接算法性能的最关键指标[20].准确率指标则可以被视为重要的参考指标,按照惯例,对于实体链接结果中不在当前知识库中的实体指称项集合,通常在实体链接操作结束后,会对$S_2\cap T_2$集合中的实体进行聚类,然后将其加入到现有知识库中.此后在执行新的实体链接操作时,这部分(当前的$S_2\cap T_2$集合中的)实体将成为本地知识库中的成员(即可能出现在新的T_1集合中),因此在评估算法性能时不能忽略准确率的影响.

3.4 实验结果与讨论

如3.1节所述,为模拟开放域环境下的实体链接

任务场景，即训练样本对真实数据的覆盖率很低的情况，本文仅采用30条短新闻文本作为训练样本用于模型参数学习. GCCEL算法的关键参数是实体上下文相似性的判定阈值λ，通过在训练语料上进行参数调整，得到GCCEL的性能指标F_1值与λ的关系如图3所示. 从图3可以看出，当λ=0.2时，GCCEL算法的F_1值达到最优；而当λ取值过小或过大时，算法的性能均会受到影响. 这是因为λ取值过小会造成较多似是而非的噪音数据被判定为候选实体，导致算法准确率降低；而当λ取值过大时，会造成对候选实体的筛选结果过于严格，导致算法召回率下降.

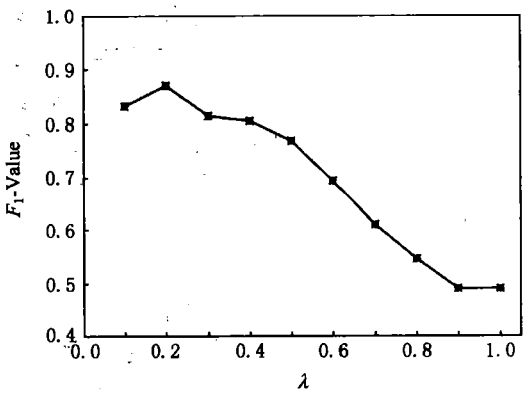


Fig. 3 F_1-values under different λ on training corpus.

图3 F_1值与参数λ在训练集上的关系

通过上述实验，采用λ=0.2作为实体相似性判定阈值，得到GCCEL在3组测试语料上的实验结果，分别如表4～6所示. 其中，表4和表5给出的是GCCEL算法与2类代表性算法和Baseline模型的实验结果对比情况，表6则给出了GCCEL算法与NLP&CC 2013竞赛优胜算法CASIA_EL的比较结果.

由实验结果可知，本文提出的GCCEL算法在短新闻语料、搜狗人名消歧语料和NLP&CC微博评测语料上的F_1值分别为88.08%，87.91%和88.53%，准确率分别为86.92%，87.50%和88.47%，均优于近期的相关工作.

Table 4 Experimental Results on Short News Corpus

表4 在短新闻数据上的实验结果 %

System	*Precision*	*Recall*	F_1	*Accuracy*
WTCosSim	78.17	81.48	79.79	78.50
Baseline	44.16	46.03	45.08	49.53
Babelfy	68.89	72.42	69.75	72.42
GCCEL	86.29	89.95	88.08	86.92

Table 5 Experimental Results on Sogou NED Corpus

表5 在搜狗人名消歧语料上的实验结果 %

System	*Precision*	*Recall*	F_1	*Accuracy*
WTCosSim	73.54	70.68	72.09	74.29
Baseline	40.35	38.79	39.63	47.14
Babelfy	62.71	66.07	64.34	66.07
GCCEL	86.21	89.69	87.91	87.50

Table 6 Experimental Results on NLP&CC 2013 Corpus

表6 在NLP&CC 2013评测语料上的实验结果 %

System	*Precision*	*Recall*	F_1	*Accuracy*
WTCosSim	79.46	81.03	80.23	80.57
CASIA_EL	86.62	84.56	85.58	88.50
Babelfy	78.06	70.42	75.93	70.42
GCCEL	89.67	87.42	88.53	88.47

与基于上下文相似度的代表性算法WTCosSim相比，GCCEL算法在3组语料上的F_1值分别提高了10.39%，21.94%和10.34%，准确率分别提高了10.73%，17.78%和9.81%. 在NLP&CC微博评测语料上，与相关工作中表现最好的CASIA_EL相比，GCCEL算法的精度提高了3.52%，召回率提高了3.38%，F_1值提高了3.45%，准确率与之相当. 该实验结果表明，GCCEL算法的性能表现一致且显著地优于经典的上下文相似度算法，但经过改良后的上下文相似度算法（如CASIA_EL）可以在性能上接近基于图模型的集成链接算法. 然而，考虑到算法模型训练所需的数据集规模，GCCEL算法对训练集样本规模和样本代表性的依赖性相对较低，因而与其他2种算法相比，算法推广性更好，这也从另外一个侧面显示出集成实体链接方法的性能优越性.

与集成链接方法的代表性工作Babelfy相比，GCCEL算法在3组语料上的F_1值分别提高了26.28%，36.63%和16.59%，准确率分别提高了20.02%，32.44%和25.63%. 分析其主要原因在于Babelfy算法的性能主要依赖于BabelNet知识库的知识含量，对于当前不在知识库中的实体指称项，该算法仅为其分配一个抽象实体（与实体指称项同名，但没有具体信息）作为链接对象，导致实体链接的精度降低. 此外，Babelfy算法没有考虑到实体间的间接关系，也没有提供增量证据挖掘机制，因此与GCCEL算法相比，其算法性能进一步恶化. 该实验结果表明，本文提出的增量证据挖掘方法和语义

关联分析方法有助于显著提高集成实体链接算法的整体性能,从而更有效地发挥集成链接的优势.该结果同时也为本文提出的增量证据挖掘方法的有效性提供了实验证据.

为进一步评估本文提出的增量证据挖掘方法的有效性,我们对从 GCCEL 算法中去除了增量挖掘机制的 Baseline 算法结果进行观察.与 Baseline 相比,GCCEL 算法在短新闻语料和搜狗人名消歧语料上的 F_1 值分别提高了 95.39%和 121.83%,准确率分别提高了 75.49%和 85.62%.该实验结果表明,通过外部证据挖掘过程,能在一定程度上弥补因本地知识库实体知识不足对实体链接算法性能所造成的负面影响,从而显著提高实体消歧的准确率.由于知识库的知识容量不足是目前制约知识库应用的核心关键问题(这也是 TAC 设定实体链接任务的主要原因之一),因此本文提出的增量证据挖掘方法和相关实验证据对于帮助解决在知识库知识容量有限前提下的实体链接问题有一定的积极意义和参考价值.

最后,通过对 GCCEL 算法输出结果中的错误部分进行人工比对和分析,归纳出导致 GCCEL 算法出错的主要原因如下:

第 1 类错误是由于实体识别错误所导致的实体链接错误.例如,对于"110 跨栏运动员刘翔若选择退役……"这句话,由于本文所采用的分词方法将句中的实体指称项"刘翔"错误地识别为"刘翔若",导致本地知识库查询时返回结果 NIL,未能将其正确链接到本地知识库中.

第 2 类错误是由于本地知识库的实体知识不足而导致的实体链接错误.例如,对于短新闻语料中的"李娜是北京奥运会跳水冠军得主",其中的实体对象李娜(跳水运动员)不在本地知识库中,但由于职业背景的相似性,GCCEL 错误地将其关联到知识库中已有的李娜(网球运动员)这一实体指称项上.

第 3 类错误是由于本文所使用的增量证据挖掘算法本身不够完善而导致的实体链接错误.例如,对于短新闻语料"在东方歌舞团,王彤认识了后来成为知名电视剧导演的刘江".由于刘江和王彤这 2 个实体在本地库中没有语义上的关联,所以 GCCEL 会调用增量证据挖掘过程,通过互动百科查找二者关系.结果在导演刘江的主页上发现了实体王彤的超链接,通过超链接访问该王彤的主页,确认其身份是摄影师,因而 GCCEL 将上述语句中提取到的实体指称项,链接到知识库中的实体对象王彤(摄影师)名下.然而,这与短新闻语料中的实体王彤身份不同.

通过对上述 3 类错误进行总结,得出如下认识:1)通过改进实体识别算法,或采用扩充单词表的方式,可以减轻和消除第 1 类错误的影响;2)第 2 类错误是当前实体链接研究面临的主要问题,主要受当前知识库的完善程度制约,随着知识库的完善,此类问题有望逐渐得到解决;3)第 3 类错误虽然发生的概率较低,但一旦发生却很难及时察觉和纠正,是本文下一步工作的重点.

4　结束语

本文提出一种基于图和增量证据挖掘的中文集成实体链接方法,该方法融合了上下文相似度、实体流行度、实体相关度等因素,并在该模型的基础上搭建了原型系统 GCCEL.对于任意给定的文本,GCCEL 综合考虑了知识库中存在的实体间的结构化关系(包括直接关系和间接关系)和从外部知识源获取的增量证据,据此构建实体相关图,然后在实体相关图的基础上利用图算法实现对文本中多个歧义实体的集成链接.在搜狗人名消歧语料、新闻语料和 NLP&CC 微博评测语料上分别取得了 87.91%,88.08%和 88.53%的 F_1 值以及 87.50%,86.92%和 88.47%的准确率,算法综合性能显著优于本领域的代表性工作.

论文的主要贡献包括如下 2 个方面:1)通过实验证明了基于图的集成实体链接方法在性能上一致地优于当前主流的基于上下文相似度的集成实体链接方法;2)实验表明本文提出的增量证据挖掘方法能够有效地弥补本地知识库的知识结构不完善的问题,显著提高基于图的集成实体链接方法的整体性能.本文提出的 GCCEL 算法具有良好的扩展性和适应性,上述成果为进一步开展大规模知识库扩容工作提供了有益的思路和方法借鉴.

在后续工作中,我们将主要从如下 2 方面着手对 GCCEL 算法进行改进.首先,改进现有的增量证据挖掘算法,通过增加深度语义分析机制和实体识别过滤机制,提高实体识别的准确性.其次,积极探索外部知识来源的融合方法,不断丰富完善本地知识库的规模和知识结构,研究并利用不同类型的外部知识,以进一步提高实体链接操作的准确率和召回率.

参 考 文 献

[1] Huai Baoxing, Bao Teng Fei, Zhu Hengshu, et al. Topic modeling approach to named entity linking [J]. Journal of Software, 2014, 9(14): 2076-2087 (in Chinese)
(怀宝兴, 宝腾飞, 祝恒书. 一种基于概率主题模型的命名实体链接方法[J]. 软件学报, 2014, 9(14): 2076-2087)

[2] Ling Xiao, Weld D S. Fine-grained entity recognition [C] //Proc of the 26th Conf on Association for the Advancement of Artificial Intelligence(AAAI'12). Menlo Park, CA: AAAI Press, 2012: 94-100

[3] Wu Fei, Weld D S. Open information extraction using Wikipedia [C] //Proc of the 48th Annual Meeting of the Association for Computational Linguistics (ACL'10). Stroudsburg, PA: ACL, 2010: 118-127

[4] Wu Fei, Weld D S. Autonomously semantifying Wikipedia [C] //Proc of the 16th ACM Conf on Information and Knowledge Management (CIKM'07). New York: ACM, 2007: 41-50

[5] Heng Ji, Ralph G. Knowledge base population: Successful approaches and challenges [C] //Proc of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL'11). Stroudsburg, PA: ACL, 2011: 1148-1158

[6] Mark D, Paul M, Rao D, et. al. Entity disambiguation for knowledge base population [C] //Proc of the 23rd Int Conf on Computational Linguistic(COLING'10). Stroudsburg, PA: ACL, 2010: 277-285

[7] Shen Wei, Wang Jianyong, Han Jiawei. Entity Linking with a knowledge base: Issues, techniques, and solutions [J]. IEEE Trans on Knowledge and Data Engineering, 2015, 27 (2): 443-460

[8] Li Xuansong, Stephanie S, Heng Ji, et al. Linguistic resources for entity linking evaluation: From monolingual to cross-lingual [C] //Proc of the 8th Int Conf on Language Resources and Evaluation(LREC'12). New York: European Language Resources Association, 2012: 3098-3105

[9] Bunescu R, Pasca M. Using encyclopedic knowledge for named entity disambiguation [C] //Proc of the 11th Conf of the European Chapter of the Association for Computational Linguistics(EACL'06). Stroudsburg, PA: ACL, 2006: 9-16

[10] Silviu C. Large-scale named entity disambiguation based on Wikipedia data [C] //Proc of 2007 Joint Conf on Empirical Methods in Natural Language Processing and Computational Natural Language Learning(EMNLP'07). Stroudsburg, PA: ACL, 2007: 708-716

[11] Yang Zhizhuo, Huang Heyan. WSD method based on heterogeneous relation graph [J]. Journal of Computer Research and Development, 2013, 50(2): 437-444 (in Chinese)
(杨陟卓, 黄河燕. 基于异构关系网络图的词义消歧研究[J]. 计算机研究与发展, 2013, 50(2): 437-444)

[12] Nguyen H T, Cao T H. Exploring Wikipedia and text features for named entity disambiguation [C] //Proc of the 2nd Int Conf Intelligent Information and Database Systems. Berlin: Springer, 2010: 24-26

[13] Zeng Yi, Wang Dongsheng, Zhang Tielin, et al. Linking entities in short texts based on a Chinese semantic knowledge base [C] //Proc of the 2nd CCF Conf on Natural Language Processing and Chinese Computing. Berlin: Springer, 2013: 266-276

[14] Zhang Tao, Liu Kang, Zhao Jun. A graph-based similarity measure between Wikipedia concepts and its application in entity linking system [J]. Journal of Chinese Information Processing, 2015, 29(2): 58-67 (in Chinese)
(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实体链接系统中的应用[J]. 中文信息学报, 2015, 29(2): 58-67)

[15] Zuo Zhe, Gjergji K, Toni G, et al. BEL: Bagging for entity linking [C] //Proc of the 25th Int Conf on Computational Linguistics: Technical Papers(COLING'14). Stroudsburg, PA: ACL, 2014: 2075-2086

[16] Han Xianpei, Sun Le, Zhao Jun. Collective entity linking in Web text: A graph-based method [C] //Proc of the 34th Int ACM Conf on Research and Development in Information Retrieval(SIGIR'11). New York: ACM, 2011: 765-774

[17] Shen Wei, Wang Jianyong, Luo Ping, et al. Linking named entities with knowledge base via semantic knowledge [C] //Proc of the 21st Annual Conf on World Wide Web (WWW'12). New York: ACM, 2012: 449-458

[18] Johannes H, Mohamed A Y, Bordino I, et al. Robust disambiguation of named entities in text [C] //Proc of the Conf on Empirical Methods in Natural Language Processing (EMNLP'11). Stroudsburg, PA: ACL, 2011: 782-792

[19] Ayman A, Robert G. Graph ranking for collective named entity disambiguation [C] //Proc of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL'14). Stroudsburg, PA: ACL, 2014: 75-80

[20] Guo Zhaochen, Barbosa D. Robust entity linking via random walks [C] //Proc of the 23rd ACM Int Conf on Information and Knowledge Management(CIKM'14). New York: ACM, 2014: 499-508

[21] Andrea M, Alessandro R, Roberto N. Entity linking meets word sense disambiguation: A unified approach [C] //Proc of the 2014 Transactions of the Association for Computational Linguistics(ACL'14). Stroudsburg, PA: ACL, 2014: 231-244

[22] Li Yang, Wang Chi, Han Fangqiu, et al. Mining evidences for named entity disambiguation [C] //Proc of the 19th ACM Int Conf on Knowledge Discovery and Data Mining (SIGKDD'13). New York: ACM, 2013: 1070-1078

[23] Zhang Wei, Su Jian, Wang Wenting, et al. Entity linking leveraging automatically generated annotation [C] //Proc of the 23rd Int Conf on Computational Linguistic(COLING'10). Stroudsburg, PA: ACL, 2010: 1290-1298

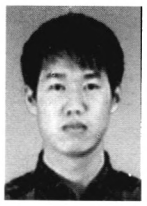
[24] Gentile A L, Zhang Ziqi, Xia Lei, et al. Semantics relatedness approach for named entity disambiguation [C] // Proc of the 6th Italian Research Conf on Digital Libraries. Berlin: Springer, 2010: 137-148

[25] Milne D, Witten H I. Learning to link with Wikipedia [C] // Proc of the 17th ACM Conf on Information and Knowledge Management(CIKM'08). New York: ACM, 2008: 509-518

[26] Johannes H, Stephan S, Nguyen D B, et. al. KORE: Keyphrase overlap relatedness for entity disambiguation [C] //Proc of the 21st ACM Int Conf on Information and Knowledge Management (CIKM'12). New York: ACM, 2012: 545-554

[27] Eneko A, Ander B, Aitor S. Studying the Wikipedia hyperlink graph for relatedness and disambiguation [DB/OL]. Ithaca: ArXiv, [2015-05-12]. http://arxiv.org/pdf/1503.01655v2.pdf

Liu Qiao, born in 1974. PhD and Associate professor. Member of China Computer Federation. His research interests include machine learning and data mining, natural language processing, and social network analysis.



Zhong Yun, born in 1990. Mater. His research interests include entity linking techniques, natural language processing and machine learning (zhongyunuestc@gmail.com).

Li Yang, born in 1990. Master. Student member of China Computer Federation. His research interests include knowledge graph, machine learning and natural language processing (kedashqs@163.com).

Liu Yao, born in 1978. PhD and lecturer. Member of China Computer Federation. Her research interests include social network analysis, machine learning, data mining, and network measurement (liuyao@uestc.edu.cn).

Qin Zhiguang, born in 1956. PhD and professor. Senior member of China Computer Federation. His research interests include information security, social network analysis, and mobile computing (qinzg@uestc.edu.cn).