DOI:10.11992/tis.201603048
网络出版地址:http://www.cnki.net/kcms/detail/23.1538.TP.20160513.0958.036.html

一种结合词向量和图模型的特定领域实体消歧方法

汪沛[1],线岩团[1,2],郭剑毅[1,2],文永华[1,2],陈玮[1,2],王红斌[1,2]
(1.昆明理工大学 信息工程与自动化学院,云南 昆明 650500; 2.昆明理工大学 智能信息处理重点实验室,云南 昆明 650500)

摘 要:针对特定领域提出了一种结合词向量和图模型的方法来实现实体消歧。以旅游领域为例,首先选取维基百科离线数据库中的旅游分类下的页面内容构建领域知识库,然后用知识库中的文本和从各大旅游网站爬取到的旅游文本,通过词向量计算工具 Word2Vec 构建词向量模型,结合人工标注的实体关系图谱,采用一种基于图的随机游走算法辅助计算相似度,使其能够较准确地计算旅游领域词与词之间的相似度。最后,提取待消歧实体的背景文本的若干关键词和知识库中候选实体文本的若干关键词,利用训练好的词向量模型结合图模型分别进行交叉相似度计算,把相似度均值最高的候选实体作为最终的目标实体。实验结果表明,这种新的相似度计算方法能够有效获取实体指称项与目标实体之间的相似度,从而能够较为准确地实现特定领域的实体消歧。

关键词:实体消歧;实体链接;Word2Vec;图模型;随机游走;维基百科
中图分类号:TP393 **文献标志码**:A **文章编号**:1673-4785(2016)03-0366-09

中文引用格式:汪沛,线岩团,郭剑毅,等.一种结合词向量和图模型的特定领域实体消歧方法[J]. 智能系统学报, 2016, 11(3): 366-375.
英文引用格式:WANG Pei, XIAN Yantuan, GUO Jianyi, et al. A novel method using word vector and graphical models for entity disambiguation in specific topic domains[J]. CAAI transactions on intelligent systems, 2016, 11(3): 366-375.

A novel method using word vector and graphical models for entity disambiguation in specific topic domains

WANG Pei[1], XIAN Yantuan[1,2], GUO Jianyi[1,2], WEN Yonghua[1,2], CHEN Wei[1,2], WANG Hongbin[1,2]
(1.School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China; 2. Key Laboratory of Intelligent Information Processing, Kunming University of Science and Technology, Kunming 650500, China)

Abstract: In this paper, a novel method based on word vector and graph models is proposed to deal with entity disambiguation in specific topic domains. Take the tourism topic domain as an example. The method firstly chooses the web-pages of the tourism category in a Wikipedia offline database to build a knowledge base; then, the tool Word2Vec is used to build a word vector model with the texts in the knowledge base and texts taken from several tourism websites. Combined with a manual annotation graph, a random walk algorithm based on the graph is used to compute similarity to accurately calculate the similarity between words within the tourism domain. Next, the method extracts several keywords from the background text of the entity to be disambiguated and compares them with the keyword text in the knowledge base that describes the candidate entities. Finally, the method uses the trained Word2Vec model and graphical model to calculate the similarity between the keywords of name mention and the keywords of candidate entities. The method then chooses the candidate entities which have the maximum average similarity to the target entity. Experimental results show that this new method can effectively capture the similarity between name mention and a target entity; thus, it can accurately achieve entity disambiguation of a topic-specific domain.

Keywords: entity disambiguation; entity linking; Word2Vec; Wikipedia; graphical model; random walking

收稿日期:2016-03-19. **网络出版日期**:2016-05-13.
基金项目:国家自然科学基金项目(61262041, 61472168, 61462054, 61562052);云南省自然科学基金重点项目(2013FA030).
通信作者:郭剑毅.E-mail:gjade86@hotmail.com.

实体链接是知识库构建的关键技术之一,其目的是将文本中已经获取到的命名实体链接到已有的

知识库中,实体消岐是实体链接的关键任务。由于海量数据中存在的实体指称通常可以对应到多个命名实体概念,这无疑对实体消歧造成了很大的障碍。实体消歧的任务就是将这些存在歧义的实体指称在众多的候选实体中匹配出对应的目标实体。目前实体消歧任务分为两种类型:实体聚类消歧和实体链接消歧[1],实体聚类消歧就是利用聚类算法来对实体进行消歧,而实体链接消歧则是借助外部知识库,将待消歧命名实体指称链接到外部知识库中对应实体来进行消歧。本文选择用后者来实现特定领域的实体消歧。

实体消歧的本质是计算实体指称项和候选实体的相似度,选择相似度最大的候选实体作为链接的目标实体[2]。针对英文实体消岐,Bunescu 和 Pasca[3]提出了一种基于余弦相似度排序的方法来实现实体消歧。Bagga 和 Gideon[4-5]等将实体指称项的上下文与候选实体的上下文分别表示成 BOW(Bag of words)向量形式,利用向量空间模型实现了人名的消歧。韩先培等[6]提出一种基于图的实体消歧方法,将指称项与实体通过带权的无向图连接起来,从而将指称项与实体、实体与实体间的语义关联通过图的形式表征出来。上述工作主要是对英文的实体消歧,相比较而言,针对中文的实体消歧工作远远落后于英文。在中文的实体消歧领域,王建勇等[7]利用一种基于图的 GHOST 算法,结合 AP 聚类算法进行相似度计算,在人名消歧方面取得了较好的实验结果。怀宝兴等[8]提出了一种基于概率主题模型的命名实体链接方法,在通用领域,通过构建歧义词表,用 LDA 基于语义层面对文档建模和实体消岐;宁博等[9]针对中文命名实体消歧问题提出了一种基于异构知识库的层次聚类方法,将维基百科和百度百科结合起来作为多源知识库,并利用 Hadoop 平台进行层次聚类,从而实现实体消歧。另外,朱敏等[10]提出了一种实体聚类消歧与百度百科词频的同类实体消歧相结合的消歧方法,通过构建同义词表、优化知识库、改进拼音距离编辑算法等方式实现对中文微博的实体消歧。

同样在旅游领域也存在着大量的实体同名现象,在维基百科中“金花”一词有 11 个同名实体,“香格里拉”一词有 12 个同名实体,这无疑对消歧工作产生很大影响,例如,给定两个句子:

1)2014 年,香格里拉县共接待国内外游客 1 080.22万人次。

2)在结束了一天的旅程后我们选择了在香格里拉酒店入住。

在上面的例子中,很明显第一句中的“香格里拉”指的是某旅游胜地,第二句指的是某著名酒店品牌,但是如何让计算机也能将实体指称项准确链接到知识库中具有特定概念的实体仍然是自然语言处理领域研究的热点和难点。

传统的消歧模型难以有效利用能反映领域特有属性的实体词特征。因此,本文针对旅游领域实体间的关系较为复杂的特征,提出了一种结合词向量和图模型的消歧方法,通过提取实体指称项背景文本的若干关键词和候选实体文本的若干关键词,利用训练好的模型对这些关键词分别进行交叉相似度计算,把相似度均值最高的候选实体作为最终的目标实体。

1　相关理论

1.1　词向量

在自然语言处理中,要将自然语言理解的问题转化为机器学习的问题,就需将自然语言的符号数学化,其中最直观和常用的方法是 One-hot 表示法。这种方法将每个词表示为一个很长的向量,其维数是词汇表大小,其中绝大多数元素为 0,只有一个维度的值为 1,这个维度就代表当前的词。

在自然语言处理中,常将 One-hot 表示采用稀疏的方式进行存储,即为每个词分配一个数字 ID。该方法因其简单易用,广泛应用于各种自然语言处理任务中,如 N-gram 模型中就采用这种词向量表示法。但这种表述方法也存在一定问题:其表示的任意两个词之间是孤立的,无法表示这两个词之间的依赖关系,从词向量上看不出两个词是否相关;采用稀疏表示法,在处理某些任务,如构建 N-gram 模型时,会引起维数灾难问题。

而在机器学习领域,一般采用分布式表示(distributed representation)的方法表示词向量,这种表示法最早由 Hinton[11]提出,通常称为 Word Representation。这种方法将词用一种低维实数向量表示,优点在于相似的词在距离上更接近,能体现出不同词之间的相关性,从而反映词之间的依赖关系。同时,较低的维度也使特征向量在应用时有一个可接受的复杂度。因此,新近提出的许多语言模型,如潜在语义分析(latent semantic analysis, LSA)模型、潜在狄利克雷分布(latent dirichlet allocation,LDA)模型以及目前流行的神经网络模型等,都采用这种方法表示词向量[12-13]。

本文利用旅游领域的丰富语料对词向量模型进行训练,从而将抽取的关键词进行向量化表示,用这

若干个关键词向量来表征一篇文档,通过计算关键词向量间的余弦相似度得出它们之间的关联程度,进而得出文档之间的相似度。

1.2 TextRank 算法

同一文档中的大多数词语都是为表达同一主题服务的,它们之间具有一定的语义关系。和词语 W 有语义关系的词语越多,词语 W 越可能是表达文档主题的重要词语,同时和词语 W 有语义关系的词语的重要性也会影响词语 W 的重要性。根据这两个特性,本节引入基于图的排序算法用于抽取多文档关键词。基于图的排序算法是决定图中点重要性的一种方法,它根据全局信息(图的结构)而不是局部信息来对节点排序。其基本理论是"投票",当图中一个点 A 和另一个点 B 之间有连线时,那么点 A 就给点 B 投票,点 B 获得的投票越多,点 B 就越重要;更进一步,投票点 A 的重要性决定了其投票的重要性,因此,点 B 的分数由其获得的投票和给 B 投票的点的分数共同决定。

Mihalcea[14]将在自然语言处理领域中应用的基于图的排序算法称为 TextRank,一般 TextRank 模型可以表示为一个加权的有向图。TextRank 的思想来源于 Google 的 PageRank 算法,通过把文本分割成若干组成单元并建立图模型,利用投票机制对文本中的重要成分进行排序,仅利用单篇文档本身的信息即可实现关键词抽取。本文采用该算法将文档表示为无向图 $G(V,E)$,由点集合 V 和边集合 E 组成,E 是 $V\times V$ 的子集,图中两点 i,j 之间边的权重为 W_j。对于一个给定的点 V_i,$\mathrm{In}(V_i)$ 为指向该点的点集合,$\mathrm{Out}(V_i)$ 为点 V_i 指向的点集合,点 V_i 的分数定义为式(2):

$$WS(V_i) = (1-d) + d \times \sum_{V_j \in \mathrm{In}(V_i)} \frac{w_{ji}}{\sum_{V_k \in \mathrm{Out}(V_i)} w_{jk}} WS(V_j) \tag{2}$$

式中:d 为阻尼因数,取值范围为 0~1,代表从图中某一特定点指向其他任意点的概率。通过这种算法我们可以获得每个词语在文档中的分数,从而可以根据分数大小来进行关键词的排序。

本文利用该算法抽取文档中的关键词,分别用抽取的关键词来表征待消歧实体指称项所在文本和目标实体所在文本。

1.3 随机游走算法

随机游走模型是在 1905 年 Karl Pearson[15]首次提出的一种数学统计模型,它是一连串的轨迹组成的,其中每一次都是随机的。它能用来表示不规则的变动形式,如同一个人酒后乱步,所形成的随机过程记录[16]。它的基本思想是,从一个或一系列顶点开始遍历一张图,在任意一个顶点,遍历者将以概率 $1-\alpha$ 游走到这个顶点的邻居顶点,以概率 α 随机跳跃到图中的任何一个顶点,称 α 跳转发生概率,每次游走后得出一个概率分布,该概率分布刻画了图中每一个顶点被访问到的概率,用这个概率分布作为下一次游走的输入并反复迭代这一过程,当满足一定前提条件时,这个概率分布会趋于收敛,收敛后,即可以得到一个稳定的概率分布。近年来,随机游走算法逐渐开始吸引机器学习研究者的目光,并开始被应用于半监督学习[17-18]、聚类分析[19-21]、图像分割[22]和图的匹配[23]等问题上。与随机游走相关的扩散核也被应用于[24-28]基于核的学习等方面。

由于实体间的关系错综复杂,可以将这种关系抽象为一种图模型,本文在这种图模型上运用随机游走算法可以将实体间的关联程度准确地表征出来。

2 领域实体消歧

2.1 系统流程

本文提出的方法由 4 个模块构成分别为关键词提取模块、词向量模块、图模型模块和空实体判断模块。

在关键词提取模块中,分别利用 TextRank 算法提取出待消歧的实体指称所在的背景文本的若干关键词和候选实体对应的知识库描述文本的若干关键词,这里提取的两组关键词用于后面的相似度计算。

在词向量模块中,抽取维基百科离线数据中旅游分类下的页面信息构建领域知识库,由于维基百科中包含大量的结构化信息,取该知识库的摘要信息作为语料对词向量模型进行训练,这时,领域实体都能通过该模型表征为一个向量,从而实现关键词之间的相似度计算。

在图模型模块中,人工构建一个领域实体关系图谱,通过在该图谱上的随机游走算法实现关键词之间相似度的计算。

在空实体判断模块中,从待消歧实体指称所在的文本中抽取若干关键词和从候选实体所在文本中抽取的关键词分别用本文提出的图模型与词向量方法相结合进行交叉相似度计算取平均值,选择其中最大的相似度平均值,因为计算结果所对应的目标实体未必在我们的知识库中存在,这时通过比对该平均值与通过大量实验确定的空实体阈值 λ 的大小,如果大于该阈值 λ,则该实体为目标实体,如果

小于 λ,则认为该实体指称在知识库中没有与之对应的目标实体,即空实体。

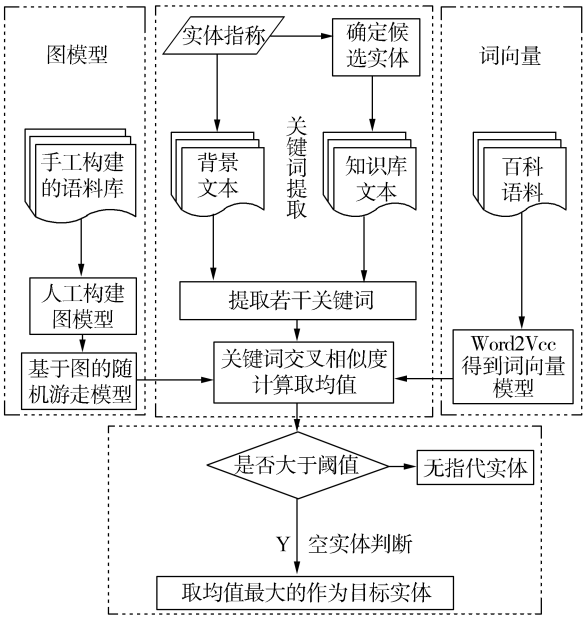


图 1　系统总体框架

Fig.1　Overall framework of system

2.2　关键词提取

关键词提取模块分为两个步骤:候选实体获取和关键词提取。候选实体获取实质上就是罗列出所有可能是待消歧的实体指称项的目标实体,由于中文语义的多样性,一个词通常有多种表达方式,同样一个实体也可能有多种形式,例如,在维基百科的重定向页面中,“驴友”与“背包客”指的是同一个实体,“虫草”与“冬虫夏草”指的也是同一实体。针对这种情况,利用维基百科离线数据库提供的 3 个 SQL 文件即可得到所有重定向的同义词,并且能得到他们对应的页面信息和链接信息。

关键词提取即在确定候选实体后,从待消歧实体所在文本中抽取 n 个关键词,然后再从所有候选实体在知识库中对应的文本中分别抽取 n 个关键词。这样做是因为本文中相似度计算的前提是假设待消歧背景文本与知识库中对应文本的主题一致,在这个前提下,本文消歧任务实质已经转变为计算待消歧实体指称所在背景文本与知识库中候选实体对应文本之间的相似度。分别抽取两个文本各 n 个关键词,这里采用 TextRank 算法抽取权重最高的 n 个关键词,具体计算方法参照本文 1.2 节。根据词与词之间在规定窗口大小内相互进行“投票”计算出每个词在文档中的权重,在使用 TextRank 算法计算图中点的权重时,需要给图中的点指定任意的初值并递归计算直到某个词语分数收敛,收敛后每个点都获得一个分数,代表该点在图中的重要性,也就是该词语在文档中的重要性。表 1 为利用该算法确定的待消歧实体文本和对应的 3 个候选实体文本中的关键词,待消歧实体和候选实体 1 指的是香格里拉(景点名),候选实体 2 指的是香格里拉(酒店名),候选实体 3 指的是香格里拉(城市名)。

表 1　用 TextRank 抽取的关键词

Table 1　Keywords extracted by TextRank

待消歧实体文本	候选实体 1 文本	候选实体 2 文本	候选实体 3 文本
香格里拉	香格里拉	酒店	藏族
心中	云南省	香格里拉	香格里拉
出发	民族	亚洲	民族
寻找	景点	集团	扎西
位置	旅游	饭店	传奇
稻城	香格里拉县	商贸	成长

2.3　词向量的训练和应用

Word2Vec 是 Google 在 2013 年推出并开源的一款将词表征为实数值向量的高效工具,其利用深度学习的思想,可以通过训练,把对文本内容的处理简化为 K 维向量空间中的向量运算,而向量空间上的相似度可以用来表示文本语义上的相似度。Word2Vec 输出的词向量可以被用来做很多 NLP 相关的工作,比如聚类、找同义词、词性分析等。如果换个思路,把词当做特征,那么 Word2Vec 就可以把特征映射到 K 维向量空间,可以为文本数据寻求更加深层次的特征表示,本文将 K 值选定为 200 维。

本文主要利用该工具来实现指称项与目标实体间的相似度计算,为了提高实验在旅游领域的准确率,在选取训练语料时有针对性地选取旅游领域文本,这样就最大程度避免其他领域文本对词向量模型的精准度产生影响,本文一方面采用维基百科的旅游分类下的文本来作为训练词向量模型的语料,同时还加入了在各大旅游网站爬取的新闻语料。训练完成后的模型能够比较准确地计算两个旅游领域词汇的相似度,效果比较理想。如表 2 所示为利用该工具计算出的背景文本中关键词“香格里拉”与知识库中目标实体文本的 7 个关键词之间的相似度,从图中可以发现其与“景点”、“旅游”等词语的相似度要明显高于“民族”、“比重”,这与现实世界中它们之间的语义关联程度相一致。通过词向量计算处理,进一步加强了实体词的领域相关性,有助于

后续环节的相似度计算。

表 2 用词向量计算出的关键词之间相似度

Table 2 Similarity between keywords calculated by Word2Vec

关键词 1	关键词 2	相似度
香格里拉	香格里拉	1.0
香格里拉	云南省	0.253 319 38
香格里拉	民族	0.117 823 526
香格里拉	景点	0.496 713 4
香格里拉	旅游	0.429 819 64
香格里拉	香格里拉县	0.496 569 78
香格里拉	比重	0.009 633 713

2.4 图模型的构建和应用

维基百科是目前世界上最大的在线百科全书，其内容每天都会由世界各地的志愿者进行编辑和更新，有着很好的时效性，另外，维基百科的页面包含有类别信息、重定向信息、外部链接信息等，这些信息无形中为实体之间建立了语义上的关联，所以本文选择维基百科作为实体消歧的知识库。由于本文是针对特定领域，本文抽取“旅游”分类信息下的所有页面作为最终的知识库来源，这样我们在很大程度上实现了消歧，例如，“香格里拉(科幻小说)”和“香格里拉(电视剧)”就自然不在知识库中，也就在一定程度上缩小了候选实体的范围。在此基础上，我们搭建了一个领域实体关系标注平台，利用图数据库 Neo4j 存储数据，这种图数据库与传统的关系型数据库相比能够更准确有效地表示各个数据项之间的复杂关系，将从维基百科中抽取到的领域实体导入该平台的图数据库，通过人工标注的方式构建了一个实体与实体之间的关系图谱，目的是通过利用在该图谱上的随机游走算法辅助计算关键词之间的相似度，目前该平台已经拥有 13 956 个实体，8 127 对关系。图 2 是部分实体及其之间的关系。

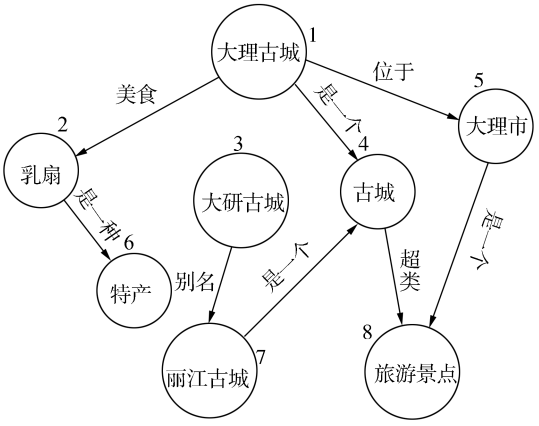
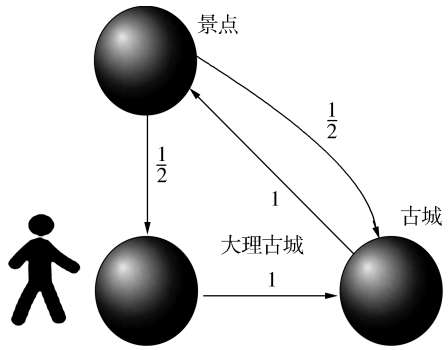


图 2 部分实体关系图谱

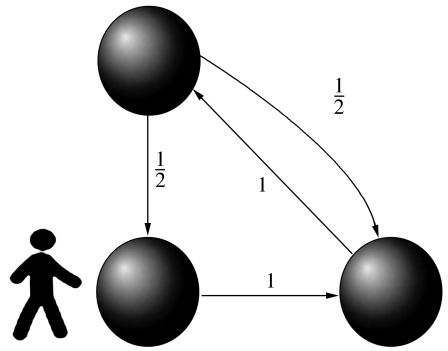
Fig.2 Part of the entity relationship mapping

为了提高关键词之间相似度计算的准确率，我们在词向量的基础上加入了利用图模型计算的相似度来综合衡量关键词之间的相似度，下面将重点介绍一种用来计算相似度的基于图的随机游走算法。

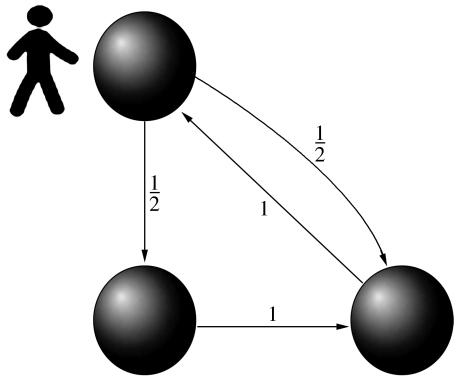
由于目前我们已经人工手动搭建了一个领域实体关系库，图 2 所示的就是一个典型的云南旅游领域相关实体的部分关系图谱，从图中我们认为“大理古城”与“大理市”之间的相似度要高于“乳扇”与“大理市”之间的相似度，因为前两者之间是“位于”的关系直接相连，而后两者之间是通过“大理古城”这个中间实体相联系起来的，所以相比较而言，“乳扇”与“大理市”之间的联系就要弱得多，同样，“特产”与“大理古城”之间的相似度要比“旅游景点”与“大理古城”之间的相似度要弱得多，因为后者之间的路径更多，这些都与现实中实体之间的联系密切程度相一致，而基于图的随机游走算法能将这种实体之间的联系定量地表示出来。



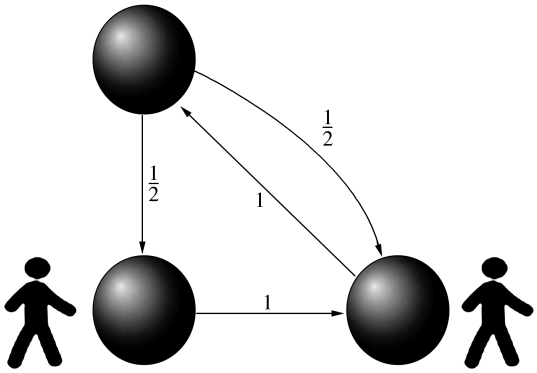
(a)从起始点出发



(b)到达第 2 个顶点后选择下一个目标



(c)到达第 3 个顶点后有两个选择



(d)依概率到达下一个目标

图 3　随机游走原理图

Fig.3　Schematic diagram of random walk

如图 3 所示我们从节点“大理古城”出发,在 3 个结点组成的图上随机游走,边上数字是转移概率,图 3(a)~(d)分别显示 4 种时刻的状态。图 3(a)中“大理古城”和“古城”之间只有一个单向的关系,箭头的方向表示关系的方向,所以“大理古城”到“古城”之间的关系在矩阵中表示为 1,图 3(c)中“景点”和其他两个实体间均有一个单向的关系,所以“景点”和另外两个实体之间的关系在矩阵中都表示为 1/2。由于实体间的关系错综复杂,可以将这种关系抽象为一种图模型,本文在这种图模型上运用随机游走算法可以将实体间的关联程度准确地表征出来。

取中心实体周围距离最近的 p 个实体构建实体关系图谱,如果待计算的两个实体都不在这个图谱中则将相似度设为 0,反之,则用随机游走算法计算相似度。假设以图 2 中的“大理古城”为中心,则令初始矩阵 $\boldsymbol{A}^{\mathrm{T}}=[1\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0]$,“大理古城”经过一步游走到 2、4、5 等 3 个节点的概率都是 0.33,这种转移概率表示成一个 $p\times p$ 的矩阵,即为 8×8 的矩阵 M,对于 α 的确定,我们参照文献[28],将 α 的值设为 0.5。

具体算法流程如下:

1)给定初始化矩阵 $\boldsymbol{A}$,并令 $\boldsymbol{B}=\boldsymbol{A}$;

2)根据图中实体间的转移概率,生成矩阵 M;

3)计算 $\boldsymbol{C}=\alpha\cdot\boldsymbol{M}\cdot\boldsymbol{B}+(1-\alpha)\boldsymbol{A}$;

4)令 $\boldsymbol{B}=\boldsymbol{C}$;

5)重复步骤 3)、4),直到 $\boldsymbol{C}$ 达到稳定状态或者迭代次数超过某个阈值。

$$\begin{bmatrix}0.5\\0.165\,0\\0\\0.165\,0\\0.165\,0\\0\\0\\0\end{bmatrix}=0.5\times\begin{bmatrix}0&0.5&0&0.33&0.5&0&0&0\\0.33&0&0&0&0&1&0&0\\0&0&0&0&0&0&0.5&0\\0.33&0&0&0&0&0&0.5&0.5\\0.33&0&0&0&0&0&0&0.5\\0&0.5&0&0&0&0&0&0\\0&0&1&0.33&0&0&0&0\\0&0&0&0.33&0.5&0&0&0\end{bmatrix}*\begin{bmatrix}1\\0\\0\\0\\0\\0\\0\\0\end{bmatrix}+0.5\times\begin{bmatrix}1\\0\\0\\0\\0\\0\\0\\0\end{bmatrix}$$

$\boldsymbol{C}$　　α　　$\boldsymbol{M}$　　$\boldsymbol{B}$　　$1-\alpha$　　$\boldsymbol{A}$

12 次迭代后矩阵 $\boldsymbol{C}$ 达到稳定状态,概率分布为

$$\begin{bmatrix}0.5716\\0.1678\\0.0052\\0.1107\\0.105\,4\\0.0270\\0.0209\\0.044\,6\end{bmatrix}=0.5\times\begin{bmatrix}0&0.5&0&0.33&0.5&0&0&0\\0.33&0&0&0&0&1&0&0\\0&0&0&0&0&0&0.5&0\\0.33&0&0&0&0&0&0.5&0.5\\0.33&0&0&0&0&0&0&0.5\\0&0.5&0&0&0&0&0&0\\0&0&1&0.33&0&0&0&0\\0&0&0&0.33&0.5&0&0&0\end{bmatrix}*\begin{bmatrix}0.571\,6\\0.167\,8\\0.005\,2\\0.110\,7\\0.105\,4\\0.027\,0\\0.020\,9\\0.044\,6\end{bmatrix}+0.5\times\begin{bmatrix}1\\0\\0\\0\\0\\0\\0\\0\end{bmatrix}$$

所以在经过多次迭代运算后,这种基于图的概率分布会趋向于一个稳定值,从 $\boldsymbol{C}$ 矩阵我们可以看出结点 1 与结点 2、结点 4、结点 5 的相似度较大,分别为 0.167 8、0.110 7、0.105 4,对应着图 2 中“大理古城”与“乳扇”、“古城”、“大理市”的相似度,而结点 1 与结点 3 的相似度最小,只有0.005 2,这与图 2 中展示的实际情况也比较相符,如此一来,我们就将这种图上的结点间的相似度实现了量化,并且实际效果与现实情况较为一致,可见该算法在辅助计算相似度时的实用价值。

2.5 相似度计算

2.2 节中已经确定出权重最高的 n 个关键词，在此基础上分别用这 n 个关键词来作为文本的特征模型：

$$v_q = (w_{q1}, w_{q2}, \cdots, w_{qn}) \quad v_e = (w_{e1}, w_{e2}, \cdots, w_{en})$$

式中：v_q 为带消歧实体指称所在背景文本的特征模型，v_e 为知识库中候选实体对应文本的特征模型，w 为利用 TextRank 算法得出的文本关键词，词与词之间的相似度用向量间的余弦值表示，具体计算如式(3)所示：

$$\mathrm{Sim}(q,e) = \alpha \cdot \frac{\boldsymbol{w}_e \cdot \boldsymbol{w}_q}{|\boldsymbol{w}_e| \cdot |\boldsymbol{w}_q|} + (1-\alpha)\,\mathrm{sim}(q,e) \tag{3}$$

式中加号的前半部分是利用词向量求关键词之间的相似度，后半部分是利用基于图的随机游走算法计算的关键词之间的相似度，其中 $\boldsymbol{w}_q$ 为背景文本中关键词的词向量，$\boldsymbol{w}_e$ 为候选实体对应文本关键词的词向量，通过参数 α 来决定这两种相似度计算方法的权重，这样我们就能得到背景文本与候选实体文本关键词两两进行计算后的相似度，一共能得到 n^2 个 $\mathrm{Sim}(q,e)$，然后对它们求均值，用这个均值来表示两篇文档的相似度，具体公式如式(4)所示：

$$\mathrm{Average} = \frac{\sum_{i}^{n}\sum_{j}^{n}\mathrm{Sim}(q_i, e_j)}{n^2} \tag{4}$$

最后利用上面计算的背景文本与候选实体文本的相似度，来对候选实体进行消歧，相似度最大的即为目标实体。

2.6 空实体判断

由于知识库不可能做到非常全面，实际消歧过程中往往会出现空链接的现象，即待消歧的实体指称项在知识库中并没有与之对应的目标实体。这种情况有两种可能：1)在获取候选实体阶段通过直接匹配和同义词匹配两种方式都没有匹配到与之对应的候选实体；2)在获取候选实体阶段匹配到至少一个候选实体，但是实际上这个候选实体并不是语义相关的。

第 1 种下情况将其直接返回 NIL。第 2 种情况下通过设定一个阈值 λ，如果最终的相似度小于 λ，则认为实体指称项与候选实体语义上不相关，同样返回 NIL。

3 实验验证与结果分析

本文利用维基百科的离线数据库实现对词向量模型的训练，并在一个小型测试集上进行测试。本文通过两个实验对所提出的方法进行了验证，实验一通过对关键词在不同个数下的对比试验，确定出消歧准确率在关键词个数为多少时达到最高；在实验二中加入了对空实体的判断，通过对空实体阈值 λ 的不断调优得出在不同关键词个数下准确率是否有所提升，提升的程度如何以及最终的消歧准确率对比。

实验步骤如下：

1)利用旅游领域的百科语料对词向量空间模型进行训练；

2)利用 2.2 中的方法在待消歧实体指称所在的文本中抽取 n 个关键词；

3)用同样的方法在所有候选实体所在文本中分别抽取 n 个关键词；

4)利用 2.3 和 2.4 中包含有丰富语义信息的模型将上面两步中的 n 个关键词分别进行交叉相似度计算，并且取平均值；

5)选取其中相似度平均值最大的作为最终目标实体。

3.1 语料的获取和模型的训练

由于本文需要利用 Word2Vec 工具对词向量空间模型进行训练，所以采用了维基百科 2014 年 12 月的中文离线数据库，并提取其中的旅游分类下的页面信息，共计 71 208 条。将这些语料经过预处理，提取页面中的摘要信息，形成一篇篇的文本。接着编制爬取程序从国内几个著名的旅游网站爬取了相关的文本，与维基文本结合，共计 75 016 篇。作为本次试验的训练语料。经过训练得到一个 131M 的实验模型文件 vectors.bin。

利用基于图的随机游走算法计算相似度时，图模型的构建是至关重要的一个环节，我们将上一个环节中得到的领域实体语料通过人工标注的方式构建了一个领域实体关系图谱，通过在这张领域实体关系网络上的随机游走算法来辅助计算关键词之间的相似度。

3.2 测试集的选取

实验所用来测试的是一个小规模的测试集，本文从某旅游网站上爬取了 596 篇旅游攻略作为测试文本，通过观察发现并不是每一篇文本中都包含有存在歧义的实体指称，所以通过人工选取符合消歧条件的文本共计 135 篇，从每一篇文本中人工标记出存在歧义的旅游领域实体指称，并将其指向的知识库中对应实体标注出来用于对实验结果进行

验证。

3.3　实验结果与分析

实验 1　本文就两种相似度计算方法的权重值 α 的确定采用了一种自动调优的方法,我们的问题可以简化为 $\boldsymbol{C}=\alpha\cdot\boldsymbol{A}+(1-\alpha)\cdot\boldsymbol{B}$,要使实验效果相对较好就是要使关键词之间的相似度值差异较大,即使 $\boldsymbol{C}$ 的方差达到最大值,这时问题又可以简化为求使得 $\boldsymbol{C}$ 方差最大时 α 的值。先给定 α 一个初始值 0.5 ,由于基于图的方法在本文中只是起到辅助作用,所以将 α 每次增加 0.05,记录取每个不同 α 值的情况下 $\boldsymbol{C}$ 的方差值,实验结果如图 4 所示。

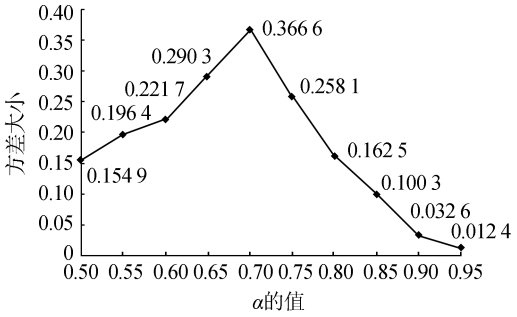


图 4　不同 α 值时对应的样本方差

Fig.4　The sample variance of different α values

根据实验结果可以得出,当 α 的值取 0.7 时,相似度样本的方差达到最大值 0.366 6,说明此时关键词之间的相似度分布最为稀疏,相似度值差异最大。

实验 2　本文就关键词个数 n 的确定做了 6 组实验,分别测试 n 在取 5、6、7、8、9、10 时对消歧准确率的影响,实验结果如图 5 所示。

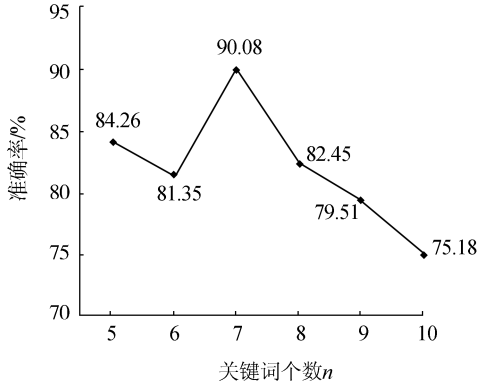


图 5　不同关键词个数时系统准确率

Fig.5　Accuracy of different number of keywords

根据实验结果发现,针对本文的测试集和知识库,将关键词个数 n 定为 7 的时候准确率达到最大值 90.08%。但是考虑到该知识库其实并不完备,并非所有的实体指称项在知识库中都有相应的目标实体与之对应,即所有的候选实体可能并不是目标实体,而判断空实体时只考虑了在知识库中是否存在,不存在则返回 NIL,如果存在,本文的方法是取相似度均值最大的候选实体,这就不可避免地增加了系统的误差。

实验 3　针对以上这种空实体,本文通过大量的实验,针对不同的关键词个数分别对其空实体阈值 λ 进行调优,最终结果如表 3 所示。

表 3　调优后的空实体阈值 λ

Table 3　The empty entity threshold λ after optimized

关键词个数 n	空实体阈值 λ	准确率/%
5	0.143 6	90.31
6	0.119 3	86.75
7	0.110 7	92.27
8	0.098 8	83.57
9	0.082 5	77.63
10	0.061 1	71.98

在加入空实体阈值 λ 后,系统准确率在关键词个数为 5、6、7、8 时都有不同程度的提高,在 9、10 时反而出现下降的趋势。经过分析发现,准确率的提升程度随着关键词的增多而下降,这是因为关键词的权重是逐渐递减的,个数的增加会使相似度均值发生不同程度的下降,这会对空实体阈值 λ 的确定造成一定影响,在判断空实体的时候容易将相似度均值较低的目标实体判断为空实体,这就反而降低了系统的准确率。实验结果如图 6 所示。

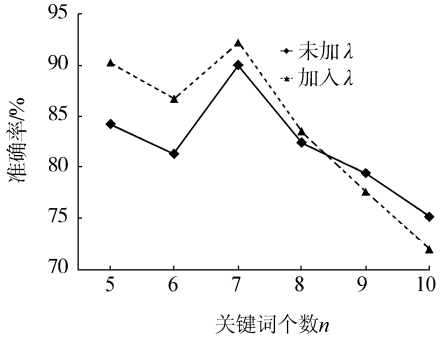


图 6　加入空实体阈值后的结果比较

Fig.6　Comparison with the result after adding an empty entity threshold

实验结果表明,在关键词个数取 7,并且加入空实体阈值判断后,系统达到了最大的准确率 92.27%,这说明本文提出的方法能够在中文旅游领域实现较为理想的消歧结果,在与现有的主流消歧方法的对比中,优势较为明显。

表 4 与主流消歧方法的比较

Table 4 Comparison with other mainstream method of disambiguation

方法名	准确率
Wikify	0.60
Cucerzan	0.71
M&W	0.83
CSAW	0.87
本文方法	0.92

4 结束语

本文针对特定领域消歧的特点,提出了一种结合词向量与图模型计算的方法,实现了特定领域实体消歧。试验结果表明,相比已有的消歧方法,本文提出的方法能在特定领域实体消歧上取得较为理想的结果。下一步的工作在关键词个数的选择方面将考虑根据词的权重动态来选择;另外对于空实体的判断方法还有待改进。本文实验结果也将应用到其他特定领域实验验证。

参考文献:

[1]赵军. 命名实体识别、排歧和跨语言关联[J]. 中文信息学报, 2009, 23(2): 3-17.
ZHAO Jun. A survey on named entity recognition, disambiguation and cross-lingual coreference resolution[J]. Journal of Chinese information processing, 2009, 23(2): 3-17.

[2]赵军, 刘康, 周光有, 等. 开放式文本信息抽取[J]. 中文信息学报, 2011, 25(6): 98-110.
ZHAO Jun, LIU Kang, ZHOU Guangyou, et al. Open information extraction[J]. Journal of Chinese information processing, 2011, 25(6): 98-110.

[3]BUNESCU R C, PASCA M. Using encyclopedic knowledge for named entity disambiguation[C]//Proceedings of the 11st conference of the european chapter of the association for computational linguistics. Trento, Italy, 2006: 9-16.

[4]BAGGA A, BALDWIN B. Entity-based cross-document coreferencing using the vector space model[C]//Proceedings of the 17th international conference on computational linguistics-volume 1. association for computational linguistics. Montreal, Canada, 1998: 79-85.

[5]MANN G S, YAROWSKY D. Unsupervised personal name disambiguation[C]//Proceedings of the 7th conference on natural language learning at HLT-NAACL 2003-volume 4. Sapporo, Japan, 2003: 33-40.

[6]HAN Xianpei, SUN Le. A generative entity-mention model for linking entities with knowledge base[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Stroudsburg, PA, USA, 2011: 945-954.

[7]FAN Xiaoming, WANG Jianyong, PU Xu, et al. On graph-based name disambiguation[J]. Journal of data and information quality (JDIQ), 2011, 2(2): 10.

[8]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.
HUAI Baoxing, BAO Tengfei, ZHU Hengshu, et al. Topic modeling approach to named entity linking[J]. Journal of software, 2014, 25(9): 2076-2087.

[9]宁博, 张菲菲. 基于异构知识库的命名实体消歧[J]. 西安邮电大学学报, 2014, 19(4): 70-76.
NING Bo, ZHANG Feifei. Named entity disambiguation based on heterogeneous knowledge base[J]. Journal of Xi'an university of posts and telecommunications, 2014, 19(4): 70-76.

[10]朱敏, 贾真, 左玲, 等. 中文微博实体链接研究[J]. 北京大学学报:自然科学版, 2014, 50(1): 73-78.
ZHU Min, JIA Zhen, ZUO Ling, et al. Research on entity linking of chinese micro blog[J]. Acta scientiarum naturalium universitatis pekinensis, 2014, 50(1): 73-78.

[11]HINTON G E. Learning distributed representations of concepts[C]//Proceedings of the 8th annual conference of the cognitive science society. Amherst, USA, 1986: 1-12.

[12]张剑, 屈丹, 李真. 基于词向量特征的循环神经网络语言模型[J]. 模式识别与人工智能, 2015, 28(4): 299-305.
ZHANG Jian, QU Dan, LI Zhen. Recurrent neural network language model based on word vector features[J]. Pattern recognition and artificial intelligence, 2015, 28(4): 299-305.

[13]MIKOLOV T, CHEN Kai, CORRADO G, et al. Efficient estimation of word representations in vector space[C]//Proceedings of the International Conference on Learning Representations. Scottsdale, Arizona, 2013: 1388-1429.

[14]MIHALCEA R, TARAU P. TextRank: bringing order into texts[C]//Proceedings of EMNLP-04and the 2004 Conference on Empirical Methods in Natural Language Processing. Spain, 2004: 404-411.

[15]PEARSON K. The problem of the random walk[J]. Nature, 1905, 72(1865): 294.

[16]郑伟, 王朝坤, 刘璋, 等. 一种基于随机游走模型的多标签分类算法[J]. 计算机学报, 2010, 33(8): 1418-1426.
ZHENG Wei, WANG Chaokun, LIU Zhang, et al. A multi-label classification algorithm based on random walk model[J]. Chinese journal of computers, 2010, 33(8):

1418-1426.

[17]SZUMMER M, JAAKKOLA T. Partially labeled classification with Markov random walks[C]//Advances in neural information processing systems (NIPS). Cambridge, 2002, 14: 945-952.

[18]ZHOU Dengyong. Learning from labeled and unlabeled data on a directed graph[C]//Proceedings of the 22nd international conference on machine learning. New York, USA, 2005: 1036-1043.

[19]TISHBY N, SLONIM N. Data clustering by Markovian relaxation and the information bottleneck method[C]//Proceedings of Neural Information Processing Systems. Vancouver, Canadian, 2000: 640-646.

[20]HAREL D, KOREN Y. On clustering using random walks[M]//HARIHARAN R, VINAY V, MUKUND M. Foundations of software technology and theoretical computer science. Berlin Heidelberg: Springer, 2001: 18-41.

[21]LUXBURG U V. A tutorial on spectral clustering[J]. Statistics and computing, 2007, 17(4): 395-416.

[22]GRADY L. Random walks for image segmentation[J]. IEEE transactions on pattern analysis and machine intelligence, 2006, 28(11): 1768-1783.

[23]GORI M, MAGGINI M, SARTI L. Exact and approximate graph matching using random walks.[J]. IEEE transactions on Pattern analysis and machine intelligence, 2005, 27(7): 1100-1111.

[24]KONDOR R I, LAFFERTY J. Diffusion kernels on graphs and other discrete structures[C]//Proceedings of the 19th international conference on machine learning. Sydney, Australia, 2002: 315-322.

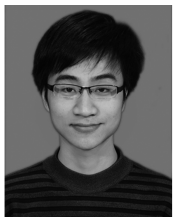
[25]BELKIN M, NIYOGI P. Laplacian eigenmaps for dimensionality reduction and data representation[R]. Chicago, USA: University of Chicago, 2002.

[26]LAFFERTY J, LEBANON G. Information diffusion kernels[C]//Advances in neural information processing systems. Cambridge, 2002: 375-382.

[27]SMOLA A J, KONDOR R. Kernels and regularization on graphs[M]//Learning theory and kernel machines. Berlin Heidelberg: Springer, 2003: 144-158.

[28]HU Jian, WANG Gang, LOCHOVSKY F, et al. Understanding user's query intent with Wikipedia[C]//Proceedings of the 18th International Conference on World Wide Web. Beijing, China, 2009: 471-480.

作者简介:



汪沛,男,1990 年生,硕士研究生,主要研究方向为自然语言处理、信息抽取。

线岩团,男,1981 年生,博士研究生,主研方向为自然语言处理、信息抽取、机器翻译、机器学习。

郭剑毅,女,1964 年生,教授,主要研究领域为自然语言处理、信息抽取、机器学习。