“语句”与“代词”的设定对指代消解的影响
——一项向心理论参数化实证研究*

上海外国语大学 许余龙 段嫚娟 华东理工大学 付相君

提要:本文采用向心理论的参数化研究方法,设计了六种指代消解算法,通过对标注语料的分析,初步探讨了“语句”与“代词”这两个参数的设定对汉语指代消解的影响。结果表明,总体来说,无论采用基于哪种确定Cf显著度排序的算法,1)语句的设定对代词指代消解的影响要比零形代词小;2)将语句设定为小句所得到的零形代词消解结果,要普遍优于将语句设定为自然句;3)汉语代词的指代消解准确率要远低于零形代词的消解准确率。

关键词:指代消解、向心理论、参数化研究

[中图分类号] H030 [文献标识码] A [文章编号] 1003-6105(2008)02-0111-10

1. 引言

向心理论 (Joshi & Weinstein 1981; Grosz *et al.* 1995; Walker *et al.* 1998)是一个关于语篇局部连贯性和代词化的理论,着重研究在一个语篇片段中,注意焦点、指称形式的选择和连贯性这三者之间的联系。该理论简单明了、易操作、易验证,因而被广泛应用于语篇回指的理解和生成研究(如见 Lappin & Leass 1994; Kehler 1997; Tetreault 2001; Beaver 2004; Kibble & Power 2004),甚至应用于作文评分研究(如 Miltsakaki & Kukich 2000)。

关于这一理论的理论基础、理论框架和分析模式,苗兴伟(2003)已作了简介。刘礼进(2005)进一步介绍了基于该理论的两种有影响的指代消解算法及其测评结果。Yeh 和 Chen (2001, 2003) 以及王德亮(2004)等也应用这一理论对汉语进行了研究。

但是,由于该理论的倡导者在提出这一理论时,致力于使其具有跨语言有效性,并且是将其作为一个抽象的、语言学的语篇理论而非一个具体的语篇生成或理解的算法规则系统提出的,所以故意对其中的一些基本核心概念不作明确的厘定。同时,不同的研究者在应用向心理论时,对其中的一些基本概念和主要论断又有不同的定义和理解,因而得出的结论难以比较。

因此,Poesio *et al.* (2004)提出了向心理论的参数化研究方法,将这些基本核心概念视为该理论中的参数,并探讨了这些参数的不同设定对该理论有效性的影响。本文将参照该研究方法,初步探讨“语句”与“代词”这两个参数的设定对汉语指代消解的影响。

2. 向心理论的基本运作模式和所涉及的主要参数

2.1 向心理论的基本运作模式和论断

向心理论假定,语篇由若干语篇片段(简称语段)构成,而语段由一组语句(utterance)[1]

* 本文第一作者得到国家社会科学基金资助项目(批准号:05BYY036)资助,第二作者得到上海外国语大学青年社科项目的资助。感谢美国宾州大学 Martha Palmer 教授允许我们免费使用 Penn Chinese Treebank 中的标注语料。

[1] “utterance”一词在苗兴伟(2003)和王德亮(2004)中译为“语段”,而在克里斯特尔(2000)中译为“话段”。这里译为“语句”是因为考虑到,“utterance”通常认为是特定语境中实际使用的一个单位,与抽象语言系统中的单位“sentence”相对立。而且,“语篇”、“语段”、“语句”这一组表示语言实际使用中从大到小的语篇结构单位的术语,可以较自然地与抽象语言系统中的“篇章”、“段落”、“句子”相对应。

万方数据

组成。一个语句中提及的所有语篇实体构成了该语句的一个前瞻中心(Cf)集,因为这些实体是语篇下文潜在的回指对象。在这些 Cf 中有两个特殊成员。一个是语句所谈论的,并回指上一语句所提到的某个实体的那个 Cf,称为回指中心(Cb)。另一个是语句中显著度最高的,并预测可能成为下一语句中的 Cb 的那个 Cf,称为优选中心(Cp)。语句中的 Cf 及其显著度排序反映了语篇局部的注意焦点(简称局部焦点)。随着语篇的展开,局部焦点以语句为单位不断更新。向心理论的基本运作模式可概括为如下三项制约条件和两条规则。

制约条件:

在由语句 U_1, …, U_m 组成的一个语段 D 中,就每个语句 U_i 而言:

1. 只能有一个回指中心 $Cb(U_i, D)$;

2. 前瞻中心集 $Cf(U_i, D)$中的每个元素都必须在 U_i 中实现;

3. 回指中心 $Cb(U_i, D)$是在 U_i 中实现的、在 $Cf(U_{i-1}, D)$中显著度最高的那个元素。

规则:

在由语句 U_1, …, U_m 组成的一个语段 D 中,就每个语句 U_i 而言:

1. 如果 $Cf(U_{i-1}, D)$中的一个元素在 U_i 中实现为代词,那么 $Cb(U_i, D)$也应实现为代词;

2. 过渡状态是有序的。延续过渡优于保持过渡,保持过渡优于流畅转换过渡,流畅转换过渡优于非流畅转换过渡。

向心理论关于语篇局部连贯性和显著性的主要论断,集中体现在制约条件 1 和规则 1 与 2 中。这三项规定构成了向心理论的三个基本论断。

关于语篇局部连贯性,向心理论的主要论断是:如果一个语段中的语句连续提及同一语篇实体,那么这一语段比那些提及不同实体的语段具有较高的连贯性(规则 2)。这一观点与许多语篇连贯理论((如 Chafe (1994)的信息流理论和 Givón (1983)的主题接续理论))的观点相似。但在向心理论中,这种连贯性进一步规定为:语段中每个语句与前一语句具有唯一的一个“主要连结点”,即 Cb(制约条件 1)。

关于语篇局部显著性,向心理论的主要论断是:语句中实现的语篇实体具有不同的显著度,并且在每个语句中只有一个最显著的实体,即 Cp。这一观点也是与许多关于语篇实体信息状态和可及性研究(如 Prince 1981;Gundel *et al.* 1993; Ariel 1990)所得出的基本结论一致的。

规则 1 进一步将上述语篇局部连贯性和显著性,与语篇中指称形式的选用联系起来,规定:在前一语句(U_{i-1})中显著度最高、充当 U_{i-1} 与 U_i 之间主要连结点的 Cb,在 U_i 中最有可能实现为代词(在汉语等语言中为零形代词)。这一论断与 Xu (1995) 和许余龙(2004)提出的高可及性标示语最有可能指称前一小句中的主题/主语的观点十分相似。也正是这一论断使向心理论成为语篇回指生成和理解的一个非常具有吸引力的研究框架,成为计算语言学中指代消解的一个重要算法基础。

然而,要检验向心理论的上述三个基本论断,以及评估将其作为语篇回指生成和理解的理论模式的可行性,首先涉及到制约 1、规则 1 和规则 2 的具体规定,其次涉及到这三条准则所包含的一些基本概念的界定,即这些论断所涉及的一些主要参数的设定,因为不同的研究者在这两个方面有着不同的观点和处理方法。

2.2 基本论断的不同理解和所涉及的参数

关于制约 1,有强制约和弱制约两种不同的观点。以 Grosz *et al.* (1995)为代表的强制约观点认为,在一个语段中,除了第一个语句之外,其他每个语句都必须只有一个 Cb;而以 Walker *et al.* (1998)为代表的弱制约观点则认为,其他每个语句最多只有一个 Cb(亦即可以没有)。制约 1 中所说的 Cb,大多

万方数据

数研究者采用 Grosz *et al.* (1995)的观点,用制约 3 来定义。要检验这一准则,涉及的参数包括:U_i 和 U_{i-1} 的确定方法、"实现"的含义以及语句中 Cf 的显著度排序(即 Cp 的确定标准)。

关于规则 1,其表述本身没有什么歧义。而要检验规则 1,所涉及的参数除了上面提到的那几个之外,还涉及该规则中所说的代词(简称 R1 代词)的定义。

规则 2 是关于连贯语篇中语句之间过渡的分类和优先排序,主要有四种不同的观点。上面 2.1 节中所引用的是 Brennan *et al.* (1987)首先提出,并由 Walker *et al.* (1998)稍作修正的排序规定,代表了主流观点。他们将过渡状态分为四类,定义如下:

	$Cb(U_i) = Cb(U_{i-1})$ 或 $Cb(U_{i-1}) = [\ ?\ ]$[2]	$Cb(U_i) \neq Cb(U_{i-1})$
$Cb(U_i) = Cp(U_i)$	延续	流畅转换
$Cb(U_i) \neq Cp(U_i)$	保持	非流畅转换

Kameyama (1986),Strube 和 Hahn (1999)及 Kibble (2000) 均提出了其他一些过渡状态和过渡模式,以及评估语篇连贯性的原则。然而,无论采用哪一种观点来理解规则 2,同样都会涉及到上面所提到的一些主要参数。

综上所述,向心理论的运作所涉及的主要参数包括:语句 U_i 和 U_{i-1} 的设定、R1 代词的确定、"实现"的含义、语句中 Cf 的显著度排序、语篇中语段的切分标准等。本文主要讨论前两个参数的设定对指代消解的影响,同时也涉及对"实现"含义的确定。语句中 Cf 的显著度排序是语言学和计算语言学中讨论最多的问题,我们将另文讨论。至于语篇中语段的确定,由于我们语料中的语篇都不是很长,而且连贯性都很强,因而我们将整个语篇作为一个语段[3]。

3. 语句和 R1 代词的不同设定方法

3.1 语句 U_i 和 U_{i-1} 的确定

在向心理论的早期研究中,由于所分析的语段几乎都是由简单句构成,因而将语句默认为句子,即 U_i 和 U_{i-1} 都是整个句子。但是在自然语篇中,句子有时会很长、很复杂,其内部可以含有不同类型的小句。因此 Kameyama (1998)提出,应该将时态小句(tensed clause)作为语句。她进一步把语句分为两类:一类是并列小句和状语从句,此类语句"永久"更新局部焦点,即既可作为 U_i,又可作为 U_{i-1};另一类是内嵌小句,此类语句只是"临时"更新局部焦点,处理完之后就弹出,不再对下一语句的局部焦点更新产生影响,即只可作为 U_i,不可作为 U_{i-1}。而 Miltsakaki (1999)则认为,局部焦点是以句子为单位更新的,而且在确定 Cb 时只需考虑那些在句子的主句中实现的 Cf。也就是说,只有句子中的主句才能成为 U_i 和 U_{i-1}。

上述语句确定的方法主要是就英语语篇而言的。由于在汉语语篇中很难区分时态与非时态小句,而且在没有出现连词的句子中有时也很难区分主句与从句,因此上述方法较难直接应用于汉语。本文主要检验下面两种不同的语句设定方法:1)将语句 U_i 和 U_{i-1} 设定为语篇中至少含有一个述谓结构,并由逗号、冒号、分号和句末标点符号断开的、结构相对独立完整的小句,这是目前汉语指代消解研究(如见 Yeh & Chen 2001, 2003;王德亮 2004)所普遍采用的语句确定标准(虽然具体做法略有差异),我们将这一设定方法称为 Udef.1;2)

[2] $Cb(U_{i-1})$=[?]适用于 U_{i-1} 是语段中的第一句时的情况,因为此时该语句中的 Cb 还不能确定。

[3] 另有两方面的原因支持这一做法:1)语篇中中心(话题)的延续和代词的回指往往是跨语段的,Poesio *et al.* (2004)的研究表明,规则 1 并不受语段大小的影响;2)在口语中很难确定语段划分的标准。

万方数据

将语句 U_i 和 U_{i-1} 设定为语篇中的自然句，即由句号、问号和感叹号断开的语符串，我们将其称为 Udef.2。

3.2 R1 代词的确定标准

规则 1 规定，语句中只要有任何一个 Cf 实现为代词，那么 Cb 也应该实现为代词。然而，向心理论没有明确说明 R1 代词可以包括哪几类代词。在将向心理论应用于英语研究时，R1 代词通常默认为第三人称单数代词。而在汉语、日语、意大利语和土耳其语等一些语言中，Cb 除了可以实现为人称代词之外，更多地实现为零形代词（如 Walker *et al.* 1994; Di Eugenio 1998; Turan 1998; Yeh & Chen 2001, 2003; 王德亮 2004）。因此，本研究将同时检验 R1 代词的这两种不同实现方式，即：1）把 R1 代词设定为第三人称代词；2）把 R1 代词设定为零形代词。

在第一种设定中，我们将第一、二人称代词排除在 R1 代词之外，原因是这两类代词典型地用于直指具体交际情境中的说话者和受话者，而不是回指语篇中提到的某个实体。而且，我们研究采用的语料是书面语料，并且按照回指理解研究的通行做法，对其中的对话部分不做分析。因为从语篇的心理表征角度来说，书面叙述部分所构建的语篇模型和对话转述部分所构建的语篇模型并不处于同一“语篇宇宙（universe of discourse）”（Givón 1992）；从向心理论的角度来说，根据 Kameyama（1998：§4.2），对话转述语段是书面叙述语段所不可及的内嵌语段。

在第二种设定中，主语和宾语控制的 PRO 被排除在需要进行指代消解的零形代词之列，因为它们的指称主要是由控制动词的语义决定的。我们根据控制动词的语义人工标注了这些空语类的指称之后，我们的指代消解算法中的句法过滤机制可以利用这些信息自动做出正确的（小）句内指代消解。例如，在下面的（1c）中，“听凭”是一个宾语控制动词，其宾语后的 PRO 只能与其宾语“这些勇猛的敌人”同指。同时，这个 PRO 又是其后面“把”字小句的主语，根据约束规则 B，“把”的宾语“他”不能与小句的主语同指，因而排除了与“这些勇猛的敌人”同指的可能性。将这一可能的先行语过滤掉之后，“他”在（1c）中所剩下唯一可能的先行语是与（1a）中“大青虫”同指的“Ø”，这也是其在该语段中最可及的先行语。

（1）a. 大青虫想挣扎，
b. Ø 却敌不过这许多只蚂蚁，
c. Ø 只好听凭这些勇猛的敌人PRO 把他拖走。

（《丁丁回家去》）

我们对关系从句中的空语类（如下例中的“Ø”）作了同样的处理。

（2）Ø 背负红布包袱的镖师已卸了下背上的兵器，……

（《书剑恩仇录》）

3.3 “实现”的含义

Grosz *et al.*（1995）认为，一个语篇实体在语句中的实现可以通过两种方式，一种是直接实现，另一种是间接实现。如在下面例（3）中（引自 Grosz *et al.* 1995：217），

（3）a. The house appeared to have been burgled.
b. The door was ajar.
c. The furniture was in disarray.

The house 在句（3a）中，以及 the door 和 the furniture 在句（3b）和（3c）中，都得到直接实现。此外，还可以认为 the house 在（3b）和（3c）中得到了间接实现，因为这两句中的 the door 和 the furniture 都间接提到了它，即此例中的 the door 实际上是指 the door of the house，the furniture 也是指 the furniture of the house。如果认为间接实现也算是实现，那么可以说句（3b）和（3c）中含有一个 Cb（=the house）；相反，如果认为只有直接实现才算是实现，那么这两句中便没有 Cb 了，从而违反了强制约 1。

如果说间接回指在英语中的识别还比较

万方数据

容易的话,那么在汉语中则要难得多。因为英语中用作间接回指的名词短语大多带有定冠词 the,有较为明显的形式标记;而汉语由于没有定冠词,语篇中用于引入一个新实体,或用于直接回指或间接回指的名词短语都可能采用光杆名词短语的形式。例如,在下面的例(4)中,

(4) a. 他左手拿着茶壶,
b. Ø 以食中两指揭开壶盖,
c. 铁莲子扑的跌入壶中。
(《书剑恩仇录》)

(4a)中的"左手"和"茶壶"都用于引入一个新的语篇实体;(4b)中的"食中两指"和"壶盖"分别间接回指(4a)中的"左手"和"茶壶";(4c)中的"铁莲子"回指前一自然句中提到的一个实体;而(4c)中的"壶"则用于直接回指(4a)中的"茶壶"。这些不同功能的名词短语都是以光杆名词短语的形式出现的。

其实,即便在英语中,也尚未有一种识别和标注所有间接回指的可靠方法,因而在 Poesio *et al.* (2004: 325)对英语的向心理论参数化研究中,只对一些较为容易识别的间接回指关系(如整体与局部的关系)作了标注。因此在本研究中,我们仅将直接实现视为实现[4]。

4. 语料与研究方法

Poesio *et al.* (2004)的参数化研究的基本方法是建立一个标注语料库和一个自动分析程序系统,分析不同的参数设定对向心理论作出的论断所产生的影响。他们的标注内容分为如下三大类:1)语篇构成属性,包括语篇的自然分节、分段、分句,句中各种形式和功能的小句;2)名词短语属性,包括名词短语的形式特征、语义属性、句法功能和线性位置;3)回指信息,包括名词短语之间的直接或间接回指关系。这些内容与许余龙(2005)的语料数据库所含内容相似,但有些内容更为详密。

在他们的研究基础上,我们选用了如下三类汉语语料进行研究:1)金庸的《书剑恩仇录》第一回;2)从中国儿童文学网(网址:http://www.61w.cn/)下载的 8 篇儿童故事;3)从美国宾州大学中文标注语料库(Penn Chinese Treebank,简介见 http://www.cis.upenn.edu/~chinese/)中随机选出的 8 篇新闻报道。我们参照了 Poesio *et al.* (2004)和宾州大学中文标注语料库的标注方法(见 Xue *et al.* 2005)对语料进行了标注,但略去了间接回指关系的标注。表 1 列出了语料的基本情况。

表 1　语料的基本情况

语料类型	字数	名词短语数	零形代词	第三人称代词	非常规指称	语句数(Udef.1)	语句数(Udef.2)
小说	14804	2955	973	151	19	1610	477
儿童故事	12029	1653	430	130	18	854	333
新闻报道	4660	540	73	6	6	212	84
总计	31493	5148	1476	287	43	2676	894

表 1 显示,在我们的语料中,共出现 1476 个零形代词和 287 个第三人称代词,其中共有 43 个用于任指(arbitrary reference)、下指、指称抽象命题或合指前面分别提及的实体等非常规指称。将这 43 例非常规指称除去后,用于明确回指语篇中提及的具体实体的零形代词和第三人称代词分别有 1442 和 278 个,这些是我们研究中试图消解的两类

[4] 这固然与更难识别汉语中的间接回指有关。但更重要的原因是,正如我们在下一节将指出,我们的研究目的与 Poesio *et al.* (2004)不尽相同。

万方数据

R1 代词。

表 1 同时显示，如果语句按 Udef.1 来确定，那么我们的语料共含有 2676 个语句；如果按 Udef.2 来确定，那么语料中的语句减至 894 个。

在研究的目的和分析方法方面，我们与 Poesio *et al.* (2004)略有不同。Poesio *et al.* (2004) 的主要目的是通过对参数的不同设定，来检验哪种设定可以使语料中出现符合向心理论论断的实例数比率更高一些。因而对他们来说，如果按 Udef.1 将上面的(3b)认定为一个语句，那么其中是否含有一个 Cb 将影响到该语句是遵守还是违反了强制约 1。而我们的研究目的则主要是直接检验哪种设定可以使以向心理论为基础的指代消解算法更有效一些。

在分析方法上，我们先设计了一个程序，将语料中标注的信息读入 MS Access 数据库。然后根据可能影响 Cf 显著度排序的不同因素，设计了 6 个不同的指代消解算法：1) Alg1 完全根据线性语序来确定显著度；2) Alg2 完全根据语法功能来确定显著度；3) Alg3 在 Alg2 的基础上进一步考虑了回指语和先行语的语法功能平行性因素，即回指语和先行语倾向于承担同一语法功能；4) Alg4 在 Alg2 的基础上进一步考虑了 $Cb(U_i) = Cb(U_{i-1})$ 的倾向，即语篇连贯性的因素；5) Alg5 在 Alg2 的基础上同时考虑了语法功能的平行性和语篇连贯性这两个因素；6)Alg6 在 Alg3 的基础上进一步考虑了主句中的回指语倾向于回指前一主句中提及的语篇实体这一因素。(这 6 种算法的具体设计我们将结合显著度的确定另文介绍，简介见 Duan 2007)

上述 6 种算法实际上代表了对 Cf 显著度这一参数的 6 种不同设定方法。这一参数与第 3 节中讨论的语句和 R1 代词这两个参数之间的互动，以及这种互动对指代消解算法有效性的影响，可以通过在对后两个参数作不同设定的情况下，分别运作上述 6 种不同算法来检验。我们将每次运作的结果读入数据库，与数据库中人工标注的回指信息进行自动比对，从而检验不同参数设定情况下的消解有效性。

5. 数据分析

检验指代消解算法有效性的衡量标准最常用的有三个，即回索率(recall rate)、准确率(precision rate)和成功率(success rate)，分别定义如下(Mitkov 2002：178-179)：

回索率=正确消解的回指语数÷系统所识别的所有回指语数

准确率=正确消解的回指语数÷系统所致力于消解的所有回指语数

成功率=正确消解的回指语数÷语料中所含的所有回指语数

由于本文的主要目的是研究向心参数的不同设定对指代消解的影响，而不是计算语言学和人工智能研究所致力于研究的语篇回指自动处理，因而我们的研究没有设计和采用代词和零形代词自动检索程序，语料中的所有代词和零形代词都是人工标注的，从而我们研究中的回索率大致相当于成功率。因此，下面的数据分析主要采用准确率和成功率这两个衡量标准。两者在本研究中的区别是，在成功率的计算基数中包含了我们语料中所出现的所有回指语，而准确率的计算则在基数中扣除了表 1 所列的那些用于非常规指称的回指语。

5.1 语句的设定对指代消解的整体影响

表 2 列出了语句的两种设定对 6 种指代消解算法的整体准确率和成功率的影响，消解的百分比越高，说明有效性越大。

该表显示，总的来说，无论是就准确率而言，还是就成功率来说，如果将语句设定为至少含有一个述谓结构的小句(Udef.1)，那么 6 种指代消解算法所得到的消解结果，都要比将语句设定为语篇中的自然句 (Udef.2)好。其中的主要原因是，按 Udef.2 定义的语句通

万方数据

表 2　语句的设定对指代消解的整体影响

		Alg1	Alg2	Alg3	Alg4	Alg5	Alg6
语句	准确率(%)	81.9	86.9	87.4	85.3	87.3	91.3
(Udef.1)	成功率(%)	79.9	84.7	85.3	83.2	85.2	89.1
语句	准确率(%)	66.1	73.3	76.2	72.0	76.2	80.5
(Udef.2)	成功率(%)	64.5	71.5	74.4	70.2	74.4	78.5

常较长,不能及时反映语篇中话题和 Cf 集的更新,从而会出现较多的违反规则 1 的语句,影响指代消解的准确率和成功率,这一结果与 Poesio *et al.* (2004)对英语的研究结果是一致的。而按 Udef.1 定义的语句较短,在大多数情况下不存在这个问题。比如,在下面的例(5)中,

(5) a. 陆菲青手下留情,
b. 这一掌蕴劲回力,
c. Ø 去势便慢,
d. 焦文期明知对方容让,
e. Ø 竟然趁势直上,
f. Ø 乘着陆菲青哈哈一笑,手掌将缩未缩、前胸门户洞开之际,
g. Ø 突然左掌"流泉下山",
h. 五指已在他左乳下猛力一截。

(《书剑恩仇录》)

如果按 Udef.1 来定义语句,那么从(a)到(h)构成 8 个语句。无论采用 6 种算法中的哪一种,(c)中的"Ø"都可以消解为回指(b)中的"这一掌";而(e)到(g)中的"Ø"都可以消解为回指(d)中的"焦文期"。但是,如果按 Udef.2 来定义语句,那么从(a)到(h)只构成一个语句,不仅会使当前语句中"Ø"的指代消解变得更为复杂,而且还可能影响到下一语句中回指语的消解,因为根据 Udef.2 定义的语句不能及时正确地反映出此时语篇的话题已从"陆菲青"转为"焦文期"。

5.2 语句的设定对两种不同 R1 代词消解的影响

由于准确率直接反映了语篇中用作回指(而不是任指或下指)的回指语的实际消解情况,本节将只用准确率来分析语句设定和 R1 代词设定在指代消解过程中的相互作用。图 1 反映了语句和 R1 代词的设定对指代消解影响的两大概貌。首先,无论是就代词还是零形代词的指代消解来说,在将语句按 Udef.1 设定的情况下,6 种算法所得出的指代消解准确率,都要比将语句按 Udef.2 设定为高。

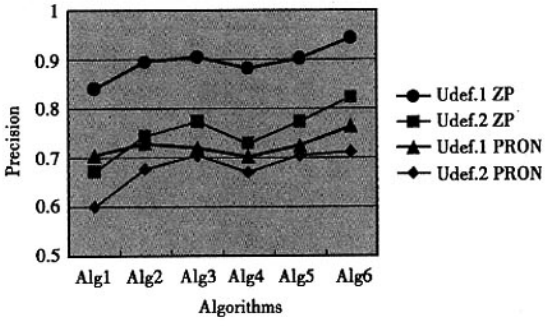

图 1　两种不同 R1 代词指代消解的准确率比较

其次,无论采用 6 种算法中的哪一种,汉语零形代词(ZP)的指代消解准确率都要比代词(PRON)高,最高的要高出将近 20%。其中一个可能的原因是,汉语中的零形代词和代词分属两种不同的可及性标示语,零形代词所标示的指称对象的可及性要比代词高(Ariel 1990, 2006; 许余龙 2000)。进一步的分析表明,如果按 Udef.1 来定义语句,那么我们语料中只有 5%的零形代词的先行语既不在本语句(U_i)中,也不在上一语句(U_{i-1})中;但有 26%的代词的先行语都出现在上一语句之前。基于向心理论模型的指代消解算法是以语句之间的中心(话题)延续和转换为

万方数据

主要依据的，总是将最显著、最可及的语篇实体作为代词和零形代词的指称对象，所以对于指称距离较远的代词来说，其消解结果自然不如零形代词好。

再让我们来分别具体考察语句和 R1 代词的设定对 6 种指代消解算法的准确率的影响。图 1 直观地显示，语句的设定对零形代词指代消解的影响十分相似：在将语句按 Udef.1 设定的情况下，6 种算法的指代消解准确率普遍要比按 Udef.2 设定为高。但是，语句的设定对代词的指代消解准确率的影响则因算法不同而异。图 1 显示，在采用 Alg3 和 Alg5 这两种算法时，语句的设定对代词指代消解准确率的影响很小。表 3 进一步列出了语句的设定对两类 R1 代词指代消解影响的显著性分析结果。

该表可以说明两个问题。首先，该表清楚地显示，在运用 6 种算法对零形代词进行指代消解时，语句的设定都具有显著影响；而在对代词进行指代消解时，语句的设定仅在运用 Alg1 时才有（在 .05 水平上的）显著影响。其次，虽然总体上来说，将语句按 Udef.1 设定的消解准确率要比按 Udef.2 设定为高，但是无论采用哪一种算法，都会出现按 Udef.1 设定不能正确消解而按 Udef.2 设定可以正确消解情况。仔细具体分析这些案例，可以帮助我们进一步提高消解的整体准确率（我们将结合具体算法另文摘要讨论），这也是向心理论参数化研究的优点所在。

表 3 语句的设定对两类 R1 代词指代消解影响的显著性分析

	零形代词		代词	
	正确消解频数	显著度	正确消解频数	显著度
Alg1' vs. Alg1"[5]	Alg1" < Alg1':299[6]	0.000	Alg1"<Alg1':58	0.003
	Alg1">Alg1':57		Alg1">Alg1':29	
	Alg1"=Alg1':1086		Alg1"=Alg1':191	
Alg2' vs. Alg2"	Alg2"<Alg2': 13	0.000	Alg2"<Alg2':42	0.069
	Alg2">Alg2':56		Alg2">Alg2':26	
	Alg2"=Alg2':1373		Alg2"=Alg2':210	
Alg3' vs. Alg3"	Alg3"<Alg3':246	0.000	Alg3"<Alg3':36	0.282
	Alg3">Alg3':59		Alg3">Alg3':32	
	Alg3"=Alg3':1137		Alg3"=Alg3':210	
Alg4' vs Alg4"	Alg4"<Alg4':274	0.000	Alg4"<Alg4':40	1.000
	Alg4">Alg4':56		Alg4">Alg4':30	
	Alg4"=Alg4':1110		Alg4"=Alg4':208	
Alg5' vs Alg5"	Alg5"<Alg5':247	0.000	Alg5"<Alg5':36	0.625
	Alg5">Alg5':61		Alg5">Alg5':31	
	Alg5"=Alg5':1134		Alg5"=Alg5':211	
Alg6' vs Alg6"	Alg6"<Alg6':214	0.000	Alg6"<Alg6':35	0.082
	Alg6">Alg6':33		Alg6">Alg6':21	
	Alg6"=Alg6':1195		Alg6"=Alg6':222	

[5] Alg1'和 Alg1"分别表示在将语句分别按 Udef.1 和 Udef.2 设定的情况下运行 Alg1，其余类推。

[6] Alg1"<Alg1':299 表示 Alg1'能正确消解而 Alg1"不能的有 299 例；Alg1">Alg1':57 表示 Alg1'不能正确消解而 Alg1"能正确消解的有 57 例；Alg1"=Alg1':1086 表示两者都能正确消解的有 1086 例；其余类推。

万方数据

6. 小结

本文采用 Poesio *et al.* (2004)提出的向心理论的参数化研究方法，设计了 6 种指代消解算法，通过对标注语料的分析，初步探讨了“语句”与“代词”这两个参数的设定对汉语指代消解的影响。结果表明，总体来说，无论采用基于哪种确定 Cf 显著度排序的算法，1)语句设定对代词指代消解的影响要比零形代词小；2)将语句设定为小句所得到的零形代词消解结果，要普遍优于将语句设定为自然句，这说明目前汉语指代消解研究所通常采用的语句确定方法在总体上是可行的；3)汉语代词的指代消解准确率要远低于零形代词的消解准确率，这说明目前汉语指代消解研究主要关注零形代词的做法是不够的，应该加强对代词的指代消解研究。

这一参数化研究的意义并不局限于运用向心理论进行指代消解研究本身，更重要的是可以通过这种方法发现影响回指理解的各种因素，以及这些因素之间的相互作用，从而推动和深化语篇分析和语篇回指的研究，检验并实质性地改进现有的一些理论，或提出更符合语言事实的理论，因为无论采用哪一种理论框架，都要涉及到对这些因素的综合处理。

参考文献

Ariel, M. 1990. *Accessing Noun-phrase Antecedents* [M]. London: Routledge.

Ariel, M. Accessibility theory [A]. In K. Brown (ed.). *Encyclopedia of Language and Linguistics* (2nd ed.), Vol. 1 [C]. Oxford: Elsevier, 2006: 15-18.

Beaver, D. 2004. The optimization of discourse anaphora [J]. *Linguistics and Philosophy* 27:3-56.

Brennan, S., M. Friedman & C. Pollard. 1987. A centering approach to pronouns [P]. *In Proceedings of ACL-87*, 155-162.

Chafe, W. 1994. *Discourse, Consciousness, and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing* [M]. Chicago: The University of Chicago Press.

Di Eugenio, B. 1998. Centering in Italian [A]. In M. A. Walker, A. K. Joshi & E. F. Prince (eds.). *Centering Theory in Discourse* [C]. Oxford: Oxford University Press, 115-138.

Duan, M. J. 2007. Centering in Chinese anaphor resolution: A parametric study [P]. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*, 350-358. Available from http://mandrake.csse.unimelb.edu.au/pacling2007/files/final/7/7_Paper_meta.pdf. (Accessed on 30 Oct. 2007)

Givón, T. 1983. Topic continuity in discourse: An introduction [A]. In T. Givón (ed.). *Topic Continuity in Discourse: Quantitative Cross-Linguistic Studies* [C]. Amsterdam: John Benjamins, 1-42.

Grosz, B. J., A. K. Joshi & S. Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse [J]. *Computational Linguistics* 21:203-225.

Gundel, J. K., N. Hedberg & R. Zacharski. 1993. Cognitive status and the form of referring expressions in discourse [J]. *Language* 69:274-307.

Joshi, A. K. & S. Weinstein. 1981. Control of inference: Role of some aspects of discourse structure-centering [P]. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 435-439.

Kameyama, M. 1986. A property-sharing constraint in centering [P]. In *Proceedings of ACL-86*, 200-206.

Kameyama, M. 1998. Intra-sentential centering: A case study [A]. In M. A. Walker *et al.* (eds.). *Centering Theory in Discourse* [C]. Oxford: Oxford University Press, 89-112.

Kehler, A. 1997. Current theories of centering for pronoun interpretation: A critical evaluation [J]. *Computational Linguistics* 23: 467-475.

Kibble, R. 2000. *A Reformulation of Rule 2 of Centering Theory* [R]. Technical report,

万方数据

University of Brighton, ITRI. GNOME project internal deliverable.

Kibble, R. & R. Power. 2004. Optimizing referential coherence in text generation [J]. *Computational Linguistics* 30:401-416.

Lappin, S. & H. Leass. 1994. An algorithm for pronominal anaphora resolution [J]. *Computational Linguistics* 20: 536-561.

Miltsakaki, E. 1999. Locating topics in text processing [P]. In *Computational Linguistics in the Netherlands: Selected Papers from the Tenth CLIN Meeting*, 127-138.

Miltsakaki, E. & K. Kukich. 2000. The role of centering theory's rough-shift in the teaching and evaluation of writing skills [P]. In *Proceedings of the ACL-2000.* Available from http://acl.ldc.upenn.edu/P/P00/P00-1052.pdf. (Accessed on 30 Sept. 2007)

Mitkov, R. 2002. *Anaphora Resolution* [M]. Edinburgh: Pearson Education Limited.

Poesio, M. *et al.* 2004. Centering: A parametric theory and its instantiations [J]. *Computational Linguistics* 30:309-363.

Prince, E. F. 1981. Toward a taxonomy of given-new information [A]. In P. Cole (ed.). *Radical Pragmatics* [C]. New York: Academic Press, 223-255.

Strube, M. & U. Hahn. 1999. Functional centering—grounding referential coherence in information structure [J]. *Computational Linguistics* 25:309-344.

Tetreault, J. R. 2001. A corpus-based evaluation of centering and pronoun resolution [J]. *Computational Linguistics* 27: 507-520.

Turan, U. 1998. Ranking forward-looking centers in Turkish: Universal and language-specific properties [A]. In M. A. Walker *et al.* (eds.). *Centering Theory in Discourse* [C]. Oxford: Oxford University Press, 139-160.

Walker, M. A., M. Iida & S. Cote. 1994. Japanese discourse and the process of centering [J]. *Computational Linguistics* 20: 193-232.

Walker, M. A., A. K. Joshi & E. F. Prince. 1998. Centering in naturally occurring discourse: An overview [A]. In M. A. Walker *et al.* (eds.). *Centering Theory in Discourse* [C]. Oxford: Oxford University Press, 1-28.

Xu, Y. L. 1995. Resolving third-person anaphora in Chinese narrative discourse [D]. Ph. D. dissertation, Hong Kong Polytechnic University.

Xue, N. W., F. Xia, F. D. Chiou & M. Palmer. 2005. The Penn Chinese TreeBank: Phrase structure annotation of a large corpus [J]. *Natural Language Engineering* 11: 198-207.

Yeh, C. L. & Y. J. Chen. 2001. An empirical study of zero anaphora resolution in Chinese based on Centering Theory [P]. In *Proceedings of ROCLING*, Tainan, Taiwan, China.

Yeh, C. L. & Y. J. Chen. 2003. Zero anaphora resolution in Chinese with partial parsing based on Centering Theory [P]. In *Proceedings of NLP-KE03*, Beijing, China.

克里斯特尔编, 2000, 现代语言学词典(沈家煊译) [M]。北京:商务印书馆。

刘礼进,2005,中心理论和回指解析计算法 [J]。外语学刊(6):23-27。

苗兴伟,2003,语篇向心理论述评 [J]。当代语言学(2):149-157。

王德亮,2004,汉语零形回指解析—基于向心理论的研究 [J]。现代外语(4):350-359。

许余龙,2000,英汉指称词语表达的可及性 [J]。外语教学与研究(5):321-328。

许余龙,2004,篇章回指的功能语用探索——项基于汉语民间故事和报刊语料的研究 [M]。上海:上海外语教育出版社。

许余龙,2005, 语篇回指实证研究中的数据库建设 [J]。外国语(2):23-29。

收稿日期:2007-11-09;
作者修改稿,2008-02-23;
本刊修订,2008-03-18

通讯地址: 200083 上海市虹口区大连西路 550 号
上海外国语大学语言研究院
<xuyulong@shisu.edu.cn> (许)

万方数据