# 鲁棒性的汉语人称代词消解*

王厚峰+, 梅 铮

(北京大学 计算机科学技术系,北京 100871)

## Robust Pronominal Resolution within Chinese Text

WANG Hou-Feng+, MEI Zheng

(Department of Computer Science and Technology, Peking University, Beijing 100871, China)

+ Corresponding author: Phn: +86-10-62753081 ext 106, E-mail: wanghf@pku.edu.cn, http://www.icl.pku.edu.cn

Received 2004-06-27; Accepted 2004-08-10

**Wang HF, Mei Z. Robust pronominal resolution within Chinese text. *Journal of Software*, 2005,16(5):700−707.**
DOI: 10.1360/jos160700

**Abstract**: Anaphora Resolution is playing more and more important role in Natural Language Processing. There is an increasing need for the development of effective and robust strategies of anaphora resolution to meet the demands of practical applications. However, traditional approaches to anaphora resolution rely heavily on multilevel linguistic knowledge, such as syntactic, semantic, contextual and domain knowledge. It is undoubtedly difficult to acquire such knowledge at present. This paper presents a two-step approach with limited knowledge to resolve pronominal anaphora within Chinese text, which only uses number features, gender features and the features of grammatical roles. In this approach, a filter is firstly used to eliminate those expressions whose features are inconsistent with the pronoun, and thus form a set of potential antecedent candidates; then, a scoring algorithm is employed to calculate score of the candidates, and the candidate with the highest score is selected as the resultant antecedent. The algorithm does not examine each candidate in the set, but automatically determine whether to end the calculation or not by dynamically testing a termination condition, therefore the computational complexity is low. In addition, the approach does not need a deep analysis of the text, and can easily be implemented. Experiment shows the result is satisfactory.

**Key words**: pronominal anaphora resolution; antecedent; feature; filter; score algorithm

**摘 要**: 指代消解在自然语言处理中起着越来越重要的作用.许多自然语言处理应用系统都需要高效、鲁棒的指代消解策略.然而,传统的指代消解方法需要用到句法知识、语义知识、上下文知识,甚至领域知识等多级知识,在目前的自然语言处理水平下,要有效获取这些知识是相当困难的.结合汉语的特点,提出了一种弱化语言知识的人称代词消解方法,仅仅用到了单复数特征、性别特征和语法角色特征.该方法主要分为两步,首先,利用这3 种特征的简单约束关系,过滤与人称代词特征不一致的词,并形成可能的先行语候选集;然后,使用一个权值算

* Supported by the National Natural Science Foundation of China under Grant Nos.60173005, 60473138 (国家自然科学基金); the National High-Tech Research and Development Plan of China under Grant No.2001AA114210-05 (国家高技术研究发展计划(863))

作者简介: **王厚峰**(1965－),男,湖北天门人,博士,副教授,主要研究领域为自然语言处理;**梅铮**(1978－),男,硕士生,主要研究领域为自然语言处理.

法,计算候选的权值,并将最高权值的候选作为代词最终的先行语.权值算法并不是枚举式地计算每个候选的权值,而会通过动态评测机制,在合适的条件下自动终止计算,因而有效地控制了计算复杂度.此外,该方法不需要对文本进行深层的分析处理,实现起来也很容易.测试结果表明,该方法达到了满意效果.

**关键词**: 人称代词消解;先行语;特征;过滤;权值算法

**中图法分类号**: TP18 **文献标识码**: A

指代(anaphora)是自然语言中广泛存在的一种现象.文本的概念关联性在很大程度上就是通过指代关系来刻画的.指代消解过程实际上是建立概念关联的过程,是文本处理的核心问题之一.由于指代关系的重要性,著名的 MUC(message understanding conference)将指代消解列为评测的子任务之一.当然,指代消解不仅在信息提取 IE(information extraction)中起着重要的作用,在文本总结、问答系统和多语处理等方面的作用也越来越明显[1−3].

然而,指代消解是困难的.无论是对潜在先行语的识别,还是从先行语候选集中挑选出最合适指代语(也称照应语 anaphor)的先行语(antecedent),都需要用到多级语言知识,包括句法知识、语义知识、上下文知识,甚至领域知识.在当前自然语言处理的水平下,要有效地得到所需的这些知识仍然不是件容易的事情.

20 多年来,不少研究者一直在努力寻求指代消解的计算方法,并取得了鼓舞人心的成果.Hobbs 曾提出人称代词消解的两种算法,一种是基于完全句法树算法,另一种是基于语义知识的算法.基于完全句法树的方法也称为简单方法(naive approach)[3].手工测试表明,简单方法得到的结果似乎更好.然而,完全句法树的自动分析技术仍然很不成熟,因此,Hobbs 所提出的方法在实际系统上很少使用,该方法只能算作一种理论模型.

在具体实现上,Lappin 和 Leass 于 1994 年提出了消解第三人称代词的算法[4],该算法事先使用槽文法分析器(slot grammar parser)分析句子的结构,再通过一系列的约束规则过滤不可能的候选,最后通过计算先行语候选的突显值(salience)确定最可能的候选.该方法使用权值思想来评判先行语的合理程度,但该方法对句法分析器仍有很强的依赖性.

另一种指代消解方法是基于语料库的方法,包括统计方法或有指导的机器学习方法[5].但这类方法需要大规模的训练语料,在篇章一级上加工语料是耗时、费力的.目前,在同指消解中,无指导的机器学习方法也开始受到重视[6].

相对而言,计算权值的方法在实现上更容易,而且最具有鲁棒性.Mitkov 用此方法在处理英语文本时,取得了较好效果[1].所谓权值,是人们基于先行语特征的重要性所赋的一个值.Mitkov 主要使用了如下特征:距离(先行语候选与指代语之间相隔多少句子),先行语语法角色(主语,直接宾语,间接宾语,介词宾语,…),先行语是否嵌套在一个更大的名词短语中,是否在章节的标题上,是否是术语,是否有特殊的搭配模式等[1].当然,特征的获取并不十分容易,有些特征仍然需要对文本进行较深层次的处理才能得到.特别是,从汉语中获取一些特征明显比英语(印欧语系)要困难得多,如,名词的单复数信息、人名的性别信息以及人称代词的语法角色等在汉语中都不明显.

汉语人称代词消解的研究比英文的人称代词消解研究要少得多,而且,所报道方法的鲁棒性和可实现性并不强.本文针对汉语的特点,提出了一种基于简单特征的汉语人称代词消解策略,基本不需要进行复杂的语言分析.方法主要分为两个步骤:首先,过滤掉与人称代词特征明显不一致的词和短语,形成潜在的先行语候选集;然后,通过权值计算策略,计算候选的优先性权值,并选择权值最大的候选作为最终先行语.其中,优先性权值的计算使用了一种动态策略,在合适的时机,能够自动停止权值计算过程,避免了因枚举性地计算而导致的计算复杂性问题.

## 1 基本概念

指代消解的过程就是确定指代语所对应的先行语的过程.先行语和指代语是相互依赖的两个概念,先行是相对指代语而言的,是指具有明确的指称意义,且被指代语指向的词或短语.具有明确指称意义的短语如果没有

指代语指向,也就无所谓是先行语.此外,当指代语被消解后,指代语的指称意义也是明确的,此时,该指代语又可以作为后面新的指代语的先行语.如果按线性顺序从前向后依次消解指代,前面的指代语可以作为后面指代语的先行语候选.

指代语对先行语的依赖存在多种关系,如,等价关系、上下位关系、整体和部分关系等.与指代消解经常交叉使用的另一类问题——同指(coreference)消解,则只考虑等价关系,即,只考虑两个词(或短语)是否指称"现实世界"的同一对象(或实体)的问题.同指不一定是指代.例如,在相同的语篇中,某个人名字,如"张三",多处以相同全称出现,两者指称同一个人,属于同指,但不必考虑究竟谁指代谁的问题.反过来,指代关系也不一定属于同指关系,如,上下位关系及整体和部分的关系就超出了同指的范围.本文所讨论的人称代词的指代关系属于同指关系.

指代消解是在一个语篇(text 或 discourse)的范围内进行的.为了在一个合理的复杂度内考虑指代语与先行语之间的约束关系,我们将语篇看成是句子的有序集合,并将句子进一步划分为子句序列,这样,只需要在一定范围的句子内考察即可.句子界定标准如下:

**定义 1.1**. 设标点符号集 Punc_Mark={逗号,句号,分号,问号,惊叹号,冒号},mark∈Punc_Mark 标示了右边一个句子(或子句)的开始;mark∈Punc_Mark-{逗号,冒号},则标示左边一个句子的结束;逗号则标示了左边一个子句的结束.

由上述定义可以看出,Punc_Mark-{逗号,冒号}既表示左边句子的结束,自然也表示右边句子的开始;但逗号",",则为子句的形式分界符,而冒号":"只引导右边,并不一定结束左边.区分句子和子句,主要是为了进一步处理句子的层次关系.

虽然分号有别于句号和逗号,但为简化起见,我们没有再作区别.

**定义 1.2**. 设 Rw 是"现实世界"的对象集,如果自然语言的一个表示式(在本文中指词和短语)$x$ 指称一个对象 $r$(∈Rw),则记为 denote($x$)=$r$.

Rw 是一个庞大的集合.这里所谓"现实世界"的对象,并不一定意味着现实存在,可以是虚幻的,如"鬼、神"等.

**定义 1.3**. 当自然语言的两个表示式 $x1$,$x2$ 指称 Rw 中的同一个对象 $r$ 时,即,denote($x1$)=denote($x2$)=$r$,便称 $x1$ 与 $x2$ 同指(coreference),记为 IsCoref($x1$,$x2$)=True(简记为 IsCoref($x1$,$x2$)),否则,IsCoref($x1$,$x2$)=False(简记为 ~IsCoref($x1$,$x2$)).

**定理 1.1**. 同指关系是等价关系.

这是显然的,因为同指关系满足自反、对称和传递的规律.

**定义 1.4**. 在语篇 $T$ 中,如果存在自然语言表示式集合 $S$,对于∀$x1$,$x2$∈$S$,都有 IsCoref($x1$,$x2$)成立,则称 $S$ 是同指集,记为 Coref($S$).

由于同指关系就是一种等价关系,因此:

**定理 1.2**. 如果将语篇 $T$ 中指称 Rw 内元素的所有语言表示式用一个集合 $U$ 表示,那么,同指关系就将 $U$ 划分成若干个互不相交的等价类子集,$S_1$,$S_2$,…,$S_n$,即,$S_1 \cup S_2 \cup \ldots \cup S_n = U$ 且 $S_i \cap S_j = \varnothing$,其中,$i,j \in [1..n]$且 $i \neq j$.每个子集是一个同指集.

同指集可以根据其所在语篇中的先后顺序用链式结构表示,此时的链称为同指链(coreference-chain).同指链实际上表示了一个有序的集合,在不必特别区分顺序的情况下,也用集合 Coref($S$)表示同指链.

**定义 1.5**. 同指消解就是将语篇中的 $U$ 按同指关系划分成若干个等价类子集的过程.

人称代词与所指代的先行语之间的关系一定属于同指关系.

**定义 1.6**. 能够被人称代词指向的自然语言表示式集 Obj-Human-like=表示人的普通名词集合∪人名集∪地名集∪机构组织名集∪人称代词集.

在一个确定的语篇中,上述集合 Obj-Human-like 是确定的.因此,人称代词的消解就是对集合 Obj-Human-like 按同指集划分的过程.在这里,之所以将地名和机构组织名也纳入进来,是因为有的人称代词,如"他们",可能会指代地名或机构组织名,如下例:

例 1.1:*山西大同市三医院*是建国初期兴建的一所较大的人民医院.*他们*以病人为中心…

单复数信息和性别信息是人称代词(主要是第三人称代词)消解的重要依据,但在汉语中,很多指称人的名词的单复数信息和性别信息是不明显的(即很难从形式上判断).因此,我们都引入了 3 类:

Gender_Type={m,f,u},其中,m 表示男性,f 表示女性,u 表示不确定;

Num_Type={s,p,u},其中,s 表示单数,p 表示复数,u 表示不确定.

**定义 1.7**. 设 gender($x$)表示 $x$ 的性别,若(gender($x$)=m∧gender($y$)=f)∨((gender($x$)=f∧gender($y$)=m)成立,则称 $x$ 和 $y$ 在性别上不相容,否则,表示相容.$x$ 和 $y$ 的性别不相容记,为 G_agreement($x$,$y$)=False(简记为~G_agreement($x$,$y$)),相容记为 G_agreement($x$,$y$)=True (简记为 G_agreement($x$,$y$)).

**定义 1.8**. 设 num($x$)表示 $x$ 的单复数,若(num($x$)=s∧num($y$)=p)∨((num($x$)=s∧num($y$)=p)成立,则称 $x$ 和 $y$ 在单复数上不相容,否则,表示相容.$x$ 和 $y$ 的单复数不相容,记为 N_agreement($x$,$y$)=False(简记为~N_agreement($x$,$y$)),相容记为 N_agreement($x$,$y$)=True(简记为 N_agreement($x$,$y$)).

由上面两个定义可知,无论是单复数信息,还是性别信息,如果 $x$ 和 $y$ 中有一个的值为 u,则总是相容的.显然,减少 u 的取值,可以提高消解准确度和效率.

**定理 1.3**. 对于语篇 $T$ 中的人称代词 $p$,如果有一个等价类子集 $S$ 和一个自然语言表示式 $x$∈$S$,满足(~G_agreement($p$,$x$))∨(~N_agreement($p$,$x$)),那么,称 $x$ 和 $p$ 是不一致的,记为~IsCoref($p$,$x$).

**推论 1.1**. 对于语篇 $T$ 中的人称代词 $p$,如果能够确定某自然语言表示式 $x$ 和等价类 $S$,$x$∈$S$,且满足:~IsCoref($p$,$x$),那么,∀$y$∈$S$,都有~IsCoref($p$,$y$).

上述概念用来过滤不可能的先行语候选是非常重要的.这也是本文所给的过滤规则所基于的基本准则.

## 2 汉语人称代词消解的特殊处理

与人称代词消解密切相关的许多特征在汉语中很不明显,下面是比较突出的 3 个方面:

(1) 人称代词的语法角色并不明确.同一个人称代词,可以充当主语,宾语,定语,但英语中,它们通过不同词明确表征.见表 1.

(2) 指称人的绝大部分名词的性别信息并不明确,即使是人的名字,也不像很多外国人名那样可以判断.表 2 是对 4 万多个中国人名统计后选出的有代表性的最后用字在不同性别名字中出现的次数.

(3) 指称人的绝大部分普通名词的单复数信息并不明确,如“老师”.即使有明确数量词限定,在位置上也不一定相邻,如:“一位名叫秦秀英的老人”.

**Table 1** Comparison between Chinese-English personal pronouns

**表 1** 汉英人称代词形式比较

| | Singular | Singular possessive | Plural | Plural possessive |
|---|---|---|---|---|
| 1st person | 我(I,me) | 我+的(my, mine) | 我们(we, us) | 我们+的(our, ours) |
| 2nd person | 你(you) | 你+的(your,yours) | 你们(you) | 你们+的(your, yours) |
| 3rd person (male) | 他(he,him) | 他+的(his) | 他们(they, them) | 他们+的(their, theirs) |
| 3rd person (female) | 她(she,her) | 她+的(her, hers) | 她们(they, them) | 她们+的(their,theirs) |

**Table 2** Statistics on the last Chinese characters some Chinese personal names specific to gender

**表 2** 中国人名的最后用字与对应性别统计(选)

| | 波 | 春 | 宏 | 红 | 华 | 君 | 娥 | 龙 | 强 | 宇 |
|---|---|---|---|---|---|---|---|---|---|---|
| Male | 608 | 239 | 147 | 115 | 926 | 152 | 2 | 390 | 787 | 538 |
| Female | 116 | 100 | 74 | 478 | 604 | 156 | 10 | 7 | 11 | 119 |

除了上面(1)所述的人称代词之外,指称人的名词的语法角色也是不明确的.为了在一定程度上解决这些问题,我们设计了一个预处理器,专门识别代词和先行语的语法角色、先行语的性别和单复数信息,基本方法见文献[7].预处理结果表明,对人称代词的语法角色的识别,正确率可以超过 95%,但对于指称人的名词的语法角色的识别,正确率却只有接近 80%,指称人的名词的单复数信息和性别信息,在绝大部分情况下不一定能够获得.

显然,利用预处理的结果去排除先行语候选,其信息明显不足.为了避免将正确的候选过滤掉,我们弱化了过滤处理,只用到了 3 条最基本的过滤规则:

过滤规则 1. 如果先行语候选的单复数信息或性别信息与当前人称代词不相容,该候选将被过滤.

过滤规则 2. 对于同一个子句,如果人称代词不是定语,则,非定语性的名词不能作为人称代词的先行语,该名词应被过滤.

如果一个句子 $A$ 的谓语动词又引导一个子句 $B$,那么,$A$ 中的主语和 $B$ 中的主语、宾语就不属于同一个子句,此时,即使 $B$ 中的主语或宾语是人称代词,也不应过滤 $A$ 中的主语.此外,兼语句中的宾语和主语之间也不具有被过滤的特点.

虽然指称人的名词(短语)语法角色识别的正确率并不是很高,但规则 2 的过滤要求并不强,因此,并未导致错误过滤发生.

过滤规则 3. 如果一个句子的动词引导一个子句作为宾语,那么当句子的主语是人称代词时,被引导的子句的名词不能作为该代词的先行语.

能够作为人称代词先行语的词(短语)有,人名(如张三)、指称人的普通名词(如工人)、机构和组织名(如中国足球队)、地名(如北京)和已被消解的人称代词.由同指关系的基本属性可知,如果某个先行语候选被上述 3 条过滤规则之一排除,那么,与该候选具有同指关系的其他候选,也应该被过滤.

名词的单复数信息和性别不容易直接得到,但由于人称代词本身带有单复数和性别信息,一旦确定了人称代词所指的名词,该名词及其与该名词有同指关系的其他名词的单复数和性别信息也就明确了.随着人称代词消解的逐步展开,指称人的名词的性别信息和单复数信息会逐步明确.这也将有利于使用过滤规则.

## 3 先行语的权值计算

汉语人称代词与先行语的角色、性别、单复数和距离的关系十分密切.先行语的语法角色分为主句主语、一般主语、兼语、宾语和定语 5 种.单复数特征包括单数、复数和未知 3 种,性别信息包括男性、女性和未知 3 种.我们特别引入“未知”特征,主要是因为其特征值并不容易直接分析得到.

为了比较各个先行语候选的重要程度,我们分别对角色、性别、单复数和距离 4 个因素进行量化处理.角色权值表示为 Roleweight,距离为 DistanceWeight,性别为 GenderWeight,单复数为 NumWeight.每个先行语的总权值= RoleWeight+GenderWeight+NumWeight−DistanceWeight.总权值最大的候选为最终的先行语.权值计算的基本算法如下:

**算法 1**.

(a) 确定代词所在的句子层次(以标点符号作为标记);

(b) 对该层次内每个候选,计算值 $T$=(RoleWeight+GenderWeight+NumWeight);

(c) 记录每个候选的与代词的距离 $W$=DistanceWeight 值(间隔的子句数目);

(d) 从上面确定的可能候选中,找出使($T$−$W$)的最大值候选,作为同层的唯一候选;

(e) 为进一步向左边句子扩展做准备.此时,置 DistanceWeight=1;并将该候选的总权值调整为 $TW$=$T$−DistanceWeight(即,变成以句子为单位设置距离权值);

(f) 如果左边候选的总权值不可能大于 $TW$(最大 RoleWeight+GenderWeight+NumWeight−DistanceWeight−1),则结束,当前候选为最终候选;否则继续(g);

(g) 如果当前句子中没有候选,则向左移动一个句子单位,直至找到有候选的句子;

(h) DistanceWeight=DistanceWeight+1;然后,在当前句子中,从左向右计算每个候选的权值 RoleWeight+GenderWeight+NumWeight−DistanceWeight,并与当前保留的最大权值 $TW$ 作比较.较大的放入 $TW$ 中.

(i) 重复(f)~(h).

目前,我们使用了如下权值设置:

(1) 角色权值 Roleweight:如果人称代词与先行语在同一层次(子句之间由逗号分隔),则候选的角色与人称代词的角色兼容时权值为 2(相同角色总是兼容,此外,兼语既与兼语相容,又与主语、宾语相容),否则为 0.如果超出的同层子句的限制(候选与人称代词之间有句号、分号、问号和惊叹号分隔),则候选为主语,或者与人称代词的角色相同时,权值为 1,否则为 0;如果人称代词出现在小句中,则当候选为主句主语时,权值为 2.

(2) 距离权值 DistanceWeight:在同一个层次内(可能有多个逗号隔开),如果两个候选在相同的子句中,则距离值相等;每向左移动一个子句,且存在没有被排出的候选,则距离值增加 1.但如果超出的与代词相同层次的子句,则不再以子句计算距离,而是以句子(由两类主语关系和标点符号联合判定)计算距离.此时,相同层次的所有子句的距离值设置为 1,然后每左移动一个句子,而且存在候选时,距离增加 1.

(3) 性别权值 GenderWeight:相同时为 1,当特征值为"未知"时,权值为 0.

(4) 单复数权值 NumWeight:分两种情况:其一,对单数代词(他/她),如果先行语候选是:相同的代词,或者人名,或者有表示单数数词限定的普通词,其权值设为 7,表示单数头衔(如总理),其权值为 5,当未知时,权值为 0;其二,对复数代词,如果先行语候选是:相同代词,或者机构名,或者有明确复数信息的普通词,权值设为 5;地名设置为 3,未知时,权值为 0.

上述权值的设定基于如下思想:其一,先行语候选的性别信息相对比较模糊,性别信息确定与不确定,在权值设定上不宜有太大差距.其二,与距离相关的权值主要基于平行性原则,在句子范围内,子句之间是平行关系,当超出子句范围时,句子之间才是平行关系.因此,距离的比较应以对等的关系为考察对象.其三,角色的优先性同样也需要分层考虑,在同层的子句范围内,角色的选择一般满足平行优先,但是,当超出同层子句范围时,角色的影响因子可能发生变化,此时,前面的主语具有的影响力更大一些.最后,我们将单复数的权值设置较大,一方面,因为在这 4 类因素中,单复数信息最重要;另一方面,单复数信息相对容易确定,因而副作用不大.

当人称代词不是定语时,其消解将采用上述算法 1.如果人称代词是定语,算法需作如下修改:(1) 取消语法角色的权值;(2) 单复数明确时,权值为 3,"未知"时,权值为 0;(3) 距离的计算严格递增.即每向左找到一个候选,距离值增加 1.

由上述算法可知,权值计算的结束条件不是事先可确定的.每当计算到某一时刻,算法将保存到当前为止最高权值的候选,并判断其后剩余候选的权值是否可能超出当前最高值,如果不可能,便结束,并选取最高的候选为最终先行语,否则,继续选择剩余候选中距代词最近的一个候选处理,直至满足条件为止.在计算过程中,候选与代词间的距离在不断扩大,距离因子也因此隐含地影响着权值的计算.

## 4 测 试

指代测试有多种评测标准,但大都采用了准确率(precision)和召回率(recall).定义如下:

$$Precision=\frac{\text{正确消解的指代词数目}}{\text{系统识别的指代词数目}}\times 100\%,$$

$$Recall=\frac{\text{正确消解的指代词数目}}{\text{总的指代词数目}}\times 100\%.$$

对人称代词消解而言,每个人称代词都是可以识别的,因此,上述两个公式实际上具有相同的含义.我们只使用 precision 作为评测指标.

我们选择了 1998 年 1 月份的"人民日报"作为评测语料.语料事先被切分成词,并标注词性.标注集使用的是北京大学计算语言学研究所的扩展标注集.其中,与先行语候选有直接关系的标注包括:

人名:一般人名(没有强调姓的名字)nr,姓 nrf,名 nrg

其中,单复数特征:单数(s),性别特征初始值:未知(u).

地名:ns;

其中,单复数特征:复数(p),性别:未知(u).

组织机构名:nt;

其中,单复数特征:复数(p),性别:未知(u).

指称人的普通名词:可数类 nap,单复数特征:未知(u),性别:未知(u).

其他 n*p,单复数特征:复数(p),性别:未知(u).

测试有两种方法:针对算法的手工测试方法和针对实现系统的自动测试方法.严格讲,对算法性能的测试应该以手工测试更为合理,因为实现系统中算法的功能并不一定能够完全正确实现.但是,我们采用了后者.我们

设计了一个实验系统.输入是带有切分标注的“人民日报”语料.系统先对人称代词和上述 4 类名词进行预处理,得到语法角色、性别和单复数信息[7],然后,依次处理每个人称代词,包括先行语过滤处理和权值计算处理,最终得到代词的先行语.整个处理完全自动进行,没有人工干预.我们主要测试了汉语人称代词“他”、“她”和“他们”的消解情况.这 3 个代词具有明显的性别特征和单、复数特征.测试语料共包含 139 篇文章.

在评价一个人称代词的消解是否正确时,有两个标准:其一,正确地找到所指“实体”;其二,正确地找到先行语.假如有两个相同的代词 *P*1 和 *P*2,*P*1 先于 *P*2 出现在文本中,且 *P*1 是和 *P*2 在一条同指链的最近语言单元,那么,在第 1 个标准下,只有当 *P*1 被正确消解时,才能认为 *P*2 与 *P*1 同指的判断是正确的.基于这一标准,我们得到了表 3 所示的测试结果.

**Table 3** Results of pronominal resolution based on evaluation standard I

**表 3** 人称代词消解结果(1)

| | Total number of pronoun | Number of pronoun being correctly resolved | Precision (%) |
|---|---|---|---|
| 他 | 480 | 406 | 84.6 |
| 他们 | 117 | 79 | 67.5 |
| 她 | 113 | 91 | 80.5 |

在第 2 个标准下,无论代词 *P*1 的消解是否正确,只要能够确定 *P*2 的先行语是 *P*1,就认为对 *P*2 的消解是正确的.表 4 是按标准之二统计出的结果.

**Table 4** Results of pronominal resolution based on evaluation standard II

**表 4** 人称代词消解结果(2)

| | Total number of pronoun | Number of pronoun being correctly resolved | Precision (%) |
|---|---|---|---|
| 他 | 480 | 440 | 91.7 |
| 他们 | 117 | 84 | 71.8 |
| 她 | 113 | 100 | 88.5 |

从应用的角度看,第 1 个标准更具有意义,但从理论上看,第 2 个标准似乎更为合理.主要基于如下两点原因:其一,人称代词一旦被消解,该人称代词的语义指向就明确了,此后,其他人称代词能够以此代词为先行语.其二,人称代词与较近的先行语的关系比与较远的先行语的关系更明确,确定两者的关系要容易得多,即使较近的是代词;相反,如果每次一定要找名词(包括专有名词和普通名词)作为先行语,在距离上可能会相当远,计算复杂性会增大.在此条件下,要评价消解是否正确,更合理的界定标准应该是其所指代的先行语是否正确.

表 3 和表 4 所得到的准确率的差距在 4%~8%之间,是比较大的.这也表明,要提高标准之一下的准确率,需要提高前面人称代词和名词同指的准确率.

由实验结果可知,该方法达到了较好的结果.特别是对“他”的处理,与我们早期的方法相比[7],准确率有较大的提高(约提高 12%).对“他们”的处理,其准确度低一些,主要原因是先行语识别不准确.一方面,有些先行语不是紧缩出现的,因而需要将多个分离的表示人的词组合,这种现象在此次测试的语料中增多,但我们目前尚未处理先行语分离的情况;另一方面,有一些表示机构的普通词没有被识别为先行语候选,也对结果有影响,如“一家鞋业公司”.关于人称代词“她”的错误,与“他”相比,多出了非指称人的现象,如,“母爱,作为人类一种崇高的爱,是一棵人类精神大树,她永久地枝繁叶茂.”

## 5 结束语

本文给出了汉语人称代词消解的方法,该方法的实现非常容易,并可以达到较高的准确度.在权值计算方面,我们使用了一种动态权值计算的方法,既不必限制先行语搜索的范围,又能在合适条件下自动终止计算,这与目前在英语中普遍用到的方法是不同的.当然,如果能进一步加强较长短语的捆绑处理,如表示数量关系的名词短语捆绑的处理(“一/mx 位/qe 四十/mx 多/mzh 岁/qt 的/ud 农妇/nap”)和性别关系的分析处理(“女/b 个体户/nap 李/nrf 少英/nrg”),将会有效降低单复数和性别不一致导致的错误,从而提高人称代词消解的准确率,这也是我们下一步要重点解决的问题.

**References**:

[1] Mitkov R. Anaphora Resolution. London: Longman Press, 2002.

[2] Stuckardt R. Design and enhanced evaluation of a robust anaphor resolution algorithm. Computational Linguistics, 2001,27(4): 479−506.

[3] Wang HF. Survey: Computational models and methods of anaphora resolution. Journal of Chinese Information Processing, 2002, 16(6):9−17 (in Chinese with English abstract).

[4] Lappin S, Leass H. An algorithm for pronominal anaphora resolution. Computational Linguistics, 1994,20(4):535−561.

[5] Soon WM, Ng HT, Lim CY. Machine learning approach to coreference resolution of noun phrases. Computational Linguistics, 2001,27(4):521−544.

[6] Ng V, Claire C. Improving machine learning approaches to coreference resolution. In: Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002). 2002. 104−111. http://acl.ldc.upenn.edu/P/P02/

[7] Wang HF, Mei Z. An empirical study on pronoun resolution in Chinese. In: Gelbnkh A, ed. Proc. of the 5th CiCLing Conf. Lecture Notes in Computer Science 2945, Heidelberg: Springer-Verlag, 2004. 213−216.

**附中文参考文献**:

[3] 王厚峰.指代消解的方法和实现技术.中文信息学报,2002,16(6):9−17.

## 第 12 届全国图像图形学学术会议

征 文 通 知

由中国图像图形学学会主办的“第 12 届全国图像图形学学术会议”将于 2005 年 10 月 12 日~10 月 14 日在北京召开。

**一、会议征文范围**

A) 图像处理和编码: 图像采集、获取及存储，图像重建，图像变换、滤波、增强、校正、恢复/复原，图像/视频压缩编码，图像数字水印和图像信息隐藏等

B) 图像分析和识别: 边缘检测、图像分割，目标表达、描述、测量，目标颜色、形状、纹理、空间、运动等的分析，目标检测、提取、跟踪、识别和分类等

C) 图像理解和计算机视觉: 立体成像，图像配准、匹配、融合、镶嵌，3-D 表示、建模、重构、场景恢复，图像感知、解释、推理等

D) 计算机图形学: 图形学基本理论和算法，真实感图形生成，图形系统、GIS 及图形数据库，几何造形基础理论和算法，自然景物模拟，计算机动画等

E) 虚拟现实和增强现实: 虚拟环境构建，基于图像的建模和绘制，场景建模和漫游，分布式虚拟现实，人工生命等

F) 数字娱乐与游戏设计: 游戏设计方法，游戏开发平台，游戏中的人工智能，游戏引擎，手机游戏，三维游戏等

G) 多媒体技术: 人机交互与用户界面，人体生物特征提取和验证，基于内容的图像和视频检索等

H) 图像图形技术应用: 图像图形技术在通信广播，生物医学，文档处理，遥感、雷达、测绘，工业自动化等领域的应用

**二、论文投稿格式** 详见会议网址 http://www.csig.org.cn/ncig2005

**三、重要日期** 2005 年 5 月 15 日：全文投稿　　2005 年 6 月 15 日：发录取通知

2005 年 7 月 15 日：上交最终稿　2005 年 10 月 12 日~14 日：召开会议

**四、邮件地址** ncig2005@ia.ac.cn