文章编号:1000-5862(2015)02-0111-06

中文跨文本人名同名同指消解研究

陈 晨,王厚峰*
(北京大学信息化建设与管理办公室,北京大学计算语言学教育部重点实验室,北京 100871)

摘要:跨文本命名实体同指是指出现在多个文本中的相同名字指称相同对象.同指消解则是判断相同的名字是否指称相同对象的过程.跨文本同指消解对于多文本摘要和信息融合等具有重要作用.针对中文中最典型的命名实体——人名,研究了使用层次聚类方法在进行跨文本同指消解中的 2 个重要问题:特征选择和聚类停止条件判断.

关键词:人名同指消解;层次聚类;特征选择;停止条件
中图分类号:TP 391 **文献标志码**:A **DOI**:10.16357/j.cnki.issn1000-5862.2015.02.01

0 引言

命名实体的识别和解析是自然语言处理中的一类典型问题,是信息抽取,信息检索等应用领域的关键技术[1].命名实体有多种类型,如人名、地名、机构名、影视剧名、小说歌曲名等.相比而言,人名的出现更频繁,相关的应用也更多.跟踪社会名人、搜索专家成果等都与人名相关.与此同时,人名的歧义现象又特别突出.例如,仅在“百度百科”上,收录的王刚(2015 年 1 月 12 日)就达 85 个.上世纪 90 年代的一次世界人口普查,名为“约翰.史密斯(John Simth)”的人有 50 124 个.2015 年 1 月 12 日在国内的同名同姓网上显示,“李娜”有 258 581 个(http://www.tongmingtongxing.com/index.asp),这还不是该网上同名数量最多的.可见,消解命名实体歧义,尤其是人名同名歧义,是一项挑战性极大的工作.

所谓人名同指消解就是根据其所在的上下文确定人名具体指称意义的过程.跨文本人名同指消解已经是近年来自然语言处理领域研究的热点[2-4].前几年举办的 WePS(Web People Search)评测和近几年正在开展的 KBP 评测都是围绕这一内容展开的或者与此内容密切相关[5-6];在中文信息处理方面,最近几次连续举办的 SIGHAN 评测也均设置了相应的任务[7-8].

为了消解命名实体的歧义,也可以事先为不同的命名实体建立知识库,在知识库中对实体加以定义和描述.这样,对实际文本中出现的命名实体消歧,就是将其链接到知识库中对应的实体上.也就是说,通过链接关系确定实体的指称语义.这类命名实体消歧任务也称为实体链接(Entity Linking).TAC(Text analytics conference)中的 KBP(Knowledge based Population)评测属于这一类任务.这一任务类似于“词义消歧”,而上面的跨文本命名实体同指消解任务类似于“词义发现”.当然,由于命名实体是完全开放的,较难对所有的命名实体全部定义.因此,即便是 Entity Linking 任务,也需要处理“空链接”的问题[1,9].

在汉语中,命名实体同指消解面对的问题更多.由于汉语的命名实体有可能与普通词相同,在消歧之前,还需要先识别命名实体.例如,“白云”可以是普通词,也可以是人名;而“金山”可以是人名,地名以及机构名,而且还可以不是名字.此外,有些名字在特殊的情况下可能不完整,只是其他名字的一部分.例如,“林海”,在一些文本中可以是一个完整的人名,而在“林海峰”,“谢林海”中只是名字的一部分.CLP-SIGHAN2010 和 CLP-SIGHAN2012 的评测均覆盖了这一系列的问题[7-8,10].

本文主要探讨跨文本人名同名歧义消解中一些关键的问题.对于命名实体而言,假定没有为实体提供知识库.

收稿日期:2015-01-24
基金项目:国家自然科学基金(61370117,61333018)和国家社科重大课题(12&ZD227)资助项目.
通信作者:王厚峰(1965-),男,湖北天门人,教授,博士生导师,主要从事语言信息处理、计算语言学、文本信息处理、指代消解和机器翻译等方面的研究.

1 问题描述

所谓的跨文本人名同指消解,是指当某个人名 $Pname$ 出现在若干个文本中时,判断哪些文本中出现的人名 $Pname$ 具有相同的指称语义(指向同一个人),并根据指称是否相同,将文本划分成若干个子集合.

可以更形式地将这一问题描述为:假设有 1 个文本集合 $D = \{d_1, d_2, \cdots, d_n\}$,其中的每个文本 $d_i (1 \leqslant i \leqslant n)$ 都含有字串 $Pname$,这一字串可能是人名.在集合 D 中 $Pname$ 所指称的真实实体的个数 k 事先并不确知,各个真实实体事先也不清楚.跨文本人名歧义消解就是根据 $Pname$ 所在文本中的信息确定指称语义,根据指称语义的相同与否将 D 划分成若干子集合,使得同一子集 D_j 中每个文本所含的 $Pname$ 都有相同的指称对象,而不同的子集合中的文本所含的 $Pname$ 具有不同的指称语义,即 $D_j \subseteq D, D_i \cap D_j = \varnothing$,其中,$1 \leqslant i, j \leqslant k, i \neq j$.

由于没有 $Pname$ 的任何描述,这一任务通常才用聚类方法求解,即根据 $Pname$ 的语义信息的相同与否,将文本集合 D 通过划分成若干个子集合.

2 基于聚类方法的同名歧义消解

跨文本同名歧义消解也称为人名同指(Personal name coreference).最常用的方法是聚类方法.但采取聚类方法面临一系列的问题.本文重点考察了特征选择与特征值计算以及类划分的终止条件判断这 2 个基本问题.

2.1 特征选择与特征值计算

跨文本人名同指消解的关键问题是从文本中选出最能代表人名 $Pname$ 所指称的实体的特征.通过这些特征,将不同的实体区分开来.

2 个不同的人即便有相同的名字,但是仍然存在大量的本质上有区别的属性.一般而言,有下面一些特征:(i) 个人的基本信息.不同的人有不同的个人信息,如,性别、出生地、出生时间、民族、学历、从事的行业(专业)或常参与的活动等,所属机构或所属地;(ii) 个人的社交圈.不同的人有不同的社交网络,包括经常出入的场所,时常关联的人和机构等.针对(ii),笔者曾经基于社会关系网络研究了同名实体的同指消解问题[4].基本思想是将文本中同现的实体作为待消解的人名的社会关系网络,以此为基准,进行社团的划分.而针对(i),本身又是一个很难的问题.首先,并不是每个包含有人名 $Pname$ 的文本都会有个人的基本信息;其次,即便有基本信息,如何识别并提取出来也并不是一件容易的事情,这也是信息抽取问题的核心和难点.

为了降低特征抽取的难度和复杂度,现有的方法大部分直接在窗口范围内取词为特征.本文依然沿用这一思想,但重点考察了如下 3 种情况:(i) 窗口大小的设置.主要考虑了 2 种窗口:以 $Pname$ 所在的段落为窗口,以及不设窗口大小限制,而是取全文;(ii) 选取窗口内对人名有贡献的词作为特征.有贡献的词主要为实词,包括名词、动词、形容词、缩略语和其他命名实体;(iii) 扩展已选特征词的同义词,减少因数据稀疏产生的问题.

基于选取的特征,需要考虑特征值计算方法.目前最为典型的方法是 $tf \cdot idf$ 式子,但实验结果表明,在 $tf \cdot idf$ 基础上,融入互信息得到的效果更好.一个词和 $Pname$ 之间的互信息计算公式为

$$MI(Pname, w_i) = \frac{P(Pname, w_i)}{P(Pname) \cdot P(w_i)} = \frac{df(Pname, w_i)/N}{df(Pname) \cdot df(w_i)/N^2} = \frac{df(Pname, w_i) \cdot N}{df(Pname) \cdot df(w_i)},$$

其中 $df(Pname, w_i)$ 表示名字 $Pname$ 和词 w_i 同时在多少文本中出现,N 为含有 $Pname$ 汉字串的文本的个数;$df(Pname)$ 表示 N 个文本中,真正以 $Pname$ 为完整名字的文本数,$df(w_i)$ 表示含有词 w_i 的文本数.于是,融合互信息后,特征词在特定文本中的特征值计算为

$$Weight(w_i, Pname, d_j) = (1 + \log(freq(w_i, d_j))) \cdot \log(N/df(w_j)) \cdot \log(1 + MI(Pname, w_i)).$$

2.2 聚类终止条件

聚类算法的一个难题是如何确定类的个数或者如何确定划分的终止条件.经典的 K-means 算法直接假设类的个数 K,这对于跨文本的人名同指消解显然不合适.从已有的大量的评测(如 CLP-SIGHAN2010/2012 和 WePS)可以看出,不同的人名指称的真实实体数目相差非常大,CLP-SIGHAN2010 中一个名字少的指称 4 个不同的人,多的指称 160 多个不同的人.无论是自顶向下的划分,还是自底向上的聚合,都涉及到终止条件的判断问题,下面是几种典型的方法.

2.2.1 固定阈值法 在运用层次聚类算法时,一种简单而实用的方法是使用固定阈值.固定阈值主要有 2 种:(i) 类别个数,即当聚类结果达到预先设定的类别数后便停止进一步聚合;(ii) 最低相似度阈值,即当类与类相似度低于固定阈值时就停止继续

聚合. 对于第(i) 种方法,面临着与 K-means 相同的难题. 对于第(ii) 种情况,如何确定阈值也需要大量的经验,阈值的选择是否合适直接影响结果的好坏.

2.2.2　准则函数　聚类问题可以看成优化问题: 类内元素之间相似值和最大,类间元素的相似值和最小. 优化条件可以看成停止聚类的准则. 基于这一思想,可以引入准则函数. 可以从 3 个角度考虑: 类内准则函数(Internal Criterion Functions) 、类间准则函数(External Criterion Functions) 和混合准则函数(Hybrid Criterion Functions) . 类内准则函数应确保聚类的结果满足类内相似度最大或者类内距离最小; 类间准则函数应确保结果满足类间相似度最小或者类间距离最大; 混合准则函数将类内准则函数和类间准则函数结合起来.

G. W. Minlligan 等在 1985 年对 30 种停止条件准则函数进行对比实验后,发现 R. Calinski 等提出的 C&H 指标表现最佳[11-12]. C&H 指标的计算方法为

$H(k) = (E(k)/(k-1))/(I(k)/(n-k))$,

其中 k 为类别个数,n 为文本总数,$E(k)$ 为 k 个类别中心点与全局中心点之间的距离,$I(k)$ 计算 k 个类中的任意向量与类别中心向量的距离之和. R. Calinski 等用欧式距离计算 E 和 I,并指出使得函数 $H(k)$ 取最大值的 k 对应最佳聚类类别数. 假设 $\boldsymbol{C}$ 为文本全集中心向量,$\boldsymbol{C}_r$ 为第 r 类的中心向量,S_r 为划分属于第 r 类的文本集合,N_r 表示类别 r 中的文本数量,C&H 方法将准则函数计算 $E(k)$ 和 $I(k)$:

$$E(k) = \sum_{r=1}^{k} N_r L_2(\boldsymbol{C}_r, \boldsymbol{C}),$$

$$I(k) = \sum_{r=1}^{k} \sum_{d_i \in s_r} L_2(\boldsymbol{d}_i, \boldsymbol{C}_r),$$

其中 $L_2(\cdot,\cdot)$ 表示欧氏距离.

2.2.3　自动阈值法　上面 2 种方法都有明显的缺点,准则函数的停止条件无法处理极端情况(如 $k=1$ 或 $k=n$) ; 而固定阈值的停止条件将统一的阈值使用到所有的人名消解过程中,但不同人名歧义程度不同,阈值的确定很难具有普遍适用性. 针对上述情况,本文提出了在聚类过程中,为每个类自动估计相似性阈值的方法.

以人名消解为目的的文本聚类与普通文本聚类最大的区别就是,前者蕴含着 1 个前提: 不包含某个人名的文本无法与包含该人名的文本聚为 1 类. 基于这一思想,可以将这两者的相似度作为阈值. 即在人名消解过程中,当待聚合的文本相似度低于其一与不包含该人名的文本集合的相似度时就停止聚类. 不包含该人名的文本较容易获得.

不包含某人名的特征向量权值计算方式与前面介绍的方法一致. 但是 tf 的计算方法改为用在所有不含该人名的文本中出现的平均次数. 假设 $\boldsymbol{D}_{-name}$ 表示不包含歧义人名的文本集合中心向量,便有计算公式

$$New\_Weight(w_i, \boldsymbol{D}_{-name}) = (1 + \log(freq(w_i, \boldsymbol{D}_{-name}) / |\boldsymbol{D}_{-name}|)) \log(|\boldsymbol{D}| / df(w_i)) \cdot \log(1 + MI(w_i, name)).$$

将 $\boldsymbol{D}_{-name}$ 的特征向量作为计算相似度阈值的参考向量,在层次聚类过程中,寻找待聚合的 2 类文本,需要满足下列条件: (i) 2 类之间的相似度大于这 2 类与参考向量之间的相似度; (ii) 在满足条件(i) 的情况下,类间相似度大于等于其他任意 2 类.

无法满足条件(i) 时,就停止聚类,从而完成由自动生成的阈值判断聚类是否停止的过程. 当 2 个类合并为一类后,系统需要计算新类与参考向量 $\boldsymbol{D}_{-name}$ 之间的相似度,作为新类的相似性阈值. 在计算新类的参考阈值时,如表 1 所示.

表 1　自动生成阈值的计算方式

函数	定义
最大值	集合中样本与 $\boldsymbol{D}_{-name}$ 的最大相似度
最小值	集合中样本与 $\boldsymbol{D}_{-name}$ 的最小相似度
平均值	集合中样本与 $\boldsymbol{D}_{-name}$ 的平均相似度
中心值	集合中样本中心向量与 $\boldsymbol{D}_{-name}$ 的相似度

3　实验数据与实验结果

3.1　实验数据

本文采用 CLP 2010 的中文人名消歧任务训练数据作为实验测试集. 该训练数据共有 32 个中文人名的文档集. 每个人名对应的文档集含有 100 ~ 300 个文档. 丢弃文本数的比例从 0 ~ 94% 不等(如果文本中不含有待消解的人名,则选择丢弃,如“高明”可能是形容词不是人名) . 在 32 个人名中,人名指称的真实实体总数从 4 ~ 166 不等,随机性较强. 语料包括下面 3 种情况:

(i) 给定的汉字串在一些文本中的确表示名字,而且这个串正好表示一个完整的名字,如下面文档中“杨波”正好是选定的人名: (记者朱国贤) 1990 年浙江省“十佳”运动员评选结果今天揭晓. 在世界杯体操赛上获得平衡木金牌的杨波名列榜首.

(ii) 给定的字串在文本中表示人名的一部分. 例如,给定了“何海”,文本中的真实名字是“何海

霞”:研究会的会长由著名教授袁晓园担任,副会长有董寿平、周怀民、何海霞、宫达非、孙轶青、张旭、欧阳中石等.

(iii) 给定的字串在文本中不是人名或人名的一部分,这一字串很可能就是一个普通的词,甚至连普通词都不是. 例如,下面的“高军”:现年 69 岁的贝穆德斯原是尼加拉瓜前独裁者索摩查的国民卫队中的一名上校. 桑地诺民族解放阵线执政期间,他是反对桑解阵政府的武装部队的最高军事领导人. 1990 年桑解阵在大选中失败,查莫罗夫人当选总统.

3.2 评测实验

CLP 2010 使用了 B_Cubed 和 P_IP2 种评测指标,详见评测中对 2 种指标的说明.

3.2.1 特征选择的对比实验 本文主要从窗口大小选择和语义扩展 2 个方面做了对比实验. 在窗口大小上,考虑了 2 种情况,即全文或者名字所在的段落. 在特征选择上,选择了窗口范围内的名词、动词、形容词、命名实体、缩写词,因为这些词具有更强的意义表达能力. 在语义扩展方面,使用了 LTP 工具,对预处理后词义消歧中对词义的解释(哈尔滨工业大学信息检索实验室的“语言技术平台 LTP”),即 wsdexp 的内容,如图 1 所示.

```
cont = “军事” wsd = “Di11” wsdexp = “军队_战争” pos = “n”ne
cont = “法院” wsd = “Dm02” wsdexp = “司法机关_监狱” pos = ''
```

图 1 基于 LTP 的语义扩展

实验结果如表 2 所示,“文本”表示从整篇文本中提取特征,“文本 + 语义扩展”表示从整篇文本中提取特征,并将扩展后的词义加入特征集合;同理,“段落”代表从段落中提取特征,“段落 + 语义扩展”代表从段落中提取特征,并加入语义扩展特征. 对 32 个人名分别进行试验,给出 B_Cubed 和 P_IP 下各个评测指标平均值,B_Cubed 的“#Pre”为精确率,“#Rec”为召回率,P_IP 的“#Pre”为 Purity,“#Rec”为 InversePurity,“F”为 F 值. 使用向量的余弦作为相似性计算,对 32 个人名中的每个人名按照聚合式层次聚类的最佳结果作为最终结果,再计数按平均值,得到表 2 所示的结果.

表 2 不同特征情况下的实验结果比较 %

	B_Cubed			P_IP		
	#Pre	#Rec	#F	#Pre	#Rec	#F
文本	88.92	87.47	87.52	90.99	91.05	90.76
文本 + 语义扩展	88.87	87.56	87.51	90.98	90.97	90.69
段落	89.73	86.69	87.77	91.61	90.92	91.01
段落 + 语义扩展	89.78	87.09	88.01	91.70	91.12	91.15

从表 2 可以看到,用“段落 + 语义扩展”的方法获得的特征人名消解实验达到了本实验的最好效果. 主要有以下几个方面的原因:(i) 从整篇文本中提取特征会引入过多的噪音,尤其当人名并非文本讨论的中心时,而从人名所在段落提取特征;(ii) 在文本中提取特征后,加入语义扩展特征效果差,可能的原因是带有噪音的特征经过语义扩展引入了更多噪音;(iii) 从段落提取特征后,加入语义扩展特征效果较好,是因为引入了有用信息,降低了数据的稀疏性.

因此,实验得出了这样的结论:在进行人名消解特征选择时,与人名共现距离较近的词相对较远的词带有更多的信息,同时,只有当特征信息噪音较少时,加入语义扩展特征才能改进实验结果. 笔者进一步对层次聚类的终止条件计算作了比较,结果如表 3 所示.

表 3 不同终止条件下的实验结果比较:P_IP 的 F 值 %

	最佳结果	准则函数 C&H	自动生成阈值			
			最大值	最小值	平均值	中心值
高军	94.43	11.26	93.97	93.04	93.04	93.97
高明	72.70	70.86	72.83	46.20	74.10	72.83
高伟	88.96	62.50	78.68	53.80	65.65	83.94
郭华	97.90	41.53	95.03	90.77	95.03	95.76
郭伟	89.74	67.90	84.57	73.68	85.65	87.08
郭勇	98.30	26.95	85.21	68.66	98.30	67.74
何海	97.40	53.54	94.69	92.33	95.19	94.69
何涛	97.83	19.01	56.96	97.04	80.20	34.12
胡刚	98.51	31.45	95.82	96.12	95.82	95.82
胡明	98.95	47.81	97.91	96.26	98.16	97.91

表 3(续)

	最佳结果	准则函数 C&H	自动生成阈值			
			最大值	最小值	平均值	中心值
黄海	59.56	59.56	58.15	58.15	58.15	58.15
黄明	91.72	89.00	81.60	75.97	79.83	82.36
李刚	88.87	56.25	75.37	61.29	85.92	64.07
李军	88.58	87.95	78.38	35.87	73.11	79.05
梁伟	94.08	56.08	93.75	84.75	93.08	93.82
林海	95.87	42.08	94.62	94.21	95.46	94.62
刘海	85.11	31.92	69.63	75.44	78.23	67.53
罗杰	87.92	86.29	79.43	27.54	70.54	80.21
马杰	95.90	57.62	90.45	82.88	90.12	90.45
马强	100.00	93.33	40.00	40.00	40.00	40.00
孙海	91.91	72.17	87.85	71.80	86.76	88.05
孙明	94.12	74.80	88.94	85.79	88.78	89.36
孙涛	92.94	30.93	92.01	86.57	89.03	92.01
唐海	70.97	69.59	70.33	50.63	68.89	70.33
王华	89.14	88.65	76.20	63.70	74.39	77.54
徐明	97.49	90.98	90.87	84.99	87.88	90.87
杨波	92.94	60.44	89.10	45.33	77.65	87.00
杨伟	92.35	53.19	79.25	83.67	89.04	76.99
张建军	95.56	33.09	42.66	92.61	55.86	42.66
张志强	91.82	49.34	87.21	80.78	82.75	87.21
赵伟	95.73	65.51	79.70	83.36	87.53	89.72
朱建军	89.39	25.80	75.98	84.23	85.62	60.89
平均	91.15	56.48	80.54	73.67	81.87	78.96

表 3 中的最佳值,是指根据黄金标准,选取聚类得到的最佳结果作为最终结果. 表 4 是平均结果值的比较. 在表 3 和表 4 中,特征为“段落 + 语义扩展”,并使用 *tf* · *idf* · *mi* 公式计算特征权值,用向量夹角的余弦作为相似度计算,使用组平均的层次聚类方法. 得到的最佳结果平均 P_IP 的 F 值能达到 91.15%. 但是,在不知道标准人名消解结果时,需要定义停止条件,尽量选择与最佳阈值接近的划分作为人名消解的结果. 从表 3 和表 4 可以发现以下现象: (i) C&H 准则函数停止条件得到的平均 F 值(P_IP) 只有 56.48%,远远低于其他停止条件下的结果. 特别是在不同人名下,效果表现差异非常大. 虽然能够在 7 个人名中表现最好,但是在“高军”、“何涛”等数据下表现非常的差,仅有 11.26% 和 19.01% 的 F 值; (ii) 在自动生成阈值实验中,“平均值”方法得到的平均结果相对较好,且对所有的人名的测试结果比较稳定,“最大值”其次,“最小值”和“中心值”方法得到的平均结果略差,但是远好于 C&H 准则函数停止条件的结果; (iii) 可以精心选择固定阈值(相似值不小于某个阈值) 提升聚类效果. 但是,这需要大量的实验,经过反复观察才能确定,表中未列相应的结果.

表 4　层次聚类停止条件实验平均结果比较　%

		B_Cubed			P_IP		
		#Pre	#Rec	#F	#Pre	#Rec	#F
最佳结果		89.78	87.09	88.01	91.70	91.12	91.15
准则函数: C&H		93.09	39.94	48.93	93.80	46.67	56.48
自动阈值	最大值	77.90	78.85	74.43	81.95	84.27	80.54
	最小值	54.50	92.95	65.01	63.98	95.24	73.67
	平均值	72.84	85.37	75.55	78.48	89.89	81.87
	中心值	78.95	76.56	72.90	82.79	81.91	78.96

4 结论

跨文本的人名同指消解在信息检索、多文本摘要和信息融合等与多个文本相关的应用具有重要作用. 本文选择 CLP-SIGHAN2010 的评测任务,对中文人名同指消解问题做了讨论和研究.

人名实体的特征在同指消解中起着重要作用,但提取好的特征是非常困难的. 本文通过提取人名所在的上下文(context)信息,表征实体的特征;然后在特征表示基础上进行聚合式层次聚类,并提出了停止条件的自动判断方法.

特征表示是机器学习的一个难题. 同样,停止条件又是聚类算法中的一个开放式问题. 虽然本文对此作了一些试验,但仍有大量的问题需要研究,笔者将进一步对此问题进行深入探讨.

5 参考文献

[1] 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.

[2] Bagga Amit, Breck Baldwin. Entity-based cross-document coreferencing using the vector space model [C] // Proceedings of the 36th Annual Meeting of the ACL and the 17th International Conreference on Computational Linguistics (COLING-ACL),1998:79-85.

[3] Wang Houfeng, Zheng Mei. Chinese multidocument personal name disambiguation [J]. High Technology Letters, 2005,11(3):280-283.

[4] 陈晨,王厚峰. 基于社会网络的跨文本同名消歧[J]. 中文信息学报,2011,25(5):75-82.

[5] Javier Artiles, Julio Gonzalo, Satoshi Sekine. The SemEval-2007 WePS Evaluation: Establishing a benchmark for the Web People Search Task [C] // Proceedings of the 4th International Workshop on Semantic Evaluations (Semeval-2007),2007:64-69.

[6] Heng Ji, Ralph Grishman, Hoa Trang Dang, et al. An overview of the TAC2010 knowledge base population rrack [C] // Proceedings of Text Analytics Conference (TAC2010),2010.

[7] He Zhengyan, Wang Houfeng, Li Sujian. The task 2 of CIPS-SIGHAN 2012: named entity recognition and disambiguation in chinese bakeoff [C] // Proceedings of the second CIPS-SIGHAN Joint Conference on Chinese Language Processing,2012:108-114.

[8] Chen Ying, James Martin. Towards robust unsupervised personal name disambiguation [C] // Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL),2007:190-198.

[9] Suzanne Tamang, Chen Zheng, Ji Heng. CUNY_BLENDER TAC-KBP2012 entity linking system and slot filling validation system [C] // Proceedings of Text Analytics Conference (TAC2012),2012.

[10] Chen Ying, Jin Peng, Li Wenjie, et al. The Chinese persons name diambiguation evaluation: exploration of personal name disambiguation in Chinese news [C] // Proceedings of the first CIPS-SIGHAN Joint Conference on Chinese Language Processing,2010.

[11] Milligan G W, Coope M C. An examination of procedures for determining the number of clusters in a data set [J]. Psychometrika,1985,50:159-179.

[12] Calinski R, Harabasz J. A dendrite method for cluster analysis [J]. Communications in Statistics,1974(3):1-27.

The Chinese Cross-Document Personal Coreference Resolution

CHEN Chen, WANG Houfeng*

(Office of Informatization, Peking University; Key Laboratory of Computational Linguistics (Peking University) Ministry of Education, Beijing 100871, China)

Abstract: Cross-document named entity coreference resolution is the process of determining if an identical name occurring in different texts refers to the same object. With the increasing need for multi-document applications, for example, multi-document summarization and information fusion, cross-document name entity coreference resolution has drawn much attention. The paper focuses on multi-document personal coreference resolution, and realizes an agglomerative clustering approach for personal coreference resolution, in which feature selection and stopping measures of the clustering to estimate the number of entities are discussed in detail.

Key words: personal coreference resolution; agglomerative clustering; feature selection; cluster-stopping measure

(责任编辑:冉小晓)