情 报 学 报 ISSN 1000 - 0135
第 34 卷 第 1 期 105 - 112, 2015 年 1 月

JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION ISSN 1000 - 0135
Vol. 34 No. 1, 105 - 112 January 2015

doi:10.3772/j.issn.1000-0135.2015.001.011

实体链接研究综述[1)]

陆 伟 武 川[1]

(武汉大学信息资源研究中心,信息检索与知识挖掘研究所,武汉 430072)

摘要 实体链接是指将文本中的实体指称链向知识库实体的过程,它能够丰富文本语义信息,在自然语言处理、信息检索等领域有着广泛的应用前景。本文详细介绍了实体链接的概念和步骤,回顾了实体链接发展过程中的相关研究,在总结现有实体链接研究的基础上,对实体链接研究框架、实体链接步骤及各阶段所采用的技术方法进行了综述。最后对实体链接在国际评测会议中的现状进行了总结,分析了未来的研究方向。

关键词 实体识别 实体消歧 实体链接

Literature Review on Entity Linking

Lu Wei and Wu Chuan

(*Center for Studies of Information Resources, Wuhan University, Wuhan* 430072)

Abstract Entity Linking means the process of linking entity mentions to corresponding entities in Knowledge Base. Entity Linking can enrich the semantic information of text, having wide potential application in research areas such as Natural Language Processing and Information Retrieval. This paper introduces the concept and steps of Entity Linking in detail, reviews related research in the development of Entity Linking. Based on existing entity linking research, research framework and methods adopted in each step of Entity Linking is summarized. The current status of Entity Linking in international evaluation conference is introduced, and future research direction is analyzed.

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。为了获取目标信息,我们需要处理大量无用的信息。这一问题源于自然语言表达的多样性,具体说来,即是同一实体可用不同的文本表达(多词一义),而同一文本可能表达多个不同的实体(一词多义)。通过进行实体链接(Entity Linking),也即将文本中的实体指称与知识库中的实体进行链接,能够将文本数据转化为带有实体标注的文本,进而帮助人和计算机理解文本的具体含义。它一方面能够为人带来更好的阅读体验,帮助人们更好地理解所浏览信息的含义,另一方面也能辅助构建以实体为核心的信息网络,推动语义网的发展。语义网的核心是通过为互联网上的文档添加能够被计算机所理解的语义数据,使互联网的信息交流变得更有效率。实体链接对文本的实体

收稿日期:2014 年 7 月 26 日

作者简介:**陆伟**,男,1974 年生,武汉大学信息管理学院,教授,博士生导师,主要研究方向:信息检索、知识挖掘、情报方法与技术等。E-mail:reedwhu@gmail.com。**武川**,男,1989 年生,武汉大学信息管理学院信息管理科学系,博士研究生,主要研究方向:信息检索。E-mail:wu.chuan@whu.edu.cn。

1) 本文系国家自然科学基金面上项目"基于语言模型的通用实体检索建模及框架实现研究"(项目编号:71173164);教育部人文社会科学基地重大项目"面向细粒度的网络信息检索模型及框架构建研究"(项目编号:10JJD630014)的研究成果之一。

标注,使计算机能够对实体而非文本进行处理,从而更好地理解文本的含义。

实体链接是指将文档中出现的文本片段,即实体指称(entity mention)链向其在特定知识库(Knowledge Base)中相应条目(entry)的过程,有时也称命名实体链接(Named Entity Linking)[1]。在实体链接研究中所使用的知识库包括TAP[2]、维基百科[3]、Freebase[4]、YAGO[5]等。随着维基百科的快速发展,以及维基百科转储的发布和持续更新,大多数现有研究都使用维基百科作为实体链接的知识库。实体链接能够利用知识库丰富文本的语义信息,在文本分类和聚类、信息检索[6]、知识库构建[7]等领域都有着重要的理论意义和应用前景。

本文首先介绍实体链接的相关研究基础,阐述实体链接与相关研究的联系。随后,本文根据实体链接步骤对现有研究进行分类,论述实体链接研究各步骤主要采用的方法和技术。通过介绍实体链接在国际评测会议中的现状,呈现通过评测促进实体链接研究的发展趋势。最后总结全文,并指出实体链接研究中存在的不足及未来的研究方向。

2 相关研究基础

实体链接包含两项关键技术:指称识别、实体消歧。指称识别旨在从文档中识别出可能链向知识库中特定条目的实体指称。由于自然语言中普遍存在的一词多义和别名现象,所识别的实体指称在大多数情况下并不能唯一确定其所指向的实体。因此,需要利用实体消歧技术,根据给定实体指称所在上下文,确定其所指向的实体。实体链接主要基于如下研究领域:命名实体识别、词义消歧等。因此,在介绍实体链接的主要技术方法之前,先对上述相关研究进行简要介绍。

2.1 命名实体识别

实体链接的第一步是识别出文档中的实体指称,也即可能指向实体的词或短语;它与命名实体识别较为相似。命名实体识别(Named Entity Recognition)的任务是识别出文本中的人名、地名等专有名称和有意义的时间、日期等数量短语并加以归类[8]。二者的相同之处是都要识别出指代实体的文本片段,而区别在于:

①对实体范围的界定:命名实体识别的目标是识别出所有的实体,而指称识别则致力于识别出在知识库中存在相应条目的实体。

②目标:命名实体识别的最终目标是识别出实体,进而对实体进行归类;而指称识别的目标是以尽可能高的召回率识别出实体,为后续的实体消歧做准备。

传统的命名实体识别所采用的主要技术方法包括:基于规则和词典的方法、基于统计的方法、二者混合的方法等[8]。随着知识库的发展,也出现了利用维基百科进行命名实体识别的方法[9]。实体链接中采用的指称识别方法主要借鉴基于规则和词典的方法,通过从维基百科中抽取实体别名,丰富词典信息,从而提高指称识别的召回率。例如,Toral等[10]通过综合利用维基百科和WordNet,来自动创建和维护地名日志,以进行命名实体识别。也有研究者将知识库信息用于有监督的命名实体识别。例如,Kazama等[11]抽取句子中的实体指称,检索其在维基百科中对应的实体,抽取该实体对应文章的第一句话,并从中抽取类别标签;然后以此为特征采用基于CRF(Conditional Random Field)的命名实体标注器识别实体指称,提高了识别效果。

2.2 词义消歧

词义消歧(Word Sense Disambiguation)是指根据给定关键词的上下文信息,分析其可能的多种含义,判断其在当前上下文中的词义的过程。实体消歧是指给定实体指称及其所在上下文、候选实体,判断其在当前上下文中所指向实体的过程。二者较为相似,词义消歧中的词与词义对应于实体链接中的实体指称与知识库实体。两者的区别主要在于词义与知识库实体间的区别:词义消歧中的候选词义是词汇知识库中的词义,而实体消歧中的候选实体是知识库中的实体。相比之下,实体链接的知识库中包含的信息更为丰富。例如,维基百科是一个自由、免费、内容开放的网络百科全书,参与者来自世界各地,这意味着任何人都可以编辑维基百科中的任何文章及条目[12]。随着维基百科的发展,它既能够通过广泛的参与涵盖对歧义词多种词义的解释,又能通过其非结构化文本中丰富的链接结构提供不同词义的上下文信息。

传统的词义消歧研究采用的方法主要包括:①基于知识的方法;②数据驱动的算法[13]。基于知识的方法依赖从词典中获取的词义信息。例如,Navigli等[14]通过利用丰富的在线词汇知识库,例如WordNet、SemCor,构造了基于图的词义表示法,并

在此基础上提出了 SSI 算法进行词义消歧；Mihalcea[15]在对词义间依赖关系建模时，需要用到已标注词义的语料库或是从纯语料库中获取基于定义的相似度。基于数据的方法则构建基于词义标注语料库的机器学习分类器进行词义消歧。例如，Pedersen[16]构造了一个基于语料库的决策树，根据歧义词附近出现的双字母组进行词义消歧；Gliozzo 等[17]提出了一个有监督的消歧方法，用核方法对词义间的区别进行建模。

知识库的发展也在一定程度上促进了词义消歧的发展，部分研究者探索利用维基百科的标注数据、链接结构等进行词义消歧[13,18-20]。例如，Mihalcea 等[13]受 Lesk 算法启发，通过计算词项各词义对应的维基百科文章与词项所出现的上下文（以该词项所出现的段落为其上下文）之间的上下文重叠度，选择重叠度最大的词义，实现了基于知识的方法；同时，将局部特征和主题特征整合到朴素贝叶斯分类器中，实现了数据驱动的方法。最后，实现了一种投票模式，以平衡两种方法的判断，过滤不正确的预测。

3 主要技术方法

3.1 实体链接框架

传统的实体链接框架大多包括两步：指称识别、实体消歧。虽然有研究者的步骤划分方式不同，但是本质是一样的。例如，Cucerzan[21]虽然将实体链接划分为文档分析、实体识别、实体消歧三步，但是其中文档分析是预处理步骤，主要研究的是实体识别和实体消歧，与传统实体链接框架无本质区别。根据实体链接时是否给定指称，可以将实体链接研究分为两类：未给定指称的实体链接和已给定指称的实体链接；根据上述两个步骤整合方式的不同，可进一步将未给定指称的实体链接研究分为两类：先识别后消歧、识别消歧联立求解。

先识别后消歧类研究是指以顺序方式完成指称识别和实体消歧，用指称识别的输出作为实体消歧的输入，用实体消歧的输出作为最后的指称-实体映射结果。例如，Bunescu 与 Pasca[3]首先进行实体检测，鉴别给定名称是否指向字典中的实体，然后进行实体消歧，对给定指称可能指向的多个实体（候选实体）进行消歧；Mihalcea 与 Csomai[13]先利用所构建的基于维基百科的受控词表进行关键词抽取，得到文档中所有的实体指称，然后综合利用基于知识的方法和朴素贝叶斯分类器进行实体消歧；Medelyan 等[22]用类似的方法构建了受控词表以检测实体，然后用机器学习方法进行消歧。Zhang 等[23]首先利用维基百科搜索引擎对实体别名集进行扩展，以提高实体检测召回率，然后用 SVM（Support Vector Machine）分类器进行消歧。

识别消歧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歧阶段，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，从而在实体消歧阶段对所有可能的“实体指称-实体”组合进行消歧，得到最可能的指称-实体映射。Stern 等[24]构建了一个同时考虑指称识别和实体消歧的系统，一方面指称识别阶段所识别的错误指称可能在实体消歧过程中被鉴别为“非实体”，从而消除部分错误指称；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 等[25]将实体链接与实体发现任务整合为联合实体解析任务，构造了实体及其属性的树形结构，并基于此提出层次实体解析模型，利用马尔可夫链蒙特卡罗（Markov Chain Monte Carlo）算法进行推理，得到实体链接和实体发现结果。Sil 和 Yates[26]提出了对指称识别和消歧联合求解的实体链接框架，首先用基础 NER 系统得到尽可能多的指称，然后用基础链接器进行消歧，最后对所产生的候选指称-实体对进行重排序，得到最终的实体链接结果，从而避免了传统实体链接结构中指称识别错误对实体消歧的影响。

已给定指称的实体链接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歧，因此专注于解决实体链接中的实体消歧问题，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[27]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[28]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。Han 等[29]、Han 和 Sun[30]、Shen 等[31]在定义实体链接问题时，将实体指称及其所在上下文视为问

题的给定信息,仅关注实体消歧过程。

鉴于上述三类研究存在重合之处,本节接下来组织如下。首先,介绍实体指称识别的方法,它在先识别后消歧、识别消歧联立求解两类研究中有所体现。其次,介绍实体消歧方法,它是先识别后消歧、实体消歧两类研究的关注重点。

3.2 指称识别方法

传统的研究大多利用维基百科中的信息构建实体别名词典,得到实体指称与其候选实体之间的一对多映射关系。Bunescu 与 Pasca[3] 指出,在自然语言中,名字和实体之间是多对多的关系,而这种关系体现在维基百科的重定向页面、消歧页面、类别信息和超链接信息中。利用这些信息,Bunescu 与 Pasca[3] 把实体的标题、重定向名称、消歧名称作为实体的名称集合,将从名称到实体的一对多映射关系整合成字典结构,以进行指称识别。在指称识别时,用基于字符串匹配的方法在实体别名词典中搜索给定候选指称;如果该候选指称在字典中有相应条目,则返回该指称的候选实体。Sil 和 Yates[26] 同时利用维基百科和 Freebase 抽取别名,其中 Freebase 中的实体有名字和别名属性,可以通过其内置的 API 获取给定指称的候选实体。

除了上述方法外,研究者也探索了其他不同的利用维基百科信息的方式。Mihalcea 与 Csomai[13] 利用维基百科中文章的标题构造了一个受控词表,并用词表中词的不同形态对该词表进行了扩展;随后使用无监督方法从文档中解析候选指称,并利用其在维基百科中为正文锚文本的概率(等于该词作为锚文本出现次数/该词在维基百科中出现的次数)对候选指称进行排序,得到最可能的指称。Milne 与 Witten[32] 对上述研究进行了扩展,用分类器整合了链接概率、上下文相关度、消歧置信度、一般性、位置和延展等特征,从而识别与文档主题较为相关的指称。Zhang 等[23] 采用了传统的别名抽取方法,同时考虑了拼写错误和不常用别名问题:当对给定别名,用传统别名抽取方法返回的候选实体集为空时,利用维基百科的“Did You Mean”特征和“Wikipedia Search Engine”返回一系列潜在相关的实体;如果排序靠前的实体与文档上下文密切相关,则该指称可能拼写错误或是不常用别名。此外,部分研究关注对实体别名集的扩展。Zhang 等[33] 探索了首字母缩略词扩展问题,首先用模式匹配的方法抽取首字母缩略词的候选别名,然后用 SVM 分类器找出有效的别名。

3.3 实体消歧方法

大多实体链接系统采用有监督的方法进行消歧,包括分类方法、机器学习排序方法、基于图的方法、模型集成方法等。分类方法将实体消歧视为二元分类问题,在训练阶段将指称与其所指向实体间的指称-实体对作为正实例,与该指称其他候选实体间的指称-实体对作为负实例,训练得到分类器;在评测阶段对每个“指称-实体”对进行分类,判断其是否为真。由于可能有多个指称-候选实体对被标记为真,分类方法还需要应用其他技术来选择最有可能为真的指称-候选实体对。Zhang 等[23] 利用词法特征、词-类别对、实体类型等特征,采用 SVM 分类器进行实体消歧;当多个候选实体被标记为所指向实体时,构造基于词袋和实体共现的特征向量,计算指称与候选实体间的余弦相似度,选择得分最高的候选实体。Pilz 和 Paaβ[34] 构造了基于主题的实体表示,以计算实体指称所在上下文与候选实体上下文之间的主题距离,将其作为特征之一利用 SVM 分类器进行二元分类;并在英语、德语、法语维基百科数据集上的实验验证了该方法的效果。

机器学习排序方法是一种利用训练数据自动构建排序模型的有监督方法。在实体链接问题中,训练数据是某一实体指称的所有候选实体在给定上下文中的排序列表,排在第一位的则是该指称在上下文中所指向的实体。Ratinov 等[35] 构建了两大类特征,也即局部特征和全局特征,利用 ranking SVM 进行训练得到排序模型。Shen 等[31] 对实体流行度、语义联系度、语义相似性、全局主题一致性等特征进行线性组合,利用最大间隔技术训练获取各特征的权重,得到最终的排序模型。Zheng 等[28] 分别实现了 pairwise 和 listwise 两种机器学习排序方法,前者将给定查询的结果排序列表中的各项两两组成一对,根据项与项之间的相对位置关系构建成训练实例,将排序问题转化为分类问题进行训练,利用得到的排序感知机进行排序;后者则将给定查询的结果排序列表视为一个训练实例,利用 ListNet 算法训练得到排序模型;在 TAC 2009 数据集上的实验表明这两类方法相比传统的分类方法性能更好。

基于图的方法将文档中的实体指称及其候选实体构建为图结构,利用实体指称间、候选实体间、实体指称与候选实体间的关联关系进行协同推理[29,36,37,38]。Han 等[29] 构建了以实体指称和候选

实体为节点，包含指称-实体、实体-实体关系的图，提出类似主题敏感的 PageRank 协同推理算法，得到实体指称所指向的实体。Hoffart 等[36]在构建指称-候选实体图结构的基础上，用实体流行度、文本上下文相似度等对“实体指称-实体”边加权，用映射实体一致性对“实体－实体”边加权，然后计算对每个指称只包含一条指称-实体边的稠密子图，得到指称-实体映射结果。

集成方法整合多种实体消歧模型，利用不同消歧模型得到多个消歧结果，找到最好的消歧结果。Zhang 等[39]开发了三类方法，包括基于信息检索的方法、机器学习排序方法、分类方法，并将它们整合到一个系统中，通过训练得到一个 SVM 分类器来判断哪一个系统的消歧结果更可信。Chen 和 Ji[40]采用多数投票和加权平均的方法，整合了四种无监督方法和四种有监督方法，实验表明该模型集成方法相比单独的消歧模型效果更好。

无监督方法在实体消歧中也有所应用，包括传统的基于 VSM 的方法和基于信息检索的方法。基于 VSM 的方法首先定义实体指称与候选实体的向量表示，并计算二者间的向量相似度，然后根据相似度得分进行排序，选择得分最高的候选实体。Cucerzan[21]抽取出实体指称的上下文实体和候选实体在维基百科中的上下文实体及其类别标签，构建二者的向量表示；随后通过最大化实体指称与候选实体间的相似度，以及所有实体指称的候选实体间的类别一致性，来选择最可能的候选实体。Gottipati 和 Jiang[41]分别利用 Dirichlet 平滑后的极大似然估计和词的经验分布来估计候选实体和实体指称的语言模型，然后通过计算实体指称与候选实体之间的 KL 距离对候选实体进行排序。

4 实体链接的评测

随着实体链接研究的发展，如何对比不同的实体链接方法也成为研究者关注的重点。Hachey 等[1]指出，虽然 TAC 任务的流行催生了众多具有创新性的实体链接系统[32,42,43]，但是鉴于所有参与者都各自尽可能提高准确率，这些系统各有不同，因此难以判断系统的哪些部分是高性能所必要的、哪些方面需要改进。基于此，Hachey 等[1]构造了统一的实体链接框架，在这一框架下实现了三种方法，也即基于特征线性组合的最优化方法[3]，基于文档级实体消歧的方法[21]，基于文本相似度的方法[44]，分别对实体链接的不同步骤进行详细的分析和评价。有研究者认识到，很少有作者将其方法的源代码公布出来，或者提供其系统的 Web 服务接口，如 REST API 等，这使得不同技术间的比较变得困难[45]。基于此，实体链接研究者构建了开源的实体链接框架[45,46]，便于实现不同的实体链接策略，以进行性能对比。

此外，国际评测会议对实体链接的评测给予了一定的关注，如 INEX 会议中的“Link the Wiki”任务、TAC 会议的 KBP 任务、TREC 会议的 KBA 任务等。INEX 的“Link the Wiki”任务探索了如何在 Wikipedia 文章中自动发现应当被创建链接的文本[47]。该任务通过面向知识库的实体链接丰富知识库的链接结构，促进知识库的快速更新。TAC 的 KBP 任务中，主办方向参赛者提供了一个从维基百科中抽取出来的知识库，每一个查询包括一篇文本文档。参赛者需要开发一个实体链接系统，确定文本中的实体指称是否与知识库中的某些条目相匹配；如果匹配，计算出该实体指称具体是指哪一个条目[48]。TREC 会议自 2012 年起新加入的 KBA 任务则与实体链接密切相关，该任务要求识别出文档流中与特定实体相关的文档，并标注文档与实体的相关程度，包括极为重要的(vital)、有用的(useful)、中性的(neutral)、垃圾(garbage)等[49]。这需要 KBA 系统识别出文档中的实体指称，进行实体消歧，并鉴别出该指称所指向的实体。

5 结 语

实体链接是一个新兴的研究领域，对自然语言处理、信息检索等领域有着重要的潜在价值，是语义网技术的重要基础。本文对实体链接的相关研究领域、实体链接框架及其关键技术、实体链接评测进行了详细的综述。尽管实体链接领域已有多年的研究，但依然存在一些局限：①目前尚没有一个受到广泛认可的实体链接评测框架，不同实体链接研究在问题定义、基本假设、评测数据集等多个维度均存在较大差异，难以进行有效的比较；②目前已有的研究大多专注于英文实体链接，对非英语语言的实体链接关注较少。展望未来，实体链接呈现出如下发展趋势：

(1)通过开放的公共评测促进不同研究之间的直接对比

从国际会议的发展趋势来看，越来越多的国际

会议顶级研讨会以挑战赛的形式举办:给出定义明确的问题,指定数据集、评测指标等,让各参与者在一定时间内利用各自的算法解决问题,最终通过参加研讨会进行交流。例如,由微软和谷歌赞助的实体检测与消歧(Entity Recognition and Disambiguation)挑战赛[50]提供了开放的 web service 接口,供参赛者评测其实体链接系统的性能;参赛者可以参加相应的研讨会,交流经验,共同推动实体链接研究的发展。

(2)跨语言实体链接

现有的实体链接研究大多研究某一语言的实体链接,而未利用数据集之间的跨语言关联,实现跨语言实体链接。某一语言的实体链接能够将文本与该语言的知识库相联系,而跨语言实体链接能够将不同语言的知识库相联系,从而实现跨语言实体间的连接,对各项跨语言自然语言处理任务有极为重要的潜在价值。Wang 等[51]提出了跨语言知识链接问题,利用维基百科中的跨语言链接来挖掘新的跨语言实体关联;同时,利用所提出的链接因子图模型挖掘百度百科与英文维基百科之间的实体关联,取得了较好的效果。

(3)利用实体链接促进自然语言处理任务

随着实体链接性能的不断提升,实体链接对基于文本的自然语言处理任务的潜在价值将逐渐凸显。利用实体链接丰富文本的语义信息,推动自然语言处理任务性能的提升,将会成为未来的发展趋势。

参 考 文 献

[1] Hachey B, Radford W, Nothman J, et al. Evaluating entity linking with wikipedia[J]. Artificial Intelligence, 2013, 194(4): 130-150.

[2] Dill S, Eiron N, Gibson D, et al. SemTag and Seeker: Bootstrapping the Semantic Web via Automated Semantic Annotation[C]// Proceedings of the 12th International Conference on World Wide Web, Budapest, Hungary, 2003: 178-186.

[3] Bunescu R C, Pasca M. Using Encyclopedic Knowledge for Named entity Disambiguation[C]//Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, Trento, Italy, 2006: 9-16.

[4] Bollacker K, Evans C, Paritosh P, et al. Freebase: a Collaboratively Created Graph Database for Structuring Human Knowledge[C]//Proceedings of the 2008 ACM SIGMOD international Conference on Management of Data, Vancouver, BC, Canada, 2008: 1247-1249.

[5] Suchanek F M, Kasneci G, Weikum G. Yago: a Core of Semantic Knowledge [C]//Proceedings of the 16th International Conference on World Wide Web, Banff, Alberta, Canada, 2007: 697-706.

[6] Pantel P, Fuxman A. Jigs and Lures: Associating Web Queries with Structured Entities[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, Portland, Oregon, USA, 2011: 83-92.

[7] Lin T, Mausam Etzioni O. Entity Linking at Web Scale [C]//Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction, Montreal, Canada, 2012: 84-88.

[8] 孙镇,王惠临. 命名实体识别研究进展综述[J]. 现代图书情报技术,2010,193(06): 42-47.

[9] Balasuriya D, Ringland N, Nothman J, et al. Named Entity Recognition in Wikipedia[C]//Proceedings of the 2009 Workshop on The People's Web Meets NLP, Suntec, Singapore, 2009: 10-18.

[10] Toral A, Munoz R. A Proposal to Automatically Build and Maintain Gazetteers for Named Entity Recognition by Using Wikipedia[C]//Proceedings of the Workshop on NEW TEXT Wikis and Blogs and Other Dynamic Text Sources, Trento, Italy, 2006: 56-61.

[11] Kazama J I, Torisawa K. Exploiting Wikipedia as External Knowledge for Named Entity Recognition [C]//Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague, 2007: 698-707.

[12] 维基百科_百度百科[EB/OL]. [2014-12-19]. http://baike.baidu.com/view/1245.htm.

[13] Mihalcea R, Csomai A. Wikify!: Linking Documents to Encyclopedic Knowledge [C]//Proceedings of the sixteenth ACM Conference on Conference on Information and Knowledge Management, Lisboa, Portugal, 2007: 233-241.

[14] Navigli R, Velardi P. Structural semantic interconnections: a knowledge-based approach to word sense disambiguation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(7): 1075-1086.

[15] Mihalcea R. Unsupervised Large-Vocabulary Word Sense Disambiguation with Graph-based Algorithms for Sequence Data Labeling [C]//Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Vancouver, 2005: 411-418.

[16] Pedersen T. A Decision Tree of Bigrams is an Accurate Predictor of Word Sense [C]//Proceedings of the

Second Annual Meeting of the North American Chapter of the Association for Computational Linguistics, Pittsburgh, 2001: 79-86.

[17] Gliozzo A, Giuliano C, Strapparava C. Domain Kernels for Word Sense Disambiguation[C]//Proceedings of the 43rd Annual Meeting of the ACL, Ann Arbor, 2005: 403-410.

[18] Mihalcea R. Using Wikipedia for Automatic Word Sense Disambiguation [C]//Proceedings of NAACL HLT, 2007: 196-203.

[19] Fogarolli A. Word Sense Disambiguation Based on Wikipedia Link Structure [C]//Proceedings of the third IEEE International Conference on Semantic Computing, Berkeley, CA, 2009: 77-82.

[20] Li C, Sun A, Datta A. A generalized method for word sense disambiguation based on wikipedia[J]. Advances in Information Retrieval. 2011, 6611: 653-664.

[21] Cucerzan S. 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, 2007: 708-716.

[22] Medelyan O, Witten I H, Milne D. Topic Indexing with Wikipedia [C]//Proceedings of the AAAI 2008 Workshop on Wikipedia and Artificial Intelligence, Chicago, 2008: 19-24.

[23] Zhang W, Su J, Tan C L, et al. Entity Linking Leveraging: Automatically Generated Annotation [C]//Proceedings of the 23rd International Conference on Computational Linguistics, Beijing, China, 2010: 1290-1298.

[24] Stern R, Sagot B, Bechet F. A Joint Named Entity Recognition and Entity Linking System [C]// Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data, Avignon, France, 2012: 52-60.

[25] Wick M, Singh S, Pandya H, et al. A Joint Model for Discovering and Linking Entities[C]// Proceedings of the 2013 Workshop on Automated Knowledge Base Construction, San Francisco, CA, USA, 2013: 67-71.

[26] Sil A, Yates A. Re-ranking for Joint Named-Entity Recognition and Linking[C]// Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management, San Francisco, CA, USA, 2013: 2369-2374.

[27] Kulkarni S, Singh A, Ramakrishnan G, et al. Collective Annotation of Wikipedia Entities in Web Text [C]// Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, 2009: 457-465.

[28] Zheng Z, Li F, Huang M, et al. Learning to Link Entities with Knowledge Base[C]// Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Los Angeles, California, 2010: 483-491.

[29] Han X, Sun L, Zhao J. Collective Entity Linking in Web Text: a Graph-Based Method[C]//Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, Beijing, China, 2011: 765-774.

[30] Han X, Sun L. An Entity-Topic Model for Entity Linking [C]//Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Jeju Island, Korea, 2012: 105-115.

[31] Shen W, Wang J, Luo P, et al. LINDEN: Linking Named Entities with Knowledge Base via Semantic Knowledge [C]//Proceedings of the 21st International Conference on World Wide Web, Lyon, France, 2012: 449-458.

[32] Milne D, Witten I H. Learning to Link with Wikipedia [C]//Proceedings of the 17th ACM Conference on Information and Knowledge Management, Napa Valley, California, USA, 2008: 509-518.

[33] Zhang W, Sim Y C, Su J, et al. Entity Linking with Effective Acronym Expansion, Instance Selection and Topic Modeling[C]//Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, 2011: 1909-1914.

[34] Pilz A, Paaβ G. From Names to Entities using Thematic Context Distance [C]//Proceedings of the 20th ACM International Conference on Information and Knowledge Management, Glasgow, Scotland, UK, 2011: 857-866.

[35] Ratinov L, Roth D, Downey D, et al. Local and Global Algorithms for Disambiguation to Wikipedia [C]// Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, Portland, Oregon, USA, 2011: 1375 - 1384.

[36] Hoffart J, Yosef M A, Bordino I, et al. Robust Disambiguation of Named Entities in Text [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, Edinburgh, Scotland, UK, 2011: 782-792.

[37] Hachey B, Radford W, Curran J R. Graph-based named entity linking with wikipedia [J]. Web Information System Engineering, 2011, 6997: 213-226.

[38] Guo Y, Che W, Liu T, et al. A Graph-based Method for Entity Linking[C]//Proceedings of the 5th International

Joint Conference on Natural Language Processing, Chiang Mai, Thailand, 2011: 1010-1018.

[39] Zhang W, Sim Y C, Su J, et al. Nus-i2r: Learning a Combined System for Entity Linking[C]//Proceedings of Text Analysis Conference 2010 Workshop, Gaithersburg, Maryland, USA, 2010.

[40] Chen Z, Ji H. Collaborative Ranking: A Case study on Entity Linking[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, Edinburgh, Scotland, UK, 2011: 771-781.

[41] Gottipati S, Jiang J. Linking Entities to a Knowledge Base with Query Expansion[C]//Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, Edinburgh, Scotland, UK, 2011: 804-813.

[42] Ferragina P, Scaiella U. TAGME: On-the-fly Annotation of Short Text Fragments (by Wikipedia Entities)[C]// Proceedings of the 19th ACM International Conference on Information and Knowledge Management, Toronto, Ontario, Canada, 2010: 1625-1628.

[43] Daiber J, Jakob M, Hokamp C, et al. Improving Efficiency and Accuracy in Multilingual Entity Extraction[C]// Proceedings of the 9th International Conference on Semantic Systems, New York, NY, USA, 2013: 121-124.

[44] Varma V, Bysani P, Reddy K, et al. HIT Hyderabad at TAC 2009 [C]//Proceedings of Text Analysis Conference 2010 Workshop, Gaithersburg, Maryland, USA, 2009.

[45] Ceccarelli D, Lucchese C, Orlando S, et al. Dexter: an Open Source Framework for Entity Linking [C]// Proceedings of the Sixth International Workshop on Exploiting Semantic Annotations in Information Retrieval, San Francisco, USA, 2013: 17-19.

[46] Cornolti M, Ferragina P, Ciaramita M. A Framework for Benchmarking Entity-Annotation Systems [C]// Proceedings of the 22nd International Conference on World Wide Web, Rio de Janeiro, Brazil, 2013: 249-259.

[47] Huang D, Xu Y, Trotman A, et al. Overview of INEX 2007 Link the Wiki Track[C]//Pre-Proceedings of the INEX 2007 Conference, Dagstuhl, Germany, 2007: 373-387.

[48] Ji H, Grishman R, Dang H. Overview of the TAC2011 Knowledge Base Population Track[C]//Proceedings of Text Analysis Conference 2011 Workshop, Gaithersburg, Maryland, USA, 2009.

[49] Wang J, Song D, Lin C. BIT and MSRA at TREC KBA CCR Track 2013 [C]//Proceedings of. the Twenty-Second Text REtrieval Conference, Gaithersburg, Maryland, USA, 2013.

[50] Entity Recognition and Disambiguation Challenge[OL]. [2014-03-28]. http://web-ngram. research. microsoft. com/erd2014/Default. aspx

[51] Wang Z, Li J, Wang Z, et al. Cross-Lingual Knowledge Linking Across Wiki Knowledge Bases [C]// Proceedings of the 21st International Conference on World Wide Web, Lyon, France, 2012: 459-468.

(责任编辑 魏瑞斌)