

文章编号:1007-5321(2017)06-0115-05

DOI:10.13190/j.jbupt.2016-220

基于重启随机游走的实体识别与链接方法

谭咏梅¹, 郑迪¹, 刘姝雯¹, 吕学强²

(1. 北京邮电大学 智能科学与技术中心, 北京 100876;

2. 北京信息科技大学 网络文化与数字传播北京市重点实验室, 北京 100101)

摘要: 提出基于重启随机游走的实体识别和链接方法,在知识库部分实体构成的图结构中进行随机游走,从而获得实体和指称的分布式表示,并由此计算出相似度最高的实体作为链接实体. 该方法在 2015 年 Tri-Lingual Entity Discovery and Linking 评测任务中的 F 值为 0.665,高于其他参赛系统. 实验结果表明,本方法可以有效克服特征稀疏问题,并减轻流行度差异对实验结果造成的影响.

关键词: 实体链接; 语义相似度; 随机游走

中图分类号: TN911.22

文献标志码: A

Entity Discovery and Linking Approach Based on Random Walk with Restart

TAN Yong-mei¹, ZHENG Di¹, LIU Shu-wen¹, LÜ Xue-qiang²

(1. Intelligence Science and Technology Center, Beijing University of Posts and Telecommunications, Beijing 100876, China;

2. Beijing Key Laboratory of Internet Culture and Digital Dissemination Research, Beijing Information Science and Technology University, Beijing 100101, China)

Abstract: An entity discovery and linking approach based on random walk with restart was presented. Unified semantic representation for entities and documents—the probability distribution obtained from a random walk on a subgraph of the knowledge based was adopted. According to this distributed representation, the entities that are similar with mentions as the linking results was obtained. This method achieved 0.665 F value on entity linking section of TAC 2015 TEDL task, it performs better than other participating systems. It is illustrated that the method can overcome the feature sparsity issue and is less amenable to feature sparsity bias.

Key words: entity linking; semantic relatedness; random walk

实体链接任务是将文本中的指称链接到知识库中一个无歧义实体的过程^[1]. 该任务是信息抽取以及许多其他应用的关键技术. 然而,自然语言的多样性和歧义性使实体链接任务面临很大的挑战^[2].

通过分析研究前人的工作,发现目前大部分方法使用指称及其上下文信息来进行实体链接,对于流行度较高的指称,上下文信息较为丰富,实体链接性能较好,而对于流行度较低的实体,上下文信息比

较匮乏,链接性能较差. 笔者将指称和候选实体统一表示为概率分布形式,通过比较分布之间的差异来计算语义相似度.

1 相关工作

目前实体链接方法主要分为 2 种:单一实体链接和协同实体链接^[3].

单一实体链接仅考虑当前指称与实体的链接关

收稿日期: 2016-12-12

基金项目: 网络文化与数字传播北京市重点实验室开放课题(ICDD201703); 国家自然科学基金面上项目(61671070)

作者简介: 谭咏梅(1975—),女,副教授, E-mail: ymtan@bupt.edu.cn.

系,而不考虑文本中其他指称的链接信息. Mihalcea 等^[4]通过词袋模型计算指称与实体间的相似度. 谭咏梅等^[5]通过 ListNet 排序模型对候选实体进行排序,取 top1 作为链接实体. 单独对每个指称进行链接没有考虑到同一篇文章中指称的语义一致性,影响了实体链接的效果.

协同实体链接方法考虑了当前处理的文本中指称之间的语义一致性关系,从而对同一篇文章中的所有指称进行协同实体链接. 怀宝兴等^[6]提出了一种基于概率主题模型的命名实体链接方法.

Cucerzan^[7]使用维基百科中的分类条目来衡量指称间的全局一致性. Han 等^[8]使用重启随机游走算法来获得候选实体的相关性向量,并将该向量的相应分量作为指称和候选实体之间的相关度. 李茂林^[3]在 Han 的基础上加入了主题信息.

2 实体识别与链接

基于重启随机游走的实体识别与链接方法包括预处理、指称发现、指称扩充、候选实体生成、候选实体排序、指称聚类 6 部分,流程图如图 1 所示.

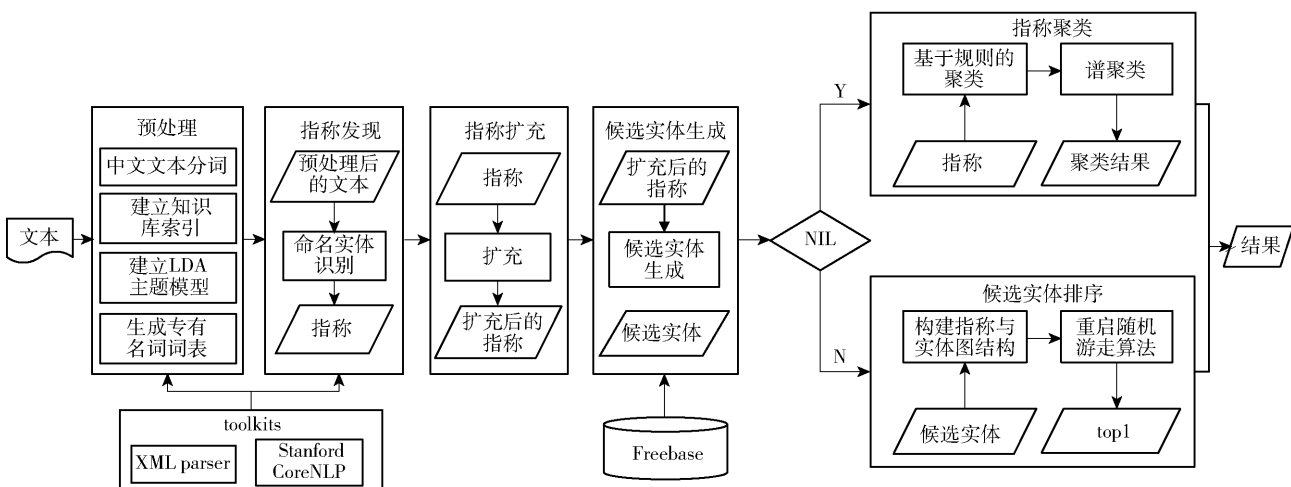


图 1 基于重启随机游走的实体识别与链接方法流程图

2.1 预处理

预处理包括建立知识库索引、生成专有名词词表和建立 Latent Dirichlet Allocation 主题模型.

2.1.1 知识库索引

笔者使用 Elasticsearch^① 对知识库 Freebase 构建索引,提高了搜索候选实体的效率.

2.1.2 生成专有名词词表

知识库中蕴含着丰富的实体信息,如实体的别名、简写、缩写、昵称等,提取以上信息形成专有名词词表.

2.1.3 建立 LDA 主题模型

主题相似的文本集中,指称之间的关联程度较高,有助于指称的消歧. 笔者使用 LDA 主题模型将主题相似的文本划为一簇.

2.2 指称发现

指称类型包括人名、组织机构名、地名、地理政治名和设施名. 笔者使用 Stanford NER 工具识别文本中的部分指称,使用预处理部分构造的专有名词词表来识别缩写、简写、别名、昵称等工具难以准确

识别的指称.

2.3 指称扩充

对指称进行扩充可以减少指称的歧义性. 笔者主要针对别名和简写形式的指称进行扩充.

2.4 候选实体生成

在生成候选实体时,为避免候选实体数量过多,影响实体链接的性能,根据指称类型和知识库中实体类型的对应关系进行筛选.

同时,在知识库中的多个域,如别名、昵称、缩写、简写等搜索指称,避免遗漏候选实体. 此外,对于中文指称,利用汉语拼音查找候选实体来缓解由于书写错误、繁体字和简体字不统一以及外语音译词书写不一致等造成的指称难以识别的问题.

2.5 候选实体排序

当指称存在多个候选实体时,根据相似度算法对候选实体进行排序.

① <https://www.elastic.co/>

2.5.1 指称与实体图结构的构建

通过预处理构建 LDA 主题模型得到主题分布相似的文本集 D , 从 D 中发现指称集合 M , 则最初的实体链接图结构 $G = (V, E)$ 的构造方式为: V 包括指称集合和候选实体集合, E 包括指称到候选实体的边以及候选实体之间的边. 然后对图进行扩充, 增加与候选实体相连的实体以及边, 如图 2 所示.

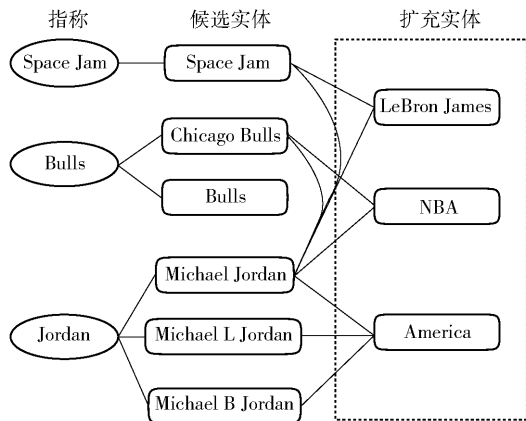


图2 指称与实体图结构

图2中,最左边一列表示指称,中间表示候选实体,最右边一列表示额外添加的与候选实体相连的实体. 为了避免引入噪声,只保留与2个或2个以上候选实体都直接相连的实体结点.

2.5.2 重启随机游走

随机游走是遍历图的一个随机过程,该过程收敛后会对应每个结点生成一个概率分布值,表示该结点被访问到的可能性^[3,9]. 用 T 表示转移矩阵, T 中的元素 T_{ij} 表示从结点 e_i 到结点 e_j 的概率. T_{ij} 的计算公式为

$$T_{ij} = \frac{w_{ij}}{\sum_{e_k \in N(e_i)} w_{ik}} \quad (1)$$

其中: $N(e_i)$ 为与 e_i 直接相连的结点集合, w_{ij} 为结点 e_i 和 e_j 之间的权重(这2个结点在知识库中的共现次数). 用 r^t 表示迭代 t 次时的概率分布,重启随机游走模型可表示为^[3]

$$r^{t+1} = (1 - \lambda)r^t T + \lambda s \quad (2)$$

其中: s 为初始向量, λ 为重启概率. 最终达到平稳分布^[3]:

$$r = \lambda(I - cT)^{-1}s, c = 1 - \lambda \quad (3)$$

2.5.3 语义特征

1) 实体语义特征的计算

对任一实体 e_i , 重启随机游走初始向量 s 的设

置为: 初始向量 s 中第 i 个结点对应的分量为1, 其余结点为0 ($s_i = 1, s_j = 0 (j \neq i)$), 这样保证从结点 e_i 开始随机游走并且每次重启也是从结点 e_i 开始. 之后进行重启随机游走得到的平稳分布即为实体的语义特征.

2) 指称语义特征的计算

计算指称的语义特征时需要将指称的候选实体 E_m 设置为初始结点并计算候选实体 e_i 在初始向量 s 中的对应权重 $s_i = p(m \rightarrow e_i)$. $p(m \rightarrow e_i)$ 表示指称 m 指代实体 $e_i \in E_m$ 的概率, 计算方法为^[3]

$$P(m \rightarrow e_i) = \frac{S(m, e_i)}{\sum_{e_k \in E_m} S(m, e_k)} \quad (4)$$

其中 $S(m, e_i)$ 表示 m 和 e_i 之间的语义相似度. 利用 tf-idf 为词语赋予权重并根据词袋模型将指称和实体的上下文分别转换为向量 m 和 e_i , 则 $S(m, e_i)$ 计算方式为^[3]

$$S(m, e_i) = \frac{m \cdot e_i}{|m| |e_i|} \quad (5)$$

然后执行随机游走, 得到指称 m 的语义特征向量.

2.5.4 指称和候选实体之间的语义相似度计算

Guo 等^[10]借助 KL (Kullback-Leibler) 散度来进行语义相似度的计算. KL 散度是2个概率分布 A 和 B 之间差别的非对称性的度量, 可以表示为

$$D_{KL}(A \parallel B) = \sum_i A_i \lg \frac{A_i}{B_i} \quad (6)$$

考虑到 B_i 可能为0的情况, Guo 等^[10]使用 KL 散度公式为

$$D_{KL}(A \parallel B) = \sum_i A_i \begin{cases} \lg \frac{A_i}{B_i} & B_i \neq 0 \\ \gamma & B_i = 0 \end{cases} \quad (7)$$

其中 γ 为一个实数.

然而, 由于 KL 散度是非对称的度量, 参数 γ 依据经验值设定, 不具备一般性.

针对以上不足, 笔者使用 Hellinger 距离来度量2个概率分布的差异. 概率分布 A 和 B 的 Hellinger 距离为

$$H(A, B) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^K (\sqrt{A_i} + \sqrt{B_i})^2} \quad (8)$$

由式(8)可以看出, Hellinger 距离具有非负性和对称性, 可以很好地满足实验需求. 指称和候选实体之间的语义相似度 $\psi(m, e_i)$ 计算公式为

$$\psi(m, e_i) = \frac{1}{H(C(e_i), C(m))} \tag{9}$$

其中： $C(e_i)$ 表示候选实体 $e_i \in E_m$ 的语义特征， $C(m)$ 表示指称的语义特征。将指称 m 的候选实体按语义相似度得分从大到小排序，取 top1 作为指称 m 在知识库中的目标实体。

笔者使用迭代方法进行实体链接，先对歧义性较小的指称进行链接。在每一轮迭代中，通过在图中执行重启随机游走，获得指称和实体的语义特征，然后计算指称 m 和候选实体 e_i 之间的语义相似度，找出相似度得分最大的实体作为指称的链接实体。在下一轮迭代过程中，会将 m 链接到 e_i 这一结果考虑进去，对图进行剪枝，去掉无关的候选实体并且重新计算概率转移矩阵。然后对剩余指称进行链接，直到所有的指称都链接完毕。如果指称没有候选实体或者指称与候选实体的相似度得分低于一定的阈值，则将该指称表示为 NIL。

由于图结构是根据知识库中的实体来构建的，如果流行度较低的实体与其他实体有更多的关联，则该实体也会获得更高的权重，因此对流行度低的实体也有较高的链接准确率。此外，指称的语义特征是根据指称对应的知识库中的实体得到的，从而保证指称和实体相似度计算的一致性和有效性。

2.6 指称聚类

在实体链接系统中，如果存在多个指称在知识库中都无法找到与其对应的实体，则将指代同一实体的指称聚为一簇。

指称聚类模块使用了规则和谱聚类相结合的方法：先使用规则对指称进行粗划分，即如果 2 个指称 m_i 和 m_j 的命名实体类型相同且有相同的表面字符串时将其划为一簇。然后利用谱聚类算法对划分后的指称进行聚类。

3 实验

3.1 实验数据

笔者使用 TAC Knowledge Base Population (KBP) 2015 年的 Tri-Lingual Entity Discovery and Linking (TEDL) 任务的实验数据。知识库为 Freebase，版本为 2015 年 1 月份，包括上千万个实体。文本包括中文文本 166 篇、英文文本 167 篇和西班牙文文本 167 篇。总共发现的指称有 32 533 个，其中在知识库中可以找到其目标实体的有 8 817 个，在知识库中没有其对应的目标实体的有 23 716 个。在

所有的指称中，15 014 个来源于新闻文本，8 702 个来源于论坛。

3.2 评价方法

对于实体链接部分要求指称在文本中的位置、指称类型以及指称对应的知识库中的实体 id 和官方所给的标准答案完全一致，则该指称的链接结果正确。笔者所用的准确率(P)、召回率(R)和 F 值的计算方法为^[11]

$$P = \frac{|N_{\text{gold}} \cap N_{\text{sys}}|}{|N_{\text{sys}}|} \tag{10}$$

$$R = \frac{|N_{\text{gold}} \cap N_{\text{sys}}|}{|N_{\text{gold}}|} \tag{11}$$

$$F = \frac{2PR}{P + R} \tag{12}$$

其中： N_{gold} 为正确结果， N_{sys} 为实体链接系统得到的结果。

3.3 实验结果及分析

为了验证方法的有效性，笔者重现了文献[9]以及文献[3]中的实验方法，并将实验结果进行对比，如表 1 所示。

表 1 不同实体链接系统在 KBP2015 数据集上的实验结果

方法	KBP2015					
	实体链接			指称聚类		
	P	R	F	P	R	F
文献[9]	0.549	0.485	0.515	0.659	0.551	0.600
文献[3]	0.651	0.614	0.632	0.687	0.549	0.610
本文	0.697	0.636	0.665	0.695	0.556	0.618

由表 1 可以得出，在 KBP2015 数据集上，笔者所用实体链接方法优于其他 2 种实验方法。

经过分析，Alhelbawy 等^[9]和李茂林^[3]提出的算法在构建图结构的时候仅考虑了指称的候选实体，当这些候选实体之间没有直接边相连时，算法效果较差（在极端情况下，任意 2 个候选实体间都不存在直接边相连，算法的性能等同于单一实体链接）。由于知识库中的实体有一部分是间接相连的，笔者使用的方法在图结构中引入与 2 个或 2 个以上候选实体都有直接边相连的实体，使图结构的构建更加合理，实验结果证明该扩充方法的有效性，与 Guo 等^[10]的结论一致。此外，笔者利用主题模型对主题相似的文本进行多文本协同实体链接，由于主题相似的候选实体联系更加密切，这样构造的图结构中边的信息便更加丰富，从而提高了算法的性能。

另一方面,笔者采用 easy-first 思想,即从歧义性较小的指称进行实体链接(歧义性较小的指称链接准确率相对较高),之后利用已有的链接信息,从剩余的指称中找到歧义性较小的指称进行链接,直到所有指称链接完毕.与李茂林^[3]提出的方法相比,笔者的算法流程更加合理、清晰,因而效果更好.

KBP2015 前 3 名参赛队伍的实验结果如表 2 所示.

表 2 KBP2015 前 3 名参赛队伍的实验结果

队伍 编号	实体链接			指称聚类		
	<i>P</i>	<i>R</i>	<i>F</i>	<i>P</i>	<i>R</i>	<i>F</i>
1 ^[11]	0.692	0.632	0.661	0.693	0.554	0.616
2	0.659	0.528	0.586	0.646	0.589	0.616
3	0.549	0.528	0.539	0.683	0.462	0.551
所提方法	0.697	0.636	0.665	0.695	0.556	0.618

由表 2 中的数据可以看出,笔者提出方法的实验结果排名第 1,再次证明了此方法的有效性.

4 结束语

笔者提出了一种基于重启随机游走的实体识别与链接算法,包括预处理、指称发现、指称扩充、候选实体生成、候选实体排序、指称聚类 6 部分.在 KBP2015 的 TEDL 评测任务中实体链接的 *F* 值达到了 0.665,指称聚类部分的 *F* 值达到了 0.618,高于其他参赛队伍.同时与前人的工作进行对比,实验结果也证明了本方法的有效性.然而,该算法中仍存在以下不足:①在指称扩充部分,只利用了表面字符串进行扩充,没有考虑指称间的语义关联,虽然对指称类型和在文本中的位置做了限制,还是有可能出现指称扩充错误的情况;②论坛语料中,由于用语随意以及网络新词的使用,对指称的识别和链接造成干扰,需要对互联网信息加以搜集和利用,以便解决这部分问题.

参考文献:

[1] Roth Dan, Heng Ji, Chang Mingwei, et al. Wikification and beyond: the challenges of entity and concept grounding [R]. ACL 2014 Gaithersburg, MD, 2014: 7-18.

[2] 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.

Tan Yongmei, Yang Xue. An named entity disambiguation algorithm combining entity linking and entity clustering[J]. Journal of Beijing University of Posts and Tele-

communications, 2014, 37(5): 36-40.

[3] 李茂林. 基于主题敏感的重启随机游走实体链接方法[J]. 北京大学学报(自然科学版), 2016, 52(1): 17-24.

Li Maolin. An entity linking approach based on topic-sensitive random walk with restart [J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2016, 52(1): 17-24.

[4] Mihalcea Rada, Csomai Andras. Wikify! linking documents to encyclopedic knowledge [C] // Sixteenth ACM Conference on Information and Knowledge Management. Lisboa, Portugal: ACM, 2007: 233-242.

[5] 谭咏梅, 王睿, 李茂林. 基于上下文信息和排序学习的实体链接方法[J]. 北京邮电大学学报, 2015, 38(5): 33-36.

Tan Yongmei, Wang Rui, Li Maolin. An entity linking approach based on context information and learning to rank[J]. Journal of Beijing University of Posts and Telecommunications, 2015, 38(5): 33-36.

[6] 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.

[7] Cucerzan Silviu. Large-scale named entity disambiguation based on wikipedia data [C] // Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Prague, Czech Republic: EMNLP-CoNLL, 2007: 708-716.

[8] Han Xianpei, Sun Le, Zhao Jun. Collective entity linking in web text: a graph-based method [C] // International ACM SIGIR Conference on Research and Development in Information Retrieval. Beijing: ACM, 2011: 765-774.

[9] Alhelbawy Ayman, Gaizauskas Robert. Graph ranking for collective named entity disambiguation [C] // Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, Maryland: ACL, 2014: 75-80.

[10] Guo Zhaochen, Barbosa Denilson. Robust entity linking via random walks [C] // ACM International Conference on Information and Knowledge Management. Shanghai: ACM, 2014: 499-508.

[11] Ellis J, Getman J, Fore D, et al. Overview of linguistic resources for the TAC KBP 2015 evaluations: Methodologies and results [C] // Proceedings of TAC KBP 2015 Workshop. Gaithersburg, Maryland: National Institute of Standards and Technology, 2015: 16-17.