

# 基于微博文本的层次化实体链接方法

李禹恒<sup>1,2,3</sup>; 宋俊<sup>1,2,3</sup>; 黄宇<sup>1,2,3</sup>; 付琨<sup>1,2</sup>; 吴一戎<sup>2</sup>; 陈昊<sup>1,2,4</sup>

(1.中国科学院空间信息处理与应用系统技术重点实验室,北京 100190; 2.中国科学院电子学研究所,北京 100190; 3.中国科学院大学,北京 100049; 4.北京空间信息中继传输技术研究中心,北京 100094)

**摘要:** 命名实体链接是自然语言理解的重要研究内容,同时也是知识图谱构建及实体搜索的基础。本文提出一种针对微博文本的层次化实体连接方法(HEL)。基于用户偏好一致性假设,该方法首先对所有提及根据信息函数进行排序,得到歧义最小的提及利用消歧算法消歧,并将返回的确认实体包含进消歧函数。通过这种迭代策略让正确的结果正向传递给下一层更模糊的消歧任务。在人工标注测试集上的实验表明,本文的方法表现出良好的性能。

**关键词:** 实体链接 文本消歧 数据挖掘

中图分类号: G202

## Hierarchical Entity Linking based on Microblogs

LI Yu-heng, SONG Jun, HUANG Yu, FU Kun, WU Yi-rong, CHEN Hao

(1. Key Laboratory of Technology in Geo-spatial Information Processing and Application System, Institute of Electronics, Chinese Academy of Sciences, Beijing 100190; 2. Institute of Electronics, Chinese Academy of Sciences, Beijing 100190; 3. University of Chinese Academy of Sciences, Beijing 100049; 4. Beijing Space Information Relay Transmission Technology Research Center, Beijing 100094 )

**Abstract:** Named Entity Linking is a crucial research of Nature Language Understanding, and the foundation of Knowledge Graph construction and Entity Searching. This paper provides a Hierarchical Entity Linking method (HEL) based on Twitter content. Considering the assumption of user preference consistency, the method first rank all the candidates mentions based on proposed Information Function, and then assign the most familiar candidate to the given mention by adopting a Scoring Function. This procedure will iterate by incorporating disambiguated entities into the Scoring Function, which consequently passes on the certainty from previous linking results to the following rounds of more abstract linking tasks. Experiments on human-annotated dataset showed that this method outperforms others' work.

**Key words:** Entity linking Text disambiguation Data mining

## 0. 引言

近年来,以推特、微博为代表的自媒体(以下简称微博)在互联网时代的大背景下迅速发展成为当今社会最为重要的信息来源之一,其中仅推特一项每一天即产生超过四十亿条消息,其中包含了大量重要的实体信息和知识。互联网上,诸如维基百科、百度百科等一系列基于用户产生式内容(UGC)

构建的百科类网站的迅猛发展为人类构建大规模通用知识库提供了便利,这种以实体为单位的知识结构具有较强的语义特征,可以准确的描述现实世界中客观存在的对象,同时通过建模实体之间的关系,可以进一步形成完备的知识图谱。现在较为成熟的知识库包括: DBpedia<sup>[1]</sup>, YAGO<sup>[2]</sup>, Freebase<sup>[3]</sup>和 Probable<sup>[4]</sup>。这些现有的知识库可以用来理解海量的微博文本,作为真实的语料库,这些微博文本反过来可以补充和增强现有的知识库结构。

自然语言表达具有多样性和歧义性,因此其中涉及的命名实体往往是模糊的。如图 1 中 t1 中用户提到的 SF 既可以表示一种文学体裁(Science Fiction)又可以表示地理位置(San Francisco)。此外,相较于传统实体链接任务中的文档,作为一种非结构化的自由文本,对于微博的实体链接任务还

收稿日期: 2015-02-07

基金项目: "863"国家高技术研究发展计划(2012AA011005)

作者简介: 李禹恒(1989-),男,硕士研究生。研究方向: 文本挖掘。

E-mail: liyuheng12@mails.ucas.ac.cn

通信作者: 付琨(1974-),研究员,博士生导师。研究方向: 计算机视觉与遥感图像理解,地理空间信息挖掘与可视化。Email: fukun@mail.ie.ac.cn

受制于长度限制和口语化表达。

我们将微博实体链接定义为对微博文本中的实体名词（指称项）与知识库中唯一的实体之间的映射，若实体名词对应的实体在知识库中不存在，则被映射到一个空实体（NIL）。

实体链接任务是知识图谱构建的基础，在基于知识库的自动问答系统中，对于自然语言的消歧和链接是影响系统效力的关键因素；另一方面，在微博文本挖掘中，实体链接可以用来检测新闻动态、监视舆情、品牌管理以及个性化的推荐<sup>[5,6]</sup>。比如通过对用户微博中提及的实体进行链接分析，可以得到用户的兴趣偏好<sup>[7,8]</sup>，从而根据这些信息来进行有针对性的博文推荐<sup>[9,10]</sup>，或者提供精确地用户检索服务<sup>[6]</sup>。

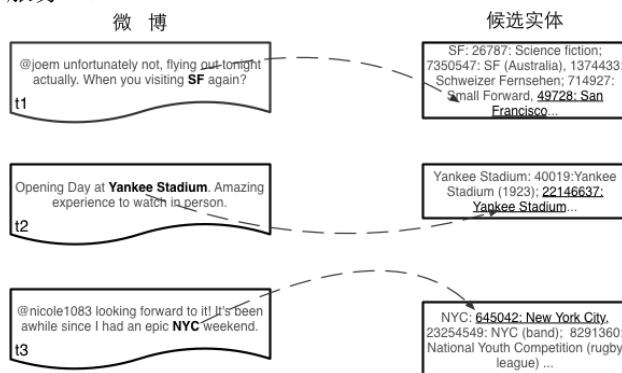


图 1 实体链接任务示例。

Fig.1 Example of entity linking

表 1 微博中不规范的用语。

Table 1 Irregular language usage in microblogs

用户	编号	微博正文
50888543	77	RT @MarcCarig: RT @BryanHoch: RT @DKnobler: And <u>Yankee Stadium</u> remains only current <u>AL</u> park where <u>Justin Verlander</u> has never won.
50888543	111	Here we go. E...A...G...L...E...S.....EAGLES!!!
101935227	27	Opening Day at <u>Yankee Stadium</u> . Amazing experience to watch in person. http://t.co/jiEyI6kE
4081481	4	Evan Turner's personal war with the <u>Bulls</u> front line ended with a foul.

针对网络文本，前人在实体链接问题上进行了较为深入的研究<sup>[11,12,13,14]</sup>，这些研究的主要的思路是通过规定一个指称项与实体页面之间的相似性度量准则来对候选实体进行打分排序，从而返回相似度最大的候选实体链接到该指称项上。然而基于上述对于微博文本的特性，不规范的用语和有限的上

下文信息令这种基于本文内部上下文静态特征关系的方法在处理微博文本链接问题上效果并不理想。如图 1 所示 t1 中 SF 的上下文并没有与地点 San Francisco 相关的信息，因此链接算法很难将指称项 “SF” 与实体 “San Francisco” 正确地链接起来。

另外一类研究则通过基于图的协同推断<sup>[15,16]</sup>，综合考虑了指称项与实体间的文档内相似度，以及指称项间、实体间的文档间相似度，在实际应用中，构建实体关系图需要消耗大量的时间，同时图中大量的非相关候选实体会为权重传递引入负面的影响。

针对上述方法的不足，本文提出了一种层次化的实体连接方法。该方法通过迭代策略让正确的结果正向传递给下一层更模糊的消歧任务，即根据指称项的模糊程度层次化的实现链接任务。

## 1. 系统构架

本文方法认为任务中的指称项既不是独立的，也不是并列，而是层次化的。该方法首先对同用户的所有提及根据信息函数进行排序，得到歧义最小的指称项利用消歧算法消歧，并将输出地确认实体包含进消歧函数，歧义较小的提及比较容易返回正确的结果，通过这种迭代策略让正确的结果正向传递给下一层更模糊的提及的消歧任务。如表 1 中用户 50888543 的第 77 条微博中 Justin Verlander 代表美国知名棒球选手，是一个歧义极小<sup>1</sup>的指称项，故首先将其链接到维基百科中编号为 3616702 的词条 Justin Verlander，进而当对指称项 Eagles、Yankee Stadium、AL 等进行消歧的时候，会倾向于选择与 Justin Verlander 关系更近的候选实体，从而将他们链接到 Philadelphia Eagles、Yankee Stadium 和 American League。方法框架如图 2 所示。

### 1.1. 预处理

预处理主要面向两部分数据，其中微博数据作为测试集，需要进行数据清洗；知识库作为训练集需要进行数据清、实体页面特征提取以及提及-实体映射表的构建。

我们首先去掉用户微博中的标点符号、“@”及后面的用户名、超链接 URL 以及转发微博标志符 RT，对非英文字符编码问题进行修正。测试集中部分人工标注的链接实体在知识库中不存在，或标注有误，同样在此阶段进行修正。

1. 维基百科中共出现 139 次 “Justin Verlander”，均指向棒球运动员 Justin Verlander (3616702)。

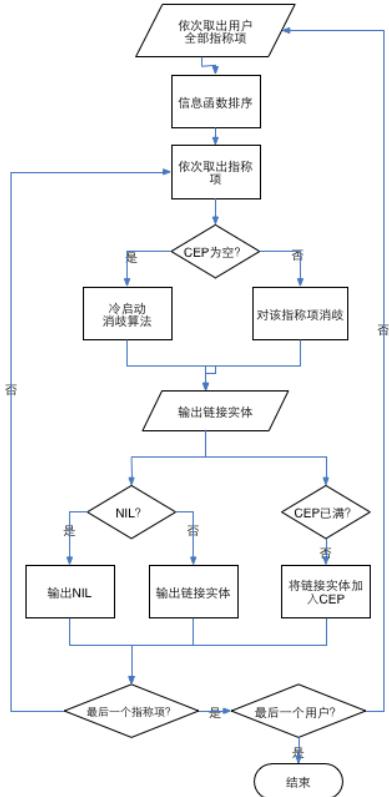


图 2 层次化实体链接算法框架。

Fig.1 Framework of HEL

本文使用 2014 年 3 月 4 日的维基百科作为训练集来训练算法模型。维基百科中每篇文章都对应一个唯一的实体，实体的不同名称经过超链接、重定向页面和歧义页面指向对应的实体本身。通过对实体页面文本的建模，我们可以得到丰富的语义信息，而通过对其中链接关系的梳理，可以进一步得到提及与实体的完整映射关系以及实体之间的关系。

## 1.2. 特征提取

根据前人的研究并结合本文模型的特点，我们选择实体流行度、上下文、临近上下文、主题关联度作为主要消歧特征。其中实体流行度与上下文特征在独立消歧模型中被广泛用来刻画实体的先验概率及语义，主题关联度常用于协同消歧算法用来描述实体间关系强度，此外，针对微博文本字数受限的问题，我们创新性引入临近上下文以进一步挖掘实体的词性特征。

### (1) 流行度

实验表明，超过 70% 的指称项链接到流行度最高的候选实体。对于某待消歧的命名指称项  $o_j^i$ ，其备选实体集合为  $E_j^i$ ，对于其中的备选实体  $e_{j,q}^i \in E_j^i$ ，其实体流行程度定义如下：

$$Pop(e_{j,q}^i) = \frac{count(e_{j,q}^i)}{\sum_{k=1}^n count(e_{j,q}^i)} \quad (1)$$

式中  $count(e_{j,q}^i)$  为  $o$  对应实体  $e_{j,q}^i$  的指向次数。

### (2) 上下文相似度

本文模型中利用上下文相似度来比较指称项与候选实体的语义相似性。通常在计算文本向量空间的相似度时采用向量之间的余弦夹角进行计算，但对于本研究课题，有些长尾实体在整个维基百科中出现的次数较少，上下文信息相对较少，在用余弦法计算其向量和命名指称项的文本向量相似度时，很容易出现相似度为零的情况，和同样其他非主流的实体不太具有区分性。因此，本文使用唐博蓉<sup>[17]</sup>提出的一种基于朴素贝叶斯的加权相似度计算方法。

对于某待消歧的命名指称项  $o_j^i$ ，其上下文向量空间表示为  $D(o_j^i) = \{d_1, d_2, d_3, \dots, d_m\}$ ，对于其中的备选实体  $e_{j,q}^i \in E_j^i$ ，其实体的上下文相似度为：

$$Sim_{ct}(e_{j,q}^i) = \frac{\sum_{k=1}^m (\log P(d_k | e_{j,q}^i)) \times \log \frac{|E_j^i|}{|\{t: d_k \in e_t\}|}}{|E_j^i|} \quad (2)$$

式中  $|E_j^i|$  为命名指称项  $o_j^i$  的候选实体个数， $|\{t: d_k \in e_t\}|$  为实体集合  $E_j^i$  中，上下文词汇表中包含词语  $d_k$  的实体个数， $\min(sim_{ctx})$  为上下文词汇表中包含  $d_k$  的实体的相似度最小值， $P(d_k | e_{j,q}^i)$  使用  $m$ -估计方法求得：

$$P(d_i | e_j) = \frac{n_q^{(k)+1}}{n_j + v} \quad (3)$$

式中  $n_q^{(k)}$  表示在备选实体  $e_{j,q}^i$  的上下文词汇表中词  $d_k$  出现的次数， $n_q$  表示备选实体  $e_{j,q}^i$  的上下文词汇表中词的总数（包括全部重复的词）， $v$  表示整个文档集中无重复的词的个数。

### (3) 临近上下文相似度

由于微博文本长度较短，传统的上下文相似度捕捉到的语义信息十分有限，甚至有的微博会出现无上下文的极端情况（如表 1 中第 111 条微博）。为了克服微博实体链接的这一问题，本文对临近上下文特征进行建模。

我们将命名实体或指称项的前一个词和后一个词分别称为临近上、下文。通过观察我们发现这些与实体名词位置上紧密相连的词包含着丰富能够反映名词性质的信息，比如微博中提到“*How did we get a New Benz? I'll show you.....*”<sup>2</sup>，文中的指称项既可以表示人名 Karl Benz，也可能是汽车品牌 Mercedes-Benz。显然文中提供的上下文并不足以支

<sup>2</sup> UID35619, Index85

撑模型做出正确的判断，但直观上，“new”这个词更多用来形容汽车而非人物，实际上“new”在“Mercedes-Benz”的上文词典中出现过 23 次，而从未在“Karl Benz”前使用过。

我们基于知识库中的文本，为命名实体建立了临近上下文词典，进而通过将待消歧的命名指称项  $o_j^i$  的临近上下文  $D(o_j^i) = \{d_l, d_r\}$  与候选实体  $e_{j,q}^i \in E_j^i$  的临近上下文词典作比较，得到指称项与候选实体的临近上下文相似度：

$$Sim_{ac}(e_{j,q}^i) = \frac{\sum_{k=l,r} \log PP(d_k | e_{j,q}^i) \times \log \frac{|E_j^i|}{|\{t: d_k \in e_t\}|}}{2} \quad (4)$$

式中  $|E_j^i|$  为命名指称项  $o_j^i$  的候选实体个数， $|\{t: d_k \in e_t\}|$  为实体集合  $E_j^i$  中，临近上下文词汇表中包含词语  $d_k$  的实体个数， $\min(sim_{ctx})$  为临近上下文词汇表中包含  $d_k$  的实体的相似度最小值， $P(d_k | e_{j,q}^i)$  利用公式 3 求得。

#### (4) 实体相关度

维基百科中的超链接蕴含丰富的实体之间的关系信息。通过对维基百科进行锚文本挖掘，可以对这种实体相关性进行建模，得到关系矩阵以辅助完成实体链接任务。

本文实体之间关联度计算方法使用 Milne 和 Witten 提出的维基百科概念之间的语义关联度计算方法 WLM<sup>[18]</sup>，这种方法基于维基百科的链接结构，其基本思想是：如果两个实体拥有更多的共享实体，那么这两个实体就越相关。对于两个实体  $e_1$  和  $e_2$ ，其语义关联度计算公式如下：

$$TR(e_1, e_2) = 1 - \frac{\log(\max(|E_1|, |E_2|)) - \log(|E_1 \cap E_2|)}{\log(|WP|) - \log(\min(|E_1|, |E_2|))} \quad (5)$$

式中  $E_1, E_2$  是分别有链接指向  $e_1$  和  $e_2$  的文档的集合， $WP$  为全部知识库实体。该公式为语义更相关的实体对赋予更高的值， $TR(e_1, e_2)$  的取值范围为 [0.0, 1.0]。

### 1.3. 信息函数

在人们试图去理解一篇微博中的不同指称项的含义的时候，如果遇到不能确定的模糊名称，会倾向于先去理解那些容易理解的名称，然后带着从确定实体中得到的先验知识去理解那些模糊的提及。如下微博：“Jordan is a super star in the field of Machine Learning!!!”，指称项 Jordan 是一个十分常见的人名<sup>3</sup>，然而 Machine Learning 则是一个具体的学科，几乎没有歧义，因此我们认为这里的 Jordan 与 Machine Learning 领域关系密切。

<sup>3</sup> 根据我们对维基百科的学习，“Jordan”可以代 282 个不同的实体

据此，我们提出信息函数的概念来衡量实体名词模糊的模糊程度，并在接下来的消歧任务中，按照信息函数的打分对不同模糊程度的指称项进行层次化的链接。

$$Info(o) = \log \left( \frac{Len(o_j^i)}{Count(o_j^i)} \right) \quad (6)$$

式中  $Len(o_j^i)$  为提及  $o_j^i$  字符串长度， $Count(o_j^i)$  为提及  $o_j^i$  的候选实体数量。

该函数的提出基于假设：1. 候选实体少的提及信息量较高；2. 字符串较长的提及信息量较高。

### 1.4. 层次化实体消歧算法

在对训练样本和测试样本提取特征之后，我们依图 2 所描述的算法框架对候选实体进行消歧。

#### (1) 冷启动

在算法起始阶段，由于确定实体池(CEP)为空，消歧主要依赖数据的静态特征进行。我们将微博  $t_i$  中的第  $j$  个指称项  $o_j^i$  的第  $q$  个候选实体表示为  $e_{j,q}^i$ 。首先对于该用户的全部指称项  $O_u$ ，根据信息函数利用指称项形式上的特征进行排序得到  $O'_u$ ，排在前面的指称项具有更清晰表达和较低的歧义。针对每一个指称项的全部候选实体  $E_j^i$ ，从概率层面上，其流行度能够有效的反应其被人们所熟知的概率；语义上层面上，候选实体与指称项之间的上下文相似度可以捕捉其语义关联；词性层面上，二者的临近上下文相似度可以为描述候选实体的属性提供更多的参考。这三项特征在候选实体间相互独立，因此可以作为静态独立特征来描述候选实体与指称项之间的亲疏程度。基于以上特征，我们对每个候选实体  $e_{j,q}^i \in E_j^i$  利用静态相似方程加权打分得到  $p_{j,q}^i$  如下：

$$p_{j,q}^i = \alpha Pop(e_{j,q}^i) + \beta Sim_{ct}(e_{j,q}^i) + \gamma Sim_{ac}(e_{j,q}^i) \quad (7)$$

式中  $Pop(e_{j,q}^i)$  代表实体流行度， $Sim_{ct}(e_{j,q}^i)$  代表实体与指称项上下文的相似度， $Sim_{ac}(e_{j,q}^i)$  代表实体与指称项临近上下文的， $\alpha, \beta, \gamma$  分别为各项系数，通过实验学习得到， $\alpha + \beta + \gamma = 1$ 。

#### (2) 层次消歧

微博文本与传统文档的不同之处在于具有用户属性，这里我们假设同一用户发布的微博具有一定的主题相关性，那么已经识别出的用户微博中的实体可以作为先验知识来辅助后继的消歧任务。

为了保证确认实体池的效力，我们首先根据对数据的分析，制定流行度阈值  $\theta$  来限制哪些实体可以

进入确定实体池，并设置实体池容量  $V$ ，以在保证信息量充足的前提下降低算法复杂度。具体参见算法 1。

当确定实体池不为空的时候，如图 2 所示模型将选择使用层次消歧算法。算法会根据公式 6 逐一计算指称项的候选实体与确认实体池  $C$  中每个实体  $c_s (s \in V)$  的实体相关性，并返回其平均实体相关性  $\overline{TR}_{j,q}^i$  作为特征补充进公式 8。得到如下公式：

$$d_{j,q}^i = p_{j,q}^i + \mu \overline{TR}_{j,q}^i \quad (8)$$

式中  $p_{j,q}^i$  代表指称项  $o_j^i$  与其候选实体  $e_{j,q}^i$  的静态相似性得分， $\overline{TR}_{j,q}^i$  为该候选实体与确定实体池中实体相关性的平均值， $\mu$  为权重。

#### 算法 1：层次消歧算法

*Algorithm 1: Hierarchical disambiguation algorithm*

**输入：** 用户  $u$  的待链接指称项  $o_j^i \in O$   
**输出：** 每个指称项  $o_j^i$  对应的实体  $e_j^i \in E$

```

1. Initialize the CEP      //CEP.count=0
2. foreach o_j^i in O
3.     计算 Info(o_j^i)
4.     O_rank=O.sort(info(o_j^i)) //排序得到O_rank
5.     foreach o in O_rank
6.         if(CEP != 0)
7.             foreach (can in E)
8.                 Score = d_{j,q}^i //by 公式 9
9.         else
10.            foreach (can in E)
11.                score = d_{j,q}^i //by 公式 8
12.            res={e|score = max(score)}
13.            If(res.pop>θ && CEP.count<V)
14.                CEP.Add(res)

```

## 2. 实验结果及分析

本章主要介绍针对我们算法进行实验分析，以检验算法的效果，论证可行性及存在的问题。

### 2.1. 实验数据

根据我们的调研，目前尚没有针对微博文本的统一公开数据集，我们使用清华大学 Wei Shen 提供的数据集进行实验。该数据集利用 Twitter API 从随机采样得到的 71,937 名 Twitter 用户的微博中提取 3,200 条最新微博，并进一步从中随机选择 20 名用户，每名用户 200 条微博（不足 200 则全部纳入）进行人工标注，形成标准实验集。关于该数据集的

详细情况见表 2，从中可以看出 3,818 条微博中有 1,721 (45.08%) 至少含有一个实体指称项。另有 241 条指称项难以对其作出准确判断，标记为 **uncertain**；437 条指称项被判定无任何知识库中的实体与之对应而链接到 **NIL**。最终，我们过滤掉不确定指称项，保留不可链接指称项，得到总计 2,677 条测试用指称项。

我们下载了 2014 年 3 月 4 日的维基百科作为训练集（见 1.1），为 1.2 提取流行度、上下文相似度等特征，并建立指称项到实体的映射表。

表 2 数据集概况

Table 2 Summary of the data set

用户数	20
微博数	3818
至少含有一个实体指称项的微博数	1721
命名实体指称项总数	2918
不确定指称项数	241
测试用指称项总数	2677
可链接指称项总数	2239
不可链接指称项总数	438

### 2.2. 评测指标

实体链接通常使用 **NIL 准确率** 和 **InKB 准确率** 分别衡量实体链接对于空实体的判别能力以及多候选实体的消歧能力。其中

$$\text{NIL 准确率} = \frac{\text{被链接到 NIL 的不可链接指称项个数}}{\text{不可链接指称项个数}} \quad (9)$$

$$\text{InKB 准确率} = \frac{\text{被链接到 NIL 的不可链接指称项个数}}{\text{不可链接指称项个数}} \quad (10)$$

此外，算法效力的综合评价指标参考 TAC KBP 实体链接评测任务的主要评测指标，使用 **Micro-averaged accuracy**，即所有链接结果的平均准确率：

$$P_{\text{Micro}} = \frac{\sum_{q \in Q} \sigma(L(q), C(q))}{|Q|} \quad (11)$$

式中  $Q$  是所有 query 的集合， $L(q)$  是实体连接系统给出的 query  $q$  的目标实体 ID， $C(q)$  是 query  $q$  的准确目标实体 ID， $\sigma(L(q), C(q))$  用于判断  $L(q)C(q)$  是否相同，相同为 1，否则为 0。

### 2.3. 实验结果

实体链接算法效果分析：本文采用 2.1 所述数据集进行算法效果测试。由于本文所述方法基于用户特性，因此测试中以用户为单位进行实验。试验中我

们分别比较引入不同特征时静态消歧算法的消歧效果以及引入层次化迭代策略后的影响。实验结果如

表 3 所示，其中基于公式 8 的独立特征消歧方法及层次消歧方法分别用 LOCAL、HEL 表示。

表 3 实验结果分析

Table 3 Experimental results over the data set

特征	可链接		不可链接		整体	
	正确链接数	准确率	正确链接数	准确率	正确链接数	准确率
LOCAL <sub><math>\beta=0, \gamma=0</math></sub>	1775	79.28%	311	71.00%	2086	77.92%
LOCAL <sub><math>\gamma=0</math></sub>	1887	84.28%	311	71.00%	2198	82.11%
LOCAL <sub><math>\beta=0</math></sub>	1944	86.82%	311	71.00%	2255	84.24%
LOCAL <sub>full</sub>	1988	88.79%	311	71.00%	2299	85.88%
HEL <sub><math>\beta=0, \gamma=0</math></sub>	1783	79.63%	312	71.23%	2095	78.26%
HEL <sub><math>\beta=0</math></sub>	1891	84.46%	312	71.23%	2203	82.30%
HEL <sub><math>\gamma=0</math></sub>	2012	89.86%	313	71.46%	2325	86.85%
HEL <sub>full</sub>	2019	90.17%	313	71.46%	2332	87.11%

试验中我们发现，对于微博文本，其上下文相似度对算法准确度的提升不及临近上下文相似度，这也符合 1.2 中我们的假设。具体来看，表 1 第四个例子中指称项 “Yankee Stadium” 利用本文算法在不同特征下的打分如表 4 所示。在 Wikipedia 中指称项 “Yankee Stadium” 被指向电影 “Yankee Stadium(1923)” 的次数要高于运动场 “Yankee Stadium”，同时微博文本较短，能够提供的上下文信息也十分有限，而且两个候选实体描述的对象实际上也很相似，因此上下文相似性特征并未起到很强的区分作用。然而通过计算二者的临近上下文相似度，我们发现介词 “at” 出现在地点名词 Yankee Stadium 前的概率大大高于出现在电影之前，因此通过这种词性层面的特征，我们的算法可以对指称项链接成功。

表 4 静态链接方程消歧分析

Table 4 Analysis on static linking function disambiguation

特征	Yankee Stadium	Yankee Stadium (1923)
Pop	0.45	0.54
Pop+Ct	-1.58	-1.38
Pop+Ct+Ac	-6.26	-7.34

此外，确认实体池在一定程度上能够反映用户的特征，因为里面捕捉到那些用户切实关注到的，比较具体的实体，而通过计算这些实体与其他实体的相关性，又可以增强很多协同相关的长尾候选实体的打分，从而增强链接准确性。表 5 是通过本文方法得到的确认实体池，可以看出其捕捉到了不同用户所关心的歧义很小的实体，显然用户 14473492

对科技主题更感兴趣，而用户 4081481 热衷于篮球赛事，结合 1.2 中提到的实体相关性，那些同主题的模糊实体消歧效果可以得到增强，如表 1 中 4081481 的第 4 条微博中的指称项 “Bulls” 会被链接到 Chicago Bulls (篮球队) 而非 Bulls (rugby union) (橄榄球队)。

表 5 确认实体池举例

Table 5 Example of certain entity pool

**信息函数准确率-覆盖度：**利用信息函数计算 1.3 中微博提及的信息量，结果如表 6 所示，实验结果

用户	14473492	4081481
CEP	OS X	LeBron James
	Android (operating system)	Kobe Bryant
	San Francisco Bay Area	Utah Jazz
	Google	Michael Jordan
	Nokia	YouTube
	.....	.....

符合预期：经过信息函数排序后，约前 50% 的提及的链接准确度可以达到 95% 以上，这一点保障了后文中确认实体池的效力。例如用户 347276428 的所有命名指称项经过信息函数排序后，从熵最低的一条开始依次进行实体消歧，当累计链接准确率降至 95% 时，其处理覆盖率达到 67.44%，当累计链接准确率降至 90% 时，其处理覆盖率达到 72.09%，以此类推，当累计链接准确率降至 79% 时，其处理覆盖率达到 100%。试验中我们将 CEP 大小设置为 10，即将排序后符合条件的前十个链接结果纳入 CEP。

表 6 命名指称项的信息量  
Table 6 Information of mentions

Mention	len(o)	C(o)	Info(o)
ML	2	30	-1
Machine Learning	16	2	0.90
Jordan	6	282	-1.67

### 3. 结 论

本文针对传统方法在微博实体链接任务中存在的问题，创新性提出一种层次化的实体链接方法，解决了微博上下文信息不足导致的链接准确率不高的问题。主要通过临近上下文捕捉指称项词性来补充上下文知识，并结合信息函数和层次化的实体消歧策略，产生确定实体池来支持下一轮的实体链接任务。通过在用户标注数据集上的实验表明，本文方法切实可行，并具有较高的链接准确率。然而本文算法能力倾向于 InKB 实体消歧，对于空实体判别问题仍有较大进步空间。此外，目前实体链接算法都是通过学习知识库来提取实体特征，然而微博文本与百科型知识库，无论遣词造句还是主题分布都会有很大的差别，知识库中的流行度并不可以完全表征微博词频。综上所述，我们下一步的研究将集中在生成优质候选集，以及微博文本的特征提取。

#### 参考文献：

- [1] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, and Z. Ives. Dbpedia: A nucleus for a web of open data. In ISWC'07.
- [2] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: A core of semantic knowledge unifying wordnet and wikipedia. In WWW'07.
- [3] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In SIGMOD'08.
- [4] W. Wu, H. Li, H. Wang, and K. Q. Zhu. Probase: a probabilistic taxonomy for text understanding. In SIGMOD'12.
- [5] J. Chen, R. Nairn, L. Nelson, M. Bernstein, and E. Chi. Short and tweet: experiments on recommending content from information streams. In CHI'10.
- [6] J. Weng, E.-P. Lim, J. Jiang and Q. He. Twitterrank: finding topic-sensitive influential twitterers. In WSDM'10.
- [7] M. Michelson and S. A. Macskassy. Discovering users' topics of interest on twitter: a first look. In AND'10.
- [8] Z. Xu, L. Ru, L. Xiang, and Q. Yang. Discovering user interest on twitter with a modified author-topic model. In WI-IAT'11.
- [9] J. Chen, R. Nairn, L. Nelson, M. Bernstein, and E. Chi. Short and tweet: experiments on recommending content from information streams. In CHI'10.
- [10] K. Chen, T. Chen, G. Zheng, O. Jin, E. Yao, and Y. Yu. Collaborative personalized tweet recommendation. In SIGIR'12.
- [11] R. Bunescu and M. Pasca. Using Encyclopedic Knowledge for Named Entity Disambiguation. In EACL'06.
- [12] S. Cucerzan. Large-Scale Named Entity Disambiguation Based on Wikipedia Data. In EMNLP-CoNLL'07.
- [13] X. Han, L. Sun, and J. Zhao. Collective entity linking in web text: a graph-based method. In SIGIR'11.
- [14] J. Hoffart, M. A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, and G. Weikum. Robust disambiguation of named entities in text. In EMNLP'11.
- [15] W. Shen, J. Wang, P. Luo, M. Wang. LINDEN: Linking Named Entities with Knowledge Base via Semantic Knowledge. In WWW'13.
- [16] W. Shen, J. Wang, P. Luo, M. Wang. Linking Named Entities in Tweets with Knowledge Base via User Interest Modeling. KDD'13.
- [17] 唐博蓉 基于维基百科的命名实体消歧研究
- [18] D. Milne and Ian H. Witten. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links [J], In Proc. of AAAI, 2008