基于语义结构平行的汉语人称代词消解*

臧翰芬[1,2] 韦向峰[2] 张 全[2]

([1]中国科学院研究生院 北京 100039；[2]中国科学院声学研究所 北京 100190)

zanghf@163.com, wxf@mail.ioa.ac.cn, zhq@mail.ioa.ac.cn

摘要：如何让计算机根据自然语言的语义表示消解句子乃至段落篇章中的人称代词，一直是自然语言处理的一大难题。本文依据HNC理论的句类表达式和语义块构成的相关知识，提出了一种基于语义结构平行的人称代词消解算法，通过定义句子语义块的层次结构，制定相关的人称代词消解规则和算法，实现了段落中人称代词的指代消解，经开放测试表明该方法具有较好的消解效果。

关键词：人称代词 指代消解 HNC理论 消解规则

Chinese Pronominal Anaphora Resolution Based on Parallel Semantic Structure

Hanfen Zang[1,2] Xiangfeng Wei[2] Quan Zhang[2]

([1]Graduate University of Chinese Academy of Sciences [2]Institute of Acoustics, Chinese Academy of Sciences)

zanghf@163.com, wxf@mail.ioa.ac.cn, zhq@mail.ioa.ac.cn

Abstract: It is a difficult problem in Natural Language Processing to resolve the ambiguity of Pronominal Anaphora in a sentence or paragraph by computer according to semantic expression. This paper is mainly centered on the knowledge of semantic categories of sentences and the formation of semantic chunks in the HNC theory. It brings forward a resolution of pronominal anaphora arithmetic based on parallel semantic structure, with the definition of the levels of semantic chunks and summarizing the rules of pronominal anaphora resolution. It realizes the resolution of pronominal anaphora in a paragraph. The experiment in open corpus shows that the arithmetic based on parallel semantic structure has a satisfied effect.

Keyword: Pronominal Anaphora, Anaphora Resolution, HNC Theory, Resolution Rules.

一、引言

指代是语言中常见的一种现象，由于“语言尚简”，故常使用指示代词代替已经出现过的能照应的语言单位（先行语）。指代消解就是确定指示代词指向哪一个先行语的过程，其中研究得最多的就是人称代词的消解。人称代词消解是自然语言处理中的一个重要问题。在信息抽取中，若抽取出的信息是人称代词，则会产生信息模糊和理解困难，必须进行指代消解，找出人称代词

* 本文承国家 973 项目“自然语言理解的交互引擎研究”(2004CB318104)、中科院声学所知识创新工程项目“句群理解处理理论及其应用”(0654091431)、中国科学院声学研究所“所长择优基金”(GS13SJJ04)、中国科学院青年人才领域前沿项目（O754021432）的资助。

作者：臧翰芬 (1976 -)，女，硕士研究生，研究方向：自然语言理解。 韦向峰：助理研究员。

张全：研究员，博导。

E-mail: zanghf@163.com

的真正含义。在文本倾向性判别中，需要得到指示代词和名词的同指关系，特别是人称代词和人名之间的同指关系，从而确定文本中各对象之间的褒贬关系。

本文主要研究汉语中人称代词的消解问题。对这一问题国内已有许多研究者提出了一些规则和算法，如鲁棒性的动态权值算法[1]、利用语言成分的层次结构和谓词的语义进行规则消解[2]、基于语篇表述理论 DRT 的消解规则[3]、基于语料库统计与学习的方法[4][5]等，他们都是利用语法或语义类别、性别、单复数、代词与先行语的距离等信息作为规则依据或统计特征。

本文将结合概念层次网络（HNC）理论，研究语义块内部和语义块之间的同指关系，在一个段落内研究人称代词的消解。本文在前人研究的基础上，定义了句子语义块的层次和结构，总结了基于语义结构平行原则的人称代词的一些常用消解规则，并给出了实现人称代词消解的具体步骤和方法。在实际语料中对人称代词“他、她、我”的消解规则进行了测试，结果表明本文使用的方法和规则具有较好的消解效果。

二、HNC 的语义结构观

HNC 理论认为，自然语言理解是从语言空间到语言概念空间的映射过程，人类感知语言的过程是对概念的激活、扩展、浓缩、转换和存储的全过程运作[8]。HNC 理论认为，自然语言语句句类表达式的基本类型是有限的，是可以穷尽的，总共有 57 组。这 57 组表示式可分为七大句类及其子类，称为基本句类。最基本的七大句类为：作用句（X）、过程句（P）、转移句(T)、效应句（Y）、关系句（R）、状态句（S）、判断句（D）。每种基本句类都有自己的句类表达式，基本句类之间可以混合，形成不同的语句表达式，共有 3192 组。

语义块是语句的下一级语义构成单位。这意味着两点：第一，语义块是句子的语义构成单位，而不是句法的构成单位；第二，语义块是句子语义的直接构成单位，而不是间接构成单位。第二点是必需的，因为词义也是句子的语义构成单位，却不是句子语义的直接构成单位。要建立句子的语义结构表示式，语义块的概念是必需的。

根据对句子基本语义信息的重要程度，语义块分为主辅两大类，主语义块（简称主块）是句子语义的必要的、主干的成分，辅语义块（简称辅块）是句子语义的可选的、次要的成分，辅块的符号统称 fK。而建立句子的语义结构表示式基本上只需要主语义块。主语义块有 4 种基元类型，分别是对象语义块 B、作用语义块 A、特征语义块 E 和内容语义块 C。其中 B、A、C 统称为广义对象语义块，用符号 GBK 表示。

句类表达式描述了句子的基本语义构成，由主语义块构成，例如作用句“XJ=A+X+B”。在句子中，语义块可以块扩（语义块扩展为句子）、句蜕（语义块由句子蜕化而来），而块扩（句蜕）中又可以含有语义块，这样就会形成一种语义块的嵌套层次关系。

以上关于句类和语义块的知识对于人称代词的指代消解具有非常重要的指导作用，可以确定一个句子的层次和语义块内部的层次，从而根据语句层次结构和语义块平行原则消解人称代词。

三、人称代词的消解规则

在人称代词的消解中，运用较多的是性数一致原则、就近原则、句法平行原则、焦点理论和语义平行原则。性数一致是指人称代词及其所指名词必须在性别和单复数上保持一致，如“她”

不能指男性人名，“他们”不能指单个人名，但是由于汉语的人名难以确定性别、也没有专门的单复数标志，这给计算机获取性别和单复数信息造成了困难。就近原则是指在一般情况下把与人称代词最近的人名作为其所指，这条原则可以作为没有规则可使用的情况下的选择。句法平行原则是指处于相同句法功能位置的代词和人名形成同指。焦点理论主要记录每个句子的中心，根据中心获得人称代词所指。而语义平行原则借助句子的语义结构和相关规则，找到人称代词对应的人名，具有较高的准确性，但前提是计算机能正确地分析出句子的语义结构。

本文主要研究一个段落中人称代词的指代消解，使用 HNC 理论及其分析技术对段落及句子的语义结构层次进行划分，在划分出的结构层次的基础上结合具体的句类给出了相关的人称代词消解规则。

3.1 语义结构及其层次的定义

定义 1：段落 Paragraph(P) 由一个或多个“句子”S 构成，即 $P=\sum_{i=1}^{n} S_i$ *。其中的“句子”是按照 HNC 理论定义的句子语义类别的句类表达式，即 57 组基本句类或混合句类作为句子的划分依据，而不是以逗号、句号、问号或者叹号等形式标记作为句子划分的唯一依据。*

定义 2：“句子” Sentence(S) 由一个或多个语义块 K 构成，即 $S=\sum_{i=1}^{n} K_i$ *。*

定义 3：语义块 Chunk(K) 分为辅助语义块(fK)、特征语义块 Eigen chunK(EK) 和广义对象语义块 General oBject chunK(GBK)。fK 为时间、手段、方式等辅助成分，EK 相当于句子的谓语，而 GBK 相当于句子的主语或宾语。

定义 4：当广义对象语义块 GBK 具有简单构成（不嵌套句子或句子的变形）时，语义块分为要素 YS 和说明 SM 两个部分，即 GBK=SM+YS。要素是广义对象语义块的中心，说明部分是块中心的修饰或说明。

定义 5：句子语义结构的一般表示式为：S=[fK]+GBK1+EK+GBK2+GBK3，其中 fK 称为辅块，实际句子可以没有 fK，但是必须有其它主语义块中的至少一个。

定义 6：当 GBK2 或 GBK3 为一个句子的原形时，而且该句子的 EK 属于 HNC 定义的 8 个块扩句类或 2 个条件块扩句类时，称该 GBK 语义发生块扩（语义块扩展为一个句子），即 GBK=[#Sk#]。

定义 7：当 GBK 或 fK 中包含句子或句子的变形时，且 GBK 不发生块扩，称其中的句子或变形后的句子为句蜕（由一个句子蜕化而来），即 K=St。句蜕分为三种类型：原型句蜕 St1、要素句蜕 St2 和包装句蜕 St3。原型句蜕中的句子为句子的原来形式，要素句蜕是句子变形后某要素变为语义块的末尾中心，包装句蜕是用原型句蜕或要素句蜕去修饰某成分。

根据以上定义，一个段落被划分为若干个句子，句子和句子之间的语义块可形成平行对应关系。如果一个句子的 GBK1 为人称代词，它的前一个句子的 GBK1 是人名，那么按照语义块平行原则这两个 GBK1 形成照应，完成了人称代词的消解。由于语义块可以包含句子，包含的句子中的语义块又可以嵌套句子，因此形成复杂的多层嵌套结构，但最终一个语义块总是能分解为某层的某个语义块的要素或说明部分。

定义 9：构成段落的句子的语义块定义为第一层次；第一层次的语义块如果发生块扩或包含句蜕，那么块扩或句蜕中的语义块定义为第二层次；如果第二层次的语义块又嵌套块扩或句蜕，那么其中的语义块定义为第三层次，依此类推。

定义 10：如果某个词语（如人称代词或人名）直接属于某个语义块的说明或要素部分，那么该词语的层次数定义为所属语义块的层次数。

3.2 人称代词消解的基本步骤和方法

以文本中的自然段落为处理对象，首先找出段落中的人称代词、人名，将句子中的语义块及其在句子中的层次划分清楚，然后依据语义结构平行原则和就近原则，结合具体的句类知识，得到人称代词所指向的人名。具体步骤如下：

1) 先使用最大熵模型对段落中的人名进行自动识别[7]，然后人工校正。
2) 划分出句子的语义类别和语义块内部构成情况，确定人称代词所处的层次。
3) 确定人称代词的位置类型：语义块的 SM 部分，还是 GBKn(n=1,2 或 3)、fK。
4) 根据排除规则[6]，在句子内找出人称代词不可能指向的人名，并记入排除列表。
5) 以人称代词所在的最低层次句类为基准，如果同层次前一句类的相同位置类型是人名，那么完成消解。否则，按先要素（YS）后说明(SM) 的顺序寻找人名，若找到人名则完成消解。若还找不到人名，则继续按此方式往前追溯，直到段落中同层次的第一个句类。如果完成消解，转 9)。
6) 将人称代词提升为上一层次的位置类型，以新的层次和位置类型向前追溯，直到段落中同层次的第一个句类。如果完成消解，转 9)。
7) 改变人称代词的位置类型，如 GBK2 改为 GBK1，继续向前追溯，若完成消解，转 9)。
8) 如果段落中有人名，那么将距离人称代词最近的人名作为先行语。
9) 消解结束。

以上是处理人称代词所指人名的基本流程，在这一过程中可能还需要根据第一二三人称、有无引号、是否特殊句类等制定一些特殊的约束规则，根据句子的具体条件改变追溯的方向或位置，下面是在人称代词消解过程中制定的一些规则。

3.3 人称指代消解的基本规则

为了便于规则的使用和理解，本文将规则分为语义结构平行规则、变换规则、排斥规则和特殊规则，限于篇幅关系，以下只给出了几个具体的规则和示例。

3.3.1 语义结构平行规则

在所有的规则中，语义结构平行规则是最基本的规则。如果人称代词与人名处于同一语义层次的同一位置，那么它们的语义结构类型是一致的，因此可以形成同指的语义关系。例如，关于语义块 GBK1 的语义平行规则指代消解规则如下：

规则 1：IF S(0).GBK1=Ppnoun，S(-1).GBK1= PersonName, THEN Anaphora[Pp,PN].

其中，S(0)表示当前句子，S(0).GBK1 表示当前句子句类表示式中的第一个广义对象语义块，S(-1)表示当前句子的上一个句子；Ppnoun 表示人称代词，简记为 Pp，PersonName 表示人名，简记为 PN。Anaphora[Pp,PN]表示人称代词和人名形成同指，如例 1 所示。

例 1 卓贤麟//表示//，[#希望//[#黄珊汕的这枚铜牌//能够促进//<中国蹦床运动/的发

展>，+让中国||~在２００８年奥运会时~||出现||更多的蹦床人才#]#]。他||认为||，[#中国在这个项目的难度上还和俄罗斯等强国||有||微小差距，++但我们的优势||在于||高度和动作完成质量#]。

在例 1 中，“卓贤麟”和“他”都是两个句子第一层次的 GBK1，符合规则 1，因此人称代词“他”优先指向“卓贤麟”，而不是第三层次的“黄珊汕”。规则 1 可以在语义结构类型和句子距离上进行扩展，即 GBK1 可以用 GBK2、GBK3、语义块的 YS、语义块的说明等进行替换，而当 S(-1).GBK1 不是人名时，可以用 S(-2)替换，同理用 S(-n)也成立。

3.3.2 变换规则

在实际语句中，并非所有的人称代词与所指人名都直接满足这种语义结构的平行关系，由于句类和格式等方面的不一致，人称代词与所指人名可能会是不同的语义类型，也可能处于不同的语义层次。这时候必须通过规则的变换使得它们处于同一层次的同一类型，才能运用语义平行原则进行人称代词的消解。以下是一些常见的变换规则：

规则 2： *IF S(0).fK=St,S(0). fK.St.GBK1=Pp，S(0).GBK1!=(PN and Pp),*
THEN S(-1).GBK1=S(0).fK.St.GBK1, S(0).GBK1 →... →S(-n).GBK1.

规则 2 表示，当前句子的 GBK1 不是人名也不是人称代词，而其辅块句蜕中的 GBK1 为人称代词，那么辅块中的 GBK1 可以作为当前句子的 GBK1 人称代词向前追溯消解。*“S(0).GBK1 →... →S(-n).GBK1”* 表示从当前句子的 GBK1 开始在前面句子的 GBK1 中逐个消解人称代词所指的人名。

规则 3： *IF S(0).GBK1.SM=Pp, ¬Anaphora[Pp,S(-1).GBK1], S(-1).GBK2=Sk, Sk.GBK1=PersonName, THEN Anaphora[Pp,S(-1).GBK2.Sk.GBK1]*

规则 3 表示，如果当前句子的人称代词是 GBK1 的说明部分，该人称代词无法在前一句的 GBK1 中得到消解，而前一句的 GBK2 为块扩，且块扩的 GBK1 为人名，那么当前句子的人称代词与前一句的 GBK2 中的人名形成照应，人称代词得到消解。如例 2 所示。

例 2 *虽然意大利媒体||预言||[#刘翔||[会]在本届奥运会上~||有||出色表现#]，[但]他的夺冠||还是在||\<意大利人|的意料>之外/。*

在例 2 中，GBK1 中的人称代词可以变换到前一句的 GBK2 中进行消解，找到所指人名。同理，GBK2 中的人称代词也可以变换到前面句子的 GBK1 中进行消解，寻找人名。

3.3.3 排斥规则[6]

除语义平行规则和变换规则外，还有句内共指的排斥规则，这些规则描述了哪些语义块之间或语义块内部构成成分之间不能形成共指。文献[5]总结了这些排斥规则，规则 5 是其中之一。

规则 5： *设 JKi 与 JKj 分别是同一个句子的二个广义对象语义块，则有：*

(Const (JKi)=Const (JKj)=1)∧(i≠j) → ¬Anaphora(Ωh (JKi)，Ωh (JKj))

在规则 5 中，JK 与 GBK 的定义相同，只是符号表示上有所不同。*“Const (JKj)=1”* 表示广义对象语义块不含句蜕或块扩，只包含简单的要素或说明部分。规则 5 表明，在同一个句子中具有简单构成的两个广义对象语义块之间不能形成共指，如例 3 的“她”和“席尔瓦”。

例 3 *她||以 8-5~||淘汰了||巴西姑娘席尔瓦。*

3.3.4 特殊规则

特殊规则是针对特殊句类或特殊人称等情况制定的一些规则。如信息转移句句类，它的常见的两种表现形式为：T31J=TA+T3+T3C 和 T3J=TA+T3+TB+T3C，它们的广义对象语义块 T3C

都会发生块扩，即 T3C=[#Sk#]。T3C 可以是多个句子，甚至是一大段话，并且 T3C 可以是打引号的直接引语或不打引号的句子。“说、表示、称”等动词引导的句子容易形成这种复杂的信息转移句句类。

例4：[illegible]教练||~赛后~||称||：“[#我||对[illegible]的失利||感到十分遗憾。他||错失了||\{与[illegible]在决赛会面}的机会/，这||本可成为||他拳击生涯中的一大亮点。#]”

例5：说起||<结束不久|的惊心动魄的比赛>，刘翔||又来了劲头。“我||真的没有想太多，我||只是想发挥||自己的水平。[说实在]，拿金牌||，我||是||没有想到，我||只是想||放开跑，+拼出去。”

对于直接引语中的第一人称代词“我”，需要到引号外寻找语义块 TA，人名 TA 可以在引号前面也可以在引号后面出现，如例 4。如果在当前句子的引号外的部分也找不到 TA，那么就到前面句子中寻找人名，进行人称代词的消解，如例 5。而直接引语（引号中的内容）中的第三人称代词一定不能指向人名 TA。

规则6：*IF　S(0).SC=(T31 or T3), S(0).T3C=[#Sk#], Sk.GBK1=FirstPpnoun,*

Quotation{T3C}, S(0).GBK1=TA=PersonName,

THEN　Anaphora[FirstPpnoun,PersonName].

规则 6 表示，如果当前句子为信息转移句的 T31J 或 T3J，且 T3C 为块扩类型的直接引语 Quotation{T3C}，那么 T3C 中的第一人称“我”与引号外的人名 TA 形成共指。

对于没有引号的间接引语，如果第三人称代词“他”或“她”在T3C的块扩Sk中充当GBK1，那么优先选择人名TA作为第三人称代词的照应语，如例7和例8。

例7：[illegible]||告诉||记者，[#当{刘翔|已经走过|她身边}很久了~||，她||[还在]为这句话||而激动。#]

例8：{回到|匈牙利}后~||，[illegible]||通过他的经纪人||表示||，[#他||因为卷入||兴奋剂丑闻，+感觉||{受到了|“羞辱”}，+所以将结束||运动生涯。#]

四、实验结果及分析

本文用经过人名识别处理和语句标注的 2004 年雅典奥运会的新闻报道作为研究语料，从约 30 万字的语料中选择了 40 段文本进行规则的提取与训练，又从语料中随机选择了其它 40 段文本进行手工测试，测试结果如表 1 所示。

表 1　人称代词消解测试结果

人称代词	人称代词出现的次数	正确消解数	错误消解数	正确率
他	41	35	6	85.4%
她	28	25	3	89.3%
我	56	51	5	91.1%

在测试的 40 段语料中，出现错误的原因主要有：（1）段落中没有出现人名；（2）人称代词所指的不是人名，而是“记者”等人类概念，或者指的是国家、机构；（3）未利用性别信息；（4）两个人名并列，如例 9。

例9：[illegible]与[illegible]都是 1 7 岁，已经拥有雅典奥运会的女子双人 1 0 米台桂冠。[illegible]||笑着说：“[#[illegible]今天竟然没有一个动作失误，真是少见！但愿明天决赛也能发挥成这样！”#]

为了和文献[1]的结果进行比较，本文也从 1998 年 1 月的人民日报语料中随机选取了 50 段文本进行测试，测试结果如表 2 所示。

表 2　用人民日报 199801 语料测试结果

人称代词	人称代词出现的次数	正确消解数	错误消解数	正确率
他	43	39	5	88.4%
她	26	25	1	96.2%
我	22	17	5	77.3%

与文献[1]不同，本文的测试标准是必须找出人称代词所指的人名，而不是某个同指链上的名词或人称代词。本文没有测试文献[1]中测试了的“他们”，但本文测试了“我”。另外文献[1]使用全自动的处理，而本文主要还是人工测试。测试中出现错误的原因大多数还是在段落中找不到人名，其次是“我”作为我党、我国等的简称，第三是“我”指的是作者，文中并未出现。

本实验没有消解“你、我们、你们、他们、她们”这些人称代词，这是因为“你”的用法比较灵活和宽泛，常出现在对话或讲话当中，有时候甚至不指向特定的某个人。而其他几个人称代词都是复数形式，需要更复杂的消解规则，还有待深入研究。

五、结语

本文使用 HNC 理论的语句分析技术和句子语义块结构的知识，对句子语义层次和结构做了定义和划分，根据语义平行规则、就近规则、排斥规则、转换规则等规则和相关算法来处理人称代词的消解问题。采用在封闭语料中总结规则和开放语料人工测试的方法，测试结果表明规则是有效的，对“他、她、我”的人称代词的消解正确率分别为：85.4%、89.3%和 91.1%。本文还从 1998 年 1 月的人民日报语料也选取了 50 段进行测试，结果正确率分别为：88.4%、96.2%、77.3%。对于“我”的消解正确率下降主要是因为“我”所指的是国家机构或作者，而且段落中往往没有出现“我”所指的人名。本文认为，先对语料进行 HNC 的句类层次结构划分是十分重要的，因为这样能较为准确地使用语义结构平行原则，更有利于人称代词的消解。

本文下一步的工作是，不断扩大语料分析的范围并总结出新的规则，通过测试进一步完善规则和处理算法，结合人名自动识别和人称代词自动消解技术，使计算机自动实现整个人称代词的消解过程，再进一步验证其准确率与召回率。

参考文献

[1] 王厚峰, 梅铮. 鲁棒性的汉语人称代词消解. 软件学报[J], 2005,16(5):700-707.

[2] 曹军, 周经野, 肖赤心. 基于语义结构分析的汉语零代词消解. 湘潭大学自然科学学报[J], 2001,23(4):28-33.

[3] 王晓斌, 周昌乐. 基于语篇表述理论的汉语人称代词的消解研究. 厦门大学学报(自然科学版) [J], 2004,43(1):31-35.

[4] 庞宁, 杨尔弘. 基于统计模型与规则的指代消解研究. 应用技术[J], 2007(5):61-62.

[5] 冯元勇, 孙乐等. 基于分类信心重排序的中文共指消解研究, 中文信息学报[J], 2007,21(6):22-28.

[6] 王厚峰. 汉语指代消解与省略恢复研究. 博士后出站报告[D], 中国科学院声学研究所, 2000 年 11 月.

[7] 贾宁, 张全. 基于最大熵模型和规则的中文姓名识别. 计算机工程与应用[J], 2007,43(35):1-4.

[8] 黄曾阳. 语言概念空间的基本定理和数学物理表示式[M]. 海洋出版社, 2004 年 7 月.