汉语语篇理解中元指代消解初步*

张 威, 周昌乐

(浙江大学 信息学院 计算机科学与工程系,浙江 杭州 310027);
(浙江大学 人工智能研究所,浙江 杭州 310027)
E-mail: hdstsysgn@cise.zju.edu.cn
http://www.zju.edu.cn

摘要: 指代消解是语篇机器理解中的重要一环.研究发现,由于表示语篇本身某一部分而非语篇内容的元指代现象普遍存在,语篇元指代消解也就成为困扰着语篇机器理解实现的困难之一.对语篇中的元指代现象进行了分析,提出句焦点的概念,并在句焦点集的基础上,用优先和过滤算法实现了元指代的消解.在使用自然语料的实验中表明,句焦点集的作用对于元指代机器消解有重要作用.它丰富了语篇分析和表述理论,对汉语语篇分析理解中寻找元指代关系,从而完成连贯语篇意义具有重要意义.

关 键 词: 自然语言理解;汉语语篇分析;指代消解;元指代;句焦点
中图法分类号: TP18 文献标识码: A

汉语计算机处理在当前信息社会具有广阔的应用前景.以前的汉语研究注重对词、词组、句子的分析和处理,对汉语语篇(篇章或话语)的分析研究较为薄弱.但汉语的机器理解最终要落实到语篇一级的理解上来.

作为语篇衔接与连贯的重要手段之一,指代(anaphora)是指在语篇中用一个指代词回指某个以前说到过的语言单位.指代词的使用使语篇的表述不显累赘、简明清晰.同时,指代反映了语篇中各语句之间的语义联系,是语篇成其为语篇的重要特征.一般来说,指代可以分为代词性指代、名词性指代和零指代 3 种.从意义层次上来分类,还可分为指代和元指代.

作为自然语言的汉语,普遍存在着语言与元语言混用的现象,比如"'焦点'这个词是一个名词"这句中就是如此.语篇中也存在指代和元指代之分.而所谓元指代是指语篇中存在的这样一种指代,其所替代的对象并非是语篇表述内容,而是语篇本身某一部分.比如像"你好这两个字使用频率很高"、"记住一句话:'学无止境.'"、"本章我们论述了…"、"上文谈到…"等中的"这两个字"、"一句话"、"本章"、"上文"等等就属于元指代类属.其他像"说了...,就该说"虽没有显式的指代词,但也同样存在着元指代意味.而对语篇意义理解而言,很明显,只有确实完成了这些元指代的确认和消解,才真正谈得上整幅语篇意义的理解贯通.

在国外,有关语篇表述和指代消解的研究工作开展得较为普遍,并取得了不少成绩[1~3].就我们所关心的方面来看,主要有 Grosz,Sidner,Walker 等人提出、发展并加以完善的指代消解焦点理论.所谓焦点,就是单句当前注意力所在.一般来说,每句话都有焦点,一段篇章由焦点链组成,焦点链体现了篇章的脉络和结构.Grosz 将焦点空间设计为局部的语义网,运用到一个会话理解系统中.Sidner 对焦点的确定、转移作了大量研究,结合到 PAL 系统中.Walker,M.A.对 Grosz 和 Sidner 提出的针对言语中指代消解的 Focus Stack 模型进行了修改,根据长时记忆和短时记忆的类比,加入了代表短时记忆的 Cache 模型.

国内,针对汉语语篇,在必须解决人称指代问题时(如汉语的整篇机译),一般都遵循"最近匹配原则".谌志群

* 收稿日期: 2000-09-03; 修改日期: 2000-11-20
基金项目: 国家自然科学基金资助项目(69983006)
作者简介: 张威(1974－),男,广东惠阳人,博士生,主要研究领域为计算语言学,软件工程;周昌乐(1959－),男,山东文登人,博士,教授,博士生导师,主要研究领域为计算语言学,认知逻辑学,神经动力学,人工智能.

对人称指代消解进行了研究,采用“关注焦点”集的计算方法.在此基础上,提出了一种基于“关注焦点”集计算的人称指代消解算法[4].至此,还未见有关元指代消解的研究.

本文的工作主要是针对汉语书面语,设计算法找到元指代词所指的对象,并评价该元指代消解实验.

1 语篇表示方法

要进行指代消解,首先我们选择一种方便在计算机中表示语篇的结构的方法.我们使用复杂特征集来进行语篇中名词、动词、句、段等语篇成分的特征描述.给出特征集描述如下:设 A 为一特征集,当且仅当 A 可表示为

$$A=\{f_1=v_1,f_2=v_2,...,f_n=v_n\}^T,n>0;n\in Z.$$

其中 f_i 为原子,表示特征名,v_i 为原子,表示特征值,$f_i=v_i$ 表示特征名 f_i 的值为 v_i.A 称为“属性/值”对集合.进一步地,复杂特征集定义如下:设 A' 为一复杂特征集,当且仅当 A' 可表示为

$$A'=\{A_1,A_2,...,A_n\}^T,n>0;n\in Z.$$

这里,A_i,i=1,…,n,为特征集或复杂特征集.由此可见,复杂特征集是一种嵌套结构,有利于表示复杂的词组、句子、段落、语篇的结构.

具体模型我们采用我国计算语言学家冯志伟提出的一种汉语句子的多叉多标记树模型(multiple-branched and multiple-labeled tree model,简称 MMT 模型)[5],该模型以一棵多叉树图形来表示句子的层次结构关系,并且多叉树的每个节点都是多标记的.所谓“多标记”也就是“复杂特征”.但 MMT 主要针对句子分析,描述级别不够丰富.为了对语篇进行分析,需要对 MMT 模型进行扩展,故引进句间的关系、段间的关系.在扩展的 MMT 模型中,我们采用的标记有 9 个方面的信息,包括汉语语篇中词、词组、句读、段落各个成分的类型、句法功能、语义关系、逻辑关系[5].

2 汉语元指代分析策略

有了汉语语篇的复杂特征集表述,我们就可以采用如下策略来进行元指代分析.

2.1 分析元指代词组的构成

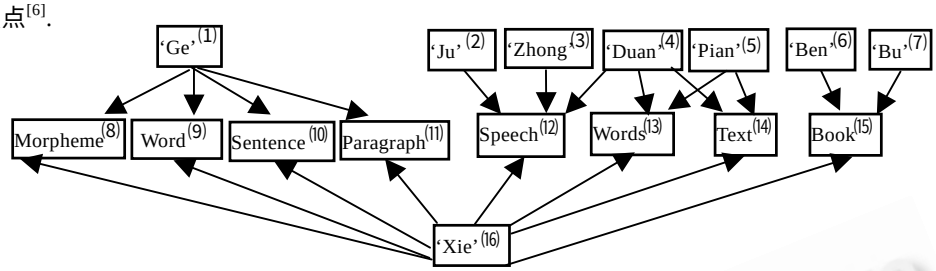
元指代词组可以看成是定指名词词组中的一类.有些名词在语篇中或在话语环境中能找到具体的一个或一类事物与之对应,称为可定指的名词,简称定指名词.Li,Yan-hui[6]认为:汉语的普通名词就其自身而言,都表示一类事物的抽象名称,是无指的.量词把抽象名称个体化(individuate),数词把个体数量化,而指别词(或指别词前的成分)把整个结构和话语环境中的具体事物联系起来,这几项组合起来,形成定指名词.

定义元指代表示模型为:设 V 为词汇集,L 为 V 上的语言.n 元组<W_1,W_2,…,W_n>为超距相关的词组成的 n 元组,W_i 包含于 V,可以为空集.M 包含于 L,定义函数 f:$W_n^+\rightarrow M$,其中 W_n^+表示 W_i(i 从 1 到 n)连乘,且其中至少有一个 W_i 不能为空.M 的元素为 k 元组,k 的秩小于等于 n.函数的约束条件表现在具体实现中,函数源 W_n^+中最后一个不为空的 W_j={字,词,词语,话,句,句子,段,段落,文字,章,文章,书},该 W_j 称为元指代标志词集.其他的 W_i 包括指别词集={这,那};量词 a 类集={个,句,种,些,段,篇,本,部};形容词 a 类集={表形状、色彩的形容词};专有名词集;人称代词集;方位词集等.对于量词和元指代标志词的组合,人们有一定的习惯,在具体实现中,为了简化处理,我们选择一些基本的组合约束(如图 1 所示).最后,我们得到的 M 中的元素就称为元指代词组.我们初步把元指代形式定义成一个封闭集,因为虽然存在例外实例,但由于数量较少,因此我们可以在一些规则中予以单独考虑.

2.2 区别话题和焦点

在语用层面,句子有话题和焦点之分.相对而言,动作、行为对施事的影响比对受事的影响要小,也就是说,施事本身的状态比较稳定,而受事通常在受到动作、行为作用后本身状态发生变化.人类认知倾向于以稳定的事物为话题去引导出多变的、往往也就是新的信息.典型的宾语是受事和常规焦点的结合.人类语言之所以倾

向于选择受事作焦点,可能是因为受事直接受动作影响而改变状态,往往是表达中的新信息,而新信息倾向于成为交流的焦点[6].



(1)个,(2)句,(3)种,(4)段,(5)篇,(6)本,(7)部,(8)字,(9)词,(10)句,(11)段、段落,(12)话,(13)文字,(14)文章,(15)书,(16)些.

Fig.1 Compositive habit between Chinese quantifier and meta_anaphoric symbol

图 1 汉语量词与元指代标志词组合习惯

2.3 引入EE和句焦点

由于汉语的句子构成复杂,我们引入小句——EE 的概念[7].EE 即 Elementary Event,代表一个最小但意思完整的语句结构.EE 作为我们关于句子元指代消解算法的基本处理单元.例如:我们听说他走了.该句中"我们听说某事"主谓宾完整,构成一个 EE,"他走了"这个主谓结构形成子句,构成另一个 EE.在多叉树图形中我们用椭圆表示句中的个体成分——名词词组,用方框表示段、句、EE 等句子层面成分.在不引起混淆的情况下,为方便起见,我们把方框表示的段、句、小句等句子层面成分统称为"EE 成分".在多叉树图形中,椭圆表示的名词词组是叶子节点,"EE 成分"是中间节点."EE 成分"的类型从大到小包括段、句、tell_event(具有句子结构的讲话内容)、preposition_event(具有句子结构的前置修饰语)、quote_event(由引号对括起的引语)、EE.在消解算法中,我们对"EE 成分"进行识别,得到的句子层面焦点称为句焦点.

建立多叉树图时,我们对"EE 成分"节点标记上数字,以便于区别.上层与下层的 EE 成分节点之间的关系属于前面提到的汉语句读之间的语义关系,用复杂特征集的属性描述.在此,我们把同层节点原本的句读之间的关系用上层与下层 EE 成分之间的关系来表示,且上层 EE 成分节点与它的下层中第 1 个 EE 成分节点的关系一般为 Condition,即状态描述关系,语义关系还包括 Cause 关系,即因果关系、Conjunction(连接)关系、Content(内容)关系、Explanation(解释)关系、Disjunction(转折)关系,或是直接由引号标记的 Reference(引用)关系.虽然句读之间的关系非常复杂,但如果对其进行简化,还是可以满足消解算法的要求的.

3 元指代消解算法

元指代消解算法建立在焦点理论之上.我们在扫描语篇时对某个 EE 中的组成对象进行识别,把焦点对象放置在一组焦点寄存器中,方便进行指代消解.我们对"EE 成分"本身组成的集合也进行焦点的识别,得到的焦点称为句焦点,同时使用一组规则来实现元指代消解算法.

判断某个 EE 是否是段中的句焦点,我们根据实验语料发现主要有两个规则:段内第 1 个"EE 成分"可以缺省地认为是本段的句焦点;如果后续的"EE 成分"是前个"EE 成分"的 Content(内容)关系、Explanation(解释)关系、Disjunction(转折)关系,或是直接由引号标记的 Reference(引用)关系,则后续的"EE 成分"成为新的句焦点,如果后续的"EE 成分"与前个"EE 成分"关系不是前 3 种关系,则句焦点不变.

消解算法中我们应用两个原则:

(1) 优先:焦点优先,选择焦点集的对象为候选对象.

(2) 过滤:对不符合性、数、格等词语属性特征的词语进行过滤.

从简到繁,我们对元指代按字词、句、段、篇分别进行分析.在此,我们只考虑句子的处理.字词、段的处理原理类似,此处不再赘述.

具体采用下列规则来确定指代对象,规则建立的依据是实验中获得的经验.

规则 **1**. 当扫描器发现元指代词组时,如果元指代词组＝{(专有名词)+(人称代词)+(方位词)+(指别词)+数

量词+引号对–形容词+元指代标志词},则引号对中的内容是指代对象.

规则 **2**. 当扫描器发现元指代词组时,如果该词组所在的 EE 中,出现元指代词组短语与普通名词呈并列句子成分(本实验中,认为同为主语成分或宾语成分)或元指代词组短语与由引号对所括起的短语呈并列句子成分,则普通名词或由引号对所括起的短语是元指代的指代对象.

规则 **3**. 当扫描器在某个 EE 中发现元指代词组时,如果不满足规则 1 和规则 2 的前提条件,且该元指代词组有数量词出现而无指别词同时出现,则暂时不进行元指代词组的消解,直到扫描器扫描下一句的 EE.如果这两句之间的标点符号是冒号,表示后一句的 EE 是前一句 EE 的内容,则该元指代词组的指代对象为句焦点寄存器 EE_CurrentFocus 中的对象.若无冒号,再看该元指代词组所处 EE 中的位置;若处于宾语位置,则输出无元指代现象(是指代现象);若处于主语位置,则选择在 EE_CurrentFocus 中的对象作为该元指代词组的指代候选对象,如果通不过属性特征的过滤算法,则对 EE_OldFocusStack 堆栈(它保存老的句焦点)由上而下依次选择堆栈中的候选对象进行属性匹配;若无匹配成功的对象,则输出无元指代现象(是指代现象).

规则 **4**. 当扫描器在某个 EE 中发现元指代词组时,如果无规则 1~3 的前提出现,则在 EE_CurrentFocus 中的对象作为该元指选择代词组的指代候选对象,如果通不过属性特征的过滤算法,则对 EE_OldFocusStack 由上而下依次选择堆栈中的候选对象进行属性匹配,若无匹配成功的对象,则输出无元指代现象(是指代现象).

为了更好地解释消解算法,我们通过例 1 来阐明算法和寄存器的使用方法.这一例句选择于语料库 1999.7.19《解放日报》“轻松制作镜像字”一文.

例 1:有人抱怨说 Word 样样都好,可就是没法制作出镜像字.这话太不公道了,Word 可算是此中“高手”.

初始设置多叉树图.段:=｛句 1|句 2｝.

EE11:有人抱怨说 tell_event.

段首的 EE 初始化寄存器后,使用设置寄存器的规则,将当前焦点寄存器 CurrentFocus(它保存当前 EE 的焦点)设置为宾语部分“tell_event”,话题寄存器 Theme(它保存当前 EE 的话题对象)设置为施事“有人”.话题集堆栈 ThemeStack(它保存本段中前几句 Theme 中的对象,按一定顺序排列)和焦点堆栈 FocusStack(它保存本段中提到的焦点对象)暂时为空.从句子层面看还有一组寄存器保存句子的层次结构,句焦点寄存器 EE_CurrentFocus(它保存当前的句焦点)中原本为“句 1”,“句 1”的下层“EE11”与它的关系是状态描述,句焦点不变,再扫描到下层“tell_event”与上层的关系是 Content,即内容,则句焦点转变成“tell_event”,原来的句焦点压入 EE_OldFocusStack.寄存器内容如下:

CurrentFocus＝tell_event;Theme＝有人;ThemeStack=NULL;FocusStack=NULL;

EE_CurrentFocus=tell_event;EE_OldFocusStack=句 1.

EE111:Word 样样都好.

这个 EE 未出现指代和元指代现象,由于当句子的动词是不及物动词或系动词时,焦点可以认为就是主语部分,即与话题重合[7],所以话题“Word”成为新的焦点.“EE111”是它的上层“tell_event”的第 1 个 EE 成分,关系是状态描述(condition),它不能成为新焦点.寄存器结果是:

CurrentFocus=Word;Theme＝Word;ThemeStack=有人;FocusStack=tell_event;

EE_CurrentFocus=tell_event;EE_OldFocusStack=句 1.

EE112:可就是没法制作出镜像字.

这个 EE 缺少主语,引入 it 对象,表示零指代,it 是指代词,应用消歧规则进行指代消歧.由于此处出现了零指代,而零指代产生的原因是前句中的焦点对象距后句中的同一对象很近,从而对后句的对象进行省略,形成零指代现象.我们采用指代消解规则,用分析上一个 EE 得到的 CurrentFocus 内容消解 it,由于零指代现象的研究较少,这种规则的正确率我们将在其他实验中与予评价.EE112 中出现转折词组“可就是”,与前面“EE 成分”呈转折关系,可以设为句焦点.寄存器结果是:

CurrentFocus=镜像字;Theme＝Word(it);ThemeStack=Word|有人;FocusStack=Word|tell_event;

EE_CurrentFocus=EE112;EE_OldFocusStack=tell_event|句 1.

EE21:这话太不公道了.

这个 EE 中发现了元指代词组“这话”,位置为句 2 的起始位置.对元指代词组“这话”进行消解,运用规则

3,EE112 就是指代对象.当扫描到“不”字时,作为否定系动词,焦点为主语部分“这话”.“EE21”的上层节点“句 2”与“段”呈 Conjunction 关系,即连接关系,“句 2”不能成为新焦点.同时,“EE21”是它的上层“句 2”的第 1 个 EE 成分,关系是状态描述(condition),它不能成为新焦点.寄存器内容为:

CurrentFocus=这话;Theme＝这话;ThemeStack=Word|有人;FocusStack=镜像字|Word|tell_event;
EE_CurrentFocus=EE112;EE_OldFocusStack=tell_event|句 1.

经过这个 EE21 的分析,对元指代词组“这话”进行了消解,EE112 就是指代对象.

EE22:Word 可算是此中“高手”.

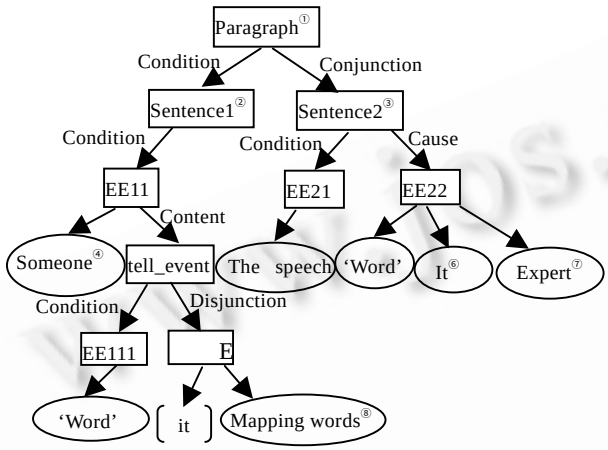
扫描此 EE 时发现 3 个名词.其中“此”字作为辅助成分限定了宾语的范围,它作为指代词可以是一般指代或语篇指代,情况比较复杂,为了集中注意力,我们对此类没有元指代标志词的留待以后讨论.由于此 EE 中的动词是系动词,焦点与话题为同一名词“Word”.寄存器内容变为

CurrentFocus=Word;Theme＝Word;ThemeStack=这话|Word|有人;FocusStack=这话|镜像字|Word|tell_event;

由于 EE22 与 EE21 是 Cause(解释关系),EE22 可以作为句焦点.

EE_CurrentFocus=EE22;EE_OldFocusStack=EE112|tell_event|EE11.

例句分析完成后,相应的 MMT 树图也建立起来了(如图 2 所示).



①段,②句 1,③句 2,④有人,⑤这话,⑥此,⑦高手,⑧镜像字.

Fig.2 The multiple-branched and multiple-labeled tree of example 1

图 2 例 1 的 MMT 树

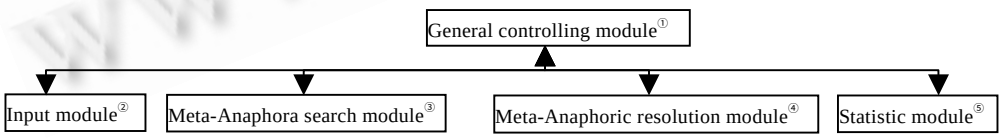
可以看出采用句焦点集,使“句”元指代词组的消解不再简单依靠相邻句子优先的选择方法,而是从句焦点候选集中挑选,使之更有针对性,消解更准确.

上述的消解原则也同样适合段、章节、篇、本文、本书等元指代现象的分析,这里从略.

4 实验分析

采用真实语料来分析.语料库选择为《解放日报》1999.1.1~1999.7.29 的全文库,体裁包括新闻、述评、小说、散文等.《青岛日报》1999.8.1~1999.8.30 的全文库,体裁包括新闻、述评、小说、散文、文摘等.挑选语料基于以下几个原则:(1) 代表性.所选语料均有一定的代表性,能反映典型的汉语语言现象.(2) 广泛性.所选语料力争能较全面地反映汉语元指代的各种用法.(3) 真实性.语料都从语料库中选出,没有我们自己随意组织或改动的语段.

我们对“字词”、“句”、“段”等语段分别进行实验.实验系统的粗框图如图 3 所示.



①总控模块,②输入模块,③元指代词发现模块,④元指代对象发现模块,⑤统计模块.

Fig.3 A raw drawing of the system

图 3 系统框图

系统采用 VisualC++6.0 实现.总控模块提供本系统的主界面,并可调用其他模块完成各项功能.输入模块功能是提供人-机交互手段,输入组成语篇的各个语句和输入解决人称指代问题时需要的各种语义信息.由于我们没有现成的 MMT 汉语句子分析系统,无法通过计算机自动分析得到语句的 MMT 表示,因此只能手工输入有关的语义信息.一个完整的多叉多标记树是比较复杂的.考虑到实验系统重在说明原理,因此本模块输入的并不是

各个语句完整的 MMT 表示,而只输入了后续模块需要的语义信息.

实验 1. 选择有关“字”的语段 100 段,由两名本科生对其中的元指代现象进行识别,答案包括有元指代现象、非元指代、拿不准是元指代还是非元指代 3 种.再把甲乙共同选出的元指代 48 项、非元指代 70 项、拿不准的 5 项的语段,作为进行计算机元指代消解的语料的标准答案,来判断计算机消解的正确率.同样处理有关“句”的语段 100 段,有关“段”的语段 60 例.

实验 2. 识别元指代词.对“字”类元指代的识别结果见表 1.

Table 1 Recognizability of ‘Zi’ type of meta-anaphora

表 **1** “字”类元指代词的识别结果

	Meta-Anaphora (48 items)①	Non-Meta-Anaphora (70 items)②	Uncertainty (5 items)③
Items found by the algorithm④	55	72	0
Items match the criterian⑤	44	64	0
Recall factor⑥(%)	91.7	91.4	…
Pertinency factor⑦(%)	80	89	…

①元指代(48 项),②非元指代(70 项),③拿不准(5 项),④算法找到的元指代项数,⑤算法找到的与标准答案相同的项数,⑥相对于标准答案判断正确率(查全率),⑦查准率.

其中人工判断中拿不准是否为元指代的词组都被算法判断成为元指代.如果考虑把人工拿不准的归入到元指代词中,元指代的查全率为 92.5%,查准率可达到 89%.结论是,由于元指代形式标记比较明显,机器识别时,正确率比较高.

实验 3. 寻找指代对象.

算法找到“字”类元指代词的指代对象有 55 项,其中正确的有 39 项,查全率为 81.3%,查准率为 71%.应该说不是很高.但这种误差是由识别元指代词时的误检误差扩大而造成的.我们可以看到,在识别元指代正确的 44 项中,找到指代对象 39 项,成功率为 88.6%.算法正确找到“句”类元指代词的指代对象 35 项,查准率为 87.5%.算法正确找到“段”类元指代词的指代对象 38 项,查准率为 84.4%.

基于句焦点的消解算法在元指代消解中效果还是令人满意的.通过实验我们得到如下结论:

句焦点的对于元指代对象的识别有很大帮助.之所以有这样的效果,是由于人们在使用句子表达思想时,不自觉地会对句子结构的安排有所侧重,使句子更好地为他人所理解.这可能是句焦点产生的原因.

由实验发现,语篇中的未定指元指代词组(数量词+元指代标志词组成的元指代,不包含指别词),往往在其后面可以找寻到它的指代对象,即有较强的定位期待.这在一般的未定指名词中比较少见.这种语言现象可能有一定的心理机制,对语篇元指代现象的产生有心理研究价值.

对错误识别元指代的解释:

(1) “鱼杆一字排开”、“德意志联邦共和国大十字勋章”这类句子中出现的“一字”、“十字”、“人字”是以字的形状为描述对象,人工判断时拿不准它们是不是元指代.算法把它们判断为元指代.我们按照有否指代语篇本身这一标准,认为它们不是元指代.

(2) 宝钗为拥林派所诟病的地方是一个假字,或者还有一个冷字.

此句中,“假”、“冷”字是元指代词组“一个字”的指代对象.由于“假”字、“冷”字在句中未被引号括起,误判为句中没有元指代词.由于文本中存在大量“是一个汉字”、“有一个红字”、“是个错字”等非元指代词,我们有理由认为此类字不用引号括起易造成歧义,人们可以根据上下文和语调进行分辨,而机器只能一筹莫展.

在对章节、篇、本文等元指代现象进行分析时,很多情况由于元指代发现策略不太好,造成指代对象分析错误.尤其在书评文章中,经常出现“在这本书中”、“开头的两篇”等字样,显然不是指代书评本身,而是指代被评论的文章.

总之,通过实验,我们一方面肯定了所采用元指代消解算法的思路及其算法实现的有效性,但同时也发现了一些分析错误.这对于我们将来进一步完善修改元指代发现规则和算法也指明了方向.

5 结束语

从实践中我们认识到,指代消解作为自然语言处理中不可或缺的一环,在语篇理解中有其重要的作用,它在机器翻译、文摘生成、情报检索中都将得到应用.元指代作为指代的一种,对它的分析,可以更深入地理解语篇的复杂结构,对语言本身有更完整的认识.目前,元指代消解系统还只是一个初步的系统,它所使用的规则还有很大的改进余地.比如就章节、篇中“本文”、“本书”等元指代的处理研究以及对无明显形式标记的元指代现象,对包括其他语种的元指代现象的研究,无论在理论上还是在实际应用上,都有十分重要的价值,也是我们今后继续努力的方向.

References:

[1] Grosz, B.J., Sidner, C.L. Attentions, intentions, and the structure of discourse. Computational Linguistics, 1986,12(3):175~204.
[2] Grosz, B.J., Joshi, A.K., Weinstein, Scott. Centering: a framework for modeling the local coherence of discourse. Computational Linguistics, 1995,21(2):203~225.
[3] Huls, C., Claassen, W., Bos, E. Automatic referent resolution of deitic and anaphoric expressions. Computational Linguistics, 1995,21(1):59~80.
[4] Chen, Zhi-qun. FB-Model of Chinese discourse semantic representation. Journal of Hangzhou University (Natural Science Edition), 1998,25(Suppl.):29~33 (in Chinese).
[5] Feng, Zhi-wei. Natural Language Processing. Shanghai: Shanghai Foreign Language Teaching and Education Press, 1996 (in Chinese).
[6] Lu, Bing-fu. Analysing syntactic relation using semantics and pragmatics. Journal of Chinese Philology, 1998,5:353~367 (in Chinese).
[7] Azzam, S., Humphreys, K., Gaizauskas, R. Evaluating a focus-based approach to anaphora resolution. In: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics. Montreal: Morgan Kaufmann Publishers, Inc., 1998. 74~78.

附中文参考文献:

[4] 谌志群.汉语语篇语义表示的 FB 模型.杭州大学学报(自然科学版增刊),1998,25(增刊):29~33.
[5] 冯志伟.自然语言的计算机处理.上海:上海外语教育出版社,1996.
[6] 陆丙甫.从语义、语用看语法形式的实质.中国语文,1998,5:353~367.

Study on Meta-Anaphoric Resolution in Chinese Discourse Understanding*

ZHANG Wei, ZHOU Chang-le

(*Department of Computer Science and Engineering, College of Information Science and Technology, Zhejiang University, Hangzhou* 310027, *China*);

(*Institute of Artificial Intelligence, Zhejiang University, Hangzhou* 310027, *China*)

E-mail: hdstsysgn@cise.zju.edu.cn

http://www.zju.edu.cn

Abstract: Anaphoric resolution is a hot field in computational linguistics. According to the research, it is found that a kind of anaphora, meta-anaphora, which represents the form of sentence, paragraph or discourse (not the subject or object of the sentence), plays an important role in discourse. So meta-anaphoric resolution affects the computer’s ability of discourse understanding. This paper focuses on the meta-anaphoric resolution. A new concept called Elementary Event focus (EE_focus) is introduced. And some meta-anaphoric resolution algorithms based on the concept are designed. The algorithms examined by a large Chinese natural language corpus. The experimental results show that the EE_focus set can do good job in meta-anaphoric resolution. The work enriches the theory of discourse’s structural representation. It is important in confirming the cohesion and coherence of the Chinese discourses.

Key words: natural language understanding; Chinese text analysis; anaphoric resolution; meta-anaphora; EE_focus

* Received September 3, 2000; accepted November 20, 2000
Supported by the National Natural Science Foundation of China under Grant No.69983006