面向实体链接的多特征图模型实体消歧方法*

高艳红[1], 李爱萍[1,2], 段利国[1]
(1. 太原理工大学 计算机科学与技术学院, 太原 030024; 2. 武汉大学 软件工程国家重点实验室, 武汉 430072)

摘 要: 实体链接技术是将文本中的实体指称表述项正确链接到知识库中实体的过程,其中命名实体消歧的准确性直接影响实体链接的准确性。针对中文实体链接中命名实体的消歧,提出一种融合多种特征的解决方案。首先,以中文维基百科为知识库支撑,从实体指称表述项的上下文和候选实体在维基百科的内容描述两个方面抽取多种语义特征并计算语义相似度;然后将语义相似度融合到构建的图模型中,基于 PageRank 算法计算该图模型的最终平稳分布;最后对候选实体排序,选取 top1 实体作为消歧后的实体链接结果。实验通过与仅围绕名称表述特征进行消歧的基线系统相比,F 值提升了 9%,并且高于其他实体链接技术实验的 F 值,表明该方法在解决中文实体链接技术的命名实体消歧问题上取得了较好的整体效果。

关键词: 中文实体链接; 实体消歧; 语义特征; 图模型
中图分类号: TP391.1 **文献标志码**: A **文章编号**: 1001-3695(2017)10-2909-06
doi:10.3969/j.issn.1001-3695.2017.10.007

Entity disambiguation method based on multi-feature fusion graph model for entity linking

Gao Yanhong[1], Li Aiping[1,2], Duan Liguo[1]
(1. *School of Computer Science & Technology, Taiyuan University of Technology, Taiyuan* 030024, *China*; 2. *State Key Laboratory of Software Engineering, Wuhan University, Wuhan* 430072, *China*)

Abstract: Entity linking is the task of linking name mention in a document with their referent entities in a knowledge base. The accuracy of the named entity disambiguation affects the accuracy of the entity linking directly. According to the named entity disambiguation in the technology of Chinese entity linking, this paper proposed a disambiguation method based on multi-feature fusion. Firstly, it used the Chinese Wikipedia as the knowledge base. It made full use of Wikipedia's rich structural information, such as the abstract, the category, the ambiguity page, the anchor text, and so on. After that, it extracted varieties of the semantic features to measure the semantic similarities between the context of entity mention and the information of the candidate entities in Wikipedia. And then, it modeled a graph which represented the relationship between the name mention and the candidate entities with these similarities. At last, it used the PageRank algorithm to rank the candidate entities and chose the top1 entity as a result of the entity linking. Compared with the baseline system which focused on expression characteristics of the name mentions, the value of F increased by 9%. The proposed approach can improve the entity linking system's performance.

Key words: Chinese entity linking; entity disambiguation; semantic features; graph model

0 引言

2009 年美国国家标准与技术研究院(National Institute of Standards and Technology,NIST)在文本分析会议(Text Analysis Conference,TAC)的知识库扩充(knowledge base population,KBP)任务中提出了实体链接(entity linking)这一子任务[1],是将文本中的实体指称表述项正确链接到知识库中实体的过程,即对具有歧义的实体指称表述项进行实体消歧[2]。

实体链接在知识库的发展历程中起到关键性的作用。一方面,它是知识库的应用入口,可以有效地将知识库应用到很多自然语言处理的研究领域;另一方面,通过实体链接技术可以快速准确地获取目标实体信息,对知识库进行不断扩充和更新。实体链接面临的最主要挑战是解决实体的歧义现象。实体的歧义现象可以概括为多样性和歧义性两类,即多名问题和重名问题。例如"孙中山"的别名包括孙逸仙、孙大炮和孙文等;"公牛"可以指代动物界的公牛以及芝加哥公牛队。总而言之,同一实体名在不同的环境所对应的真实世界的实体可能不同,所以解决歧义现象就需要根据实体名的上下文来确定所具体指代的实体。

目前大多数实体链接技术针对的语言为英文,面向中文的相关研究还比较少,主要原因是缺乏统一且权威的面向中文的实体链接语料库,例如 TAC 在处理跨语言(包含中文)实体链接任务时,采用的方法是将中文实体表述链接到英文知识库上[3],而且对于中文的研究工作仅仅围绕在名称表述层面[3]。因此,就现有的中文实体链接技术,都是建立在构造的语料库和其对应的知识库基础之间进行研究的。

最传统的消歧方法是通过建立词袋(bag of words,BoW)子模型来获取实体表层的特征,认为相同的实体在一定程度上具

收稿日期: 2016-07-12; **修回日期**: 2016-09-05 **基金项目**: 国家自然科学基金资助项目(61572345)

作者简介:高艳红(1991-),女,山西大同人,硕士研究生,主要研究方向为自然语言处理(751248552@qq.com);李爱萍(1974-),女,副教授,博士,主要研究方向为软件形式化与智能语言处理;段利国(1970-),男,副教授,博士,主要研究方向为自然语言处理.

有相似性[4],该方法仅仅建立在衡量词的共现上,没有考虑到实体之间的语义信息。所以如何有效利用实体指称表述上下文和知识库的信息,充分地抽取语义信息是解决实体消歧的有效办法。

目前大多数研究工作在文本表层特征的基础上增加了文本语义信息的获取。文献[5]从文本中抽取多种语义特征,对每个候选实体对依次进行排序过滤;文献[3]在此基础上,对从上下文文本中选取的多种语义特征进行加权融合,一定程度上减少了单特征过滤的偶然性;为了更好地体现文本的语义性,文献[6]基于维基百科的链接结构构造图模型,通过图节点的路径长度反映实体与实体之间的相关性;文献[1]在图方法上进行随机游走算法捕捉实体之间的语义相似度。

基于以上相关工作,本文研究的中文实体链接技术是建立在现有的英文实体链接技术之上,以中文维基百科作为知识库支撑构造图模型,从实体指称表述项的上下文和候选实体在维基百科的内容描述两个方面,不仅只考虑维基百科的锚文本链接,而且充分利用维基百科中的摘要、类别、消歧页面等结构信息,充分地抽取多种语义特征并计算语义相似度,将这些语义信息融合到图模型中进行随机游走,选取图中概率分布的top1实体作为最终的消歧结果。

1 实体链接任务定义

从实现的角度,实体链接任务是结合实体指称表述项的上下文和实体在知识库中的信息,通过一定的方法识别出该实体指称表述项在知识库中所对应的实体;如果知识库中不包含该指称项所对应的实体,则系统返回NIL结果,其中NIL代表空链接[1]。

实体链接任务包括预处理、候选实体集合生成和候选实体消歧。实体链接过程中用到的所有符号定义如表1所示。

表1 实体链接的符号定义

标记名	含义	备注
D	包含实体指称表述项的文档,即查询文本	
K	知识库中所有的实体集合	$K \subset E$
E	真实世界中所有的实体集合	
e	知识库中的实体名	$e \in K$
m	待链接的实体指称表述项	$m \in D$
Em	m 的候选实体集合	$Em \subset K$
e_i	候选实体集合 Em 中一个实体名	$e_i \in Em$
NIL	未链接实体标记	

即
$$\text{query}(m,K)=\begin{cases}\text{id}(e) & \text{如果 } K \text{ 包含 } e\\ \text{NIL} & \text{如果 } K \text{ 不包含 } e\end{cases} \tag{1}$$

其中:e 表示 m 在知识库中的链接实体,$\text{id}(e)$ 表示 e 在 K 中的ID。

1.1 预处理

预处理包括对知识库的预处理、评测文本预处理和查询文本的预处理。

1)知识库 K 的预处理

由于维基百科信息质量较高、领域广泛而且持续更新,大多数现有的知识库(YAGO、DBpedia等)都是从维基百科中的词条里撷取出结构化信息。维基百科已经成为实体链接评测任务的重要资源,由于其涵盖了对歧义词的解释,并且非结构化的锚文本链接包含了丰富的链接结构等一系列特点,可以为实体提供丰富的上下文信息。首先下载官方的中文维基百科离线数据包,使用维基百科提供的抽取器(Wikipedia extractor)抽取正文信息,然后解析所有中文维基百科的页面获取实体概要、类别、锚文本链接、重定向信息等,得到的中文维基百科数据库作为本次实体链接工作的知识库支撑。

2)评测文本的预处理

本文使用TAC提供的评测集,然后从中抽取面向中文的评测文本。该评测文本包含了实体指称表述项信息的XML文件,其中包括实体指称表述项的ID值、名称、所在文档的doc_id等。通过DOM4J解析该评测文本获取表述项的信息,作为整个实体链接技术的输入项。

3)查询文本的预处理

根据评测文本预处理得到的doc_id映射到TAC所提供的对应的查询文本,并对该XML查询文档解析获取实体指称表述项的全文信息,如下所示为实体指称表述项“乔丹”的查询文本信息:“前NBA篮球巨星乔丹尚未决定是否复出,参加即将在一星期内就要举行的NBA球员练习赛。最近外界一直期待,前芝加哥公牛队篮球巨星乔丹会正式宣布复出,不过乔丹的经纪公司SFX今天却仍未发表乔丹复出的声明。”通过IKAnalyzer进行中文分词,最后借助维基百科数据库对名称进行维基化得到查询文本中的所有规范化实体名。

1.2 候选实体集合生成

候选实体集合的生成是基于词典的方法[1,2,7],在预处理后的维基百科知识库中采用模糊匹配和精确匹配相结合的方法生成候选实体集合,其中生成的候选实体均为维基概念。例如实体指称表述项“乔丹”通过上述方法得到如表2所示的实体—表述映射表。

表2 实体—表述映射表

实体指称表述项 m	候选实体 Em
乔丹	迈克尔·乔丹 邓不利多的军队 乔丹·克尔 乔丹

1.3 候选实体消歧

实体消歧的准确性直接影响实体链接的准确性。本文采用图模型的方法从候选实体集合中选取最有可能和实体指称表述项相似的实体作为实体链接结果。下面详细介绍具体的实体消歧过程。

2 实体消歧

实体消歧的实现过程如下:

a)在实体—表述映射表2的基础上,将表中的实体指称表述项 m 与所有的候选实体 Em 作为图的节点,从上下文以及知识库中抽取多种语义特征,进行量化后建立节点之间的连接,构造图模型;

b)基于图的随意游走算法计算概率分布,根据最终的平稳分布衡量实体的重要性,最后对候选实体进行排序,得到图中紧凑性最强的实体,为与实体指称表述项的上下文最相似的实体,即链接结果。

2.1 图模型的构造

实体消歧阶段中,首先构造无向图模型 $G(V,E)$,从表2

实体—表述映射表得到图节点集合 $V=\{$q_乔丹,迈克尔·乔丹,邓不利多的军队,乔丹·克尔,乔丹$\}$(为了防止实体指称表述项和候选实体出现重名现象,本文对实体指称表述项加前缀"q_"),构造的图模型如图 1 所示,其中节点"q_乔丹"为实体指称表述项,背景为灰色的节点为候选实体。

由于构造的图模型为无向图,所以根据图中的节点性质可以将边分为两类,分别为实体指称表述项 q_乔丹与其他候选实体之间的边(本文为了区别通过实线连接)、候选实体之间互相连接的边(通过虚线连接),如图 2 所示。

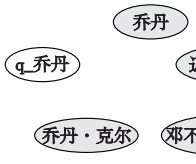

图1　图模型节点

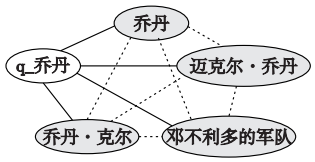

图2　图模型边连接

边的权重表示节点之间的语义相似度。由于边的连接节点不同,边的赋值方法也不同。边的值分为两类,分别命名为指称相似度和实体相似度。其中实体指称表述项与各个候选实体$\langle m,e_i\rangle$之间的相似度为指称相似度(即实线连接的边权值),候选实体$\langle e_i,e_j\rangle$之间的相似度为实体相似度(虚线连接的边权值)。

本文采用图模型的方法进行实体消歧,图方法认为同一文本中的实体往往基于同一主题[4]。为了充分获取节点之间的语义关系,从实体指称表述项的上下文以及知识库中的信息的各个角度抽取节点之间的多种特征。对这些特征进行总结归纳将其分为两大类:局部特征和语义特征。

目前局部特征的相似度计算方法有很多可供选择,如 Dice 系数、最小编辑距离、Jaccard 距离等。语义特征的选取方法也有很多,如基于流行度的相似度方法[2,3,7],通过 TD-IDF 方法选择最"流行"的实体,该方法可以有效地获取流行实体,但如果需要链接的实体并非流行度最高,就会对链接结果产生很大的噪声;Milne 等人[8]提出的一种基于维基百科的语义相似度计算方法;在此基础上,Shen 等人[9]采用 Wikipedia link-based measure(WLM)相似度来衡量候选实体与实体指称表述项的匹配度等。通过研究目前现有的相似度特征方法,与本文基于维基百科的链接页面信息、类别信息以及实体指称表述项上下文的自身特点相结合,提出了如表 3 所述的三大类特征,通过下列特征可以全面地获取实体的局部特征以及语义特征。

表3　特征总览表

特征概述	特征名称
字符串特征	指称部分匹配、指称完全匹配、基于编辑距离的匹配
文本特征	上下文相似度、上下文类别相似度
实体特征	实体类别相似度、基于维基百科的链接的实体相似度

2.1.1　指称相似度

指称相似度表示实体指称表述项与候选实体之间的相似度,融合了局部字符串特征、文本特征以及文本类别特征。

从实现的角度,本文把对这些特征的计算分别映射为编辑距离意义上的指称相似度 simEditDistance(m,e)、上下文指称相似度 simContext(m,e)、上下文类别指称相似度 simCategoryQuery(m,e)的计算,然后将这些特征进行线性融合以实现对$\langle m,e_i\rangle$之间指称相似度的衡量,下面详细介绍算法的实现过程。

a)编辑距离意义上的指称相似度 simEditDistance(m,e)。

编辑距离(edit distance,ED)是两个字符串之间,由一个字符串通过替换、插入和删除等一系列操作转换成另一个字符串所需的最少编辑操作代价[3]。

对编辑距离进行归一化处理得到归一化编辑距离(normalized edit distance,NED),字符串 x、y 的归一化编辑距离计算公式如下:

$$\mathrm{NED}(x,y)=\frac{\mathrm{ED}(x,y)}{\max\{m,n\}} \tag{2}$$

其中:m 为 x 的字符串长度,n 为 y 的字符串长度。当 x 和 y 完全相同时,NED = 0;当 x 与 y 完全不相同时,NED = 1,即 $\mathrm{NED}(x,y)\in[0,1]$。于是,将归一化的编辑距离转换为词语间的语法相似度方法如式(3)所示。

$$\mathrm{simEditDistance}(m,e)=1-\mathrm{NED}(m,e) \tag{3}$$

式(3)中所表达的转换后的编辑距离体现了两个词语之间在编辑距离意义上的接近或相似的程度,即值越大的两个词被认为越靠近,反之则越不相似。按照式(3)计算的结果,将其作为编辑距离意义上的指称相似度,如图 3 所示。

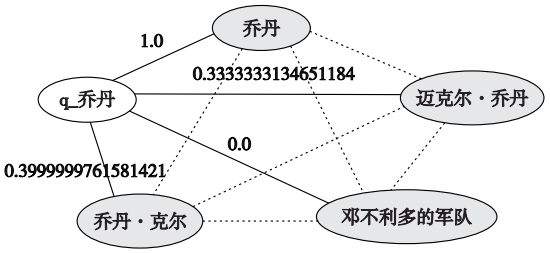

图3　赋值后的编辑距离意义上的指称相似度的图模型

b)上下文指称相似度 simContext(m,e)。

图 3 中的边值只反映了$\langle m,e_i\rangle$之间的字符串特征,还没有包含任何语义特征。对于$\langle m,e_i\rangle$之间的文本特征,采用经典的向量空间模型(vector space model,VSM)进行计算,通过空间上的相似度直观易懂地表达语义的相似度。

计算过程为:将维基百科知识库与文本中特定指称项表示成其上下文的文本向量,通过文本向量间的距离衡量$\langle m,e_i\rangle$间的相似程度。首先将查询文本预处理后获得的上下文转换为实体向量 $T_1(x_1,x_2,x_3,\cdots,x_n)$,以及候选实体在数据库中对应的摘要实体集合作为另一个向量 $T_2(y_1,y_2,y_3,\cdots,y_n)$,式(4)所示为文本向量的余弦相似度的计算公式。

$$\cos(\theta)=\frac{\sum_{i=1}^{n}(x_i\times y_i)}{\sqrt{\sum_{i=1}^{n}(x_i)^2}\sqrt{\sum_{i=1}^{n}(y_i)^2}} \tag{4}$$

其中:x_i、y_i 为向量 $T_1(x_1,x_2,x_3,\cdots,x_n)$、$T_2(y_1,y_2,y_3,\cdots,y_n)$的值,通过式(4)得到的空间距离作为$\langle m,e_i\rangle$上下文相似度特征,如式(5)所示。

$$\mathrm{simContext}(m,e)=\cos(\theta) \tag{5}$$

如果余弦值越接近 1,则表示两个向量越相似。将上下文指称相似度赋值于图模型上,得到图 4 的模型。

c)上下文类别指称相似度 simCategoryQuery(m,e)。

维基百科中实体的类别反映了实体的所属领域和范围,通过类别信息可以衡量实体之间的主题一致性。与上述的上下文相似度不同,类别相似度的文本对象是每个实体的所属类别,利用知识库获取上下文体的类别信息,构成两个文本向量 $T_1(x_1,x_2,x_3,\cdots,x_n)$、$T_2(y_1,y_2,y_3,\cdots,y_n)$,然后使用式(4)计算类别文本向量的相似度。

d)指称相似度 weightQuery(m,e)。

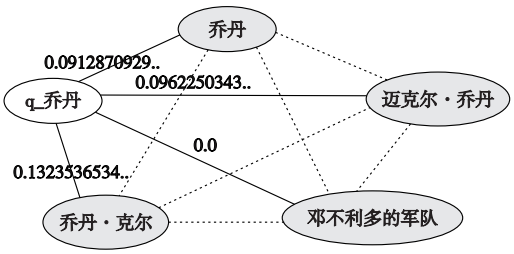

图4 赋值后的上下文相似度的图模型

综合考虑上述三种相似度,本文提出实体指称表述项与各个候选实体$\langle m,e_i\rangle$之间的指称相似度为上述三种相似度的线性组合,如式(6)所示。

$$\begin{aligned}\text{weightQuery}(m,e) &= \alpha\times\text{simEditDistance}(m,e) + \\ &\beta\times\text{simContext}(m,e) + \gamma\times\text{simCategoryQuery}(m,e)\end{aligned} \tag{6}$$

其中参数设定如下:$\alpha=0.23$、$\beta=0.41$、$\gamma=0.36$。

将式(6)得到的结果赋值于图上,得到最终指称相似度赋值之后的图模型,如图 5 所示。

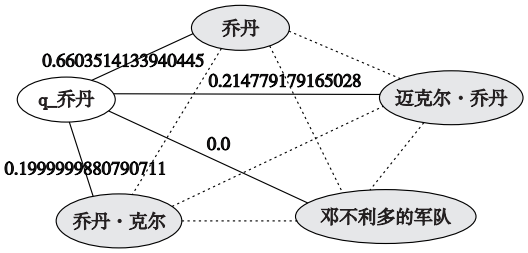

图5 赋值后的指称相似度的图模型

2.1.2 实体相似度

实体相似度是指候选实体$\langle e_i,e_j\rangle$之间的相似度,可以通过实体在维基百科的链接和类别关系来衡量。

本文通过基于维基百科的链接实体相似度 Wikipedia link-based measure(e_1,e_2)、上下文实体类别相似度 simCategoryEntity(e_1,e_2)这两个算法的结合来反映实体之间相似度的语义特征。

a)基于维基百科的链接实体相似度 Wikipedia link-based measure(e_1,e_2),简称 wlm(e_1,e_2)。

WLM 是一种基于规范化的谷歌距离(normalized Google distance)的算法,将 WLM 归一化处理后,转换成为实体相关度,计算公式如下:

$$\text{wlm}(e_1,e_2)=1-\frac{\log(\max(|E_1|,|E_2|))-\log(|E_1\cap E_2|)}{\log(|W|)-\log(\min(|E_1|,|E_2|))} \tag{7}$$

其中:E_1 和 E_2 分别表示实体 e_1 和 e_2 链接到知识库的实体概念的集合,W 是所有维基概念的总和。

b)上下文实体类别相似度 simCategoryEntity(e_1,e_2)。

与计算指称相似度中的上下文类别相似度的处理方法相同,区别在于处理的对象,在指称相似度中处理的对象为实体指称表述项与候选实体之间的相似度,而在本节中,针对的对象是候选实体。首先基于知识库,从中获取每个候选实体的类别,构成向量空间模型利用式(4)得出实体之间的类别相似度。

c)实体相似度 weightEntity(e_1,e_2)。

通过将以上两种算法相结合构成实体相似度,即基于维基百科的链接实体相似度 Wikipedialink-based measure(e_1,e_2)和上下文实体类别相似度 simCategoryEntity(e_1,e_2)的线性组合,建立实体相似度 weightEntity(e_1,e_2)如式(8)所示。

$$\begin{aligned}\text{weightEntity}(m,e) &= \alpha\times\text{WLM}(e_1,e_2) + \\ &\beta\times\text{simCategoryEntity}(e_1,e_2)\end{aligned} \tag{8}$$

其中:参数 $\alpha=0.43$、$\beta=0.57$。图 6 所示是实体相似度赋值之后的图模型。

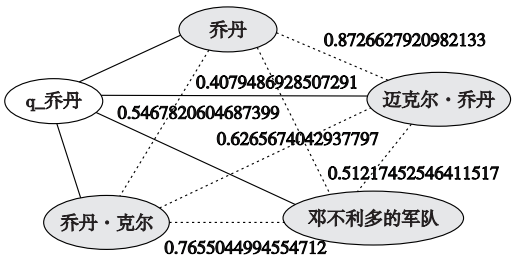

图6 赋值后的实体相似度的图模型

2.1.3 综合两种计算方法得到的图模型

结合$\langle m,e_i\rangle$之间的指称相似度和$\langle e_i,e_j\rangle$之间的实体相似度计算后,完成对边的赋值过程,得到的完整的图模型结构如图 7 所示(其中对指称相似度和实体相似度进行背景颜色区分)。

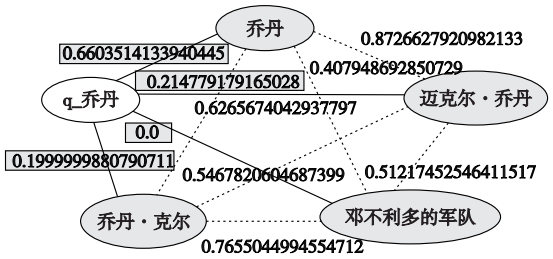

图7 图模型

2.2 PageRank 算法

PageRank 算法是基于实现网页重要性排序的一种算法。本文将模型中的节点对应为实体概念,然后通过 PageRank 算法捕捉图模型中两个节点之间的相似度,选取图中紧凑性最强的实体作为链接目标。

通过前面图模型的建立过程,已经将实体节点的相似度特征融合到图模型中,下面通过 PageRank 算法来获取实体指称表述项与候选实体之间的相似度,并进行排序。

PageRank 算法中所有的符号定义如表 4 所示。

表 4 PageRank 算法中所有的符号定义

标记名	含义
$G(V,E)$	以节点的集合 V 以及边的集合 E 构成的图
P	转移概率矩阵
v_0	初始分布向量
v	随机游走过程中的分布向量

首先通过初始化$(n+1)\times(n+1)$的矩阵 W 来表示该图模型,其中 n 为实体指称表述项 m 的候选实体数目,1 为实体指称表述项本身。为了方便表示矩阵内容,对节点的实体信息数字化,针对表 2 的实体—表述映射表,假定:q_乔丹 =0,迈克尔·乔丹 =1,邓不利多的军队 =2,乔丹·克尔 =3,乔丹 =4,那么矩阵中对应元素 w_{0j}表示节点 0 到 j 的指称相似度,w_{ij}表示节点 i 到 j 的实体相似度(其中 $i\neq0$)。

根据图 7 所示的图模型,通过矩阵方式显示该图(为了方便表示,矩阵中的小数值保留小数点后两位),如式(9)所示。

$$\begin{bmatrix}0.0 & 0.21 & 0.0 & 0.20 & 0.66\\ 0.0 & 0.0 & 0.51 & 0.63 & 0.87\\ 0.0 & 0.51 & 0.0 & 0.77 & 0.41\\ 0.0 & 0.63 & 0.77 & 0.0 & 0.55\\ 0.0 & 0.87 & 0.41 & 0.55 & 0.0\end{bmatrix} \tag{9}$$

定义从节点 i 到 j 的转移概率为 $P(i|j)$,经过对边值的归一化处理后,得到如下所示的计算公式:

$$P(i|j)=\frac{w_{ij}}{\sum_{k=0}^{n-1}w_{ik}} \tag{10}$$

其中:n 为矩阵的维数,该公式得到的转移概率矩阵 P^{T}(同式(9),矩阵中的小数值保留小数点后两位),如式(11)所示。

$$\begin{bmatrix} 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.20 & 0.0 & 0.30 & 0.32 & 0.48 \\ 0.0 & 0.25 & 0.0 & 0.39 & 0.22 \\ 0.19 & 0.31 & 0.45 & 0.0 & 0.30 \\ 0.61 & 0.43 & 0.24 & 0.28 & 0.0 \end{bmatrix} \tag{11}$$

图模型中概率初始分布向量 $v_0[1/n,1/n,\cdots,1/n]$,如式(12)所示。

$$[0.20 \quad 0.20 \quad 0.20 \quad 0.20 \quad 0.20] \tag{12}$$

基于上述计算获得的转移概率矩阵 P^{T} 和概率初始分布向量 v_0 进行图的随机游走算法,直到概率分布达到平稳状态。达到平稳状态的概率分布值表示了图模型中的节点之间语义相似度,图模型的 PageRank 算法如下所示。

算法 1　图的 PageRank 算法

输入:图模型的初始化概率分布向量 v_0,转移概率转移矩阵 P^{T},衡量达到平稳分布的阈值 distance。

输出:图的稳定状态分布向量 v。

a)初始化 $v=v_0$,设置阈值为 0.0000001;

b)do while 达到平稳状态,即 distance 小于或者等于阈值

c)计算 $v_{\mathrm{new}}=\alpha*P^{\mathrm{T}}*v+(1-\alpha)*v_0$;

d)distance = v 与 v_{new}的向量差;

e)$v=v_{\mathrm{new}}$;

f)end while

其中:参数 α 是在随机游走的过程中节点返回初始节点的概率,此处参数 α 的值设为 0.5[2]。

通过 PageRank 算法之后,得到了最终的概率分布向量 v,即实体之间的重要性排序,如式(13)所示。

$$\begin{bmatrix} 0.0300000000004.. \\ 0.2590419237851.. \\ 0.2162845070032.. \\ 0.2492431191423.. \\ 0.2454304500600.. \end{bmatrix} \tag{13}$$

从中获取的最大概率为 0.2590419237851,即候选实体序号 1 所对应的“迈克尔·乔丹”为最终的消歧结果,最后返回 id(迈克尔·乔丹) = 19339。

3　实验

针对中文实体链接的研究,都需要一个中文实体链接的基准语料库来对研究方法的性能进行评估。由于目前缺乏统一的面向中文的实体链接语料库,所以本文的出发点是在 NIST KBP 2013、NIST KBP 2014 的英文语料库,采用自动标注和人工校正相结合的方法,构建了一个中文实体链接语料库[1,3,7],以及 2011 年 6 月的中文维基百科的离线数据包的基础上,构建的中文实体链接语料库其对应的中文知识库。

实验过程中使用 NIST KBP 2013 官方评测数据集中的中文数据作为训练数据,NIST KBP 2014 年的数据集作为测试数据。数据集内容包括新闻文本、网络文本和论坛等,具体实验数据统计如表 5 所示。

表 5　评测数据统计　/条

数据集	所有评测数目	非空链接实体数	空链接实体数
KBP 2013	350	227	123
KBP 2014	567	363	204

3.1　参数设置

通过 NIST KBP 2013 数据集训练获取最优权值,使所有相似度特征发挥到最大作用。通过数据的训练[12],式(6)指称相似度 weightQuery(m,e)的算法中,包括三个参数 α、β、γ,具体参数设置结果如图 8 所示,从结果可得,当 $\alpha=0.23$、$\beta=0.41$、$\gamma=0.36$ 时,准确率可达到最优值。根据实验结果得到 α 值越大,准确率越低。为了方便显示,图 8 只描述了 α 从 0.1 ~ 0.5 的图像。

同样,式(8)实体相似度 weightEntity(e_1,e_2)的算法中,包括两个参数 α、β,具体参数设置结果如图 9 所示。从图 9 中分析可得,当 $\alpha=0.43$、$\beta=0.57$ 时,准确率可达到最优值。

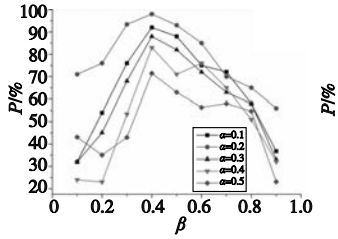

图8　指称相似度算法中的参数设置

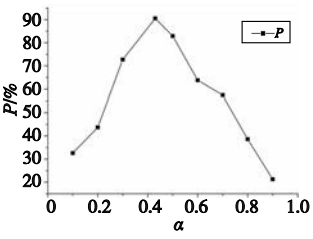

图9　实体相似度算法中的参数设置

本文采用准确率、召回率和 F 值作为衡量实体链接系统整体性能的指标[3~8,10,13],实验中分别用 P、R 和 F 表示。采用叠加的实验方法比较性能差异,其中基准系统[1,3,7](baseline system,BS)采用编辑距离意义上的指称相似度和上下文实体类别相似度的两个特征进行实体链接。

3.2　多种特征的有效性分析

本文是基于多特征融合的图模型算法实现实体链接的过程,所以为了验证这些特征的有效性,本节根据对相似度单独作用和不同相似度进行互相组合对比的方法,分别考察对最终性能的影响并分析造成差异的原因。

a)图模型中$\langle m,e_i\rangle$之间边的权值在基准系统上结合其他指称相似度后的性能比较如表 6 所示。

表 6　指称相似度之间的性能比较　/%

方法	整			空			非空		
	P	R	F	P	R	F	P	R	F
BS	70.10	100	82.42	39.71	79.41	52.94	60.29	65.08	62.60
+ simCont	80.41	100	89.14	34.61	79.41	48.21	65.38	80.95	72.34
+ simCatQ	81.44	100	89.77	34.17	79.41	47.79	65.82	82.54	73.24

其中“+”表示在基准系统之上增加此相似度,即上下文指称相似度和上下文类别指称相似度分别叠加于基准系统上进行候选实体消歧;simCont、simCatQ 分别为上下文指称相似度和上下文类别指称相似度的简称;“整”代表空链接和非空链接的性能;“空”代表其中 NIL 实体的性能;“非空”代表非链接实体的性能。

从表 6 中性能分析说明,上下文指称相似度和上下文类别指称相似度都在语义上很好地区分了不同实体,其中上下文类别指称相似度整体性能提高得稍多。

b)图模型中$\langle e_i,e_j\rangle$之间边的权值在基准系统之上结合其他实体相似度之后的性能比较如表 7 所示。

表 7 实体相似度之间的性能比较 /%

方法	整			空			非空		
	P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
BS	70.10	100	82.42	39.71	79.41	52.94	60.29	65.08	62.60
+ WLM	81.44	100	89.77	34.17	79.41	47.79	65.82	82.54	73.24

与基准系统性能相比,基于维基百科的链接实体相似度整体提高近 7%。与表 6 对比可知上下文指称相似度相对其他三种相似度来说性能较差,这主要是查询文本中不一定包含许多实体,而且有效的信息可能更少,不全面的信息导致相似度值很小,从而造成错误链接的发生。

c)图模型中$\langle m,e_i\rangle$和$\langle e_i,e_j\rangle$之间边的权值在基准系统上逐次叠加其他所有相似度的性能比较如表 8 所示。

表 8 所有相似度之间的性能比较 /%

方法	整			空			非空		
	P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
BS	70.10	100	82.42	39.71	79.41	52.94	60.29	65.08	62.60
+ simCont	80.41	100	89.14	34.61	79.41	48.21	65.38	80.95	72.34
+ simCatQ	82.47	100	90.40	33.75	79.41	47.37	66.25	84.13	74.13
+ WLM	83.50	100	91.01	33.33	79.41	46.96	66.67	85.71	75

其中“ + ”表示在上一行基础之上增加此相似度,即上下文指称相似度、上下文类别指称相似度以及基于维基百科的链接的实体相似度依次叠加于基准系统上进行候选实体消岐。

结果表明,与基准系统相比,叠加组合相似度是对$\langle m,e_i\rangle$和$\langle e_i,e_j\rangle$之间的相似度特征进行依次相加组合而成,虽然缺乏参数优化,不能最大化语义特征的作用,但它同时考虑三种相似度特征,这种做法在一定程度上克服了单个相似度特征的不足,所以相对于单个特征单独作用时性能有了明显的提高。总之,基于多特征融合的图模型方法可以有效地将语义特征应用到中文实体链接技术后 *F* 值提高近 9%,提升了整个实体链接系统的性能。

3.3 实验结果分析

为验证本文方法的有效性,重现谭咏梅等人[4]和张涛等人[2]的实验方法,其中预处理以及候选实体集合生成过程采用与本文相同的处理方法。将实验结果进行对比,如表 9 所示。

表 9 不同实体链接方法的性能比较 /%

方法	*P*	*R*	*F*
谭咏梅等人[4]	75.19	100	85.84
张涛等人[2]	80.43	100	89.15
本文方法	83.50	100	91.01

谭咏梅等人[4]在候选实体消歧过程时,只是单一地对提取的多种特征逐项排序,选择最优实体,没有充分考虑特征之间的关系,存在一定的偶然性。张涛等人[2]提出了一种基于图模型的维基概念相似度计算方法,有效地将维基百科实体之间的链接和类别关系融合到图模型中,虽然对上述问题有所改进,但在图的构建过程、充分利用多种特征以及计算候选实体之间的相似度等方面仍存在不足。本文同样采用 PageRank 算法,在此基础之上对上述不足进行一系列改进,例如,完善了图模型的构建方法,不再单一对特征进行比较、过滤,而是将所有特征充分地融合到图模型中;在抽取特征过程中,由于待链接的实体不一定为流行实体,流行度这一特征存在很大的干扰性,所以排除该特征,并且增加了其他有效特征;计算实体指称表述项与候选实体的上下文相似度时,由于每次仅处理出现待链接实体的一个文本,造成信息浪费和实体链接准确率降低,对此问题,本文不仅仅只局限于需要链接的实体,而是将对象扩展为待处理文本中的所有实体,分别对其抽取特征信息等。

通过表 9 的实验结果对比表明,本文提出的实体链接技术优于其他两种方法,有效提升了整个实体链接技术的性能,证明此方法可行有效。

4 结束语

本文提出了一种基于多特征融合的图模型算法,在现有的英文实体链接技术基础上,充分利用维基百科信息,从中抽取文本的多种语义特征融合到图模型中解决命名实体消歧问题。实验结果表明,本文采用的方法可以有效地获取上下文的语义信息并且取得了较好的整体效果。

对下一步的工作,有如下想法应用到该任务中,例如:a)本文处理空链接实体过程中,只是在候选实体集合生成的步骤中获取,接下来需要对实体消歧的最终平稳分布进行训练,确定一个阈值,如果不超过阈值,则标志空链接标记;b)本文对空链接的实体只进行标记,表示该实体在知识库中不存在,并没有进行下一步的处理过程,还需要继续对空链接的实体进行聚类和预测的工作,使得空链接的实体可以有一个趋向性的目标链接实现知识库的扩充和更新。

参考文献:

[1] 舒佳根,惠浩添,钱龙华,等. 一个中文实体链接语料库的建设[J]. 北京大学学报:自然科学版,2015,51(2):321-328.

[2] 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):58-68.

[3] 左乃彻. 基于中英文维基百科的命名实体消歧[D]. 北京:北京邮电大学,2014.

[4] 谭咏梅,杨雪. 结合实体链接与实体聚类的命名实体消歧[J]. 北京邮电大学学报,2014,37(5):36-40.

[5] 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):52-56.

[6] 郭宇航,秦兵,刘挺,等. 实体链指技术研究进展[J]. 智能计算机与应用,2014,4(5):9-13.

[7] 陈万礼,昝红英,吴泳钢. 基于多源知识和 Ranking SVM 的中文微博命名实体链接[J]. 中文信息学报,2015,29(5):117-124.

[8] Milne D, Witten I H. Learning to link with Wikipedia[C]//Proc of the 17th ACM Conference on Information and Knowledge Management. New York: ACM Press,2008:509-518.

[9] Shen Wei, Wang Jianyong, Luo Ping, *et al*. LINDEN: linking named entities with knowledge base via semantic knowledge[C]//Proc of the 21st International Conference on WWW. New York: ACM Press,2014:449-458.

[10] 怀宝兴,宝腾飞,祝恒书,等. 一种基于概率主题模型的命名实体链接方法[J]. 软件学报,2014,25(9):2076-2087.

[11] Piccinno F, Ferragina P. From TagME to WAT: a new entity annotator[C]//Proc of the 1st International Workshop on Entity Recognition & Disambiguation. New York: ACM Press,2014.

[12] Guo Zhaochen, Barbosa D. Robust entity linking via random walks[C]//Proc of the 23rd International Conference on Information and Knowledge Management. New York: ACM Press,2014:499-508.

[13] Dalvi B, Minkov E, TalukdarP P, *et al*. Automatic gloss finding for a knowledge base using ontological constraints[C]//Proc of the 8th ACM International Conference on Web Search and Data Mining. New York: ACM Press,2015:277-285.