计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2016.20160192
53(8):1696-1708,2016

基于语义一致性的集成实体链接算法

刘 峤 钟 云 刘 瑶 吴祖峰 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)

Consistent Collective Entity Linking Algorithm

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

Abstract The goal of entity linking is to link entity mentions in the document to their corresponding entity in a knowledge base. The prevalent approaches can be divided into two categories: the similarity-based approaches and the graph-based collective approaches. Each of them has some pros and cons. The similarity-based approaches are good at distinguish entities from the semantic perspective, but usually suffer from the disadvantage of ignoring relationship between entities; while the graph-based approaches can make better use of the relation between entities, but usually suffer from bad discrimination on similar entities. In this work, we present a consistent collective entity linking algorithm that can take full advantage of the structured relationship between entities contained in the knowledge base, to improve the discrimination capability of the proposed algorithm on similar entities. We extensively evaluate the performance of our method on two public datasets, and the experimental results show that our method can be effective at promoting the precision and recall of the entity linking results. The overall performance of the proposed algorithm significantly outperform other state-of-the-art algorithms.

Key words collective entity linking; information retrieval; knowledge base population; personalized PageRank; semantic correlation

摘 要 实体链接任务的目标是将从文本中抽取得到的实体指称项正确地链接到知识库中的对应实体对象上.当前主流的实体链接算法大致可分为2类:基于上下文相似度的实体链接算法和基于图的集成实体链接算法.这2类算法各自存在一些优点和不足.前者有利于从上下文语义的角度对实体进行区分,但难以充分利用知识库中已有的知识体系辅助决策;后者能够更好地利用知识库中实体间的语义关联关系,但在上下文信息不充分的情况下,较难区分概念相近的实体.提出一种基于语义一致性的集成实体链接算法,该算法能够更好地利用知识库中实体间的结构化语义关系,帮助提高算法对概念相似实体的区分度,实验结果表明:该算法能够有效提高实体链接结果的准确率和召回率,性能显著优于当前的主流算法,在对长、短文本的实体链接任务中性能表现稳定,具有良好的适应性和可推广性.

收稿日期:2016-03-21;**修回日期**:2016-05-26

基金项目:国家自然科学基金项目(61133016,61272527,61202445);国家自然科学基金青年项目(61502087);中央高校基本科研业务费专项资金项目(ZYGX2014J066)

This work was supported by the National Natural Science Foundation of China (61133016, 61272527, 61202445), the National Natural Science Foundation for Young Scholar of China (61502087), and the Fundamental Research Funds for the Central Universities (ZYGX2014J066).

关键词　集成实体链接;信息抽取;知识库扩容;个性化 PageRank;语义相关性

中图法分类号　TP391

作为互联网时代的标志性技术,Web 技术正处在快速发展和变革当中,从网页的链接(Web 1.0)到数据的链接(linked data),再到知识图谱(knowledge graph)技术,语义 Web 正在逐渐走向成熟[1].知识图谱是一种图结构的知识库,其中存储的知识元素以〈实体,关系,实体〉三元组的形式表达,也称为事实(facts)[2].知识图谱是目前智能互联网应用研究领域主要采用的知识库形式,本文主要研究知识图谱上的实体链接问题.

实体链接是知识库扩容(knowledge base population, KBP)研究领域关注的核心问题之一[3].开放域信息抽取技术的快速发展为知识库扩容带来了巨大的发展机遇,同时也带来许多挑战,其中一项关键挑战就是实体链接问题[3].实体链接任务的基本目标是将从文本中抽取得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[4].然而通过开放域信息抽取技术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),为将其正确融入到知识库中,必须首先解决实体链接问题.通过实体链接技术,可以消除知识元素在概念上的歧义,剔除冗余和错误的知识元素,从而确保知识的质量[3].

对实体链接问题的研究,当前面临的主要挑战是解决实体指称项的歧义性和多样性问题[5].所谓歧义性是指相同的实体指称项在不同的上下文环境中可能指代不同的实体对象.所谓多样性是指一个给定的实体对象可以与多个不同的实体指称项形成对应关系(如该实体的别名或缩写等).例如在如下 2 组语句样例中[6],就同时存在上述问题:

样例 1. After his departure from Buffalo, Saban returned to coach college football teams including Miami, Army and UCF.

样例 2. Saban, previously a head coach of NFL's Miami, is now coaching Crimson Tide. His achievements include leading LSU to the BCS National Championship once and Alabama three times.

在上述 2 组语境中出现的实体指称项 Saban 具有歧义性,虽然两者都是美国大学橄榄球队的教练,但前者为 Lou Saban,后者则是 Nick Saban,分别对应知识库中不同的实体对象.此外,样例 2 中的 Crimson Tide 和 Alabama 是 2 个不同的实体指称项,但实际上它们指代的是同一实体对象,即 Alabama Crimson Tide 橄榄球队.如何将上述样例中的实体指称项 Saban 和 Crimson Tide 正确地链接到知识库中的实体对象上,即解决"实体消歧"和"共指消解"问题,是实体链接研究领域当前关注的主要问题.

关于实体链接方法当前主要有 2 种研究思路[7]:1)根据文本中每个单一实体的上下文信息,通过与知识库中实体对象的已知上下文信息进行比较,选出上下文相似度高的实体对象进行链接[8-9];2)针对文本中出现的实体指称项集合,结合知识库中的已有知识构造实体相关图,批量地将其链接到知识库中.由于前者未能有效利用文本中的共现实体之间存在的天然语义相关性,因此后一种思路在近年来受到了学术界的重视,被称为集成实体链接方法[10-11].

本文提出一种新的集成实体链接算法,称为基于语义一致性的集成实体链接算法(consistent collective entity linking algorithm, CCEL).与相关工作相比,本文的贡献主要体现在以下 3 点:

1)提出了一种新的实体相关图构造方法,能够充分利用知识库中已有的知识,补全候选实体的关联关系,提高实体相关度计算结果的准确性.

2)设计了一种候选实体与输入文本语义相关性的计算方法,能有效降低错误候选实体带来的噪音影响,提高算法对概念相近的实体的区分度.

3)提出了一种基于语义一致性的集成实体链接算法原型,实验表明该算法的准确率和召回率均显著优于当前主流的相关工作.

1　相关工作

早期的实体链接研究思想是基于实体的上下文相似度进行链接消歧,即通过计算实体指称项所在文本与其相应候选实体的上下文相似度,选择相似度最大的候选实体作为目标链接对象.基本思路是首先在知识库中查找出与该指称项同名的所有实体对象,构成候选实体集合,然后使用词袋模型计算待处理文本和候选实体所在的维基百科页面之间的相似度,选择相似度最高的候选实体作为链接对象[8].研究表明,除文本内容之外,实体的类别相关性和百科页面的锚文本、重定向页面等结构信息,对于提高实体链接算法的准确性有较大帮助[9,12].基于上下

文相似度的链接算法其准确性容易受到上下文信息不足的影响,当前主要的解决方案是借助第三方知识库对候选实体的特征进行扩展,以提高候选实体之间的区分度[13],或者改为采用其他的测度(如维基概念)进行相似度计算[14].基于上下文相似度的实体链接方法的主要缺点是忽视了共现实体间天然的语义相关性,而这种语义相关性对于区分有歧义的实体通常具有帮助[15].

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当中.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目标链接对象[16];Greg 等人基于结构化条件随机场算法提出了一个实体分析的联合模型,可同时用于实体识别、实体消歧和共指消解;在 ACE2005 和 OntoNotes 等基准数据上进行了实验,取得了不错的实验效果[17].统计机器学习方法的主要缺点是算法的性能效果受制于训练语料的质量和范围,方法的移植性较差.因此,为了克服基于上下文相似度方法和统计机器学习方法的不足,Kulkarni 等学者提出了集成实体链接的算法设计思想[10].

集成实体链接方法一次性批量处理文本中所有实体指称项的链接问题,基本的思路是根据候选实体维基页面间的指向关系,建立候选实体间的语义相关图进行推理[10].常用的推理方法是采用随机游走模型,得到候选实体的排序,选择排名最高的实体作为链接对象[11,18].此外,也有学者将批量实体链接的推理问题视为图的搜索问题,通过在实体相关图上搜索包含所有实体指称项和其相应候选实体 s 的最小密集子图实现批量实体链接[19].与基于上下文相似度的链接方法类似,集成实体链接方法的性能同样易受上下文信息的影响[20].为解决该问题,Ferragina 等人通过引入概率化链接的思想,提出了一个面向短文本的集成实体链接算法 Tagme[21].

从算法性能的角度来看,Shen 等人提出的 LINDEN 算法在构造语义相关图时,综合考虑了实体所对应的维基页面间的关联关系和实体间的语义相似性,在 YAGO 知识库的支持下,LINDEN 算法在 TAC2009 数据集上实现了高达 84.32% 的实体链接准确率[22].在此基础上,Alhelbawy 等人在基于实体相关图进行推理时,基于推理结果采用一种动态选择算法对候选实体进行选择,所提出的算法在 AIDA 数据集上实现了 87.59% 的链接准确率,是目前性能表现最好的集成链接算法之一[23].

通过以上讨论可以看出,语义相关图和推理算法的质量是影响集成实体链接算法性能的主要因素.本文提出的 CCEL 算法正是从这 2 个关键环节入手进行改进:1)在构造语义相关图时,增加了对实体间语义相关度强弱的考量;2)在推理阶段,综合考虑了候选实体与待消歧文本的语义相关性.因此,CCEL 算法能够最大程度地降低错误候选所产生的噪音影响,提高算法对概念相近实体的区分度,从而显著提高实体链接的准确率和召回率.同时,与相关工作相比,CCEL 算法具有良好的适应性和推广性.

2 集成实体链接算法

CCEL 算法由 3 个步骤组成,分别是生成候选实体集合、构造实体相关图和实现集成实体链接.该算法的逻辑框架如图 1 所示.

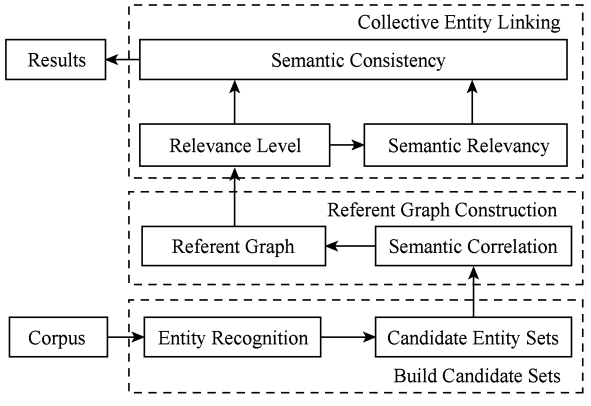

Fig. 1 The framework of the consistent collective entity linking algorithm.

图 1 基于语义一致性的集成实体链接算法框架

2.1 生成候选实体集合

对于任意给定文本 D_i 中出现的实体指称项,生成相应的候选实体集合,是实体链接的第 1 步.本文采用 Stanford NER① 对给定文本进行实体识别,并使用基于规则的方法进行共指处理,得到该文本的实体指称项集合 $M_i=\{m_{i1},m_{i2},\cdots,m_{is},\cdots\}$.然后,根据 M_i 查找本地知识库,得到与 M_i 中元素相对应

① http://nlp.stanford.edu/software/CRF-NER.shtml

的候选实体集合 N_i.本地知识库基于英文版维基百科[①](2015-08-05 打包发布版本)构造,其中包含400多万个实体页面和3 000多万条链接关系.

由于维基百科知识库中包含大量已经过人工消歧处理的同名实体,为利用这些信息,减少实体链接时候选实体的干扰项数量,首先利用维基百科的实体页面(entity pages)、消歧页面(disambiguation pages)和重定向页面(redirect pages)构造一个同名实体对象字典[22],其格式如表1所示:

Table 1　Example of Entity Dictionary

表1　同名实体对象字典示例

Key	Value
Buffalo	Buffalo Bulls Buffalo Bills Buffalo Sabres …
Miami	Miami(city) Miami Hurricans University of Miami …
UCF	UCF Knights University of Central Florida …
Michael Jordan	Michael Jordan Michael Jordan (mycologist) Michael Jordan (footballer) …

字典中键值对(key-value)的构造方法为:首先利用JPWL[②]工具对维基百科知识库中的所有实体页面和重定向页面进行遍历,以实体页面和重定向页面的页名(page title)作为字典的键(key),以相应的消歧页面中包含该页名的所有锚文本以及重定向页面所指向的页名作为与该键对应的值(value).

需要说明的是,由于字典中的键由所有实体页名和重定向页名共同组成,而字典中的值实际上对应的也是实体页名,所以在所构造的同名实体对象字典中,一个键所对应的值同时也可能是另外一个键.例如:键Jordan所对应的值包括Michael Jordan,Benn Jordan等,但Michael Jordan同时也是字典的键,与之对应的值包括Michael Jordan(basketball),Michael Jordan (mycologist)以及Michael Jordan (footballer)等.

在建立同名实体对象字典后,对于给定的实体指称项 m_{is},可以借助字典构造候选实体集合.方法是以 m_{is} 作为键在字典中进行查找,若查找成功,则将相应的值加入到 m_{is} 的初步候选实体集合 N_i' 中.若查找不成功,则将 m_{is} 的候选集合记为空集(NIL).

得到文本 D_i 的初步候选实体集合 N_i' 之后,首先对其进行噪声过滤:对每个实体指称项 m_{is},计算出 N_i' 中与之对应的所有候选实体 n_{sk} 与 m_{is} 的相似度,方法是利用词袋模型计算候选实体 n_{sk} 所在的维基页面与文本 D_i 的余弦相似度:

$$docSim(m_{is},n_{sk})=\frac{\boldsymbol{X}\cdot\boldsymbol{Y}}{\|\boldsymbol{X}\|\|\boldsymbol{Y}\|},\tag{1}$$

其中,符号 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 分别表示候选实体 n_{sk} 所在的维基页面与文本 D_i 的词向量,向量中的元素为文本中出现的所有名词和命名实体,$\boldsymbol{X}\cdot\boldsymbol{Y}$ 表示向量内积,$\|\boldsymbol{X}\|$ 表示向量长度.计算结束后,参照文献[6]的候选实体选择方法,对相似度计算结果进行降序排序,选择相似度最高的 δ 个候选实体加入到最终的候选实体集合 N_i 之中.若与 m_{is} 相对应的候选实体 n_{sk} 的个数不足 δ 个,则不对其进行过滤,将其全部加入到最终的候选实体集合 N_i 之中.δ 的取值采用交叉验证的方法经验得到(参见本文的3.2节).

2.2　构造实体相关图

CCEL算法对每篇输入文本构造一张实体相关图并据此进行实体链接,基本思路是利用共现实体间的语义相关性帮助提高实体链接的准确性并实现批量实体链接,因此,实体相关图的质量对于整个实体链接算法的性能具有关键性影响[15].本文采用无向图 $G=(V,E)$ 表达实体相关图,其中,符号 V 表示顶点集合,顶点元素为集合 N_i 中的候选实体;E 表示边集合,边元素表示顶点间的语义相关性.实体相关图的构造过程由2部分组成:顶点集合的构造和边集合的构造.首先介绍顶点集合的构造过程.

2.2.1　构造顶点集合

实体相关图中的顶点集合定义为:与给定文本 D_i 中出现的实体指称项相关的所有候选实体集合.考虑到不同指称项对应的候选实体可能存在同名的情况,为严格区分候选实体,本文采用 (m_{is},n_{sk}) 实体对来表示实体相关图 G 中的顶点,其中 m_{is} 为 D_i 中的第 s 个实体指称项,n_{sk} 表示与 m_{is} 相对应的第 k 个候选实体.顶点集合的数学定义为:

$$V=\{(m_{is},n_{sk})\,|\,m_{is}\in D_i,n_{sk}\in N_i\}.\tag{2}$$

① http://download.wikipedia.org

② https://dkpro.github.io/dkpro-jwpl/JwplTutorial/

为了利用实体指称项和候选实体的已知上下文信息,我们为图 G 中的每个顶点赋予一个先验置信度(prior confidence level, PCL).以顶点(m_{is},n_{sk})为例,$PCL(m_{is},n_{sk})$表示实体指称项 m_{is} 指向候选实体 n_{sk} 的可能性.相关工作中常用的先验置信度定义方式包括文本相似度、名字相似度和实体流行度等[11].本文 3.3 节将分别采用这 3 种定义进行实验,结合实验结果选择对 CCEL 算法最有效的定义方式.

文本相似度($docSim$)的计算方法如 2.1 节的式(1)所示.以 $docSim$ 作为先验置信度的含义是:实体指称项 m_{is} 所在的文本 D_i 与候选实体 n_{sk} 所在的维基百科页面相比,二者的上下文越相似,则 m_{is} 与 n_{sk} 直接关联的可能性越大.

名字相似度($namSim$)的计算方法为

$$namSim(m_{is},n_{sk})=1-\frac{ed(m_{is},n_{sk})}{maxlen(m_{is},n_{sk})},\quad(3)$$

其中,$ed(m_{is},n_{sk})$表示实体指称项 m_{is} 与候选实体 n_{sk} 的名字间的编辑距离(edit distance),即从字符串 m_{is} 出发,通过字符替换转化成 n_{sk} 所需的最少编辑操作次数.$maxlen(m_{is},n_{sk})$表示在字符串 m_{is} 和 n_{sk} 的长度中取较长者.以 $namSim$ 作为先验置信度的含义是:实体指称项 m_{is} 与候选实体 n_{sk} 的名字相似度越大,则二者直接关联的可能性越大.

实体流行度($popSim$)的计算方法为

$$popSim(n_{sk})=\frac{indeg(n_{sk})}{\sum_{n_{sj}\in N_{is}}indeg(n_{sj})},\quad(4)$$

其中,N_{is} 表示实体指称项 m_{is} 的候选实体集合,n_{sk} 表示 m_{is} 对应的第 k 个候选实体对象,$indeg(n_{sk})$表示维基百科中指向 n_{sk} 且锚文本内容为 m_{is} 的超链接数目.实体流行度是与语料无关的测度,以 $popSim$ 作为先验置信度的含义是:候选实体的(实体)流行度越大,则其作为目标链接对象的可能性越大.

对于输入语料中无歧义的实体指称项,本文将其对应的候选实体的先验置信度置为 1.完成顶点集合的先验置信度赋值计算后,将同一实体指称项对应的所有候选实体的先验置信度进行归一化处理.

2.2.2 构造边集合

实体相关图中边的构造(即建立图 G 中顶点间的关联关系)是算法的重要组成部分.当前建立实体(顶点)关联关系的方法主要有 2 种思路:1)基于实体对应的维基百科页面之间的超链接指向关系来定义实体间的语义相关性[10,18,23];2)采用实体间的谷歌距离(Google distance)作为其语义相关性的测度[11,19,22].

研究表明,上述的 2 种关系定义方式各有利弊,前者可以准确地反映实体间的语义相关性,但通常得到的关系并不完整[24];后者虽然能够提供更全面的关系定义,但由此建立的关系只是统计意义上的,并不能准确反映实体间真正的语义关联以及强度[25].因此,本文提出一种新的实体相关图构造方法,基本思路是将上述 2 种实体关系定义方法进行融合,以结合二者的优点,避免各自的不足.具体方法描述如下:

1) 根据顶点集合 V 中的顶点(实体)在本地知识库中的直接关联关系进行加边处理,如果集合 V 中的顶点 v_a 和 v_b 所代表的候选实体在维基百科页面(即本地知识库)具有超链接直接指向关系,则在这 2 个顶点间添加一条无向边,边的权重 w_{ab} 置为 1;

2) 根据顶点(实体)间的间接关联关系进行加边处理,方法是若集合 V 中的顶点 v_a 和 v_b 所代表的候选实体在本地知识库中均与一个以上的第三方实体存在超链接直接指向关系,则在这 2 个顶点间添加一条无向边,边的权重 w_{ab} 由实体间的谷歌距离做简单的线性变换得到:

$$w_{ab}=1-\frac{\log(\max(|A|,|B|))-\log(|A\cap B|)}{\log(|KB|)-\log(\min(|A|,|B|))},\quad(5)$$

符号 A 和 B 分别表示知识库中与顶点 v_a 和 v_b 所表示的候选实体存在超链接指向关系的实体集合,KB 表示整个知识库的实体集合,$|A|$表示集合 A 中的元素个数.w_{ab} 的取值范围是$(-\infty,1]$,当 w_{ab} 的取值为负数时,设定 w_{ab} 的值为 0,表示顶点 v_a 和 v_b 间不具有语义关联关系;当集合 A 与 B 相等时,w_{ab} 取最大值 1,表示 v_a 和 v_b 的关联实体重合度最高.

需要特别说明的是,在构造实体相关图的过程中,当顶点间存在直接关联关系时,则不考虑其间接关联关系,只有在顶点间不存在直接关联关系时,才进一步考虑其间接关联关系.此外,对于同一实体指称项对应的多个候选实体(顶点),不考虑其相互之间的关联关系,即实体相关图中同一实体指称项所对应的候选实体顶点间不存在关系边.图 2 展示了根据引言中样例 1 生成的实体相关图.

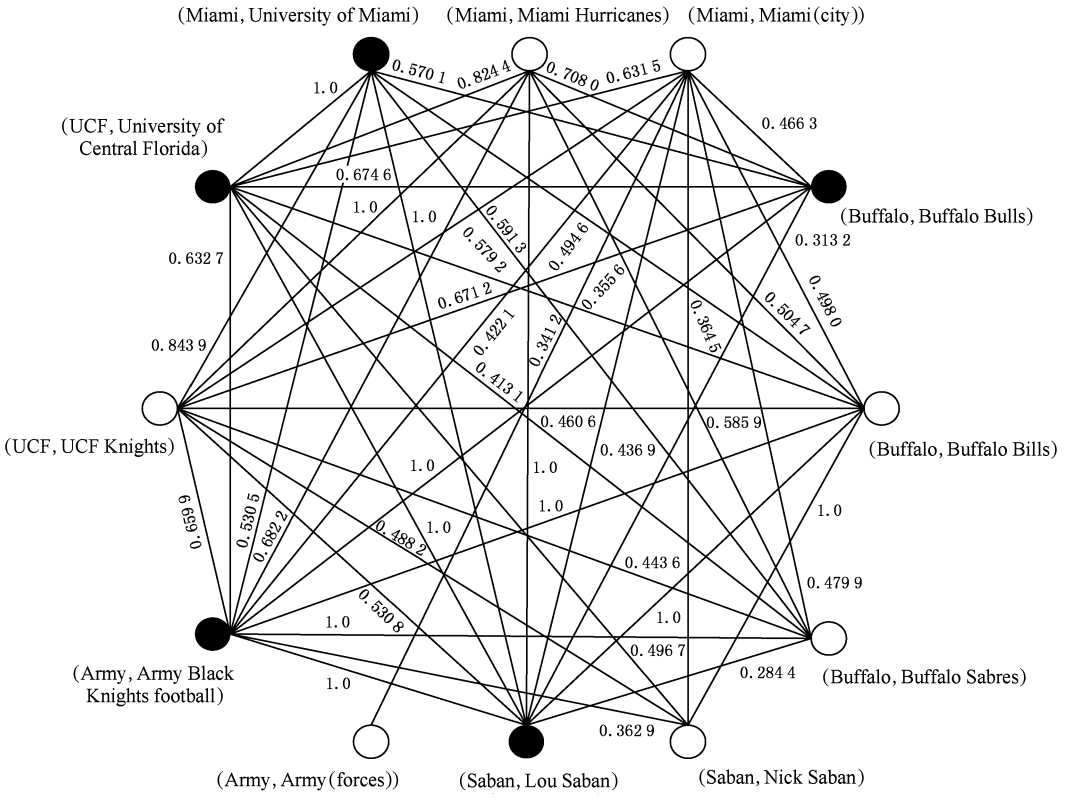

Fig. 2　The referent graph of example 1 in introduction.

图 2　样例 1 的实体相关图范例

2.3　基于语义一致性的集成实体链接

在完成实体相关图的构造之后,即可针对给定文本 D_i 进行实体链接的推理运算.本文提出一种新的推理判据,称为语义一致性判据(semantic consistency criterion, SCC).该判据由候选实体的相关度(relevance level, RL)和候选实体与输入文本 D_i 间的语义相关性(semantic relevancy, SR)2 部分组成.其中,候选实体的相关度表达的是该实体与相应实体指称项之间的语义关联证据的强度,可采用本文提出的相关度排序算法进行计算;候选实体与输入文本 D_i 间的语义相关性则采用本文提出的语义相关性计算方法得到.

2.3.1　计算候选实体的相关度

在构造实体相关图的过程中,我们为每个候选实体(顶点)赋予了一个先验置信度 PCL,然而通过观察 2.2.1 节给出的 3 种 PCL 定义方式,可以看出这些建模的方式均存在一定的片面性,均不适合直接用作实体链接的判据.例如:1)$docSim$ 的准确性依赖于文本 D_i 与候选实体 n_{sk} 所在的维基百科页面的内容质量,如果相应文本较短或包含的有效信息较少,则 $docSim$ 的值会比正常情况偏低;2)如果以 $namSim$ 作为先验置信度并据此进行实体链接操作,我们也很容易举出反例证明其局限性,例如给定实体指称项为"广电",假设有 2 个候选实体分别为"广播电视总局"和"广州电视总局",根据式(3)可知这 2 个候选实体与给定实体指称项的 $namSim$ 值是相等的;3)$popSim$ 的取值虽然与待处理的语料 D_i 无关,其质量仅取决于知识库的知识完备性,然而仅根据实体的流行度作为实体链接的判据显然是不合理的.因此,为提高实体链接的准确性,还需参考其他的已知信息,提出更有效的判据.

相关研究表明,采用实体相关图中顶点的 PageRank 值,对该顶点与相应实体指称项的余弦相似度进行加权,可以得到更好的实体链接效果[18].该结果表明实体相关图的拓扑结构对于确定候选实体的相对重要性具有积极意义.受其启发,本文提出一种新的候选实体相关度计算方法,称为相关度排序(relevance rank, RR)算法,以更好地发掘和利用实体相关图的拓扑结构信息在实体链接任务中的价值.RR 算法的数学公式如下:

$$RR(v_a) = (1-\alpha)docSim(v_a) + \alpha \sum_{v_b \in Nh(v_a)} T(b,a) \times RR(v_b), \quad (6)$$

其中,$docSim(v_a)$表示候选实体顶点 v_a 所表示候选实体与输入文本的上下文相似度,可根据式(1)计算得到,$RR(v_a)$表示候选实体顶点 v_a 与相应实体指称项在当前上下文(图 G)中的相关度,其初始值为顶点 v_a 的先验置信度.$T(b,a)$表示在实体相关图 G 中从顶点 v_b 到顶点 v_a 的带权转移概率:

$$T(b,a)=\frac{w_{ba}}{\sum_{v_k\in Nh(v_b)} w_{bk}}, \tag{7}$$

其中,w_{ba}表示图 G 中边(v_b, v_a)的权重,$Nh(v_b)$表示顶点 v_b 的邻域(neighbourhood),即图 G 中直接与 v_b 相邻的顶点集合.式(6)中的 α为阻尼因子,按照 PageRank 算法的一般惯例取值为 0.85.

由式(6)可见,当算法收敛到稳态时 $RR(v_a)$的物理意义是:在每一轮迭代过程中,随机漫步者选择当前顶点 v_a 的概率由 2 部分组成:一部分由顶点 v_a 与输入文本的上下文相似度 $docSim(v_a)$贡献,所占权重为$(1-\alpha)$;另一部分由与 v_a 相邻的顶点所贡献,所占权重为 α,其中每个相邻顶点 v_b 的贡献为 $T(b,a)\times RR(v_b)$,即把 v_b 自身的相关度按照与之相邻的边的权重进行分配.由此可见,式(6)对文本 D_i 的上下文环境(主要体现在相似度计算中)和实体间的关系(主要体现在图 G 的拓扑结构中)进行了综合考虑,通过迭代求解式(6),实现先验置信度在各顶点间的再分配,使得在当前语境下拥有较多关联证据的顶点的相关度得到强化,同时削弱拥有较少关联证据的顶点的影响.

如本节开始部分所述,当前已有的先验置信度计算方法都具有一定的片面性.通过实验我们发现,RR 算法虽然能够借助候选实体在当前上下文环境中的关联信息对其结果进行纠偏,但当 D_i 的上下文信息不足时,仍无法完全克服先入为主的偏见产生的影响.例如,在图 2 给出的样例中,黑色顶点表示在图中运行 RR 算法之后,每个实体指称项的候选实体集合中相关度最高的顶点.其中,候选实体 Buffalo Bulls(Buffalo 大学橄榄球队)虽然是实体指称项 Buffalo 的候选实体集合中具有最高相关度的顶点,但实际上正确的链接对象应为 Buffalo Bills(美国职业橄榄球队).经过分析,导致这种情况出现的原因是二者的概念相近,因此在同一上下文环境下与相关实体的关联关系也相近,导致 RR 算法对二者的区分度不高,为此本文进一步考虑从候选实体与文本 D_i 的语义相关性的角度对链接对象做进一步区分.

2.3.2 计算候选实体的语义相关性

候选实体(顶点)v_a 的语义相关性是指 v_a 与给定文本 D_i 在语义上相关的强度.由于图 G 仅由候选实体构成,因此本文采用 D_i 中的每个实体指称项对应的候选集合中相关度最高的候选实体构成子集来表示文本 D_i,该子集简记为 N_{maxR}.定义候选实体顶点 v_a 与给定文档 D_i 的语义相关性如下:

$$SR(v_a,D_i)=\sum_{v_k\in N_{maxR}} w_{ak}\times RR(v_k), \tag{8}$$

其中,w_{ak}表示图 G 中边(v_a, v_k)的权重.如 2.3.1 节所述,CCEL 算法引入候选实体的语义相关性定义的目的是帮助解决 RR 算法对某些概念相近的候选实体的区分度不高的问题,由式(8)可见,$SR(v_a,D_i)$实际上是 N_{maxR} 集合中所有候选实体(顶点)v_k 的相关度的加权和,权重因子 w_{ak} 为实体(顶点)v_a 和 v_k 之间边(v_a, v_k)的权重.由 2.2.2 节可知,v_a 和 v_k 的语义关联越少,则权重值越低.

仍然以图 2 中的指称项 Buffalo 为例,尽管干扰项 Buffalo Bulls 在相关度计算中获得了较高的值,但由于该实体与 D_i 在语义上的关联强度弱于另一个(正确的)候选项 Buffalo Bills,因此在引入语义相关性计算之后,可以提高对这 2 个概念上相近的实体的区分度,从而实现正确的实体链接.

2.3.3 语义一致性判据与实体链接

考虑如何将候选实体的相关度和语义相关性结合起来得到一个实体链接的标准,本文提出语义一致性判据:

$$SCC(m_{is},n_{sk})=\frac{RR(v_{sk})+SR(v_{sk},D_i)}{\sum_{v_j\in V_{is}} RR(v_j)+SR(v_j,D_i)}, \tag{9}$$

其中,v_{sk}是图 G 中与候选实体 n_{sk} 相应的顶点,V_{is} 表示由实体指称项 m_{is} 的所有候选实体所构成的顶点集合.由式(9)可知,SCC 判据的取值范围是(0,1),SCC 值越大,表明候选实体 n_{sk} 的语义一致性越高.由此可以得到实体链接的判据如下:

$$Link(m_{is},n_{sk})=\arg\max_{n_j\in N_{is}}(SCC(m_{is},n_j)), \tag{10}$$

其中,N_{is} 表示由实体指称项 m_{is} 的所有候选实体所构成的集合.式(10)的含义为,将文本 D_i 中的实体指称项 m_{is} 链接到本地知识库中具有最高 SCC 值的候选实体 n_{sk} 之上.

2.3.4 语义一致性判据的有效性

在 2.3.1 和 2.3.2 节的方法介绍过程中,我们多次引用了样例1的实验结果,但并未给出具体的

数值计算结果.本节以样例 1 中的实体指称项 Buffalo 和 Miami 为例,用具体的数据辅助说明所提出的 CCEL 算法的设计依据,并初步验证本文提出的 *SCC* 判据的有效性.已知实体指称项 Buffalo 应链接到本地知识库中的 Buffalo Bills 对象上(美国职业橄榄球队),实体指称项 Miami 应链接到本地的 Miami Hurricanes 对象上(美国职业橄榄球队),经归一化处理后的计算结果如表 2 所示:

Table 2　The Relevance and Semantic Consistency of Candidate in Example 1

表 2　样例 1 中候选实体的相关度与语义一致性

Mention	Candidate	*popSim*	*RR*	*SCC*
Buffalo	Buffalo Bulls	0.2547	**0.376**	0.3441
	Buffalo Bills	**0.3948**	0.3628	**0.3653**
	Buffalo Sabres	0.3505	0.2606	0.2906
Miami	Miami(city)	**0.5318**	0.2665	0.2438
	Miami Hurricanes	0.1601	0.3601	**0.3819**
	University of Miami	0.3081	**0.3735**	0.3743

表 2 的第 3 列给出的是以实体流行度为测度的实体相关图中顶点的先验置信度(PCL),可以看出,如果仅参照流行度指标,Buffalo 可实现正确链接,而 Miami 则会被链接到更为流行的 Miami(city)对象上.表 2 的第 4 列给出了经过 RR 算法修正的候选实体相关度计算结果,可以看出经过修正后的结果极差变小,特别是对于 Miami 的候选实体而言,正确的链接对象 Miami Hurricanes 的相关度得到了显著提升,该结果表明本文提出的相关度排序算法能够利用实体相关图拓扑结构中包含的实体关联信息,对实体先验置信度进行平滑修正,且修正的结果有利于降低干扰项的影响程度.同时也可以看出,单纯使用实体相关度作为链接判据并不可行,原因是 RR 算法对实体相关度的平滑效应会导致噪音放大,对实体链接的准确性造成干扰.因此有必要引入新的信息辅助判断.

表 2 的最后一列给出的是按照式(9)计算得到的语义一致性判据,可见根据该判据能够有效识别出正确的候选实体.进一步对比各候选实体的 *RR* 值与 *SCC* 值,可以看出单凭候选实体的语义相关性($SR = SCC - RR$)无法实现准确链接,由此进一步证明了本文提出的语义一致性判据的有效性.

3　实验结果分析

3.1　实验数据

为充分检验所提出的基于语义一致性的集成实体链接算法的有效性,本文采用 2 组近期发布的公开测试语料进行测试,数据详情如表 3 所示:

Table 3　Statistics of the Corpus

表 3　实验数据统计情况

Corpus	# Documents	# Mentions	# InKB	# NIL
AIDA	1393	34956	27820	7136
CSAW	107	17200	10320	6880

第 1 组是由马克斯-普朗克研究所(Max Planck Institute)的 YAGO 实验室发布的 AIDA 数据集①,该数据集包括一个训练集(train)和 2 个测试集(testA,testB),共 1393 篇文档(平均长度为 216 个字符)和 34956 个实体指称项(mentions).数据集中的所有实体指称项均通过人工标注,准确链接到了维基百科(即本地知识库)对应的实体对象上.其中在维基百科中存在对应实体对象的实体指称项共 27820 个(称为 InKB 实体),在维基百科中无对应实体对象的实体指称项共 7136 个(称为 NIL 实体)[19].

第 2 组数据集是由印度理工学院 Chakrabarti 教授领导的 CSAW 项目发布的 Annotation data(简称 CSAW 数据集)②.该数据集中所有的实体指称项均被人工消歧,其中包含 107 篇文档(平均长度为 623 个字符)和 17200 个实体指称项,InKB 实体个数为 10320,NIL 实体个数为 6880[10].CSAW 数据主要用于验证 CCEL 算法在处理短文本实体链接任务时的性能,为便于与相关工作进行比较,我们采用文献[21]介绍的方法对 CSAW 数据进行了切分,使得每篇文档的长度不超过 30 个字符.

3.2　实验方法与评估

为验证 CCEL 算法的有效性,本文从近年来的相关工作中选择了 7 个具有代表性的算法进行实验对比,相关算法的名称及简介请参见表 4 前 7 行,对相关算法细节的介绍和讨论请分别参见本文第 1 节和 3.3.1 节.除与相关工作进行比较外,本文还针对所提出的 CCEL 算法设计了 4 组对比实验(参见

① http://www.mpi-inf.mpg.de/yago-naga/aida/

② https://www.cse.iitb.ac.in/~soumen/doc/CSAW/

表4后4行),这4组实验用于讨论本文提出的实体相关图构造方法,候选实体(与输入文本)的语义相关性计算方法对于算法整体性能的影响以及进一步验证CCEL算法的有效性.

Table 4 Experimental Comparison Method

表4 参与实验比较的实验方法一览

Algorithm	Description
Cucerzan[9]	The algorithm based on context similarity
Kulkarni[10]	The seed work of collective linking algorithm
Han[11]	The collective entity linking algorithm based on random walk algorithm
Hachey[18]	The collective entity linking algorithm based on PageRank algorithm
Hoffart [19]	The collective entity linking algorithm based on Dense subgraph
Tagme[21]	The best entity linking algorithm on short text
Alhelbawy[23]	The collective entity linking algorithm based on Personalized PageRank algorithm
DWR	The algorithm used hyperlink relationship to construct referent graph
NGD	The algorithm used the google distance between entities to construct referent graph
NoSR	The algorithm does not include the process of semantic relevancy computation
Baseline	The algorithm used PCL ($docSim$,$namSim$,$popSim$) as a linking criterion

本文采用准确率(*Precision*)、召回率(*Recall*)和*F*1值等指标对算法性能进行评估.其中,准确率的含义是正确链接的实体数量占算法输出的链接总数的百分比,即算法的精确性;召回率的含义是正确链接的实体数量占测试集中已知事实总数的百分比,即算法的查全率.*F*1值的定义如下:

$$F1=\frac{2\times Precision\times Recall}{Precision+Recall}. \tag{11}$$

*F*1值受准确率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.*F*1值越大,说明算法执行实体链接任务的综合性能越好.

在统计实验结果时,仅考虑InKB实体指称项,以便与相关工作保持一致[10,11,18-19,23].对于m_{is}所指代的实体对象不在本地知识库中(即m_{is}为NIL实体)的情况,CCEL算法的处理方式如下:1)若实体指称项m_{is}的候选实体集合为空集,则判定其为NIL实体;2)若m_{is}与所有候选实体的语义一致性均低于预设的阈值λ,则判定其为NIL实体.

CCEL算法中包含2个经验参数:用于限定候选实体集合大小的阈值δ(定义参见2.1节)和用于判定实体指称项是否对应于NIL实体的阈值λ.为确定这2个参数,采用AIDA数据集中的testA作为参数验证数据集(其中包含216篇文档共4791个实体指称项),得到CCEL算法的*F*1值与δ和λ的关系分别如图3和图4所示.由图3可见,当$\delta=6$时CCEL算法的*F*1值最优,$\delta=5$时的表现与之相当,考虑到算法的计算效率,本文取$\delta=5$.由图4可见,当$\lambda\leqslant 0.2$时,CCEL算法的*F*1值最优,之后随λ增大而快速下降,因此本文取$\lambda=0.2$.

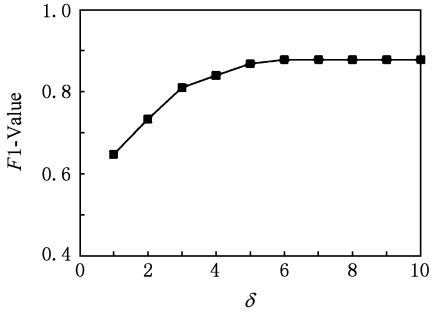

Fig. 3 *F*1-value under different δ on train corpus.

图3 参数δ的取值对算法性能的影响

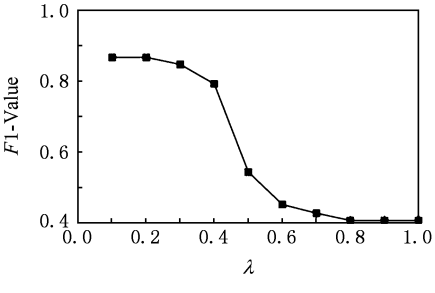

Fig. 4 *F*1-value under different λ on train corpus.

图4 *F*1值与参数λ在训练集上的关系

3.3　实验结果与讨论

3.3.1　CCEL 算法的有效性验证

为验证 CCEL 算法的有效性,首先与相关工作进行横向比较.实验分为 3 组:第 1 组在 AIDA 数据集进行测试,参与性能比较的算法主要是 Baseline(以先验置信度作为判据进行实体链接,选择先验置信度最大的候选实体作为链接对象),实体结果如表 5 所示;第 2 组同样是在 AIDA 数据集上进行测试,参与性能比较的算法主要是当前在该数据集上性能表现最好的 Alhelbawy 算法和其他 4 种综合性能表现较好的算法,实验结果如表 6 所示;第 3 组在 CSAW 数据集上测试,目的是客观评价 CCEL 算法在处理短文本实体链接任务时的性能,参与比较的工作主要是在短文本实体链接中性能表现最好的 Tagme 算法,和其他 3 种在该数据集上性能表现最好的算法,实验结果如表 7 所示,其中,Han,Kulkarni 和 Tagme 等算法的实验结果引自原文,Alhelbawy 算法的实验结果来自对原文方法的重现,由于该算法采用实体流行度作为初始置信度的实验结果最优,所以本文在重现实验时采用实体流行度作为候选节点的初始置信度.

Table 5　Experimental Result on AIDA with Baseline

表 5　CCEL 与 Baseline 在 AIDA 数据集上的实验结果 %

PCI	Baseline			CCEL		
	Precision	*Recall*	*F*1	*Precision*	*Recall*	*F*1
docSim	67.81	61.86	64.69	87.32	84.72	86.01
popSim	**72.85**	**67.27**	**69.51**	**89.94**	**87.38**	**88.65**
namSim	46.26	43.57	44.87	78.41	75.46	76.91

Table 6　Experimental Result on AIDA

表 6　CCEL 与相关算法在 AIDA 数据集上的实验结果 %

Algorithm	*Precision*	*Recall*	*F*1
Cucerzan	51.03	43.74	46.46
Hachey	81.67	79.14	80.38
Kulkarni	72.87	76.74	74.75
Hoffart	81.82	81.91	81.86
Alhelbawy	87.59	84.19	85.85
CCEL	**89.94**	**87.38**	**88.65**

Table 7　Experimental Result on CSAW

表 7　CCEL 与相关算法在 CSAW 数据集上的实验结果 %

Algorithm	*Precision*	*Recall*	*F*1
Han	69	76	73
Kulkarni	65	73	69
Tagme	64	75	69.06
Alhelbawy	67.74	69.45	68.58
CCEL	**73.77**	**78.94**	**76.54**

从表 5 可以看出,以实体流行度作为初始置信度的 CCEL 算法性能表现最优,*F*1 值达到 88.85%,在 Baseline 方法中,以实体流行度(*popSim*)作为判据的链接算法性能表现最佳,*F*1 值为 69.51%.该结果表明,以先验置信度作为判据的实体链接算法并不能有效地解决实体链接问题,而且对于不同的数据集,该类算法的性能表现会有较大的差异性.而本文提出的 CCEL 算法通过融合实体间的语义相关度,可以有效地克服前者的不足,大幅度地提升实体链接的性能,与 Baseline 中表现最好的 *popSim* 算法相比,本文提出的以 *namSim*,*docSim*,*popSim* 作为 *PCL* 值的 CCEL 算法的 *F*1 值分别提高了 10.64%,23.73%,27.53%.由于以实体流行度作为先验置信度的 CCEL 算法性能表现最佳,所以在本文下述实验部分,CCEL 算法均以实体流行度作为先验置信度.

从表 6 可以看出,CCEL 算法在所有性能指标上的表现均优于相关工作.与基于 PageRank 的集成实体链接算法 Alhelbawy 和 Hachey 相比,CCEL 算法在 AIDA 测试集上的 *F*1 值分别提高了 3.26% 和 10.29%.该结果表明,通过引入候选实体与输入文本的语义相关性,可以提高对概念上相近的实体的区分度,克服其中错误候选实体所产生的噪音影响,从而实现正确的实体链接.

与采用谷歌距离作为实体语义相关性测度构造实体相关图的算法 Hoffart 和利用实体维基页面间的超链接指向关系构造实体相关图的算法 Kulkarni 相比,CCEL 算法在 AIDA 测试集上的 *F*1 值分别提高了 8.29% 和 18.60%,该结果表明,通过对实体间的直接和间接语义关系进行区分,能够在确保实体关系完整性的同时,进一步修正实体间相关性的强度,从而显著提高实体链接准确率.

与基于上下文相似度的 Cucerzan 算法相比,CCEL 算法在 AIDA 测试集上的 *F*1 值提高了 90.91%,该结果表明,基于语义一致性的 CCEL 算法能够充分利用知识库中已有的知识结构,推理出实体间的语义相关性,从而有效弥补了基于上下文相似性的实体链接算法受到上下文信息完整性制约的缺陷,大幅度地提升了实体链接的准确性.

从表 7 可以看出,CCEL 算法在 CSAW 测试集

上的所有性能指标同样一致地优于相关工作.与基于图的算法 Kulkarni,Alhelbawy 及基于随机游走模型的算法 Han 相比,CCEL 算法在 CSAW 测试集上的 $F1$ 值分别提高了 10.92%,11.61% 和 4.85%,该结果表明 CCEL 算法能够有效解决现有集成链接方法在处理短文本时,由于上下文信息不足而导致的算法性能恶化问题,因此 CCEL 算法具有良好的适应性和推广性.

与相关工作中对短文本性能表现最优的 Tagme 算法相比,CCEL 算法在 CSAW 测试集上的准确率和召回率分别提高了 15.27% 和 5.25%.通过对 Tagme 的性能进行分析,我们发现问题的主要原因在于 Tagme 过度依赖于实体流行度和候选实体间相关性的计算准确性,当两者之一出现偏差时,算法的性能会恶化.例如,对于引言中提到的样例 1,由于受实体流行度计算结果的影响,Tagme 算法会将实体指称项 Miami 错误地链接到其最知名的候选实体 Miami(city),而非实际所指对象 Miami Hurricanes.以上结果表明,CCEL 算法通过在语义一致性算法中对实体相关度和实体流行度等信息进行综合考量,从而避免了对局部信息的过分倚重,显著提高了算法的整体性能.该实验结果同时也为证明本文提出的实体相关图构造方法和候选实体(与输入文本)的语义相关性计算方法的有效性提供了有力证据.

为进一步评估本文提出的实体相关图构造方法和候选实体的语义相关性计算方法的有效性,本文将 CCEL 算法与 3 种基准实验方法(表 4 中的 DWR,NGD 和 NoSR)进行比较,实验结果如表 8 所示:

Table 8 Experimental Result on AIDA and CSAW

表 8 CCEL 与相关算法在公开数据集上的实验结果 %

Algorithm	AIDA			CSAW		
	Precision	*Recall*	*F*1	*Precision*	*Recall*	*F*1
DWR	81.17	78.54	79.83	68.85	73.42	71.06
NGD	85.11	83.68	84.39	70.32	75.17	72.66
NoSR	84.67	82.14	83.39	70.49	75.44	72.88
CCEL	**89.94**	**87.38**	**88.65**	**73.77**	**78.94**	**76.54**

从表 8 可以看出,CCEL 算法在 AIDA 测试集和 CSAW 测试集上的所有性能指标均优于参与比较的实验方法.与 DWR 和 NGD 算法相比,CCEL 算法在 AIDA 测试集上的 $F1$ 值分别提高了 11.05% 和 5.05%,在 CSAW 测试集上的 $F1$ 值分别提高了 7.71% 和 5.34%.该实验结果表明,通过充分利用知识库的结构化语义关系(直接语义关系和间接语义关系)并且融合当前主流相关性计算方法的实体相关图构造方法(参见本文 2.2.2 节),能够在一定程度上弥补因实体相关图不完整或实体相关性不准确对实体链接算法所造成的负面影响,从而提高实体链接算法的准确率.与 NoSR 算法相比,CCEL 算法在 AIDA 测试集和 CSAW 测试集上的 $F1$ 值分别提高了 6.31% 和 5.02%.该实验结果表明,通过引入候选实体与输入文本的语义相关性,能够对 NoSR 的实验结果进一步修正,降低噪音影响,提高算法对概念上相近的实体的区分度,从而实现正确的链接.

3.3.2 CCEL 算法的错误分析

为了客观评价 CCEL 算法的性能,我们对 CCEL 算法输出结果中发生错误实体链接的部分进行了统计和人工分析,归纳出 3 种出错的场景.

错误Ⅰ.若实体指称项 m_{is} 所对应的正确候选不在本地知识库中,CCEL 将选择与 m_{is} 具有最大语义一致性的候选实体作为链接对象(前提是 SCC 的计算结果超过阈值 λ).例如,对于文本"Ijaz Ahmed is a retired Pakistani crickter….",由于指称项 Ijaz 所指的实体对象 Ijaz Ahmed(cricketer)不在本地知识库中,且它与候选实体 Ijaz Ahmed(wushu)具有较强的语义一致性,导致 CCEL 将后者视为正确的链接目标.此类错误是当前实体链接研究工作面临的共性问题,也是本文下一步工作拟重点研究解决的问题之一.

错误Ⅱ.由于本地知识库知识的不完整(如部分实体间的关系缺失)而导致的实体链接错误.例如对于测试集中的文本"Tian Liang and Zhang Liang have participated in the program….",由于文中的实体指称项 Zhang Liang 所指的实际实体对象 Zhang Liang (model)与文中的实体指称项 Tian Liang 在本地知识库中的候选实体之间不存在任何直接和间接的关联关系,导致 CCEL 算法错误地将实体指称项 Zhang Liang 链接到了具有实体流行度的实体对象 Zhang Liang (western han)上.由于这种错误受到 SCC 计算结果的制约,因此出错的概率较低,在当前算法版本中对这种情况并未做专门考虑,在下一步工作中,我们将对该问题进行深入研究.

错误Ⅲ.若候选集合中有多个候选实体与正确候选具有相同的上下文相似度,CCEL 算法从中随机选择前 5 个候选实体时(即阈值 $\delta=5$),有可能遗漏正确的候选实体.例如,对于文本"Michael Jordan

won his Second MVP award after… .",由于实体指称项 MVP 对应的候选集合中,有多个候选实体与正确的候选实体(most valuable player)具有相同的上下文相似度,例如 MVP(TV show),MVP(song),MVP(album),MVP(group),和 Montel Vontavious Porter,从而可能导致正确的候选实体被漏选.此类错误是由于 CCEL 算法的设计方式导致的,可以通过调整随机筛选的策略进行修正(如提高阈值 δ),但综合考虑到算法的计算效率和准确率,本文阈值取 $\delta=5$ 作为筛选标准.

4 结束语

本文研究了知识图谱上的实体链接问题,发现了现有集成实体链接方法的不足,并提出一种基于语义一致性的集成实体链接算法 CCEL.该算法充分利用了知识库中的结构化语义关系(直接语义关系和间接语义关系),提高了算法对概念上相近的实体的区分度.实验表明,CCEL 算法在 AIDA 和 CSAW 等公开数据集的性能表现一致且优于本领域的代表性工作.

论文的主要贡献包括 3 个方面:1)通过实验证明了本文提出的基于语义一致性的集成实体链接算法在性能上一致优于当前主流的集成链接方法,且具有较好的适应性和扩展性;2)实验表明了本文提出的实体相关图构造方法通过充分利用知识库知识,能够在一定程度上弥补因实体相关图不完整或者实体相关性不准确对实体链接算法所造成的负面影响;3)实验表明了本文提出的语义相关性(候选实体与输入文本)计算方法能够降低错误候选所产生的噪音影响,提高算法对概念上相近的实体的区分度,达成精确的实体链接.

针对 CCEL 算法产生错误的原因,我们后续的工作将主要围绕 2 个方面展开:1)改进候选实体生成方法,将实体指称项和候选实体的类别信息考虑进来,以提高候选实体召回率和算法的执行效率;2)继续丰富完善本地知识库的规模和知识结构,以进一步提高实体链接操作的准确率.

参 考 文 献

[1] Heath T, Motta E. Ease of interaction plus ease of integration: Combining Web2.0 and the Semantic Web in a reviewing site [J]. Web Semantics: Science, Services and Agents on the World Wide Web, 2008, 6(1): 76-83

[2] Dong Xin Luna, Gabrilovich E, Heitz G, et al. Knowledge vault: A Web-scale approach to probabilistic knowledge fusion [C] //Proc of the 20th Int Conf on Knowledge Discovery and Data Mining (KDD'14). New York: ACM, 2014: 601-610

[3] Ji Heng, 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

[4] Huai Baoxing, Bao Tengfei, 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)

[5] 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

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

[7] Dai Hongjie, Wu Chiyang, Tsai R, et al. From entity recognition to entity linking: A survey of advanced entity linking techniques [C] //Proc of the 26th Annual Conf of the Japanese Society for Artificial Intelligence. Berlin: Springer, 2012: 1-10

[8] 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

[9] Cucerzan S. 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

[10] Kulkarni S, Singh A, Ramakrishnan G, et al. Collective annotation of Wikipedia entities in Web text [C] //Proc of the 15th Int Conf on Knowledge Discovery and Data Mining (KDD'09). New York: ACM, 2009: 457-466

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

[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. Hong Kong: 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在实体链接系统中的应用[J].中文信息学报,2015,29(2):58-67)

[15] 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

[16] 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

[17] Greg D, Dan K. A joint model for entity analysis: Coreference, typing, and linking [J]//Trans of the Association for Computational Linguistics. 2014, 2(1): 477-490

[18] Hachey B, Radford W, Curran J R. Graph-based named entity linking with Wikipedia [C]//Proc of Int Conf on Web Information System Engineering. Berlin: Springer, 2011: 213-226

[19] Hoffart J, 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

[20] Guo Yuhang, Qin Bin, Liu Ting, et al. Microblog entity linking by leveraging extra posts [C]//Proc of the 2013 Conf on Empirical Methods in Natural Language Processing (EMNLP'13). Stroudsburg, PA: ACL, 2013: 863-868

[21] Ferragina P, Scaiella U. Tagme: On-the-fly annotation of short text fragments (by Wikipedia entities) [C]//Proc of the 19th Int Conf on Information and Knowledge Management (CIKM'10). New York: ACM, 2010: 1625-1628

[22] 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

[23] Alhelbawy A, Robert G. Collective named entity disambiguation using graph ranking and clique partitioning approaches [C]//Proc of the 25th Int Conf on Computational Linguistics (COLING'14). Stroudsburg, PA: ACL, 2014: 1544-1555

[24] Witten I, Milne D. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links [C]// Proc of AAAI Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy (AAAI'08). Menlo Park, CA: AAAI, 2008: 25-30

[25] Huang Hongzhao, Larry H, Ji Heng. Leveraging deep neural networks and knowledge graphs for entity disambiguation [DB/OL]. Ithaca: ArXiv, [2015-04-28]. http://arxiv.org/pdf/1504.07678v1.pdf

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

Zhong Yun, born in 1990. Master. Student member of China Computer Federation. His main research interests include natural language processing (NLP) and machine learning (zhongyunuestc@gmail.com).

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

Wu Zufeng, born in 1978. PhD and engineer. Member of China Computer Federation. His main research interests include machine learning, data mining, and information security (wuzufeng@uestc.edu.cn).

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