基于双词主题模型的半监督实体消歧方法研究

张 雄,陈福才,黄瑞阳
(国家数字交换系统工程技术研究中心,河南郑州 450001)

摘 要: 针对实体上下文信息主题漂移的问题,提出一种基于双词主题模型的实体消歧方法.方法考虑到实体在一定语义环境下具有不同的主题,且在同一文档中同时出现的其他实体在一定程度上能够帮助待消歧实体确定所指代内容,利用命名实体构建双词的思想,将协同实体关系融合到主题模型中,并在此基础上利用维基百科知识库,进行半监督消歧.本文最后在网络文本数据上进行了相关的实验,验证了所提算法的有效性.实验表明该方法有效的提高了实体消歧精度.

关键词: 实体消歧;维基百科;双词主题模型
中图分类号: TP393 **文献标识码:** A **文章编号:** 0372-2112 (2018)03-0607-07
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2018.03.014

Semi-supervised Entity Disambiguation Method Research Based on Biterm Topic Model

ZHANG Xiong, CHEN Fu-cai, HUANG Rui-yang
(*National Digital Switching System Engineering and Technological R&D Center, Zhengzhou, Henan* 450001, *China*)

Abstract: Aimed at the problem of theme drift of the entity context information, this paper proposes an entity disambiguation method based on biterm topic model. The proposed method considers that the entity has a different theme in a certain semantic environment and the other entity appearing in the same document at the same time can help the disambiguated entity to determine the referred content to a certain extent. Therefore, using the ideas of named entity constructing double words to incorporate collaborative entity relationship to the topic model, and on this basis, we conduct semi-supervised disambiguation using Wikipedia knowledge base. Finally, this paper conducts some relevant experiments on the web text data, and verifies the effectiveness of the proposed algorithm. The experiments show that the proposed method effectively improve the precision of entity disambiguation.

Key words: entity disambiguation; Wikipedia; biterm topic model

1 引言

命名实体消歧(Named Entity Disambiguation, NED)是将文档中提及的命名实体链接到一个无歧义的知识库中相应实体的过程,该技术广泛应用于信息抽取、知识库构建、语义搜索[1]等领域.在网络文本数据中,蕴含着大量的歧义实体,如何从这些网络数据中获得准确的实体信息是网络信息抽取亟待解决的问题.为了解决这一问题,实体消歧技术得到广泛的研究.

目前实体消歧技术的研究主要包括以下两类,一是建立文本语境模型,例如计算上下文词语与百科资源描述的文本相似性实现消歧[2~4],该方法最近的研究是从实体全局一致性的角度对文本中所有提及实体同时进行消歧.文献[5~8]通过构建词嵌入模型,从文本语料中学习词的连续向量表示,并通过向量在相对低维向量空间中获取词的语义相似性.王英帅等人[9]提出一种基于LDA主题模型的消歧方法,该方法针对网页中的人名消歧任务,在传统的上下文消歧基础上,添

收稿日期:2016-07-11;修回日期:2016-10-24;责任编辑:梅志强
基金项目:国家自然科学基金(No.61171108);国家重点基础研究发展计划("973"计划)资金(No.2012CB315901, No.2012CB315905);国家科技支撑计划(No.2014BAH30B01)

加网页主题信息,从而实现比传统聚类方法更好的实验性能. Guo 等人[10]在知识图谱上使用随机游走的方法来构造实体和文本的向量表示方法,文献[11,12]结合了实体在维基百科中的全局链接结构和上下文信息,分别用贝叶斯网络概率模型和基于图模型的 Personalized PageRank 模型进行实体消歧. 怀宝兴等人[13]使用 eLDA 概率主题模型提出一种基于链指的实体消歧方法并取得了不错的效果,也证明了主题模型对上下文信息刻画的有效性. Nguyen 等人[14]提出一种两级映射算法,将主题域作为一个重要特征进行了一级映射,然后在第一级映射的基础上,利用高信任度链接为二级映射提供了更加可靠的上下文信息,从而实现实体消歧. 二是利用现有知识库(例如:维基百科、Freebase 等)进行实体消歧[15~17],该方法能够弥补实体上下文信息稀疏、实体间结构关系等问题,知识库中实体摘要[6]能够较好的刻画实体本体语义,对实体消歧具有重要的作用.

上述的研究方法通常在计算文本相似性的时候都在待消歧词附近使用了一个固定大小的窗口,只将窗口中的词作为上下文信息,如果窗口太小会丢失很多重要信息,窗口过大又会将不相干信息引入到上下文信息中,因此如何合理地刻画上下文信息是目前研究的一个重要问题. 此外,现有的研究在对实体协同作用的分析时,并未考虑到随着实体主题迭代出现的主题漂移现象,从而导致主题模型的精度降低.

因此,本文针对全局主题信息漂移的问题,采用双词主题模型(Biterm Topic Modeling,BTM),结合维基百科知识库中的实体页面,提出基于 BTM 的半监督命名实体消歧算法. 该算法利用维基百科实体摘要信息(实体页面),构建双词集合并赋予实体主题标签,且保证双词中至少有一个为实体词,有效降低主题漂移现象,提高主题模型的精度,从而改善实体消歧效果.

2 BTM 主题模型

BTM 是一种短文本主题概率生成模型[18]. 该模型针对短文本主题模型中出现的数据稀疏的问题,通过建立共现词组合(双词)来改进传统的 LDA 主题模型[19],提高了主题中词聚合程度,有效改善了短文中主题模型的性能.

双词即同一文本中同时出现的无序词对,BTM 模型就是建立在双词共现模式的基础上,其核心思想是两个词共同出现的频率越高,则双词属于同一个主题的可能性越大. 模型假设双词中的每个词都独立地从一个主题中生成,且该主题从一个全局语料库的主题分布中生成. 本文使用的基于实体词的 BTM 模型的概率图表示如图 1 所示:

给定一个短文本集合 $D=\{d_1,d_1,\cdots,d_N\}$,对应的双词集合为 $B=\{b_1,b_2,\cdots,b_M\}$,其中 $b_i=(w_{i,1}^e,w_{j,2}^n)$ 表示每一个词对. 令 $z=\{1,2,\cdots,K\}$ 表示主题集合,$\boldsymbol{\theta}$ 表示主题分布,即 $\theta_k=P(z=k)$,$\boldsymbol{\Phi}$ 表示主题中词的多项分布,即 $\boldsymbol{\Phi}$ 为 $K\times W$ 维矩阵,其中每个元素 $\varphi_{k,w}=P(w|z=k)$. 则 BTM 产生式过程如下:

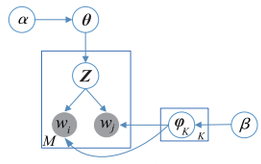

图1 BTM概率图模型

对整个语料库,采样一个主题分布:$\boldsymbol{\theta}\sim\mathrm{Dir}(\alpha)$

对每一个主题 $k\in(1:K)$

采样一个词的分布 $\varphi_k\sim\mathrm{Dir}(\beta)$

对每一个词对 $b_i\in B$

采样一个主题 $z_i\sim\mathrm{Multi}(\theta)$

独立地采样两个词 $w_{i,1}\sim\mathrm{Multi}(\varphi_{z_i})$,$w_{j,2}\sim\mathrm{Multi}(\varphi_{z_i})$

其中 $\mathrm{Dir}(\alpha)$,$\mathrm{Dir}(\beta)$ 分别为超参数为 α 和 β 的 Dirichlet 分布,$\mathrm{Multi}(\theta)$ 和 $\mathrm{Multi}(\varphi_{z_i})$ 为多项式分布.

根据上述过程,在 $\boldsymbol{\theta}$ 和 $\boldsymbol{\Phi}$ 都给定的情况下,一个双词生成的概率为:

$$\begin{aligned}P(b_i\mid\boldsymbol{\theta},\boldsymbol{\Phi}) &= \sum_{k=1}^{K}P(w_{i,1}^e,w_{j,2}^n,z_i=k\mid\boldsymbol{\theta},\boldsymbol{\Phi})\\ &= \sum_{k=1}^{K}P(z_i=k\mid\theta_k)P(w_{i,1}^e\mid z_i=k,\varphi_{k,w_{i,1}^e})\\ &\quad P(w_{j,2}^n\mid z_i=k,\varphi_{k,w_{j,2}^n})\\ &= \sum_{k=1}^{K}\theta_k\varphi_{k,w_{i,1}^e}\varphi_{k,w_{j,2}^n}\end{aligned}\tag{1}$$

在给定超参数 α 和 β 后,可以得到:

$$P(b_i\mid\alpha,\beta)=\iint\sum_{k=1}^{K}\theta_k\varphi_{k,w_{i,1}^e}\varphi_{k,w_{j,2}^n}\mathrm{d}\boldsymbol{\theta}\mathrm{d}\boldsymbol{\Phi}\tag{2}$$

则整个双词集合 B 的似然函数为:

$$P(B\mid\alpha,\beta)=\prod_{i=1}^{M}\iint\sum_{k=1}^{K}\theta_k\varphi_{k,w_{i,1}^e}\varphi_{k,w_{j,2}^n}\mathrm{d}\boldsymbol{\theta}\mathrm{d}\boldsymbol{\Phi}\tag{3}$$

本文借鉴文献[20]中使用的 collapsed Gibbs 采样算法近似求解 $\boldsymbol{\theta}$ 和 $\boldsymbol{\Phi}$. 通过不断交替地对随机变量进行后验采样,其中每次对其中一个变量的采样都基于其它随机变量的赋值. 在 BTM 中,我们需要对每一个双词 b_i 都采样一个主题,其采样后验分布为:

$$\begin{aligned}&P(z_i=k|z_{\neg i},B)\propto\\&(n_{\neg i,k}+\alpha)\frac{(n_{\neg i,w_{i,1}^e|k}+\beta)(n_{\neg i,w_{j,2}^n|k}+\beta)}{(n_{\neg i,\cdot|k}+W\beta)^2}\end{aligned}\tag{4}$$

其中 z 表示双词的主题值,下标 $\neg i$ 表示不计算当前的双词,n_k 表示主题值为 k 的双词个数,$n_{w|k}$ 表示词 w 被

赋予主题 k 的次数,$n_{\cdot|k}=\sum_{w=1}^{W} n_{w|k}$. 式(4)第 1 个因子代表代表了语料库中主题 k 所占比例,第 2 个因式则表示 $w_{i,1}^{e}$ 和 $w_{j,2}^{n}$ 属于主题 k 的概率之积.

在 BTM 中,Gibbs 采样算法首先给每一个双词随机分配一个主题作为初始状态,然后每次迭代过程都根据式(4)进行采样,并依次更新每个双词的主题. 经过足够次数的迭代之后,充分统计 n_k 和 $n_{w|k}$,利用这两个统计量可以估算出 $\boldsymbol{\theta}$ 和 $\boldsymbol{\Phi}$:

$$\varphi_{k,w}=\frac{n_{w|k}+\beta}{n_{\cdot|k}+W\beta} \tag{5}$$

$$\theta_k=\frac{n_k+\alpha}{M+K\alpha} \tag{6}$$

3　消歧算法模型

上下文信息作为实体消歧的重要参考证据,能够在一定程度上帮助文本中实体确定主题分布. 而主题漂移表示文本中的歧义实体与文本上下文信息表达的主题不一致,出现主题偏移现象. 例如词“哥伦布”可以表示航海家、城市、演员等,若不利用上下文信息,很难确定文本中“哥伦布”指代的是哪一个实体词. 但如果文本中出现词“哥伦布”的附近出现词“海洋”或者“船只”,则基本可以确定指称项“哥伦布”表示的是航海家哥伦布. 双词模型通过构建词对(例如:〈哥伦布,船只〉),首先将上下文信息以一种词联合的表示方法联系在一起,然后通过多个词对表达文本的主题信息,使之更够更加全面准确的表示上下文信息,从而减小了主题偏移的影响.

而双词模型中的词的选择则是采用实体词,实体词作为自然文本的主体词汇,对刻画文本主题具有较好的鲁棒性,因此本文选择实体词作为 BTM 双词模型〈命名实体,名词(名词短语)〉中的一个词,能够较好的抑制文本主题的偏移,文本中其他名词或者名词短语则作为双词模型中的另一个词. 例如:“迈克尔·乔丹”可以指篮球明星“乔丹”,也可以指伯克利大学教授“迈克尔·乔丹”或者其他等等(维基百科中有 8 个指称项为乔丹的实体),但如果文本中同时出现了“查尔斯·巴克利”(篮球明星、政治家等),利用实体词对〈迈克尔·乔丹,查尔斯·巴克利 〉刻画文本主题就可以表明两个实体指称更大可能是篮球明星.

3.1　模型框架

同时出现在一个文本中的实体,如果其中一个实体存在歧义,则其他实体对该实体的消歧具有辅助作用,该思想表明同一篇文档中词和实体可以映射到同一主题空间,而实体作为文本的主体,同一篇文档中同时出现的实体具有协同表达主题的能力,BTM 模型正是利用这种实体协同共现的特征,采用双词(其中一个实体,另外一个为文本名词)的思想建立主题模型,该模型保持词对中有一个为实体,确保了实体的作为文本主体的特性功能,同时可以很好的刻画实体间的协同关系,在上下文信息稀疏的情况下也能建立命名实体的主题分布,从而提高主题分布精度.

基于此,本文在 BTM 模型的基础上,提出一种半监督实体消歧算法:实体双词主题模型(Entity-Biterm Topic Modeling,EBTM),通过构建〈实体,名词(或实体)〉词对,为每一个词对分配一个主题,最终得到词对中实体的主题. 模型框图如图 2 所示.

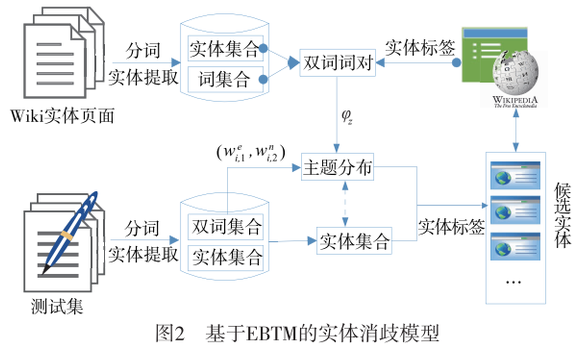

图2　基于EBTM的实体消歧模型

首先为每一个 Wiki 实体页面中分配一个单独的主题(将该主题看作一个实体标签),即对 Wiki 页面中的每一个共现词对标记该实体标签,进而通过对测试集文本构建 BTM 模型,确定文本中实体的主题分布,最后通过对测试文本中待消歧实体进行实体标注的方式,确定候选实体.

3.2　半监督算法

在 3.1 节中,所提模型算法分为两个部分,分别为训练阶段和标注阶段. 在训练阶段中,收集的 Wiki 页面集用 T 表示,对该页面中的词进行 Gibbs 采样,得到页面的词对集合,并为每个集合分配一个单独的主题 z_T(实体标签),即 z_T 表示 Wiki 页面 T 中所有共现词对 $(w_{i,1}^{e}, w_{i,2}^{n})$ 的主题分布. 在标注阶段,假设测试文本集用 D 表示,采用增量 Gibbs 采样算法对 T∪D 中的共现词对进行采样,并且对 T 中词对的主题分布保持不变,只对 D 中的共现词对进行 Gibbs 采样. 令 z_D 表示文本集 D 中共现词对的主题分布,则在每一步增量 Gibbs 采样过程中都保持 z_T 不变,只改变 z_D. 文本集 D 中每一个词对 b_i 的主题根据式(4)进行采样,且此时公式中的 $n_{\neg i,\cdot|k}$ 表示在 z_T 和 z_D 中双词分别被分配到主题 k 的总个数. 考虑到如果为每一个 Wiki 条目页面中分配一个单独的主题,将会导致主题数量的急剧增大(即主题数等于 Wiki 中实体页面个数),因此 EBTM 算法只对文本集中出现的实体对应的 Wiki 条目页面分配实体标签.

在上述 EBTM 模型中,词的主题分布完全取决于文本中词对的共现模式. 然而,随着主题数目的增加,词的主题分布空间将变得巨大,且随着迭代的进行,会出现主题漂移的现象,这时词的主题分布精度将会有所降低. 而现有的知识库(例如维基百科)具有大量的实体页面注释信息,其注释信息在一定程度上决定了实体对应主题的分布,具有保持主题中心的作用.

因此,本文在 EBTM 模型中加入半监督因子,利用维基百科的对实体词汇已有的注释提高实体标注精度. 假设 Y 表示包含带实体标签注释的词的维基百科页面,所提方法的关键思想就是使主题-词分布 φ_k 偏向于在 Y 中出现的频繁带有注释的实体词,以及使文本-主题分布 θ_k 偏向于在 Y 中页面中频繁的主题. 其概率模型图如图 3 所示.

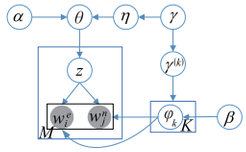

图3 EBTM半监督消歧算法概率模型

本文将维基百科注释 Y 视作一个偏向于参数 φ_k 和 θ_k 多项式观察变量,由上一节得知,θ_k 和 φ_k 分别是服从超参数 α 和 β 的狄利克雷分布. $P(\varphi_k|\beta,Y)$ 是在注释 Y 的条件下 φ_k 的后验概率,可以表示为 $\mathrm{Dir}(\varphi_k|\beta+\gamma^{(k)})$,其中 $\gamma_w^{(k)}$ 表示在 Y 中词 w 分配到主题 k 的次数. 同理也可以用超参数 $\alpha+\eta$ 表示 θ_k 的后验狄利克雷分布,其中 η_k 表示在 Y 中带注释词的主题 k 出现在文本中的个数.

根据上面的描述,在已知维基百科注释集 Y 时,最终 $z_i=k$ 的条件概率表示如下:

$$P(z_i=k|z_{\neg i},B,\alpha,\beta,Y)\propto (n_{\neg i,k}+\alpha+\eta_k)\frac{(n_{\neg i,w_{i,1}^e|k}+\beta+\gamma_{w_{i,1}^e}^{(k)})(n_{\neg i,w_{i,2}^n|k}+\beta+\gamma_{w_{i,2}^n}^{(k)})}{(n_{\neg i,\cdot|k}+W\beta+\sum_{w=1}^{W}\gamma_w^{(k)})^2} \tag{7}$$

基于此,本文通过添加两个先验观察项 $\gamma_w^{(k)}$ 和 η_k,将维基百科注释融入到了 Gibbs 采样过程中,如式(7)所示. EBTM 半监督消歧算法过程见算法 1.

算法 1 EBTM 半监督消歧算法

Input:Wiki 页面集 T、测试文本集 D、带词注释的维基百科页面 Y

Output:文本集 D 中每个 e_i 对应的实体标签.

训练阶段:

对每一个 wiki 页面中的词进行 Gibbs sampling,得到词对集合,并为该集合分配一个单独的主题(实体标签)~ z_T;

标注阶段:

1、对文本集 D 中每一个共现词对 $(w_{i,1}^e,w_{i,2}^n)$:

共现词对出现在 T 中:对词对保持 z_T 主题分布;

共现词对未出现在 T 中:统计 Y 中两个先验观察项 $\gamma_{(w}^{(k)}$ 和 η_k;

根据式(7)进行增量 Gibbs sampling,更新 z_D;

重复上述两个步骤,直到迭代次数完成或迭代收敛结束.

2、对每一个包含实体 e_i 的词对 $(w_{i,1}^e,w_{i,2}^n)$:

计算其分配到各个主题的个数 $|w_i^e\rightarrow k|$;

得到实体 e_i 的主题:

$$z_{w_i^e}^e=k,\text{where}\quad \max|w_i^e\rightarrow k|$$

3、确定 D 中每个实体 e_i 对应的候选实体.

该方法有效的结合 Y 中注释词的主题分布使得在进行 Gibbs 采样时,主题分布更加偏向于带标注词的主题分布,提高了主题分配的准确度,同时随着采样迭代不断进行,Y 中注释词不断扩展到文本集中共现的无标注的词,因此基于本文提出的 EBTM 算法是一种半监督学习方法. 在实践中这种偏向传播具有很好的性能,文献[21]已经证明了结合标注数据和大量未标注数据进行机器学习时能很大程度上提高机器学习的准确性.

3.3 复杂度分析

文章中实体消歧算法的时间复杂度主要是计算对 BTM 主题学习的时间开销,假设 N_{EM} 表示 eLDA[13] 中 EM 迭代次数,N_{ilter} 表示 BTM 模型的 Gibbs 采样次数,N_D 表示文本个数. 在 eLDA 算法中,对文档中的每个词、每个实体采样它的主题的时间复杂度为 $O(N_D(N_w+N_e))$,则总的时间复杂度为 $O(N_{\mathrm{EM}}KN_D(N_w+N_e))$. 而对于 BTM 主题学习过程,文本集组成的词对个数为 $N_DN_e(N_w-N_e)$,对每一个双词的 Gibbbs 采样时间复杂度为 $O(N_{\mathrm{ilter}}K)$,则 EBTM 的消歧算法时间复杂度为 $O(N_{\mathrm{ilter}}KN_DN_e(N_w-N_e))$. 因此两类算法的时间复杂度比值为:

$$n_b=\frac{N_{\mathrm{ilter}}N_e(N_w-N_e)}{N_{\mathrm{EM}}(N_w+N_e)}\approx\frac{(N_w-N_e)}{(N_w/N_e+1)} \tag{8}$$

由式(8)可以看出,由于每篇文档的实体个数 N_e 较小,时间复杂度的比值 n_b 趋于一个较小的常数,因此 EBTM 的消歧算法时间复杂度相比 eLDA 增加较小.

4 实验

4.1 数据和预处理

维基百科页面是目前大多数学者使用的公共数据集,广泛应用于关系抽取、实体消歧等研究领域中,本文采用 2011 年 6 月 23 日对应的维基百科中文数据资源:zhwiki-latest-pages-articles. xml,将 XML 格式数据进过处理并提取相应信息后,得到 35 万多个页面,其中重定向页面 23 万多个,实体页面 12 万多个,消歧页面 2 万多个,而重定向页面对应的 23 万多个同义词表项最终对应了 110000 多个真实的标准词项上,例如:“宝瓶口”

和“都江鱼嘴”都重定向至“都江堰”,即“宝瓶口”和“都江鱼嘴”是同义词项. 歧义词表中包括 8 千多个歧义实体,总共对应 3 万多个非歧义实体,其统计信息如表 1 所示.

表 1　同义词项和歧义实体统计信息

同义词项信息	同义词项数	230394
	标准词项数	111405
歧义词项信息	歧义实体个数	8957
	所有消歧页面中实体个数	37619
	消歧实体平均对应的实体长度	4.20

本文为了便于评测,将随机选取了五个主题:“人工智能、数据挖掘”、“素质教育、高等教育”、“体育运动”、“智能手机、安卓、苹果手机”、“流行音乐”,并在对应的 3 千多篇文章中随机抽取 800 篇文章作为本文的实验数据,其中包含词总共 1 万多个,实体 3 千多个,且每篇文章的平均实体名词数为 8,可以认为本文的实验数据是短文本. 其统计信息如表 2 所示.

表 2　实验数据统计信息

文章总数	800
命名实体总数	3216
词总数	10379
每篇文章平均命名实体数	4
每篇文章平均实体名词数	8

4.2　实验和分析

4.2.1　评价指标

实体消歧的评价指标包括召回率(recall)和准确率(precision),鉴于本文实验数据从 Wiki 页面中提取,且经过预处理后待消歧实体已经在 Wiki 页面中标注,不涉及实体抽取部分,即在实体抽取基础上进行本文的研究,因此本文只采用准确率作为实验的评价指标. 假设实验中需要进行实体消歧的总个数为 N_E,准确选择的候选实体个数为 N_{pre},则准确度 $Pre = N_{Pre}/N_E$.

4.2.2　实验对比

算法实验环境为 2.1GHz × 2 Cores CPU,4GB RAM,Ubuntu Kylin 15.10 系统. 实验参数采用文献[18]中的参数设置,$\alpha = 0.05$,$\beta = 0.01$,Gibbs 采样过程中的迭代次数设置为 1000. 考虑到本文选用的是中文实验数据,且属于基于主题模型的方法,为方便对比,实验首先选取目前在主题模型上效果最好的文献[13]中的 eLDA 算法进行对比,然后再与目前比较常见的几种算法[9,22]进行对比. 下表中涉及的 EBTM-表示 EBTM 算法不添加偏置变量 Y 时的得到的实验效果.

实验 1　主题个数对比分析

首先使用不同的主题数目对主题进行建模,验证算法的准确度,具体实验结果如表 3 所示,主题数目对算法的准确度都有较大影响.

表 3　仿真参数.

	$K=10$	$K=15$	$K=20$	$K=25$	$K=30$
eLDA	0.789	0.804	0.836	0.825	0.818
EBTM-	0.767	0.848	0.869	0.871	0.868
EBTM	0.774	0.858	0.887	0.890	0.885

表 3 结果表明基于 BTM 主题模型的实体消歧算法发挥了双词主题模型的在处理短文本(统计中每篇文中实体个数约为 4)主题的优势,通过利用实体词对不但发挥了实体的协同消歧作用,而且减小主题的偏移程度,提高了主题分类精度. 半监督消歧算法 EBTM 相比标准的 EBTM-的准确度有一定的提升,是因为 EBTM 在进行 Gibbs 采样时加入 Wiki 偏置观察量 Y 的原因.

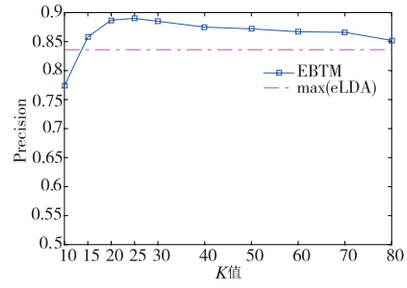

图4　EBTM算法准确度与主题数目K的关系图

eLDA 算法在主题数较小($K = 20$)时,算法取得最好性能,约为 83.6%. 而随着主题数目的增加,EBTM-和 EBTM 算法准确度不断增加,在 $K = 25$ 时,分别能够取得 87.1% 和 89.0% 的准确度. 如图 4 所示,随着主题数目的增加,EBTM 算法依然能够保证较高的准确度,在主题数目为 15 ~ 80 之间时,算法准确度都比 eLDA 最优值高,因此该算法对主题数目的敏感度较低,适用性更好.

实验 2　不同算法性能对比分析

本实验将数据预处理得到的五个主题下实验数据分别进行随机采样,取得原数据量的 1/2 数据和 1/4 数据,然后对比实验一的整体数据,实验中主题数目 $K = 25$,并与目前中文消歧几种典型算法进行对比,包括基于 LDA 主题模型[9]和基于上下文特征的 SEC-2 算法[22],结果如表 4 所示.

表 4　不同算法性能实验结果

	1/4 数据集	1/2 数据集	完整数据集
LDA	0.641	0.674	0.681
eLDA	0.798	0.819	0.836
SEC-2	0.835	0.836	0.848
EBTM-	0.808	0.853	0.871
EBTM	0.852	0.877	0.890

由表4可知,本文所提EBTM算法在性能上优于其他算法,与目前性能最好的基于上下文的SEC-2算法相比,其准确度提高了4.2%;比基于主题模型的eLDA算法提高了5.4%.由此可见本文所提算法的有效性.文献[18]中还提到BTM主题模型相对于eLDA主题模型适用于短文本、大数据集,由此表4可知,数据集相对较大时,基于主题模型的EBTM和EBTM-准确度更好,这表明了BTM主题模型能够处理较大数据集.EBTM算法相比EBTM-由于存在Wiki偏置观察量Y的原因,使得数据大小为1/4数据集时,准确度依然可以达到85.2%,比EBTM-性能下降更缓.

实验3 算法模型的主题分布

选取消歧实体:"迈克尔·乔丹",并选择其中概率最高的主题,分别对eLDA和EBTM消歧主题模型中的Top10个词和Top5个词对进行分析对比.

表5 消歧实体"迈克尔·乔丹"的主题词对分布

迈克尔·乔丹(篮球明星)		
eLDA	Top10	〈乔丹〉,〈篮球〉,〈NBA〉,〈公牛队〉,〈后卫〉,〈赛季〉,〈冠军〉,〈美国〉,〈纽约〉,〈运动鞋〉
EBTM	Top5	〈乔丹,篮球〉,〈乔丹,芝加哥公牛队〉,〈美国,NBA灌篮大赛〉,〈NCAA,NBA最佳防守球员〉,〈夏季奥林匹克运动会,NBA最有价值球员〉
迈克尔·乔丹(加州大学教授)		
eLDA	Top10	〈乔丹教授〉,〈加州大学伯克利分校〉,〈机器学习〉,〈美国〉,〈大神〉,〈计算机科学〉,〈人工智能〉,〈实验室〉,〈工程院院士〉,〈研究〉
EBTM	Top5	〈加利福尼亚大学,乔丹教授〉,〈乔丹教授,机器学习〉,〈伯克利,人工智能〉,〈加州大学伯克利分校,计算机科学〉,〈乔丹教授,人工智能〉

如表5所示,篮球明星"迈克尔·乔丹"和加州大学教授"迈克尔·乔丹"的主题分布具有较大的差异,但在Top5中某一个主题词对中的两个词有着密切的联系,共同刻画了待消歧实体的主题特征,同时词对中第一个词(命名实体,在表中用加粗表示)表示与待消歧实体频繁共现的命名实体,辅助性表明了消歧实体的真实身份.同时也可以看出,与eLDA相比,使用词对〈实体词,单词〉较好的保持了文本的主题词汇表达内容,减小了eLDA主题模型中的主题漂移现象,具有较好的鲁棒性,例如:在eLDA中,迈克尔·乔丹(篮球明星)Top10〈词汇〉中的〈美国〉,〈纽约〉和迈克尔·乔丹(加州大学教授)中的〈美国〉,〈大神〉等词分别表达迈克尔·乔丹为篮球运动员和大学教授的主题并未有太大关联;而在EBTM中,每个词对都能较为准确的刻画命名实体的相应主题.

5 结论

本文采用双词主题模型,将协同实体关系融合到主题模型中提出基于BTM的半监督实体消歧模型,通过将词对中第一个词选择为共现协同实体,有效的减小了LDA主题模型中主题偏移现象,提高了命名实体消歧准确度.在利用维基百科实体页面资源时,将实体页面注释信息看作偏置观察量,更新到Gibbs采样过程中,使得EBTM模型提高了准确度.

同时本文还具有一定的局限性,在面对文章大小长短不一的自由文本时,BTM模型性能并不比LDA模型更加优越,因此下一步计划设计更加复杂的主题模型,例如加入层次语义图关系、远程监督等,提高语义分析准确度,从而提高实体消歧准确度.

参考文献

[1] Blanco R,Ottaviano G,Meij E. Fast and space-efficient entity linking for queries[A]. Proceedings of the Eighth ACM International Conference on Web Search and Data Mining[C]. Shanghai:ACM,2015. 179-188.

[2] Mihalcea R,Csomai A. Wikify!:linking documents to encyclopedic knowledge[A]. Sixteenth ACM Conference on Conference on Information and Knowledge Management[C]. Lisbon:CIKM,2007. 233-242.

[3] Li Y,Tan S,Sun H,et al. Entity disambiguation with linkless knowledge bases[A]. Proceedings of the 25th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee[C]. Montreal:WWW,2016. 1261-1270.

[4] Habib M B,Keulen M. A generic open world named entity disambiguation approach for tweets[A]. KDIR/KMIS 2013[C]. Algarve,Portugal:Scitepress,2013. 267-276.

[5] Mikolov T,Chen K,Corrado G,et al. Efficient estimation of word representations in vector space[J]. Computer Science,2013.

[6] Mikolov T,Sutskever I,Chen K,et al. Distributed representations of words and phrases and their compositionality[J]. Advances in Neural Information Processing Systems,2013,26:3111-3119.

[7] Pennington J,Socher R,Manning C D. Glove:Global vectors for word representation[A]. Proceedings of the Empiricial Methods in Natural Language Processing[C]. Doha:EMNLP,2014. 1532-1543.

[8] Yamada I, Shindo H, Takeda H, et al. Joint learning of the embedding of words and entities for named entity disambiguation[A]. Signll Conference on Computational Natural Language Learning[C]. Lisbon: SIGNLL, 2016. 250 – 259.

[9] 王英帅,李培峰,朱巧明. 一种基于 LDA 和上下文摘要的 Web 人名消歧方法[J]. 计算机应用与软件,2011,28(7):13 – 15.
Yingshuai W, Peifeng L, Qiaoming Z. A web name disambiguation approach based on LDA and name's context snippets[J]. Computer Applications and Software, 2011, 28(7): 13 – 15. (in Chinese)

[10] Guo Z, Barbosa D. Entity linking with a unified semantic representation[A]. Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion. International World Wide Web Conferences Steering Committee[C]. Seoul: WWW, 2014. 1305 – 1310.

[11] Barrena A, Donostia B C, Soroa A, et al. Combining mention context and hyperlinks from wikipedia for named entity disambiguation[J]. Lexical and Computational Semantics, 2015: 101 – 105.

[12] Pershina M, He Y, Grishman R. Personalized page rank for named entity disambiguation[A]. Proc 2015 Annual Conference of the North American Chapter of the ACL[C]. Denver: NAACL HLT, 2015. 238 – 243.

[13] 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.
Baoxing H, Tengfei B, HengShu Z, et al. Topic modeling approach to named entity linking[J]. Journal of Software, 2014, 25(9): 2076 – 2087. (in Chinese)

[14] Nguyen D B, Theobald M, Hoffart J, et al. AIDA-light: high-throughput named-entity disambiguation[A]. Proceedings of the Workshop on Linked Data on the Web Co-Located with the 23rd International World Wide Web Conference[C]. Seoul, Korea: WWW, 2014. 1184 – 1194.

[15] 左乃彻. 基于维基百科的中英文命名实体消歧[D]. 北京:北京邮电大学,2015.
Naiqie Z. Named entity disambiguation based on Chinese and English Wikipedia knowledge base[D]. Beijing: Beijing University of Posts and Telecommunications, 2015. (in Chinese)

[16] Kulkarni S, Singh A, Ramakrishnan G, et al. Collective annotation of Wikipedia entities in web text[A]. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. [C]. Paris: ACM, 2009. 457 – 466.

[17] Li Y, Wang C, Han F, et al. Mining evidences for named entity disambiguation[A]. Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. [C]. Chicago: ACM, 2013. 1070 – 1078.

[18] Yan X, Guo J, Lan Y, et al. A biterm topic model for short texts[A]. Proceedings of the 22nd International Conference on World Wide Web[C]. Seoul: ACM, 2013. 1445 – 1456.

[19] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3(Jan): 993 – 1022.

[20] Griffiths T L, Steyvers M. Colloquium paper: mapping knowledge domains: finding scientific topics[J]. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101 (Suppl 1): 5228 – 5235.

[21] Kataria S S, Kumar K S, Rastogi R R, et al. Entity disambiguation with hierarchical topic models[A]. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. San Diego: DBLP, 2011. 1037 – 1045.

[22] 徐佳俊. 命名实体语义消歧方法的研究[D]. 上海:上海交通大学,2014.
Jiajun X. A Study of Semantic-Disambiguation Approach on Name Entities[D]. Shanghai: Shanghai Jiaotong University, 2014. (in Chinese)

作者简介

张　雄　男,1992 年生于四川万源. 现为国家数字交换系统工程技术研究中心硕士研究生. 主要研究方向为文本挖掘和信息抽取.
E-mail:979644317@qq.com

陈福才　男,1974 年生于江西高安. 现为国家数字交换系统工程技术研究中心研究员、硕士生导师. 主要研究方向为大数据处理.
E-mail:13503827650@139.com

黄瑞阳　男,1986 年生于福建漳州. 现为国家数字交换系统工程技术研究中心助理研究员. 主要研究方向为数据挖掘.
E-mail:277433109@qq.com