第36卷 第8期 计 算 机 工 程 2010年4月
Vol.36 No.8 Computer Engineering April 2010
·博士论文· 文章编号：1000—3428(2010)08—0033—04 文献标识码：A 中图分类号：TP391

指代消解综述

孔 芳[1,2]，周国栋[1,2]，朱巧明[1,2]，钱培德[1,2]
(1. 苏州大学计算机科学与技术学院，苏州 215006；2. 江苏省计算机信息处理技术重点实验室，苏州 215006)

摘 要：给出指代消解的基本概念，从指代消解的语料资源、评测系统和算法3个方面出发，介绍指代消解的国内外研究现状，分析制约指代消解的3个关键问题：结构化句法信息的自动获取和表示，深层次语义信息的自动获取和使用，跨文本指代消解，基于分析结果给出国际上指代消解的研究趋势。

关键词：自然语言处理；指代消解；信息抽取

Survey on Coreference Resolution

KONG Fang[1,2], ZHOU Guo-dong[1,2], ZHU Qiao-ming[1,2], QIAN Pei-de[1,2]
(1. School of Computer Science and Technology, Soochow University, Suzhou 215006;
2. Key Lab of Computer Information Processing Technology of Jiangsu Province, Suzhou 215006)

【Abstract】 This paper interprets the concepts of coreference resolution, and introduces the state-of-the-arts in coreference resolution from three aspects: corpus resources, evaluation measures and resolution algorithms. It analyzes the critical problems of coreference resolution: automatic capture and presentation of structured syntactic knowledge, automatic capture and usage of underlying semantic, cross-document coreference resolution. Based on the analysis, it explores its international research trend.

【Key words】 nature language processing; coreference resolution; information extraction

1 概述

随着计算机技术和互联网的迅速发展，人们步入了信息时代，各种信息呈爆炸式增长。人们在享受其提供的便利的同时，也面临着如何从海量信息中寻找自己所需内容的困境。在这一背景下，信息抽取的需求日益紧迫。指代是一种常见的语言现象，是信息抽取不可或缺的组成部分。在信息抽取中，用户关心的事件和实体间语义关系往往散布于文本的不同位置，涉及的实体通常可以有多种不同的表达方式，为了更准确且没有遗漏地从文本中抽取相关信息，必须对文章中的指代现象进行消解。

指代消解在信息抽取中起着重要作用，在自然语言接口、机器翻译、文本摘要和问答系统等应用以及篇章理解中也很关键。通常，相同信息会在同一文本中出现若干次，为了保证文本的简练、减少冗余，文本的概念关联性往往通过指代关系来刻画。因此，有必要把这些指代互相联系起来，实现相关信息的融合，以获得相应信息在该文本中的完整描述。

2 指代消解的基本概念

指代是一种常见的语言现象，广泛存在于自然语言的各种表达中。一般情况下，指代分为2种：回指(也称指示性指代)和共指(也称同指)。回指是指当前的照应语与上文出现的词、短语或句子(句群)存在密切的语义关联性，指代依存于上下文语义中，在不同的语言环境中可能指代不同的实体，具有非对称性和非传递性；共指主要是指2个名词(包括代名词、名词短语)指向真实世界中的同一参照体，这种指代脱离上下文仍然成立。回指和共指存在很大的交集，又不互相包含，目前指代消解研究主要侧重于等价关系，只考虑2个词或短语是否指示现实世界中同一实体的问题，即共指消解。本文对指代和共指并不加以严格的区分，主要讨论共指消解的关键技术。

英语中指代有多种类型，常见的包括：

(1)人称代词(pronoun)指代，例如：***Computational linguistics*** from different countries attended the tutorial. ***They*** took extensive note.

(2)别名(name alias)指代，例如：***Microsoft Corp. announced its new CEO yesterday. Microsoft said ...***

(3)同位语(apposition)指代，例如：***Julius Caesar, the well-known emperor, was born in 100 BC.***

(4)有定名词短语(definite noun phrase)指代，例如：***Computational linguistics from different countries attended the tutorial. The participants took extensive note.***

(5)指示名词短语(demonstrative noun phrase)指代，例如：Boorda wants to limit the total number of sailors on ***the arsenal ship*** to between 50 and 60. Currently, ***this ship*** have about 90 sailors.

(6)谓词别名(predicate nominal)指代，例如：***George W. Bush*** is ***the president of the United States***.

(7)其他名词短语(bare noun phrase)指代，例如：The price of ***aluminum*** siding has steadily increased, as the market for

基金项目：国家“863”计划基金资助项目(2006AA01Z147)；国家自然科学基金资助项目(60673041, 60873150)；高等学校博士学科点专项科研基金资助项目(20060285008, 200802850006)；江苏省高校自然科学基础研究基金资助重大项目(08KJA520002)；江苏省高校自然科学基础研究基金资助项目(08KJD520010)；江苏省自然基金基础研究计划基金资助项目(BK2008160)；苏州市软件专项基金资助项目(SGR0807)

作者简介：孔 芳(1977－)，女，博士研究生，主研方向：自然语言处理；周国栋、朱巧明、钱培德，教授、博士生导师

收稿日期：2009-11-10 **E-mail**：kongfang@suda.edu.cn

aluminum reacts to the strike in Chile.

中文的指代主要有3种典型的形式：

(1)人称代词(pronoun)，例如：***李明***怕高妈妈一人呆在家里寂寞，***他***便将家里的电视搬了过来。

(2)指示代词(demonstrative)，例如：***很多人都想创造一个美好的世界留给孩子，这***可以理解，但不完全正确。

(3)有定描述(definite description)，例如：***贸易制裁***仿佛成了***美国政府在对华关系中惯用的大棒***，然而，***这根大棒***果真如美国政府所希望的那样灵验吗?

3 指代消解语料资源

与其他基于有指导机器学习技术的自然语言处理问题一样，指代消解任务的完成离不开标注好的语料资源。目前较知名的指代消解标注资源有MUC(Message Understanding Conference)和ACE 2种。

MUC是美国政府支持的一个致力于真实文本理解的例会，该会议从1977年~1998年共举办了7届，负责对世界各地不同单位的消息理解系统进行系统化的评测。在1995年的MUC-6和1998年的MUC-7上，指代消解成为MUC评测的重要任务之一。

MUC中指代关系的标注引入了2个标注对，采用SGML标注方法：

(1)用<COREF ID ="I">表示实体的左边界，用</COREF>表示右边界。

(2)用<COREF ID ="J" REF ="M">表示参照表达式左边界，同样用</COREF>表示右边界，其中，I, J 表示顺序号，在一个篇章内，序号由1开始严格单调递增，而M表示J的先行语的编号。如果满足M=I，则表示编号为J的实体为指示语，指向编号为I的名词短语。如果实体I只有ID，没有REF标号，则表示该名词短语无先行词，只是被其他名词短语所引向。

图1给出了一个MUC语料中指代链的标注实例。从中可以看出，代词"it"的REF属性值为4，而名词短语"American Airlines unit"的ID属性值为4，因此，它们位于同一个指代链，其中，"it"为照应语；"American Airlines unit"为指代词。

Amr corp. 's <COREF ID = " 4 " TYPE = " IDENT " REF = " 0 " MIN = " American Airlines " > American Airlines unit </COREF> said <COREF ID = " 5 " TYPE = " IDENT " REF = " 4 " > it </COREF> has called for <COREF ID = " 6 " TYPE = " IDENT " REF = " 7 " MIN = " mediation " > federal mediation in <COREF ID = " 9 " TYPE = " IDENT " REF = " 10 " MIN = " talks " STATUS = " OPT " > <COREF ID = " 8 " TYPE = " IDENT " REF = " 5 " > its </COREF> contract talks with <COREF ID = " 11 " TYPE = " IDENT " REF = " 12 " MIN = " unions " > unions representing <COREF ID = " 19 " MIN = " pilots " > <COREF ID = " 13 " TYPE = " IDENT " REF = " 8 " > its </COREF> pilots </COREF> and <COREF ID = " 21 " MIN = "attendants" > flight attendants </COREF> </COREF> </COREF> </COREF> .

图1 MUC指代链标注实例

ACE评测从1997年开始酝酿，2000年12月开始启动，从ACE 2004开始加入了中文语料。ACE语料主要来自于广播新闻(40%)、新闻专线(40%)和网络对话(20%)，因此，语料又细分成BNEWS, NPAPER和NWIRE 3个子语料。ACE中指代信息是以实体链形式标注的，即具有相同指代关系的实体位于同一指代链，且该指代链拥有唯一的编号。语料中每篇文章的实体链独立记录在对应的XML文件中。

4 指代消解评测

随着自然语言处理研究的不断深入，在一个公开的数据集上进行公平的系统评测成为一种推动相关研究发展的方式，指代消解研究也不例外。到目前为止，指代消解相关的评测有3种：MUC评测，ACE评测和ARE评测。

MUC对指代消解结果的技术评估有3个重要标准：召回率R、准确率P和F值，其中，召回率是指代消解结果中正确消解的对象数目占消解系统应消解对象总数的百分比，它反映了指代消解系统的完备性；准确率是指代消解结果中正确消解的对象数目占实际消解的对象数目的百分比，它反映了指代消解系统的准确程度。比较2个不同系统的性能时，一般使用F值，F值是召回率和准确率这2个指标的综合值，定义如下：

$$F = \frac{(\beta+1)P \times R}{\beta \times P + R}$$

其中，P为准确率；R为召回率；β为召回率和准确率的相对权重，一般取1，因此，F值可以表示为

$$F = \frac{2 \times P \times R}{P + R}$$

MUC提供了标准的评测程序MUCScorer，它根据传递闭包算法计算准确率、召回率和F值。从1995年9月的MUC-6到1998年4月的MUC-7，MUC会议上的指代消解评测均针对英语进行。

ACE对系统的性能评测是以与标准答案的匹配程度作为衡量结果的，主要采用基于漏报(标准答案中有而系统输出中没有)和误报(标准答案中没有而系统输出中有)的方法。由于在人工标注时，对于所有对象是以描述它的最长子串作为标注对象的，因此每一个标注中包含了关于对象的修饰信息，评测时会根据匹配程度记录不同的分值。

自2003年起，ACE中开始包含中文指代消解的相关评测，至今已经开展了4次，是目前唯一的中文指代消解国际评测。评价公式如下：

$$System\_Value = \frac{\sum_{i} Value(sys\_output_i, reference_{map(i)})}{\sum_{m} Value(reference_m, reference_m)}$$

其中，sys_output_i对应抽取系统输出的第i个输出项；$reference_{map(i)}$表示标准答案中对应的匹配项；$reference_m$表示标准答案中实际标注的形式。

与MUC评测不同，ACE评测不是专门针对指代消解的评测，其中涉及很多属性信息的检查，因此，很少有论文采用这种方法进行指代消解的评测。

2006年11月-2007年3月，英国伍尔佛汉普敦大学发起了一个名为ARE的指代消解评测，它是目前最为全面的针对英语的指代消解评测，包含4项评测任务：

(1)预标注文档上的人称代词消解

文档内所有名词短语均被识别出，且需要消解的代词也被标注出。参加评测的系统需要对每个人称代词在一个不含

人称代词的名词列表中找到正确的先行词。

(2)预标注文档上的指代消解

文档内所有名词短语均被标识，参加评测的系统需要识别出所有的指代链。

(3)生语料上的人称代词消解

文档未进行任何标注，参加评测的系统需要自行识别所需信息，并识别出人称代词对应的先行词。

(4)生语料上的指代消解

文档未进行任何标注，参加评测的系统需自行识别所需信息，并识别出所有的指代链信息。

前 2 项工作在预标注文档上进行，用于评测系统的指代消解算法；后 2 项不仅评测指代消解算法，还考察了名词短语识别等预处理工作对指代消解的影响。

5 指代消解算法研究现状

5.1 国外研究现状

指代消解的研究历史悠久。许多早期的方法侧重于从理论上进行探索，运用大量手工构建的语言甚至领域知识进行指代消解。近 10 年来，由于自然语言自动处理技术的发展以及各类应用对指代消解技术的需求越来越迫切，人们转向了基于弱语言知识的方法，侧重于实用的自动指代消解技术的研究开发，并取得了一定的进展。不过，受制于弱语言知识，自动指代消解技术近年来在性能的继续提高上遇到了瓶颈，研究人员开始把焦点转向基于自动产生的深层语言知识，特别是结构化句法信息和语义信息方面的研究，以期望取得性能上的突破。

5.1.1 早期的指代消解研究

早期的指代消解研究都是利用大量手工构建的领域和语言知识形成逻辑规则进行消解的。代表性的工作包括：

(1)基于完全解析树的遍历算法

1978 年，Hobbs 提出了一种不依赖任何语义知识或语篇信息，只利用语法规则和完全解析树信息的指代消解算法。算法首先为文档中的每句句子建立完全解析树，然后采用从左到右广度优先的搜索方法遍历完全解析树，最后根据语法结构中的支配和绑定关系选择合法的名词短语作为先行语。该算法是以模型方式提出的，在实际系统中很少直接使用。

(2)基于句法知识的方法

该方法充分利用句法层面的知识，以启发式的方式运用到指代消解中。例如：

1994 年，Lappin 等提出了一种 RAP 算法，使用 McCord 提出的槽文法(Slot Grammar)获得文档的句法结构，并通过手工加权的各种语言特征(如主语和宾语)计算各先行语候选的突显性，利用过滤规则确定先行语，实现句内和句间第三人称代词和反身代词的消解。

1996 年，Kennedy 等对 RAP 算法做了修改和扩展，避免构建完整的解析树，只用自然语言处理工具预处理得到词性标注和句法功能标注等浅层信息，在此基础上确定先行语。

1998 年，Mitkov 则在词性标注的基础上，对不同的语言特征进行量化，使用计算权值的方法对可能的候选词进行排序，从中选出先行语，解决代词的指代消解。

由于上述工作中的语言特征都是手工加权的，因此为了实现语言特征加权的自动化，2005 年，Luo 等对 RAP 算法进行了改进，尝试使用最大熵模型来自动确定各种语言特征的权值。

早期指代消解方法都需要大量的人工参与，系统的自动化程度非常低，系统的可移植性也较差。

5.1.2 近期的指代消解研究

随着标注语料库的不断出现以及 Internet 的迅速发展，实验语料的获得越来越方便，目前大多数的指代消解研究已转向基于语料库的指代消解方法。其中主流的方法大致分类如下：

(1)基于规则的方法

Brennan、Strube 和 Tetreault 分别于 1987 年、1998 年和 2001 年利用中心理论，首先根据潜在或当前中心的不同对先行语候选进行分类，然后使用各种中心获取算法选出先行语。Zhou 等于 2004 年提出了基于限制规则的多代理策略，在此基础上实现的系统代表了目前基于规则的指代消解系统的国际先进水平。

(2)基于统计的方法

1990 年，Dagan 等提出优先考虑同现频率较高的先行语候选作为代词先行语，对代词“it”的消解进行了研究。1999 年，Cardie 等提出通过聚类方法进行名词短语的同指消解，其基本思想是收集篇章中的基本名词短语，根据短语的特征对名词短语聚类，判断 2 个名词是否属于同一个类。

(3)基于分类的方法

1995 年，McCarthy 把判断先行语的问题转换成分类问题，通过分类器判断指代语与每个先行语候选之间是否存在指代关系。这一思想为日后指代消解的研究开辟了一条全新的道路。文献[1]则给出了详尽完整的实现步骤，并开发出实用的系统。在此基础上，许多研究者做了不同程度的扩充，并取得了一定的进展。典型的研究成果包括：

Ng 等对 *Soon* 的研究进行了扩充，抽取了 53 个不同的词法、语法和语义特征。

文献[2]提出了一个双候选模型，直接学习各先行语候选之间的竞争关系，以更好地确定先行语。

2004 年，Zhou 等对先行语候选指代链中的语义信息在代词(特别是中性代词)指代消解中所起的作用进行了探索，并在此基础上提出了进一步使用上下文信息和网络挖掘技术自动判别代词的语义类别的方法，从而更好地解决了代词的指代消解。

目前，大多数指代消解系统都采用局部优化方法，即对于每个指代语，依据不同算法，选择最佳的先行语。为了实现全局优化。Luo 等采用贝尔树表示搜索空间，以求最优化的指代消解方案。Ng 根据不同系统各自的特点，从不同系统中选择最佳的分区方案。

5.2 国内研究现状

与国际上指代消解的长期研究相比，自然语言处理领域的中文指代消解研究才刚起步，主要集中在人称代词的消解研究方面。相关的研究可分为 2 类：

(1)引用国际上流行的研究方案进行中文指代消解的研究。代表性的研究有：文献[3]根据中文人称代词的语义角色和对应的先行语可能的语义角色，给出了消解人称代词的基本规则；王厚峰等采用近似 Mitkov 的基于弱化语言知识的方法来解决人称代词的消解。

(2)根据中文特点提出的具有中文特色的研究方案。相关的研究有：许敏等利用格框架，提出了在上下文相关语义环境中进行指代分类解决的思想，并给出了相应的算法。王厚峰提出了基于 HNC 的指代消解方法，利用各种语义块的类型特点和语义块之间的结构特点，在语义块内部和语义块之间

使用排除规则，并使用局部焦点优先的原则进行优先选择，实现语句序列之间人称代词的消解。

6 研究热点及趋势

从指代消解的国内外研究现状可以看到，随着机器学习方法的引入，结合相关的领域知识，指代消解有了长足的发展，但还存在以下3个方面的问题：

(1)结构化信息在指代消解中的应用

虽然结构化的句法信息在许多较高层次的自然语言处理研究(如句法解析、语义作用标注、语义关系抽取和指代消解)中起着关键作用，但哪些结构化的句法信息是有效的以及在具体研究中如何充分有效地表示结构化句法信息依然是悬而未决的问题。

近年来，许多研究人员对此做了大量有益的工作。早期的研究从特定的应用出发，使用基于特征的方法选择和定义一系列可以从浅层或深层解析树中获取的平面特征来表示特定的结构化信息。这种方法已被大量应用于句法解析、语义作用标注、语义关系抽取和指代消解等相关领域。但这一方法并不能有效获取、表示复杂的结构化信息。

(2)深层次语义信息在指代消解中的应用

在自然语言处理中，语义信息起了至关重要的作用。目前许多应用都是通过使用类似WordNet这样的语义字典获取语义信息的，但数据库中语义信息是有限的。

近年来，一些研究者开始尝试利用数据挖掘的方法从大量的文本语料库中寻找有效的语义信息。一种最常见的解决方法就是使用一定的模式来表示特定的语义信息。但如何自动地选择模式，以及如何评估选中模式的有效性，还在进一步的研究中。

(3)跨文本的指代消解

信息抽取除了要解决文本内的信息融合外，还要解决跨文本的信息融合。在文本来源较广的情况下，很可能多篇文本描述了同一个事件和同一个实体。此外，不同文本还会存在语义歧义，例如，相同的词具有不同的含义，不同的词代表一个意义。除了信息抽取，跨文本的信息融合在许多自然语言处理应用中非常关键。但目前国际上基于跨文本的指代消解研究极少，尚未形成有效的解决方案。

针对上述问题，指代消解的研究热点及趋势主要有3个：

(1)有效获取并表示结构化句法信息

目前指代消解研究的趋势之一是如何系统地引入结构化句法特征。文献[4]提出了一种基于卷积树核函数的代词消解方法，以更好地获取结构化句法特征。其基本思想是先从给定的完全解析树中抽取出关键部分，然后以此为对象，通过卷积树核函数直接从中自动提取各种结构化特征信息，计算2个给定对象之间的相似度，并利用基于核的机器学习方法学习形成分类器，完成指代消解。文献[5]在文献[4]的基础上对卷积树核进行了适当改进，引入了部分上下文信息以帮助完成指代消解任务。

(2)有效获取并使用深层次语义信息

获取并使用更有效的语义相关信息是指代消解研究的又一热点。2005年，Yang等分别针对语料库和Web信息做了语义方面的尝试。对于语料库，他们借助浅层句法解析树中获得的语义角色信息，统计几种固定语义模式的出现频度。对于Web页面，则借助Google和Altavista这样的搜索引擎进行固定语义模式的统计。统计得到的结果再作为特征之一参与后续的指代消解工作。文献[6]发现前人对语义类的使用仅限于否定不同语义类之间的指代关系，而对于专有名词很难确定语义类。在此基础上提出了自动推导语义类，并将得到的信息应用于指代消解：首先借助BBN实体类型语料库训练生成分类器，将名词按ACE中指定的6种语义类分类；接着生成SCA(语义类是否一致)和KS(是否为ACE中的指定语义类)2个新特征；再将这2个特征引入一般的指代消解器，完成指代消解工作，最终取得了较大的性能提升。

(3)跨文本的指代消解技术的研究

目前的跨文本信息融合研究严重受制于跨文本指代消解研究的进展，还未发现相关的系统性的研究。不过，单文本指代消解研究的不断进展和人们对跨文本信息抽取技术的强烈需求必将促进跨文本指代消解的发展，这也使得跨文本指代消解和信息融合研究成为下一步跨文本信息抽取和篇章理解研究的重点。

7 结束语

本文系统全面地介绍了指代消解的发展历史和研究现状，在分析当前制约指代消解发展原因的基础上给出了目前该领域的研究趋势，为指代消解研究工作的进一步展开做好了基础性的工作。

参考文献

[1] Soon W M, Ng H T, Lim C Y. A Machine Learning Approach to Coreference Resolution of Noun Phrases[J]. Computational Linguistics, 2001, 27(4): 521-544.

[2] Yang Xiaofeng, Zhou Guodong, Su Jian. Coreference Resolution Using Competition Learning Approach[C]//Proc. of ACL'03. Sapporo, Japan: [s. n.], 2003.

[3] 王厚峰, 何婷婷. 汉语中人称代词的消解研究[J]. 计算机学报, 2001, 24(2): 136-143.

[4] Yang Xiaofeng, Su Jian, Tan Chew Lim. Kernel-based Pronoun Resolution with Structured Syntactic Knowledge[C]//Proc. of ACL'06. Sydney, Australia: [s. n.], 2006.

[5] Zhou Guodong, Kong Fang, Zhu Qiaoming. Context-sensitive Convolution Tree Kernel for Pronoun Resolution[C]//Proc. of IJCNLP'08. Hyderabad, India: [s. n.], 2008.

[6] Vincent N G. Semantic Class Induction and Coreference Resolution[C]//Proc. of ACL'07. Prague, Czech Republic: [s. n.], 2007.

编辑 张 帆