情报科学
第34卷 第2期 2016年2月

DOI:10.13833/j.cnki.is.2016.02.029

基于上下文特征的短文本实体链接研究

武　川,陆　伟

(武汉大学 信息管理学院,湖北 武汉 430072)

摘　要: 本文构建了一个面向短文本的实体链接系统,致力于鉴别出文本中指向Freebase实体的文本片段。本文将实体链接分为两步,利用现有的实体指称识别方法鉴别出实体指称,随后利用三类特征进行实体消岐,包括:实体指称-实体相似度、实体-实体相似度、候选实体上下文指称相似度。通过考虑所有的实体指称-实体对,选择得分最高的作为实体链接结果。

关键词: 实体识别;实体消岐;实体链接

中图分类号: G254　　**文献标识码:** A　　**文章编号:** 1007-7634(2016)02-144-04

Context Feature Based Entity Linking for Short Text

WU Chuan, LU Wei

(School of Information Management, Wuhan University, Wuhan 430072, *China)*

Abstract: This paper design an entity linking system for short text, aiming at identifying all text fragments referring to an entity contained in Freebase. The task is organized in two steps. We adopt existing mention detection method to detect entity mentions and then disambiguate among candidate entities leveraging three kinds of features, including entity mention-entity similarity, entity-entity similarity, candidate entity context mention similarity. By considering all possible mention-entity pair combinations, we select the one with highest score as the linking result.

Keywords: entity recognition; entity disambiguation; entity linking

1 引　言

信息爆炸在带来海量信息的同时,也对快速准确地获取目标信息提出了挑战。互联网用户在获取信息的过程中,常常需要了解文档中部分实体的详细信息。然而,目前大多数Web文档都不包含语义标注,如果要获取文档中实体的详细信息,就需要启动检索过程。在线知识库的快速发展,为人们快速查找实体信息提供了便利。提高用户信息获取效率的方法之一,就是为文档中指代实体的文本片段添加指向知识库实体的链接,便于用户获取相关实体的信息。这一过程被称为实体链接(Entity Linking)。实体链接能够利用知识库丰富文本的语义信息,在文本分类[1]、文本标注[2]、信息检索[3]、知识库构建[4]等领域都有着重要的理论意义和应用前景。

现有的研究大多利用维基百科中的信息构建实体别名词典,以识别实体指称;然后利用实体指称所在文档的上下文和知识库实体的上下文进行消岐。在利用上下文构建相关特征时,多采用实体指称-实体相关性、同一文档中不同实体指称的候选实体间的主题一致性等特征,取得了一定的效果。

然而上述研究均未考虑上下文实体指称与其他词项上下文间的区别。本文提出利用新的上下文指称相关性指标,表征某一实体指称的候选实体与该指称所在上下文中其他指称间的相关性,以提高实体链接性能。基于公开的web service的评测验证了本文所提出的特征的效果。

收稿日期: 2014-10-21

基金项目: "十二五"国家科技支撑计划课题(2013BAG06B02);国家自然科学基金面上项目(71173164)

作者简介: 武　川(1989-),男,湖北人,博士研究生,主要从事信息检索、数据挖掘研究.

2 相关研究

传统的实体链接框架大多包括两步：指称识别、实体消岐。虽然有研究者的步骤划分方式不同，但是本质是一样的。根据这两个步骤整合方式的不同，可以将实体链接研究分为三类：先识别后消岐、实体消岐、识别消岐联立求解。

先识别后消岐类研究是指以顺序方式完成指称识别和实体消岐，用指称识别的输出作为实体消岐的输入，用实体消岐的输出作为最后的指称-实体映射结果。Bunescu与Pasca[5]将实体链接划分为检测和消岐两步，前者鉴别给定名称是否指向字典中的实体，后者则对多个可能指向的实体进行消岐；Medelyan等[6]利用维基百科构建了受控词表以检测实体，然后用机器学习方法进行消岐。

实体消岐类研究则将实体链接问题视为给定实体指称及其候选实体，鉴别该实体指称在其所在上下文中所指向实体的过程；指称识别则是由外部指称识别系统完成的。这类研究认为实体链接的难点在于实体消岐，因此专注于解决实体链接中的实体消岐问题，在实验时或采用现有的指称识别方法，或利用公开已标注数据集回避指称识别问题。Kulkarni等[7]仅在实验环节提到对输入文档进行分词，并最大程度地与知识库中实体的ID进行字符串匹配，从而找到可能的实体指称。Zheng等[8]所提出的实体链接框架包含四个步骤，其输入是实体指称，而没有考虑如何从文档中识别实体指称。

识别消岐联立求解类研究则认为指称识别的输出可能存在错误，用顺序方式进行实体链接可能会使指称识别阶段的错误传播到实体消岐阶段，影响实体链接的性能。该类研究将指称识别阶段的目标设定为“高召回率”，以尽可能多地找到可能的实体指称，从而在实体消岐阶段对不同的“实体指称-实体”组合进行消岐，得到最可能的指称-实体映射。Stern[9]构建了一个同时考虑指称识别和实体消岐的系统，一方面指称识别阶段所识别的错误指称可能在实体消岐过程中被鉴别为“非实体”；另一方面，所识别的指称可能指向库外实体（也即知识库未收录的实体），因此不被链向知识库。Wick[10]将实体链接与实体发现任务整合为联合实体解析任务，同时提升了实体链接和实体发现任务的准确率。实际上，实体发现可以视为实体链接中对库外实体的鉴别任务，既要识别可能指向实体的指称，又要确认该指称确实指向一个实体，即使该实体是库外实体。

传统的实体链接研究大多关注长文本，例如网页等。近来，短文本实体链接也受到了广泛关注。最早关注短文本实体链接的是TAGME[11]，它在Milne和Witten[12]的方法的基础上构建了投票模式，将其应用于短文本实体链接。Meij等[13]注意到了微博实体链接的重要性，尝试利用多种特征鉴别出微博信息中的概念，并将其链接到相应的维基百科文章。Liu等[14]在Meij等[3]的基础上提出了一个协同推理模型，利用了三类特征进行实体消岐。

3 基于维基百科的指称识别方法

3.1 实体指称-实体字典构建

本文采用了Bunescu与Pasca[5]提出的方法，利用维基百科的链接结构构建了实体指称-实体字典。鉴于实体在维基百科中唯一对应一个页面，后文中交替使用“实体”、“维基百科页面”。对维基百科中的每个实体，抽取如下信息作为该实体的指称：

（1）实体标题。实体标题是实体的唯一标识，所以它是实体指称之一。部分实体的标题中包含消岐信息，需要先将消岐去除，然后将其添加到实体的指称集合中。

（2）消岐页面标题。消岐页面包含一个实体列表，该列表中的所有实体共有一个别名，也即该消岐页面的标题。因此，实体消岐页面的标题被加入到该页面实体列表中所有实体的实体指称集合中。

（3）重定向页面标题。在维基百科中，存在一类页面：重定向页面。重定向页面并不包含实际的实体信息，只有链向另一个页面的链接，表示该页面的标题实际上是指另一个实体。这表达了一种别名关系，也即重定向页面的标题即为其所指向的实体的别名。重定向页面可能指向另一个重定向页面。

（4）锚文本。维基百科中的每篇文章都包含着丰富的链接，它们指向维基百科中的其他文章。这些链接的锚文本及其链向的维基百科实体构成了实体指称-实体之间的一对一关系，进而提供了丰富的别名信息。

在得到所有实体的实体指称后，构建倒排索引作为实体指称-实体字典。输入实体指称，检索得到所有该指称可能指向的实体，构成候选实体集。

3.2 指称检测过程

指称检测过程包括如下步骤：

（1）抽取n元组。从短文本中抽取所有的n元组，作为下一步的输入。

（2）检索实体指称。对每个n元组，在实体指称-实体字典中进行查找。如果一个n元组在字典中被找到，则该n元组作为候选指称进入下一步。

（3）处理可能的重叠问题。不同的指称可能会重叠。例如，对短文本“montclair elementary school”，可以检测到多个指称：montclair，school，elementary school，montclair elementary school；其中elementary school包含school。本文采用最长匹配策略，从头开始遍历给定短文本，找到最长匹配的指称；然后最长匹配得到的指称的下一个字符串开始继续检测后续指称。

（4）基于词性的指称过滤。在大多数情况下，实体指称都是名词。通过检测实体指称在短文本中的词性，我们可以过滤词性为非名词的实体指称。例如，给定短文本“Barack

Obama visit Japan”,我们能够检测到visit是一个指称,可能指向“State Visit”。此时,visit在短文本中的词性不是名词。因此可以从实体指称列表中删除该指称。

4 基于上下文特征的实体链接模型

4.1 实体链接框架

实体链接问题定义如下:给定输入短文本ST,输出为指称序列 $\vec{M}=(m_1,m_2,...,m_n)$ 及其相应的实体序列 $\vec{E}^*=(e^*_1,e^*_2,...,e^*_n)$。

计算公式如公式(1)所示:

$$\vec{E}^*=\arg\max_{\forall \vec{E}\in C(\vec{M})}\alpha\cdot\sum_{i=1}^{n}\vec{\alpha}\cdot\vec{f}(m_i,e_i)+\beta\cdot\sum_{i\neq j}b\cdot g(e_i,e_j)+\gamma\cdot\frac{1}{n-1}s(\sum_{k=1,k\neq i}m_k,e_i) \quad (1)$$

其中:

·C($\vec{M}$)是实体指称序列 $\vec{M}$ 可能对应的所有实体序列的集合;

· $\vec{E}$ 表示一个实体序列,大小与 $\vec{M}$ 相同;

· $\vec{f}(m_i,e_i)$ 是对实体指称 m_i 及其候选实体之一 e_i 间的相似度进行建模的特征向量;

· $\vec{\alpha}$ 是 $\vec{f}(m_i,e_i)$ 的权重向量,其中 $a_k\in(0,1)$,k=1,2,3,4,5,$\sum_{k=1}^{5}a_k=1$;

· $\vec{g}(e_i,e_j)$ 是对两个实体 e_i 和 e_j 间相似度进行建模的特征向量;

· $\vec{b}$ 是 $\vec{g}(e_i,e_j)$ 的权重向量,其中 $b_k\in(0,1)$,k=1,2,3,4,5,$\sum_{k=1}^{4}b_k=1$;

· $\vec{s}(\sum_{k=1,k\neq i}^{n}m_k,e_i)$ 是描述给定指称的候选实体与该指称的上下文指称间一致性的特征向量;

· $\alpha,\beta,\gamma\in(0,1)$ 是系统参数,通过训练数据得到。它们用于平衡上述三组特征之间的比例,$\alpha+\beta+\gamma=1$。

短文本中有时只有一个指称,甚至该短文本本身就是一个指称,此时无法利用任何上下文信息,也无法对候选实体进行消岐。此时,以该指称的常见度为指标,选择常见度最高的候选实体作为该短文本链向的实体。

4.2 特征

本文使用了三种上下文特征:局部特征、实体相似度相关的全局特征、上下文指称相关的全局特征。

4.2.1 局部特征

(1)先验概率。

$$f_1(m_i,e_i)=\frac{count(m_i,e_i)}{\sum_{\forall ek\in C(m_j)}count(m_i,e_i)} \quad (2)$$

其中 $count(m_i,e)$ 表示在维基百科文章中实体指称 m_i 指向实体e的频率。

(2)上下文相似度。

$$f_2(m_i,e_i)=\frac{coocurence\ number}{short\ text\ length} \quad (3)$$

其中“co-occurrence number”是既出现在包含 m_i 的短文本中,又出现实体 e_i 对应的维基百科文章中的词的数目;“short text length”表示包含 m_i 的短文本中词的数目。

(3)编辑距离相似度。

如果等式 Abs(Length(m_i)-Length(e_i))=ED(m_i,e_i)为真,则 $f_3(m_i,e_i)$ 的值为1,否则为0。Abs(.,.)表示给定表达式的绝对值,ED(.,.)表示给定参数在字符级别的编辑距离。

(4)实体指称包含实体标题。

$$f_4(m_i,e_i)=\begin{cases}1 & if\ m_i\ contains\ title\ of\ e_i\\ 0 & otherwise\end{cases} \quad (4)$$

(5)实体标题包含实体指称。

$$f_5(m_i,e_i)=\begin{cases}1 & if\ title\ of\ e_i\ contains\ m_i\\ 0 & otherwise\end{cases} \quad (5)$$

4.2.2 实体相似度相关的全局特征

(1)基于类别的相似度。

$$g_1(e_i,e_j)=\frac{|c(e_i)\cap c(e_j)|}{|c(e_i)\cup c(e_j)|} \quad (6)$$

其中c(e)是实体e相应的维基百科文章的类别的集合。

(2)基于入链的相似度。

$$g_2(e_i,e_j)=\frac{|il(e_i)\cap il(e_j)|}{|il(e_i)\cup il(e_j)|} \quad (7)$$

其中 $il(e)$ 是e相应的维基百科文章的入链的集合。

(3)基于出链的相似度。

$$g_3(e_i,e_j)=\frac{|ol(e_i)\cap ol(e_j)|}{|ol(e_i)\cup ol(e_j)|} \quad (8)$$

其中 $ol(e)$ 是e相应的维基百科文章的出链的集合。

(4)互指特征。

$$g_3(e_i,e_j)=\begin{cases}0 & if\ e_i\nleftrightarrow e_j\\ 0.5 & if\ e_i\rightarrow e_j\ or\ e_j\rightarrow e_i\\ 1 & if\ e_i\rightarrow e_j\end{cases} \quad (9)$$

公式中的箭头表示一个实体到另一个实体的指向关系。该特征帮助检测两个实体是否存在互指关系。

4.2.3 上下文指称相关的全局特征

上下文指称相关性定义如公式(10):

$$s(\sum_{k=1,k\neq i}m_k,e_i)=\frac{\sum_{k=1,k\neq i}^{n}contains(e_i,m_k)}{n-1} \quad (10)$$

如果 m_k 在实体 e_k 相应的维基百科文章中出现,则表达式 $contains(e_i,m_k)$ 等于1,否则等于0。

5 实 验

5.1 实验准备

本文选用2013年12月2日的维基百科转储作为知识库,利用JWPL处理维基百科的定义页面、消岐页面、重定向页面。鉴于评测时使用了Freebase,我们使用Freebase到维基百科的实体映射表来得到相应的Freebase Id。如果一个

实体在维基百科中被识别出来，而没有相应的Freebase实体，则标记为“NIL”，认为没有相应的实体。

本文从维基百科中抽取了标准数据集，以进行模型训练。鉴于本文专注于短文本，我们从维基百科文章从随机抽取了文本片段，作为系统输入，以该文本片段中包含的锚文本及其链向的维基百科实体作为标注数据进行训练。抽取文本片段的原则是：

(1)包含一个或多个锚文本；

(2)抽取第一个锚文本前和最后一个锚文本后的一到两个单词，以作为短文本上下文。

5.2 实验结果

本文在三种情况下进行了评测：

(1)只使用局部特征；

(2)使用局部特征和实体-实体相关度特征；

(3)使用4.2节中提到的三组特征。

我们用F1的均值作为评测指标，也即对每个短文本计算一个F1，然后对所有短文本的F1值取均值，得到最终的F1值。上述指标的计算公式如下：

$$\text{Precision}_i=\frac{|M\bigcap M^*|}{|M|}$$

$$\text{Recall}_i=\frac{|M\bigcap M^*|}{|M^*|}$$

$$\text{F1}_i=\frac{2PR}{P+R}$$

$$\text{F1}=\frac{\sum_{i=1}^{n}F1_i}{n}$$

M表示对给定短文本，系统输出的实体指称-实体对的集合；M*表示给定短文本的标注实体指称-实体对集合。

表1呈现了只使用局部特征时的结果。本文通过逐步添加新特征来评测效果。结果显示，仅使用先验概率就能得到一个合理的F1值，而添加上下文相似度和编辑距离相似度对F1值产生了负面影响。最后两个特征则对整体效果影响不大。这意味着实体指称在词项层面上的上下文信息对消岐的帮助并无帮助。

表1 只使用局部特征的结果

Local Features	Expected F1
P.P.	0.5254
+C.S.	0.5214
+E.D.S.	0.5214
+M.C.T.S.	0.5254
+T.C.M.S.	0.5274

表2 使用局部特征和实体-实体特征的结果

Global Features related to Entity-Entity Similarity	Expected F1
C.b.S.	0.5274
C.b.S.+I.b.S.	0.5294
C.b.S.+O.b.S.	0.5274
C.b.S.+M.R.	0.5374
C.b.S.+I.b.S.+O.b.S.	0.5274
C.b.S.+I.b.S.+O.b.S.+M.R.	0.5374

表2呈现了在局部特征的基础上，逐步添加实体-实体特征的结果。结果表明，基于入链的特征和基于出链的特征并未提升短文本实体链接的性能，而互指特征则较为有效。这表明互指特征能够在一定程度上揭示候选实体间的一致性。

表3呈现了使用所有的三组特征的结果。经过探索上下文指称特征，发现它对性能的影响较小。我们分析了链接结果后发现，上下文指称在短文本中较为稀疏。在大多数短文本中，只能检测到两到三个指称。因此只有一到两个上下文指称可以利用，这可能影响了该特征的效果。

表3 使用所有的三组特征的结果

Context Mention Entity Similarity	Expected F1
+C.M.E.S.	0.5374

6 结 语

本文实现了一种面向短文本的实体链接方法，提出了一些新的特征来揭示实体指称的候选实体间的一致性。在一个公开的web service上的评测，验证了本文方法的有效性。本文没有利用Freebase中的结构化信息，在未来的研究中拟采用其最终的实体别名信息。此外，还会尝试探索利用Freebase实体的其他字段信息。

参考文献

1 Ferragina, P. and U. Scaiella. TAGME: on-the-fly annotation of short text fragments (by wikipedia entities)[C]. In Proceedings of the 19th ACM international conference on Information and knowledge management, ACM,2010.

2 Milne, D. and I.H. Witten. Learning to link with wikipedia[C]. in Proceedings of the 17th ACM conference on Information and knowledge management, 2008: ACM.

3 Meij, E., W. Weerkamp and M. de Rijke. Adding semantics to microblog posts[C]. In Proceedings of the fifth ACM international conference on Web search and data mining,2012: ACM.

4 Liu, X., et al. Entity linking for tweets. in Proceedings of the 51th Annual Meeting of the Association for Computational Linguistics[C]. Association for Computational Linguistics,2013.

5 Bunescu, R.C. and M. Pasca. Using Encyclopedic Knowledge for Named entity Disambiguation[C]. in EACL,2006.

6 Ratinov, L., et al. Local and Global Algorithms for Disambiguation to Wikipedia[C]. In ACL, 2011.

7 Stern, R., et al. A joint named entity recognition and entity linking system[C]//In Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data. Avignon, （下转第160页）

Multidimensional Scaling and VOS[J].J Am Soc Inform Sci Tech,2010,61(12): 2405-2416.
5 马费成,张 勤.国内外知识管理研究热点——基于词频的统计分析[J].情报学报,2006,25(2): 163-171.
6 张果果.图书馆学近十年来研究热点分析及趋势预测[J].新世纪图书馆,2007,(1):13-15.
7 高劲松,刘延芳.国际专利信息研究热点——基于知识图谱的词频分析[J].情报杂志,2010,29(8):36-39.
8 李 品,周金元.中国图情领域2005至2009年研究热点透视——基于国内外期刊发文词频统计的分析[J].国家图书馆学刊,2010,19(4):36-40.
9 刘京京,孙 丽,徐少龙,徐宝祥.2007-2011年我国情报学研究热点分析[J].情报科学,2012,30(4):616-621.
10 王 晴.国内图书馆学学科研究热点透视与特征分布——基于2012年国家社科基金项目及课题指南的统计分析[J].新世纪图书馆,2013,(2):15-19.
11 张士靖,杜 建,周志超.信息素养领域演进路径、研究热点与前沿的可视化分析[J].大学图书馆学报,2010,(5):101-109.
12 赵蓉英,王 静.网络计量学研究热点与前言的知识图谱分析[J].情报学报,2011,30(4): 424-434.
13 袁 红,许秀玲.基于Web of Science 的信息资源管理研究的知识图谱分析[J].情报杂志,2012,31(12):58-64.
14 魏晓峰.国外专利情报领域研究论文的可视化分析[J].情报学报,2012,31(9):934-945.
15 宗乾进,袁勤俭,沈洪洲.基于VOSviewer的2010 年中国图书馆学研究热点分析[J].图书馆,2012,(4):88-90.
16 王晓光, 程齐凯. 基于NEViewer的学科主题演化可视化分析[J]. 情报学报, 2013, 32(9): 900-911.
17 程齐凯,王晓光. 一种基于共词网络社区的科研主题演化分析框架[J]. 图书情报工作, 2013, 57(8): 91-96.
18 Guimera R, Sales-Pardo M, Amaral L A N. Classes of Complex Networks Defined by Role-to-role Connectivity Profiles[J]. Nature physics, 2007, 3(1): 63-69.
19 盛小平.国内知识管理研究综述[J].中国图书馆学报,2002,28(3):60-64.
20 Kleinberg J. Bursty and hierarchical Structure in sStreams [C].Proceedings of the 8th ACM SIGKDD International Conference on Know ledge Discovery and Data Mining, 2002: 91-101.
21 杨 溢,李伟超.1990-2001年我国图书馆学情报学方法论研究统计分析[J].图书馆,2003,(5): 31-34.
22 张 力,唐健辉,刘永涛,韩松涛,潘有能,陈丽君,叶鹰.中外图书情报学研究方法量化比较[J].中国图书馆学报,2012,(3):21-27.

(责任编辑:徐 波)

(上接第147页)

France: Association for Computational Linguistics, 2012.
8 Makris, C., Y. Plegas and E. Theodoridis. Improved text annotation with Wikipedia entities[C]. in Proceedings of the 28th Annual ACM Symposium on Applied Computing, ACM,2013.
9 Nguyen, T.T. and M. Poesio. Entity disambiguation and linking over queries using encyclopedic knowledge[C]. in Proceedings of the 6th Workshop on Analytics for Noisy Unstructured Text Data. AND,2012.
10 Lin, T., Mausam and O. Etzioni. Entity linking at web scale [C]//in Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction. Montreal, Canada: Association for Computational Linguistics, 2012.
11 Medelyan, O., I.H. Witten and D. Milne. Topic indexing with Wikipedia[C]. in Proceedings of the AAAI WikiAI workshop,2008.
12 Kulkarni, S., et al. Collective annotation of Wikipedia entities in web text[C]. in Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining,2009: ACM.
13 Zheng, Z., et al. Learning to link entities with knowledge base[C]. in Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics,2010.
14 Wick, M., et al. A joint model for discovering and linking entities[C]. in Proceedings of the 2013 workshop on Automated knowledge base construction,2013: ACM.

(责任编辑:徐 波)