汉语语篇中人称指代消歧研究*

谌志群　周昌乐　郑洪

（杭州大学计算机系 58#，杭州，310028）

摘　要： 本文首先分析了汉语语篇中的人称指代规律．然后给出了“关注焦点”集的概念及“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歧算法．该算法充分利用了汉语语句级的语义信息，并反映了汉语的指代规律．在我们小规模的实验中，取得了 91.7%的消歧成功率．初步反映了该算法的可行性．

关键字： 自然语言理解，汉语，人称代词，指代消歧

中图法分类号： TP391

语篇是最高一级的语言单位，语篇一般由多个句子组成，但语篇不是多个句子的简单罗列，语篇中句与句之间必须有衔接成分，在语义上必须连贯．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．当指代词是人称代词时我们称之为人称指代．人称代词指代的一般都是名词或名词词组．消除人称代词的指代歧义是自然语言语篇理解的一项重要内容．所谓消除歧义也就是一个替换过程，即用人称代词指代的语言单位替换人称代词，使得计算机对语句和语篇的语义处理能够进行下去．本文提出了一种基于“关注焦点”集计算的人称指代消歧策略，并给出了消歧算法．实验的结果是令人满意的．

1　汉语人称指代分析

现代汉语中的人称代词主要有：我、你、他、她、它、我们、你们、他们、她们、它们等．人称代词又分为第一人称代词（我、我们）、第二人称代词（你、你们）和第三人称代词（他、她、它、他们、她们、它们）．第一人称“我”在语篇中往往指说话者．或者指代作者（比如在第一人称写法的记叙文中），或者指代语篇中的说话者（“我”出现在直接引语中）．“我们”是复数形式，指代多个说话者，有时还可将听话者包括在内．第二人称指听话一方，单数用“你”，复数用“你们”．第三人称指说话者和听话者以外的一方，单数用“他”、“她”、“它”，复数加“们”．“他”指男性，“她”指女性，“它”不指人而指物．“他们”可专指男性，也可兼指男性和女性，“她们”专指女性，如果男女兼有．可以写成“他们”．

2　“关注焦点”集及其计算

在语篇中，无论是陈述和说明的对象，还是动作的施与者与承受者都是名词、名词词组或其指代词，因此它们应成为语篇生产者(作者/说者)和语篇消费者（读者/听者）共同关注的中心．由于名词和名词词组在语句中的句法功能不同，其被“关注”的程度也就不同．名

* 浙江省自然科学基金项目资助．

词和名词词组在语句中可充当主语、宾语和其它成分（辅助语）. 为了反映名词和名词词组的被“关注”规律，我们为充当主语、宾语或辅助语的名词和名词词组分别设定一个权重序列：[5, 3, 1, 0]、[2, 1, 0]、[1, 0]. 其中，各个序列的第一个数值 5、2、1 反映的是主语、宾语和辅助语之间相对的被“关注”程度. 三个递减数值序列反映的是随着语句的推移，充当主语、宾语和辅助语的名词和名词词组被“关注”程度的递减. 这些权重序列是经验数值，可以根据实验情况进行调整.

为了量化名词和名词词组的被“关注”程度，我们分别为各名词和名词词组计算一个积分，以积分高低作为其被“关注”程度的度量. 权重序列是计算积分的基础. 当一个名词或名词词组第一次出现在一条语句中时，根据其充当的句子成分将相应权重序列的第一个值赋给它作为其积分. 如果在接下来的句子中该名词或词组不再出现，则其积分按其权重序列递减. 但如果该名词或词组又出现了，则其积分在顺序递减的基础上还要根据它在此句中的成分累加上相应的权重.

一般来说，在语篇的任一点（即语句）上总有若干个积分大于 0 的名词或名词词组. 我们将积分大于 0 的名词或名词词组称为语篇在该点的“关注焦点”. 我们把这若干个“关注焦点”按积分高低排序并组成一个有序集合，称之为语篇在该点的“关注焦点”集，记为 S_i（i 为语句序号）.

3 基于“关注焦点”集计算的人称指代消歧策略

由于积分反映的不仅是语篇生产者而且也是语篇消费者对名词和名词词组的“关注”程度，因此，只要不引起语义混乱并符合各种制约条件，积分高的名词和名词词组在它出现语句后面的语句中再出现就可以用代词替换. 这样做可以使语言表达简洁明快，更重要的是这符合认知心理，不会引起语篇生产者和语篇消费者之间产生交际障碍. 这一规律为我们解决人称指代问题提供了根据. 如果在检查语篇的第 i+1 条语句时遇到人称代词，可以在上一句的“关注焦点”集 S_i 中挑选符合语义制约条件并且积分最高的元素作为该人称代词的指代对象.

复数形式的代词（“我们”、“你们”、“他们”、“它们”等）指代的是一个名词或名词词组的集合. 如果在检查第 i+1 条语句过程中遇到复数形式的代词，扫描 S_i. 如果失败（没有找到符合制约条件的元素或根本没有复数形式的元素），可以将 S_i 中的若干个语义类别相同的元素临时组成一个集合，作为该人称代词的指代对象. 当然这个集合形式的元素应该符合制约条件.

MMT 模型是我国计算语言学家冯志伟提出的一种汉语句子语义分析和表示理论. 我们的系统用 MMT 模型来表示语篇中的单个语句.

根据第二小节和本节上面的分析我们可以给出人称指代的消歧算法. 假设一个语篇 $D(s_1 s_2 \ldots s_n)$ 由 n 条语句构成.

人称指代消歧算法：

step1 检查组成语篇的第一条语句（MMT 形式），计算 S_1.

step2 i=1

step3 若 i=n，结束. 否则检查第 i+1 条语句.

1. 若有人称代词，扫描 S_i，

(1) 若人称代词为“我”，则：

a. 若它出现在直接引语外，则将“作者”作为其指代对象

b. 若它出现在直接引语内，则将该句的深层主语作为其指代对象

（2）若人称代词为“你”，则：

若它出现在直接引语内，则将该句的深层直接宾语作为其所指对象

（3）若人称代词为“我们”，则：

若它出现在直接引语内，则将该句的深层主语和 S_i 中符合制约条件的若干个元素组成一个集合作为其指代对象

（4）若人称代词为“你们”，则：

若它出现在直接引语内，则将该句的深层直接宾语和 S_i 中符合制约条件的若干个元素组成一个集合作为其指代对象

（5）确认未定人称代词的指代对象：

a. 若人称代词为单数形式，以符合制约条件并且积分最高的元素作为该人称代词的指代对象.

b. 若人称代词为复数形式，以符合制约条件并且积分最高的元素作该人称代词的指代对象. 若失败，将 S_i 中的若干个语义类别相同的元素临时组成一个集合，作为该人称代词的指代对象.

该句的指代确认以后，以指代对象代替人称代词.

2. 计算 S_{i+1}

（1）S_i 中原有元素的积分按其权重序列递减，若它又在该句中出现，则按其在该句的句法功能累加其积分.

（2）若 S_i 中原有元素的积分变为 0，删除它.

（3）将本句出现的新的名词或名词词组加入 S_{i+1} 中，并为其赋一积分.

（4）对 S_{i+1} 中元素排序.

step4　i=i+1

step5　转 step3

说明：step3 中的制约条件包括性、数方面的一致；语义方面的一致，如名词或名词词组在语句中充当的格角色（MMT 中提供）与它或它的中心词可以充当的格角色（词典中提供）是否一致等. “作者”不能作为“关注焦点”集中的元素，直接引语内“我”、“你”指代的对象不计算积分，直接引语内出现的其它名词或名词词组均按辅助语计算积分.

4　实验及实验结果分析

根据前面提出的算法，我们在 PC 机上实现了一个实验系统. 实验系统由总控模块、输入模块、计算模块、输出模块和统计模块构成. 为了验证实验系统的效果，我们挑选了一批语料作了实验. 语料几乎都选自现代文学作品. 挑选语料基于以下几个原则：一，代表性. 二，广泛性. 三，真实性. 本文实验所用语料由 40 个语段组成. 这 40 个语段共有人称代词 121 个，各个人称代词均占一定的比例. 实验中共有 7 个语段的 10 个人称代词消歧失败. 它们或者没有找到指代对象，或者没有找到正确的指代对象. 实验总的成功率是 111/121*100%=91.7%，应该说实验结果是令人满意的.

5　结束语

面向语篇单位来进行汉语的语形分析、语义表示和语用处理正成为我国计算语言学界新的研究热点. 如何形式化地表示和处理语篇中句与句之间的语义联系是语篇理解的一个

重点和难点．本文在这方面作了一点尝试，希望我们的工作能得到专家指正．

参 考 文 献

1 胡壮麟.语篇的衔接与连贯. 上海：上海外语教育出版社,1994
2 陆汝钤.人工智能.北京：科学出版社,1996:1008-1021.
3 黄伯荣,廖序东.现代汉语(修订本).甘肃：甘肃人民出版社,1983.
4 吕叔湘.近代汉语指代词.上海：学林出版社,1985.
5 冯志伟.中文信息处理与汉语研究.北京：商务印书馆 ,1992：71-102
6 冯志伟.汉语句子描述中的复杂特征.中文信息学报,1990，4（3）：20-29
7 Carla Huls, Edwin Bos, Wim Classen. Automatic Referent Resolution of Deictic and Anaphoric Expressions. Computational Linguistics, 1995，21（1）：59-79
8 Barbara J. Grosz, Aravind K. Joshi, Scott Weinstein. Centering: A Framework for Modeling the Local Coherence of Discourse. Computational Linguistics,1995，21（2）：203-225

A Study of Personal Anaphora Disambiguating for Chinese Discourse

Chen Zhiqun Zhou Changle Zheng Hong

(Department of Computer Science and Technology, Hangzhou University, Hangzhou, 310028)

Abstract: First, this paper analyses personal anaphora of Chinese discourse. After giving concept and computing method of focus-set, a algorithm of personal anaphora disambiguating based on focus-set is presented. This algorithm takes advantage of semantic information of sentence-level and reflects the orderliness of anaphora in Chinese. For a small-scale corpus, correct disambiguating rate of 91.7% is obtained. The experimental results reveal the feasibility of the proposed algorithm.

Key words: natural language understanding, Chinese, personal pronoun, anaphora disambiguating

（上接第 115 页）

6 R.Chandra, A.Gupta, and J.Hennessy. Cool: A Language for Parallel Programming, Languages and Compilers for Parallel Computing, The MIT Press, 1990, 126-148

An User Interface Model Based on Predicate/Transition Net

Chen Huinan

(Department of Computer Science and Technology, Nanjing Institute of Posts and Telecommunications, 210003, Nanjing, PRC)

Abstract: This paper presents first an approach to specify user interface using predicate/transition net combined with object-oriented event model and then discusses some issues arising in the multi-thread dialogue, such as concurrency and synchronization, abstraction and sharing mechanism.

Key words: User interface, Multi-thread dialogue, Petri net, Event model