软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2014,25(9):2076−2087 [doi: 10.13328/j.cnki.jos.004642]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

一种基于概率主题模型的命名实体链接方法*

怀宝兴, 宝腾飞, 祝恒书, 刘 淇

(中国科学技术大学 计算机科学与技术学院,安徽 合肥 230027)

通讯作者: 刘淇, E-mail: qiliuql@ustc.edu.cn

摘 要: 命名实体链接(named entity linking,简称 NEL)是把文档中给定的命名实体链接到知识库中一个无歧义实体的过程,包括同义实体的合并、歧义实体的消歧等.该技术可以提升在线推荐系统、互联网搜索引擎等实际应用的信息过滤能力.然而,实体数量的激增给实体消歧等带来了巨大挑战,使得当前的命名实体链接技术越来越难以满足人们对链接准确率的要求.考虑到文档中的词和实体往往具有不同的语义主题(如"苹果"既能表示水果又可以是某电子品牌),而同一文档中的词与实体应当具有相似的主题,因此提出在语义层面对文档进行建模和实体消歧的思想.基于此设计一种完整的、基于概率主题模型的命名实体链接方法.首先,利用维基百科(Wikipedia)构建知识库;然后,利用概率主题模型将词和命名实体映射到同一个主题空间,并根据实体在主题空间中的位置向量,把给定文本中的命名实体链接到知识库中一个无歧义的命名实体;最后,在真实的数据集上进行大量实验,并与标准方法进行对比.实验结果表明:所提出的框架能够较好地解决了实体歧义问题,取得了更高的实体链接准确度.

关键词: 命名实体链接;概率主题模型;维基百科

中图法分类号: TP391

中文引用格式: 怀宝兴,宝腾飞,祝恒书,刘淇.一种基于概率主题模型的命名实体链接方法.软件学报,2014,25(9):2076−2087. http://www.jos.org.cn/1000-9825/4642.htm

英文引用格式: Huai BX, Bao TF, Zhu HS, Liu Q. Topic modeling approach to named entity linking. Ruan Jian Xue Bao/ Journal of Software, 2014,25(9):2076−2087 (in Chinese). http://www.jos.org.cn/1000-9825/4642.htm

Topic Modeling Approach to Named Entity Linking

HUAI Bao-Xing, BAO Teng-Fei, ZHU Heng-Shu, LIU Qi

(School of Computer Science and Technology, University of Science and Technology of China, Hefei 230027, China)

Corresponding author: LIU Qi, E-mail: qiliuql@ustc.edu.cn

Abstract: Named entity linking (NEL) is an advanced technology which links a given named entity to an unambiguous entity in the knowledge base, and thus plays an important role in a wide range of Internet services, such as online recommender systems and Web search engines. However, with the explosive increasing of online information and applications, traditional solutions of NEL are facing more and more challenges towards linking accuracy due to the large number of online entities. Moreover, the entities are usually associated with different semantic topics (e.g., the entity "Apple" could be either a fruit or a brand) whereas the latent topic distributions of words and entities in same documents should be similar. To address this issue, this paper proposes a novel topic modeling approach to named entity linking. Different from existing works, the new approach provides a comprehensive framework for NEL and can uncover the semantic relationship between documents and named entities. Specifically, it first builds a knowledge base of unambiguous entities with the help of Wikipedia. Then, it proposes a novel bipartite topic model to capture the latent topic distribution between entities and documents. Therefore, given a new named entity, the new approach can link it to the unambiguous entity in the knowledge base by calculating their semantic similarity with respect to latent topics. Finally, the paper conducts extensive experiments on a real-world data

* 基金项目: 国家杰出青年科学基金(61325010); 国家高技术研究发展计划(863)(2014AA015203); 安徽省科技专项资金(13Z02008-5); 安徽省国际科技合作计划(1303063008); 安徽省科技攻关计划(1301022064); 安徽省自然科学基金(1408085QF110)

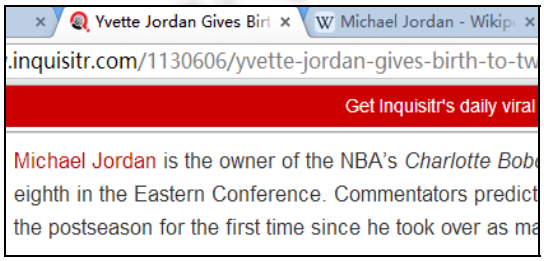
收稿时间: 2014-04-05; 定稿时间: 2014-05-14

set to evaluate our approach for named entity linking. Experimental results clearly show that the proposed approach outperforms other state-of-the-art baselines with a significant margin.

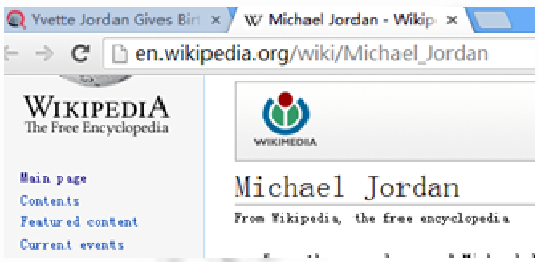
Key words: named entity linking; probabilistic topic models; Wikipedia

人们在使用互联网的过程中,接触频率最高的信息载体就是文字信息,如新闻、博客、评论等,这些文本蕴含了大量的命名实体(named entity).所谓命名实体,即包括名称(组织名、人名、地名、商品名)、表达式(日期、时间)等在内的具有明确语义信息的文本实体[1].为此,许多学者专注于研究如何在文本中高效地识别出命名实体,即进行命名实体识别(named entity recognition,简称 NER).事实上,针对 NER 的研究可以追溯到上世纪 90 年代[2],目前,NER 已是较为成熟的研究问题.近年来,学者们逐渐意识到:在 NER 的基础之上,如何有效利用识别出的命名实体,才是更有现实意义的事情[1,3,4].特别地,命名实体链接(named entity linking,简称 NEL)正是使用命名实体的流程中最为关键的步骤之一.

具体而言,命名实体链接是指将文本中已经识别出的命名实体链接到知识库中的一个具体真实实体的过程.命名实体链接能够被应用于很多现实的互联网服务场景中[5,6].举例来说,在用户兴趣建模过程中,可以通过分析用户浏览过或者分享过的历史文本(如新闻资讯等)来提取实体并进行知识库链接,这些链接过的实体可以当作关键字或者标签,为用户进行更精准的兴趣建模;再如,为了提升用户阅读体验,一些应用场景中增加了针对文本中实体的、用户可能感兴趣的内容链接,这些链接可能指向实体相关的一个商品或者指向另一篇以此实体为主题的新闻.著名新闻网站 Inquisitr(http://www.inquisitr.com/)的一篇新闻报道中提到了“Michael Jordan”(如图 1(a)所示)这个实体,同时,系统为读者提供了一条指向 Wikipedia 中“Michael Jordan”描述页面的链接(如图 1(b)所示),以帮助用户了解更多的信息.



(a) 一条关于“Michael Jordan”的新闻



(b) 新闻中命名实体链接所指向的 Wikipedia 页面

Fig.1 An example of named entity linking

图 1 命名实体链接应用实例

一种传统的命名实体链接方法是根据具体应用制定一些特殊的规则进行语义消歧[7],然而这些方法存在很大的局限性,换另外一种场景就很难取得较好的效果.也有一些研究工作是基于文本中实体和链接的特点构造语义网络[3,8],通过网络中的节点距离、出度、入度等作为特征进一步设计相似度衡量指标,从而实现语义消歧.虽然这类方法也取得相对较好的链接效果,然而却会在一些情况下存在潜在问题.如数据量规模较大的时候,网络的存储开销、训练开销都受到很大的限制.另外一些学者考虑了实体的上下文情境[9−11],把实体描述指向(链接到)知识库中的某个实体.这类方法使用的技术手段通常是计算实体与周围文本之间的语义相似度,然而,同一文本内的实体之间的语义关系却经常被忽略.事实上,随着实体数量的激增,使得当前的实体消歧等命名实体链接技术越来越难以满足人们对链接准确率的要求.

基于以上背景,本文设计了一种高效命名实体链接方法.其中,与普通文本和实体结合到一起进行语义消歧的传统链接方法不同,本文分别单独考虑了文本中词的集合与实体集合(具体细节请见第 3 节).这是因为虽然文档中的一个词或实体往往具有多个语义主题(如,“苹果”既能表示水果又可以是某电子品牌),但同一个文档中的词与实体应当具有相似的主题分布.因此,本文提出基于语义层面对文档进行建模和实体消歧.进而,本文设

计了一种完整的、基于概率主题模型的命名实体链接框架,该框架的具体流程在第 1 节介绍.值得一提的是:本文从概率主题模型的角度,以变分贝叶斯推导的方式提出了实体消歧方法,因此不仅能够取得较好的语义消歧效果,而且更容易进行并行化实现[12].

总体而言,本文的主要研究贡献如下:

- 提出了一个完整的实体链接系统框架,能够有效地把文本中有歧义的实体链接到知识库中对应的无歧义的真实实体;
- 提出使用概率主题模型进行命名实体链接的思路,它将词与命名实体映射到同一个语义主题空间,并以此对命名实体进行语义消歧;同时,该模型采用变分贝叶斯方法推导,易于实现并行化;
- 为了验证所提出的实体链接方法,本文利用真实数据进行了大量实验.实验结果表明:本文提出的方法相对其他领先的标准方法有较好的性能提升,尤其在实体链接的准确度提升较大.

1 系统框架

图 2 展示了本文设计的系统框架图,该系统主要分为 3 个模块:知识库模块、语义建模模块以及实体链接模块.本文力求知识库能够包含尽可能全面的有关实体的信息,因此,该系统的知识库基于维基百科构建,其中,同义词表和歧义词表是知识库的核心组成部分(知识库的具体细节将在第 2 节中详述).

语义建模是本文实体链接系统的关键步骤,其目的是为实体进行语义主题建模,本文采用的是基于概率主题模型(entity based latent dirichlet allocation,简称 eLDA)的语义建模模型(详细建模过程会在第 3 节中介绍).系统的最后一个模块是实体链接模块,主要功能是把文本里面的实体链接到知识库中对应的无歧义实体(该过程将在第 4 节中详细介绍).

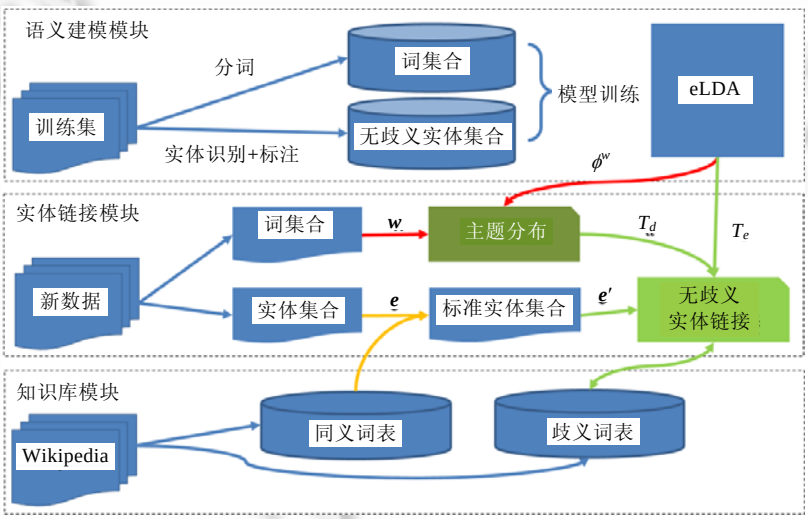


Fig.2 Framework of our novel NEL system

图 2 全新的命名实体链接系统架构图

根据该系统框架,每当给定一个待实体链接的新文档,首先用分词器把文档分词得到一组词的集合,同时使用已有的实体识别技术抽取出文本中的实体,抽取词和抽取实体是相互独立的两个过程.因为最初抽取出来的实体可能存在同义词或者缩写误写等情形,因此,首先要对命名实体进行描述形式上的归一化处理.本文通过在知识库中构建同义词表,对实体进行标准化映射来解决此问题.然后,采用概率主题模型(eLDA)计算出文档的潜在主题分布.由于本文假设同一文档中词的主题分布和实体的主题分布是一致的,因此,该主题分布其实等同于实体的主题分布.最后,结合语义消歧模型 eLDA 输出的训练集中实体主题分布以及实体歧义词表,得出标准实

体集合中每个实体对应的无歧义实体(通过计算不同实体在主题空间中的位置向量的相似度),即完成了实体链接过程.

2 知识库的构建

知识库(knowledge base,简称 KB)主要用来存储实体的信息.经过命名实体识别等技术,抽取出文本的命名实体集合.由于有些实体可能是某种简写模式(如 Mike)、缩写模式(如 MJ),或者同一个实体可能拥有几个不同的名字(如中科大、科大、中国科大等),甚至也有可能是误写(如 Mcihael Jordon),因此,这个集合中的实体还需要做进一步的处理,即需要把初步的实体映射到一种标准的表达形式.具体地,本文在知识库中构造一个如表 1 的同义词表来解决实体描述的统一性问题.该同义词表存储的是实体到实体之间的映射,关键字(key)是文本中可能出现的各种不规则形式的实体(如简写、同义词等),而键值(value)则是关键字对应的实体的标准描述.

Table 1 Examples of synonym lexicon

表 1 同义词表举例

Key (文本中实体表示)	Value (标准实体表示)
Michael Jeffrey Jordan Jordan, Michael Micheal Jordan Michael J. Jordan Michael Jeffery Jordan Michael Jordon	Michael Jordan

在实体链接的搜索阶段,需要为每个待消歧的命名实体构建一个候选实体列表,用来缩小实体链接的搜索空间.本文在知识库中构建了一个歧义词表(见表 2),这个表存储的是实体(key)及其对应的无歧义实体列表(list).这个表的 key 就是实体的标准表达形式,即,同义词表中对应的 Value.

Table 2 Examples of ambiguity lexicon

表 2 歧义词表举例

Key(标准实体表示)	List(无歧义真实实体)
Michael Jordan	Michael Jordan Michael I. Jordan Michael Jordan (mycologist) Michael Jordan (footballer) Michael Jordan (insolvency baron)

维基百科其资源的更新速度很快,通常在一个新的命名实体出现后几天内就会被更新在百科中.截至 2013 年 8 月 5 日,该数据库已经包含 443 万个实体词条.研究者能够更方便地使用它构建以实体为单位且包含实体之间的语义类别甚至相互关系的知识库[13−16].因此,本文采用如下页面结构来构建知识库中的同义词表和歧义词表:

- 重定向页面(redirect page):重定向页面一方面用来把已经改了名字的实体指向目前最新的标准实体表达方式;另一方面,对于简写或者容易拼写错误的实体指向正确的实体.如,标题为 Michael Jeffrey Jordan 的重定向页面会有一个链接指向 Michael Jordan;
- 命名实体页面(entity page):维基百科中,每个命名实体页面专门用来描述一个具体实体(即无歧义的实体),文章的标题就是命名实体的实际表示.例如,标题为 Michael Jordan(footballer)的页面内容就是描述某个足球运动员身份的 Michael Jordan.在表 1 中对应的是 value,在表 2 中对应的就是 key;
- 消歧页面(disambiguation page):消歧页面的标题一般是文本中通常出现的实体的普通表示,如 Michael Jordan,在很多文本中都只会显示成这种形式,而一般不会很明确地表明这里提到的 Michael Jordan 是篮球巨星还是机器学习著名教授.因此,消歧页面会给出一个列表,列出其对应的具体实体,这也是构造表 2 的主要途径.文本的标题可作为表 2 的 key,列表内容则可作为 value.

3 语义建模

本节介绍基于概率主题模型的命名实体语义主题建模方法.曾有研究工作提出多语言概率主题模型[17],该模型假设同一篇文档的不同语言描述遵从同一个主题分布,从而对存在用不同语言描述形式的文档进行主题建模.受此启发,本文认为:给定一篇文档 d,从表面上看 d 是由一些离散的单词 $\boldsymbol{w}$ 组成的,同时也可以认为是一组命名实体 $\boldsymbol{e}$ 组成的.特别的,抽取词和抽取实体的两个过程相互独立,互不影响.不妨假设:既然得到的这两个集合 $\boldsymbol{w}$ 和 $\boldsymbol{e}$ 是出自同一篇文档,那么二者应该符合同一个语义主题分布.从词的角度,每个词会有自己的一个主题分布,而从实体的角度来讲也是如此.根据不同实体的主题分布之间的相似性关系,即可以解决实体歧义的问题.为了学习实体的主题分布,本文提出利用概率主题模型的解决思路,图 3 展示了对应于实体语义消歧的概率图模型 Entity based Latent Dirichlet Allocation(eLDA).

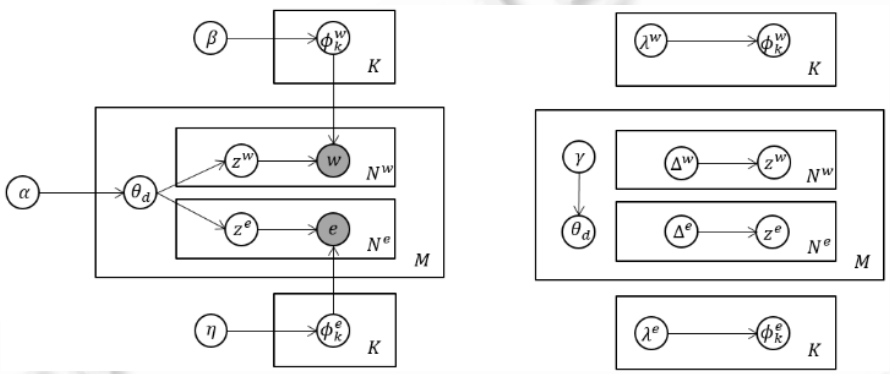


Fig.3 Graphical representations of eLDA model (left) and the Variational Inference for eLDA (right)

图 3 语义建模图模型:左侧为 eLDA,右侧为其变分形式

概率主题模型属于生成模型,对于由 M 个文档组成的文档集中的一个包含 N_d^w 个词、N_d^e 个实体的文档 d,eLDA 认为,该文档的生成过程如下:

1) 对于每个主题 k,根据狄利柯雷(Dirichlet)分布,分别生成词和实体的在主题上的分布:

$$\phi_k^w \sim Dir(\beta), \phi_k^e \sim Dir(\eta)$$

2) 对每个文档集中的文档 $d \in \{1,\ldots,M\}$,执行步骤 3)~步骤 9)
3) 为当前文档 d 生成主题分布 $\theta_d \sim Dir(\alpha)$
4) 对 d 中的每个词 $n \in \{1,\ldots,N_d^w\}$,执行步骤 5)、步骤 6)
5) 为当前词生成主题 $z_{d,n} \sim Mult(\theta_d)$
6) 生成当前的词 $w_{d,n} \sim Mult(\phi_{z_{d,n}}^w)$
7) 对 d 中的每个实体 $n' \in \{1,\ldots,N_d^e\}$,执行步骤 8)、步骤 9)
8) 为当前实体生成主题 $z_{d,n'} \sim Mult(\theta_d)$
9) 生成当前的实体 $e_{d,n'} \sim Mult(\phi_{z_{d,n'}}^e)$

上述过程中的 Dir 表示狄利克雷分布,$Mult$ 表示多项式分布.当给定参数α,β,η时,所有观察变量(即观察到的文档中的词和实体)和隐变量(如每个主题上的实体分布)的联合概率公式为

$$P(d,z^w,z^e,\Theta,\Phi^w,\Phi^e,\alpha,\beta,\eta) = P(\Theta|\alpha)P(\Phi^w|\beta)P(\Phi^e|\eta)\times\left(\prod_{n=1}^{N^w}P(w_n|z^w,\Phi^w)\right)\left(\prod_{n=1}^{N^e}P(e_n|z^e,\Phi^e)\right) \quad (1)$$

那么,文档 d 的似然值 $P(d|\alpha,\beta,\eta)$可按公式(2)计算:

$$P(d\mid\alpha,\beta,\eta)=\int p(\theta_d\mid\alpha)\int p(\Phi^w\mid\beta)\left(\prod_{n=1}^{N^w}\sum_{z_n^w}p(z_n^w\mid\theta_d)p(w_n\mid z_n^w,\Phi^w)\right)\mathrm{d}\Phi^w$$
$$\int p(\Phi^e\mid\eta)\left(\prod_{n=1}^{N^e}\sum_{z_n^e}p(z_n^e\mid\theta_d)p(e_n\mid z_n^e,\Phi^e)\right)\mathrm{d}\Phi^e\mathrm{d}\theta_d \tag{2}$$

从公式(2)可以看出:直接通过极大似然估计去求解参数(与生成过程相反,该过程是求解隐变量)是很难的,最常用的方式是通过马尔可夫链蒙特卡洛(Markov chain Monte Carlo,简称 MCMC)[18]采样来解决,其基本思想是:通过从一个其稳定统计分布为实例后验分布的马尔可夫链中进行采样模拟逼近,从而得到最优的参数[19].其中,吉布斯采样(Gibbs sampling)是 MCMC 算法的一种,被广泛用于贝叶斯模型中[19−21],其马尔可夫链是根据隐变量的条件概率分布定义的.虽然采用吉布斯采样的方法求解参数可以取得比较令人满意的结果,然而也有其不足的地方.例如:吉布斯采样过程收敛到稳定分布是难以断定的,且高数据规模的采样算法往往收敛很慢[22].

一种可替代 MCMC 的推导方法就是变分贝叶斯方法推导(variational Bayesian inference),其基于统计物理学的优化技术,通过最优化寻找接近实例后验分布的隐变量分布[23].基于变分贝叶斯推导的算法更容易被并行化实现,并且能够更好地解决维度灾难问题[12].

因此,本文设计了变分贝叶斯推导的方法进行参数估计,图 3 右侧表示了该变分近似过程.在本文中,变分近似主要是用一个简化的概率去近似后验概率 $P(\theta,z^w,z^e,\phi^w,\phi^e|d,\alpha,\beta,\eta)$,如下所示:

$$P(\theta,z^w,z^e,\phi^w,\phi^e\mid\gamma,\Delta^w,\Delta^e,\lambda^w,\lambda^e)=\prod_k Dir(\phi_k^w\mid\lambda_k^w)Dir(\phi_k^e\mid\lambda_k^e)\prod_d Dir(\theta_d\mid\gamma_d)Mult(z_{d,n}^w\mid\Delta_{d,n}^w)Mult(z_{d,n}^e\mid\Delta_{d,n}^e) \tag{3}$$

其对数似然值的下界可计算为

$$\log p(d|\alpha,\beta,\eta)\geqslant E_q[\log p(\theta,z^w,z^e,\phi^w,\phi^e|d,\alpha,\beta,\eta)]-E_q[\log q(\theta,z^w,z^e,\phi^w,\phi^e)]=L(\gamma,\Delta^w,\Delta^e,\lambda^w,\lambda^e;\alpha,\beta,\eta) \tag{4}$$

则有:

$$\log p(d|\alpha,\beta,\eta)=L(\gamma,\Delta^w,\Delta^e,\lambda^w,\lambda^e;\alpha,\beta,\eta)+D(q(\theta,z^w,z^e,\phi^w,\phi^e)\|p(\theta,z^w,z^e,\phi^w,\phi^e|d,\alpha,\beta,\eta)) \tag{5}$$

这里的 $D(q(*)\|p(*))$是 KL 距离(http://en.wikipedia.org/wiki/Kullback-leibler_divergence),用来表示变分后验概率分布和真实的后验概率分布的距离,因此,参数求解过程即是最小化 $D(\cdot)$的过程,这等价于最大化 $L(\gamma,\Delta^w,\Delta^e,\lambda^w,\lambda^e;\alpha,\beta,\eta)$,即

$$\begin{aligned}L(\gamma,\Delta^w,\Delta^e,\lambda^w,\lambda^e;\alpha,\beta,\eta)=&E_q[\log p(\theta\mid\alpha)]+E_q[\log p(z^w\mid\theta)]+E_q[\log p(z^e\mid\theta)]+E_q[\log p(\phi^w\mid\beta)]/M+\\&E_q[\log p(w\mid z^w,\phi^w)]+E_q[\log p(\phi^e\mid\eta)]/M+E_q[\log p(e\mid z^e,\phi^e)]-E_q[\log q(\theta)]-\\&E_q[\log q(z^w)]-E_q[\log q(z^e)]-E_q[\log q(\phi^w)]/M-E_q[\log q(\phi^e)]/M\end{aligned} \tag{6}$$

因此,本文进一步设计了 EM(expectation maximization)算法来估计模型参数.

- *E* 步骤:

$$\Delta_{n,k}^w\propto\exp\{E_q[\log(\theta_k\mid\gamma)]+E_q[\log(\phi_{k,n}^w\mid\lambda^w)]\} \tag{7}$$

$$\Delta_{n,k}^e\propto\exp\{E_q[\log(\theta_k\mid\gamma)]+E_q[\log(\phi_{k,n}^e\mid\lambda^e)]\} \tag{8}$$

其中,

$$E_q[\log(\theta_k\mid\gamma)]=\Psi(\gamma_k)-\Psi(\textstyle\sum_j\gamma_j) \tag{9}$$

$$\gamma=\alpha+\left(\sum_{n=1}^{N^w}\Delta_n^w+\sum_{n=1}^{N^e}\Delta_n^e\right) \tag{10}$$

- *M* 步骤:

$$\lambda_{n,k}^w=\beta+\sum_{d=1}^{M}\sum_{n=1}^{N_d^w}\Delta_{d,n,k}w_{d,n} \tag{11}$$

$$\lambda_{n,k}^e=\beta+\sum_{d=1}^{M}\sum_{n=1}^{N_d^e}\Delta_{d,n,k}e_{d,n} \tag{12}$$

其中,参数α,β,η可在每次的 M 步骤中采用牛顿-拉普森(Newton-Raphson)方法计算获得,并更新.

本文为推理过程设计了算法 1.

算法 1. eLDA 变分贝叶斯推导算法.

1) 初始化 $\Delta_{n,k}^{w}=\frac{1}{K}$,其中,$n\in[1,N^{w}],k\in[1,K]$
2) 初始化 $\Delta_{n,k}^{e}=\frac{1}{K}$,其中 $n\in[1,N^{e}],k\in[1,K]$
3) 初始化$\gamma_k=\alpha_k+(N^{w}+N^{e})/2K,k\in[1,K]$
4) 重复执行步骤 5)~步骤 13),直至收敛
5) 对所有词 $n\in[1,N^{w}]$,执行步骤 6)~步骤 8)
6) 对所有主题 $k\in[1,K]$,执行步骤 7)
7) 更新 $\Delta_{n,k}^{w}=\frac{\lambda_{n,k}^{w}}{\sum_{n'=1}^{N^{w}}\lambda_{n',k}^{w}}\exp(\Psi(\gamma_k))$
8) 归一化 Δ_{n}^{w}
9) 对所有实体 $n\in[1,N^{e}]$,执行步骤 10)~步骤 12)
10) 对所有主题 $k\in[1,K]$,执行步骤 11)
11) 更新 $\Delta_{n,k}^{e}=\frac{\lambda_{n,k}^{e}}{\sum_{n'=1}^{N^{e}}\lambda_{n',k}^{e}}\exp(\Psi(\gamma_k))$
12) 归一化 Δ_{n}^{e}
13) $\gamma_k=\alpha_k+\left(\sum_{n=1}^{N^{w}}\Delta_{n}^{w}+\sum_{n=1}^{N^{e}}\Delta_{n}^{e}\right)$

4 实体链接过程

通过 eLDA 主题建模模型,系统能够得到词和实体分别在主题空间上的分布.从而,当给定文本的词集合的前提下,可以得到文本的主题分布;最后,通过实体与文本的主题分布的相似度计算进行语义消歧.具体而言,每当给定一个文本 d,将其分词处理得到一组词的集合 $\vec{w}$;同时,将文本 d 利用实体识别技术抽取出其所包含的实体集合 $\boldsymbol{e}$,则针对这些实体的实体链接完整算法如下:

算法 2. 命名实体链接方法.

1) 根据知识库中的同义词表,对 $\boldsymbol{e}$ 中的所有实体进行同义实体合并,即进行描述标准化,得到描述标准统一的的实体集合 $\boldsymbol{e}'=\{e_1,e_2,\ldots,\}$;
2) 对 $\boldsymbol{e}'$中的每个实体 $e_i\in\boldsymbol{e}'$,根据知识库歧义词表得到实体 e_i 的候选实体 $C_{e_i}=\{e_{i,1},e_{i,2},\ldots\}$,当$|C_{e_i}|=0$ 时,说明未能在知识库中找到对应的实体,则实体 e_i 被认为是不可链接的,返回标签 Empty;当$|C_{e_i}|=1$ 时,直接返回唯一的候选实体作为最终链接实体;当$|C_{e_i}|>1$ 时,执行步骤 3)~步骤 5);
3) 通过 eLDA 中在主题上的实体分布φ^{e},可得到每个实体的主题分布 T_e;
4) 根据 LDA[24]推理过程,由 $\boldsymbol{w}$ 能得到文本的主题分布 T_d;
5) 对 $\boldsymbol{e}'$中的每个实体 $e_i\in\boldsymbol{e}'$,计算其每个候选实体 $e_{i,j}$ 的主题分布 $T_{e_{i,j}}$ 与文档的主题分布 T_d 的余弦相似度,最相近的那个便是最终要链接到的实体(与 e_i 在主题空间中的位置向量最相近的候选实体即为所求).

在算法 2 的步骤 5)中,通过简单的方法学习一个阈值τ,当一个实体的所有候选实体与文本的主题分布的相似度的最大值小于阈值τ的时候,认为没有找到合适的实体,也返回 Empty 标签.

5 数据集描述以及实验效果

5.1 知识库的建立

如前所述,本文建立知识库采用的是维基百科作为源数据,数据版本使用的是2013年8月5日的英文百科,该版本数据集包括实体页面(无歧义的实体)443万个,歧义词页面21万个,重定向页面625万个.其统计信息详见表3.

Table 3 Statistics of the Wikipedia dataset used in our experiments

表3 本文实验采用的维基百科数据集统计信息

百科数据版本	enwiki-20130805-pages-articles-multistream.xml.bz2
总页面数	13 715 114
实体页面数	4 439 671
消歧页面数	227 940
重定向页面数	6 255 904

所构建出来的知识库中,同义词表(统计信息见表4)的词条数多达625万条,最终这625万个实体映射到234万个标准形式的实体上.需要注意的是,同义词表中的value值可能是一个具体的实体,如Amaltheia→Amalthea(mythology),也有可能是一种会有歧义的形式,如Michael Jordon→Michael Jordan.歧义词表(统计信息见表5)中歧义实体有19万个,总共包括195万个非歧义实体,即平均每个歧义词条有10种解释.

Table 4 Statistics of the synonym lexicon used in our experiments

表4 同义词表统计信息

key的总数	6 255 904
value的元素总数	2 348 277

Table 5 Statistics of the ambiguity lexicon used in our experiments

表5 歧义词表统计信息

歧义实体(key)的个数	198 699
所有消歧页面包含的实体数	1 958 567
歧义实体平均对应实体长度	9.86

5.2 实验数据收集

为检验本文设计的实体链接系统,作者采用真实的数据集进行检验.相关文献[6,9,18]中使用的TAC-KB数据集不是公开数据集,故本文无法使用,但TAC-KB数据集是根据维基百科的数据库构建出来的[20],而在文献[6,19]中,研究者们也使用了维基百科页面作为数据集,因此,本文选择从维基百科数据库中抽取合适的实验数据.主要基于以下几点原因:

- 实体页面(entity page)中的每个文档是围绕其标题(即实体)而写,因此,其主题分布具有一定规律性,其性质类似于新闻文档;
- 每个实体页面都会有类别标识(category),因此可以人工制定一些简单的规则,使得抽取的数据集尽量集中在某几个类别中,这就避免了数据集文本主题分布过于分散的问题.这类似于随机从体育新闻、娱乐新闻等类别中抽取数据集一样;
- 实体页面由文本和实体共同组成,其中,实体会被双重中括号括起来,如Lincoln reached out to [[War Democrats]],这里,War Democrats就是一个实体.这样就可以直接得到文本中的一些实体集合;
- 有歧义的实体通常会被一个竖线"|"将实体与其对应的无歧义实体分开,例如Lincoln was a steadfast [[Whig Party (United States)|Whig]],则竖线右边的则是在文本中显示给读者看的实体,即Whig;而左侧的则表示当前实体的无歧义实体,在维基百科页面会有一个链接从Whig链接到Whig Party (United States)的实体页面.这样,对于一篇维基百科的实体页面,不仅能够得到里面包含的实体,而且能够知道

这个实体的真实指向,这将可以作为实验最终的真实评估标准.

为了保证随机抽取的文章在较大程度上能够在几个比较集中的主题上(便于验证),首先预先制定了几个类别.见表 6,作者拟定了四大主题,分别是“体育篮球”、“数据挖掘、人工智能”、“流行音乐、摇滚音乐”以及“电子产品、手机”.Categories 对应的是维基百科中的类别属性.

Table 6 Category information in the experimental dataset

表 6 数据集文章类别

主题	Categories
体育、篮球	National Basketball Association, All-Stars, Olympic
数据挖掘、人工智能	Artificial, Machine learning, Statistics, Scientific, Mathematical, Data mining
流行、摇滚音乐	rock, pop, Music Awards, Award-winning
电子、手机产品	Electronics, companies, Mobile, Smartphones

利用类别属性(category),在 443 万中属于这些类别的实体中分别随机选取 400 篇文章,然后从得到的总共 1 200 篇文章中随机抽取 500 篇文章作为本文的实验数据集,经过数据清理(去停用词等过程),该实验数据集平均每个文本包含 986 个词,119 个实体,具体统计情况请见表 7.

Table 7 Statistics of the experimental dataset

表 7 实验数据统计信息

文本个数	500
平均每个文本含词的个数	986
平均每个文本含实体的个数	119
词集合大小	53 203
实体集合大小	40 086

5.3 对比实验及分析

5.3.1 评估指标

本文的主要工作是把给定文本中的命名实体链接到知识库中,前提是已经确定了文本中哪些是命名实体,那么算法的优劣就体现在实体链接的准确度上,而不涉及查全率(recall)这个指标.因此,我们采用的评估指标是准确率(accuracy),这也是命名实体链接最常用、最重要的标准.如果准确进行实体链接的个数表示为 N_r,所有实体的个数为 N_a,那么 $Acc.=N_r/N_a$.文献[3,8]也使用了 *Acc.*作为评估标准,但由于这些研究工作中使用的知识库覆盖不全,使得文本中有一些实体在知识库中无法找到,因此还使用另外两个指标(*inKB*,*NIL*),其中,*inKB* 表示可链接到知识库中的准确率,*NIL* 表示无法链接到知识库中的准确率.而本文所使用的知识库是建立在完整的维基百科数据集上的,且使用的数据集来自维基百科页面,实体的覆盖很全,也使得另外两个指标在本文的实验中意义不大,因此,本文的实验主要评测算法的 *Acc.*值.

5.3.2 对比实验

本文的实验环境为 3.10GHz×2 Cores CPU,4GB RAM,参数α,β,η的初始值为 0.05,0.01,0.01.采用十折交叉验证得到最终结果.接下来称本文的实体链接方法为 eLDA,并选择两个标准算法作为对比:一个是文献[3]中提出的 LINDEN 方法,另外是文献[8]中提出的效果最好的出入度算法(oD 以及 iD 算法).这 3 种算法在其各自的文章中都被证明在 TAC-KBP 数据集上在实体链接的准确度(*Acc.*)指标上要优于 TAC-KBP 数据集官方给出的最优 3 种算法的结果.与文献[3,8]不同的是,本文在实验中使用的是完整的维基百科数据集构造知识库,这使得候选实体会比较完整.也因此,iD 和 oD 算法能够尽可能地构造出较为完整的知识网(与本文构造 KB 的方式不同),从而能够保证算法运行出最优效果.对于 LINDEN 算法,在实验中选择作者提出的 SA(semantic associativity)以及 LP(link probability)作为 LINDEN 的特征集,其权重向量 $\vec{\omega}$ 以及所有参数都采用十折交叉验证.

本文的 eLDA 算法则按照不同的主题数目(*K* 值取 15,20,25)情况进行对比实验.具体实验结果如图 4 所示,其中,柱状图表示准确率.从实验结果可以看出:iD 与 LINDEN 的准确率比较接近,iD 略好于 oD,这与文献[8]在 TAC-KBP 上的表现比较吻合.在多数 *K*(主题的个数)值设定下,eLDA 的准确率高于其他对比方法.另外,eLDA

在 K=15 时表现要弱于 K=20 或 K=25 时.eLDA 的实验效果随 K 值变化曲线如图 5 所示,可以看出:当 K=20 时,eLDA 取得较高准确率.

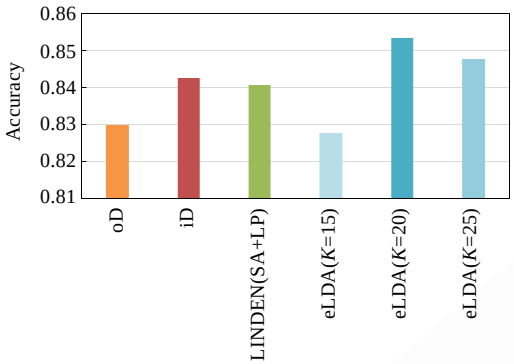


Fig.4 Performance comparison of different approaches

图 4 不同算法对比的实验结果

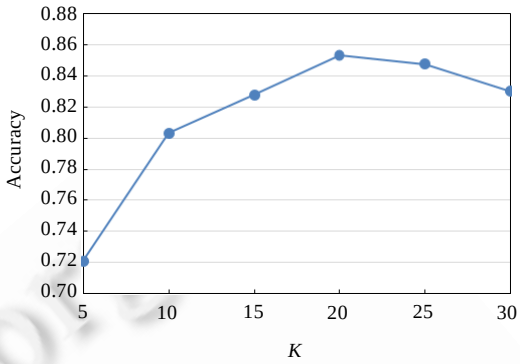


Fig.5 Performance of eLDA with respect to different number of latent topics

图 5 不同主题个数的 eLDA 实验效果

5.3.3 案例研究

本文假设同一篇文档中的词和实体可以映射到同一个主题空间中,基于此,提出使用概率主题模型来进行命名实体消歧的方法.为了更直观地验证我们假设的合理性,探究概率主题模型为何能够取得较好的效果,本文做了一个真实的案例研究.

实验效果表明:当 K=20 的情况下,本文的系统能够取得较优的准确率,所以在此设定下,在数据集中找两个有代表性的实体,分别是 Michael Jordan,Michael I. Jordan.其中,Michael Jordan 表示 NBA 篮球巨星, “Michael I. Jordan”表示加州伯克利分校(UC Berkeley)的著名机器学习教授.在 eLDA 训练完之后,分别得到了两个实体的主题分布,从中取其各自概率最高的两个主题(Rank-1,Rank-2),并列举这两个主题分布概率最高的前 5 个词(Top-5),见表 8.

Table 8 Examples of the word distributions of different topics

表 8 实体的主题词分布举例

Michael Jordan		
Rank-1 主题	Top-5 词	Jordan, American, Genesis, basketball, Lifetime
	Top-5 实体	[Basketball], [National Basketball Association Most Valuable Player Award], [Madison Square Garden], [List of National Basketball Association season scoring leaders], [Jimi Hendrix]
Rank-2 主题	Top-5 词	style, center, players, NBA, junior
	Top-5 实体	[Washington Wizards], [Boston Celtics], [college basketball] [Los Angeles Lakers], [Charlotte Bobcats]
Michael I. Jordan		
Rank-1 主题	Top-5 词	work, paper, University, research, maximum
	Top-5 实体	[University of California, Berkeley], [Machine learning], [Computer Science], [Artificial intelligence], [American Statistical Association]
Rank-2 主题	Top-5 词	Group, math, theory, Heisenberg, mathematical
	Top-5 实体	[common logarithm], [record chart], [matrix (mathematics)] [Markov process], [e (mathematical constant)]

Rank-1|Top-5 表示概率值最大的主题分布在该主题上的概率值排名前五的词以及实体.另外,为了增强可读性,表格中的命名实体用[·]括起来.从该表可以看出:Michael Jordan 和 Michael I. Jordan 两个实体概率最高的主题所对应的词和实体有着显著的差异,且都与这两个实体有着密切的关系.这也再次证明本文关于不同实体

具有不同语义主题分布的假设成立,以及从语义的层面对文档进行建模和实体消歧的思路可行.

6 结束语

基于文本中词的集合以及命名实体的集合具有相同语义主题分布的假设,本文提出了利用概率主题模型对命名实体进行语义消歧的思路,并设计了一套完整的实体链接框架;同时,所用到的知识库是采用维基百科的完整数据进行构建.真实数据上的对比实验表明:本文提出的命名实体链接框架能够较好地实现实体链接,取得了比现有工作更高的链接准确度.最后,本文还通过列举案例的方式对算法进行了细致分析.

然而,本文提出的语义消歧方法仍然有一定的局限性,例如训练数据中需要对实体进行去歧义标注;另外,当文本中实体数目很少的时候,可能并不能够很好地满足命名实体与词符合相同的主题分布的假设.因此,未来工作中将进一步探索实体扩展的方法.此外,将设计一种具体的分布式变分推导算法去实现更高效的命名实体语义消歧过程.

致谢 在此,我们向对本文的工作提出很多宝贵意见的张磊、朱琛两位同学表示感谢.

References:

[1] Hachey B, Radford W, Nothman J, Honnibal M, Curran JR. Evaluating entity linking with Wikipedia. Artificial Intelligence, 2013,194:130−150.

[2] Rau LF. Extracting company names from text. In: Proc. of the 7th IEEE Conf. on Artificial Intelligence Application. IEEE Press, 1991. 29−32. [doi: 10.1109/caia.1991.120841]

[3] Shen W, Wang JY, Luo P, Wang M. Linden: Linking named entities with knowledge base via semantic knowledge. In: Proc. of the 21st Int'l Conf. on World Wide Web. ACM Press, 2012. 449−458. [doi: 10.1145/2187836.2187898]

[4] Milne D, Witten IH. Learning to link with Wikipedia. In: Proc. of the 17th ACM Conf. on Information and Knowledge Management. ACM Press, 2008. 509−518. [doi: 10.1145/1458082.1458150]

[5] Bunescu RC, Pasca M. Using encyclopedic knowledge for named entity disambiguation. In: Proc. of the 7th Conf. of the European Chapter of the Association for Computational Linguistics. ACM Press, 2006. 9−16.

[6] Milosavljevic M, Delort JY, Hachey B, Arunasalam B, Radford W, Curran JR. Automating financial surveillance. In: Proc. of the User Centric Media. Berlin, Heidelberg: Springer-Verlag, 2010. 305−311. [doi: 10.1007/978-3-642-12630-7_38]

[7] Zheng J, Mao YH. Word sense tagging method based. Journal of Tsinghua University (Sci. & Tech.), 2001,41(3):117−120 (in Chinese with English abstract).

[8] Guo Y, Che W, Liu T, Li S. A graph-based method for entity linking. In: Proc. of the 5th Int'l Joint Conf. on Natural Language Processing. 2011. 1010−1018.

[9] Bekkerman R, McCallum A. Disambiguating Web appearances of people in a social network. In: Proc. of the 14th int'l Conf. on World Wide Web. ACM Press, 2005. 463−470. [doi: 10.1145/1060745.1060813]

[10] Cucerzan S. Large-Scale named entity disambiguation based on Wikipedia data. In: Proc. of the 2007 Joint Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. 2007. 708−716.

[11] Lin D. An information-theoretic definition of similarity. In: Proc. of the 15th Int'l Conf. on Machine Learning. Morgan Kaufmann Publishers, Inc., 1998. 296−304.

[12] Zhai K, Boyd-Graber J, Asadi N, Alkhouja M. Mr. LDA: A flexible large scale topic modeling package using variational inference in MapReduce. In: Proc. of the 21st Int'l Conf. on World Wide Web. ACM Press, 2012. 879−888. [doi: 10.1145/2187836.2187955]

[13] Bagga A, Baldwin B. Entity-Based cross-document coreferencing using the vector space model. In: Proc. of the 17th Int'l Conf. on Computational Linguistics. ACM Press, 1998. 79−85. [doi: 10.3115/980451.980859]

[14] Hasegawa T, Sekine S, Grishman R. Discovering relations among named entities from large corpora. In: Proc. of the 42nd Annual Meeting on Association for Computational Linguistics. ACM Press, 2004. 415. [doi: 10.3115/1218955.1219008]

[15] Witten I, Milne D. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In: Proc. of the AAAI Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy. AAAI Press, 2008. 25−30.

[16] Navigli R, Velardi P. Structural semantic interconnections: A knowledge-based approach to word sense disambiguation. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2005,27(7):1075−1086. [doi: 10.1109/TPAMI.2005.149]

[17] Mimno D, Wallach HM, Naradowsky J, Smith DA, McCallum A. Polylingual topic models. In: Proc. of the 2009 Conf. on Empirical Methods in Natural Language Processing: Vol.2-Vol.2. ACM Press, 2009. 880−889. [doi: 10.3115/ 1699571.1699627]

[18] Hastings WK. Monte Carlo sampling methods using Markov chains and their applications. Biometrika Press, 1970,57(1):97−109.

[19] Neal RM. Probabilistic inference using Markov chain Monte Carlo methods. Technical Report, CRG-TR-93-1, Department of Computer Science, University of Toronto, 1993.

[20] Teh YW. A hierarchical Bayesian language model based on Pitman-Yor processes. In: Proc. of the 21st Int'l Conf. on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics. ACM Press, 2006. 985−992. [doi: 10.3115/1220175.1220299]

[21] Griffiths TL, Steyvers M. Finding scientific topics. Proc. of the National Academy of Sciences of the United States of America, 2004,101(Suppl. 1):5228−5235.

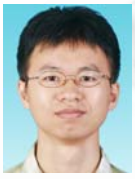
[22] Robert CP, Casella G. Monte Carlo Statistical Methods. New York: Springer-Verlag, 2004.

[23] Jordan MI, Ghahramani Z, Jaakkola TS, Saul LK. An introduction to variational methods for graphical models. Machine Learning, 1999,37(2):183−233.

[24] Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. The Journal of Machine Learning Research, 2003,3:993−1022. [doi: 10.1109/icdm.2008.75]

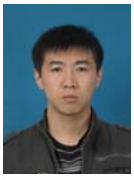
附中文参考文献:

[7] 郑杰,茅于杭.一种基于词语搭配的语义消歧方法.清华大学学报(自然科学版),2001,41(3):117−120.



怀宝兴(1987－),男,黑龙江鸡西人,博士生,主要研究领域为推荐系统.
E-mail: bxhuai@mail.ustc.edu.cn

祝恒书(1986－),男,博士生,CCF 学生会员,主要研究领域为移动智能计算,社交智能计算.
E-mail: zhs@mail.ustc.edu.cn



宝腾飞(1985－),男,博士,主要研究领域为语义计算,智能推荐.
E-mail: tengfei.bao@gmail.com

刘淇(1986－),男,博士,副研究员,CCF 会员,主要研究领域为数据挖掘与知识发现,机器学习方法与应用.
E-mail: qiliuql@ustc.edu.cn