篇章中的消解问题与消解算法：研究综述*)

李莎莎[1] 李舟军[2] 陈火旺[1]

(国防科学技术大学计算机学院 长沙 410073)[1] (北京航空航天大学计算机学院 北京 100083)[2]

摘 要 篇章消解，即识别篇章中对现实世界中同一实体不同表达的过程，包括指代消解和同指消解两个方面。作为信息抽取的重要环节，它在信息检索、自动文摘及文本挖掘等领域有着广阔的应用前景。本文分析并总结了消解过程中常用的语言知识，介绍了上世纪90年代以来具代表性的算法，并指出了篇章消解未来的发展趋势。

关键词 指代消解，同指消解，排除性特征，推荐性特征

Research on Resolution within a Text

LI Sha-Sha[1] LI Zhou-Jun[2] CHEN Huo-Wang[1]

(School of Computer, National University of Defense Technology, Changsha 410073)[1]

(School of Computer Science & Engineering, Beihang University, Beijing 100083)[2]

Abstract Resolution within a text is a procedure to identify different expressions referring to the same entity in the real world, including anaphora resolution and corefernce resolution. As an important part of information extraction, it has broad application aspects in information retrieval, automatic text summary, text mining and so on. This paper presents some necessary knowledge on resolution problem. Some typical resolution algorithms in the past 10 years are introduced and the development trend is proposed.

Keywords Anaphora resolution, Coreference resolution, Eliminating feature, Preferncial feature

1 引言

在语言表达过程中，为了避免重复，人们习惯用代词、称谓、别称等指称前面提到过的某个名词短语、句子甚至句群，正确识别这种指称关系对计算机正确理解篇章至关重要。目前，识别篇章中的指称关系已成为自然语言理解领域中的一个热点，在信息检索、自动文摘以及文本挖掘领域都有着广阔的应用前景。

篇章指称关系的识别大体上经历了两个发展阶段：指代消解阶段(Anaphora Resolution)和同指消解阶段(Coreference Resolution)。指代消解主要是为那些与上文出现过的词、短语或句子(句群)存在密切的语义关联性且单独存在时语义不明的语言单位确定指代对象。待确定指向对象的语言单位称为指代语，被指向的语言单位称为先行语。最早的指代消解系统出现于上世纪60年代，70年代大批的研究人员投入到指代消解的研究中来。经历了80年代初短暂的低潮后，指代消解又再次成为研究的热点。1997年的EACL和1999年的ACL年会都设立了指代消解的专题会议，2001年的Computational Linguistics学报还出版了指代消解的专辑，这些都推动了指代消解技术的发展[3]。而同指消解主要是针对那些单独存在时能够明确其在现实中所指对象的语言单位，它的主要任务是确定这些语言单位是否指向现实中的同一概念或实体，以便进行概念的融合或关系的融合等。同指消解的发展始自上世纪90年代，1995年MUC-6专门成立了同指消解的评测机构，负责定义任务，准备语料，评测参加测试的系统的性能。指代消解与同指消解并不是毫不关联的两个任务。MUC-6所定义的同指消解任务中就包含了部分指代消解的任务，指代消解中的很多方法和理论都可以应用于同指消解。

R. Mitkov[1]曾对上世纪90年代之前的指代消解技术做过很好的综述，本文主要介绍近十几年以来指代消解及同指消解的一些新方法和发展趋势。本文第2节概述了消解过程中常用语言知识；第3节主要介绍近十几年典型的消解算法；第4节指出了消解技术新的发展方向；最后为结束语。

2 消解过程中的常用特征

最早的指代消解系统，如STUDENT(Bobrow 1964)，主要是通过人工书写有限数量的规则，将需要消解的片断与规则进行模式匹配，满足模式的段落即可进行消解[1]。但由于规则数目有限且限制性大，查全率不高；由于规则的更改和添加工作量大，方法的可扩展性不强。因此，到了上世纪70年代，人们逐渐摒弃了这种方法，而更多地转向将指代消解(包括后来的同指消解问题)作为一个最优化问题来解决，即先确定指代语集合，对每个指代语，先确定其先行语的候选集，再根据句法、语义特征从候选集中选择最有可能的候选项作为先行语。由于句子、句群作为先行语的情况过于复杂，很少有系统涉及，因此这里只讨论先行语为名词短语(包括代词、有定代词、专有名词等)的情况。下面将介绍一些常用的语言特征[1]。

2.1 距离

最初，先行语候选集确定为在指代语之前出现的所有名词短语。但距离指代语太远的候选项成为先行语的可能性几

*)本文受到国家自然科学基金项目(60473057,60573057,90604007)的资助。**李莎莎** 硕士研究生，主要研究方向为信息抽取与文本挖掘；**李舟军** 博士，教授，博士生导师，主要研究方向为进程代数、安全协议形式化验证、数据挖掘与生物信息学；**陈火旺** 院士，博士生导师，主要研究方向为软件理论与软件工程。

乎为零，考虑这么大的搜索空间显然是不必要的。但若为了减小搜索空间，简单地将先行语候选集定义为“在指代语之前，与之距离小于等于一个句子距离的名词短语的集合”也是不适合的。R. Mitkov 的研究表明，先行语最远可在指代语之前距其 17 个句子的距离，定义的搜索空间太小会降低消解的查全率。同时，不同类型的指代语与先行语之间的距离也可能不同，例如代词指代语与先行语之间的距离与有定描述(有定冠词修饰的名词短语，如“the company”)相比要小一些。Kameyama 于 1997 年提出的同指消解算法中就将这两种类型指代语的先行语候选集的产生窗口分开设置：有定描述定义为指代语前 10 个句子，代词定义为指代语前 3 个句子。

2.2 指代语类型

根据 2.1 中所述，不同的指代语类型将会影响候选集产生窗口的大小，因此有些算法，尤其是基于机器学习的方法，通常会将指代语的类型也作为一个特征。

2.3 同义关系

若某个候选先行语与指代语之间存在同义关系，则该候选先行语更有可能与指代语属于同一指代链。

2.4 同位语关系

例如，“Microsoft, the biggest software company, is founded by Bill Gates.”“Microsoft”与“the biggest software company”处于同位语的关系，指向同一对象。存在同位语关系的两个名词短语之间往往存在指代关系。

2.5 别称关系

例如，“John Smith is a lawyer. Mr. Smith lives in New York.”中“John Smith”和“Mr. Smith”指向同一个人，使用了不同的称呼。存在别称关系的两个名词短语之间往往存在指代关系。

2.6 缩略语关系

为了避免冗长，人们在表达时往往会在提到某个专有名词之后用缩略语指代它。例如“Information Extraction”的缩略语是“IR”，“研究生学术活动委员会”的缩略语是“研委会”等等。在英文中，往往选用专有名词中每个词首字母的大写组成该词的缩略语，而中文的缩略规则就相对复杂，很难描述其具体规律。存在缩略语关系的两个名词短语之间往往存在指代关系。

2.7 词义信息

有时可以根据词义信息得到一些常识性的知识。以下面的句子为例，“Give the bananas to the monkeys although they are not ripe, because they are very hungry.”根据词义，“monkeys”是动物，可能有“饿”的特征，因此第二个“they”指代“monkey”；“bananas”是水果，可能有“生，不熟”的特征，因此，第一个“they”指代“bananas”。而这些词义特征大多可以通过本体获取。

2.8 性别

若指代语存在性别特征，则先行语必须存在性别特征且与指代语相同。否则，若指代语不存在性别特征，先行语必不存在性别特征。

2.9 单复数

先行语通常应该与先行语的单复数特征一致，但存在一些特殊情况。以英语为例，集合名词作先行语，形如单数，但往往用复数指代(如 team)。

2.10 非代词的名词短语作宾语时不会与主语指向同一对象

例如，“Hemi told them about John.”中 Hemi 与 John 不能指同一个人。

2.11 反身代词做宾语指代主语

例如，“John likes pictures of himself”中 John 与 himself 指向同一个人。

2.12 人称代词做宾语不能与主语词指向同一对象

例如，“John told Bill about him”中 John 与 him 不指向同一个人。

2.13 语义信息

名词短语的语义类型多来源于本体及命名实体标注等信息抽取过程，不同的系统定义的语义类型不同。MUC 定义的语义类型包括人、组织、地点、行政区划、设施、交通工具、武器，还可以在这几类语义类型下划分更精细的子类型，以便更准确地刻画名词短语的语义特征。先行语候选集中的候选项与先行语必须属于同一语义类型。

2.14 句法平行

即前后两句句式相同的情况下，先行语候选集中与指代语处于同一句法位置上的候选项更有可能被确定为先行语。例如，

a) The programmer successfully combined Prolog with C, but he had combined it with Pascal last time.

b) The programmer successfully combined Prolog with C, but he had combined Pascal with it last time.

第一个句子中的 it 指代 Prolog，它与 Prolog 都是做 combine 的直接宾语，处在同一个语法位置；第二个句子中的 it 指代 C，它和 C 都是与 with 组成介词短语，处在同一个句法位置。

2.15 语义对应

即指代语更有可能指代与其扮演同样语义角色的候选先行语，这是一个比句法平行约束更强的条件。例如，

a) Vincent gave the diskette to Soddy. Kim also gave him a letter.

b) Vincent gave the diskette to Soddy. He also gave Kim a letter.

第一个句子 him 指代 Soddy，him 和 Soddy 都是“gave”的作用对象；第二个句子 He 指代 Vincent，Vincent 和 He 都是“gave”的主体。

2.16 名词短语在句子中的重要度

主要是指先行语候选集中在前文语句中处于较重要位置的候选项更有可能成为先行语。该约束最早是出现在 SHRDLU(Winograd 1972)系统中，该系统假定主语的重要度大于宾语，两者都要比介词宾语更重要。1983 年前后 (Center Theory)Grosz 和 Sidner 等人提出并发展了中心理论，进一步丰富了该约束。20 世纪 80 年代中后期以来，该理论一直受到了广泛的关注(详见文[1])。

以上这些语言特征通常分为两类：一类称为排除性(eliminating)特征，即凡是不满足该特征的条件的候选项均被删除，如性别等；另一类为推荐性特征(preferncial)，即满足该特征的条件的候选项成为先行语的可能性更大，如重要度特征等。并不是所有的系统都使用了上述所有的知识，例如经典的 Hobbss 算法就只使用了上面所列的 10～12 三个句法规则进行代词指代消解并取得了较好的效果。R. Mitkov 曾经就特征使用的多少、每种特征在不同语言中的重要程度、各种特征的相互影响以及各种特征的使用顺序等问题作过深入

的研究。并不是以上所有的语言信息都使用在一个消解系统中，消解的性能就会有所提高。一方面由于每种语言的不同特点，各种特征在进行消解过程中所起的作用不同，作用较小的特征甚至可能误导先行语的选择；另一方面由于语言特征获取的困难性与不准确性，以及描述语言特征之间交互关系的困难性使得添加一些复杂特征反而降低了算法的性能。

3 主要的消解算法

文[10]中研究了算法的重要性。它使用了两种方法：一种是先用排除性特征对候选集中的候选项进行筛选，再利用推荐性特征对候选项进行评分，选择分数最高的作为先行语；另一种方法是直接在初始候选集上应用不确定推理的方法确定先行语。结果表明，同样的特征集合并不一定能够得到同样的性能，算法的选择也至关重要。如前一节所述，消解算法通常可以分为两个步骤：指代语的确定和先行语的选取。在以下章节中我们将主要从这两个方面介绍消解算法的发展。

3.1 指代语的确定

指代语的难点主要在于非指代性的名词短语的识别。近些年的研究表明，用基于机器学习的方法进行篇章消解时，影响查准率的一个很重要的原因是算法消解了许多非指代性的名词短语。因此，为了提高消解的查准率，在确定指代语时识别并剔除非指代性的名词短语成为很多指代算法关注的重点。例如，1994 年，Lappin 和 Leass 提出的算法识别并剔除了作为形式主语、形式宾语等出现的代词“it”。2000 年，Vieira 和 Poesio 提出的算法消除了不定描述，即不定冠词修饰的普通名词。2002 年，Vincent Ng 和 Claire Cardie 提出的消解算法中引入了利用分类方法识别非指代性名词短语的方法，这种方法后来被称之为局部优化方法。2004 年，Vincent Ng 和 Claire Card 又提出一种全局优化的识别非指代性名词短语的方法，这种方法将识别指示性名词短语的模型参数化，通过调整选择能够使整个消解算法性能最优的参数。同时，Vincent Ng 和 Claire Card 还讨论了将指代性特征作为排除性特征还是作为推荐性特征的问题，最终得出的结论为：将指代性特征作为排除性特征，并选用全局最优的名词短语指代性确定的算法能够更好地提高消解算法的性能[6]。

在汉语中还存在着另一困难，即零形式指代语的发现。零形式指代是指句子中本应出现的语法单位却没有出现，主要形式是承前省略。2004 年，王厚峰总结出汉语中零形式指代语主要有三种类型：谓语动词的支配成分；主谓谓语句、名词谓语句、形容词谓语句等非动词谓语句中的主语；名词所需的配价成分。零形式指代所面临的根本困难是如何定义一个完整的句子以及如何排除标点符号的干扰。具体可参考文[13]。

3.2 先行语的选取

早期的指代消解算法多是用指代语和候选先行语的语言特征结合一定的规则对候选先行语进行筛选。上世纪 90 年代随着统计方法的发展，人们逐渐开始考虑如何利用统计方法提高消解的查准率和查全率(I. Dagan 和 A. Itai 1990，R. Mitkov 1996 等)。而机器学习方法的使用又使指代消解前进了一大步。1994 年，Connolly 等人将指代消解中先行语的选择问题转化为分类问题，提出了一种基于机器学习的方法。他们将分类问题建立在指代语与它的某两个候选先行语之上。用指代语与两个候选先行语的语言特征(包括三者的比较关系)生成一个特征向量，作为分类实例。将该分类实例输入到一个二值分类器以决定两个候选先行语中哪个更有可能。反复分类即可从先行语候选集中选出最优项。

Connolly 等人提出的算法所产生的特征向量的维数很大，而且分类次数多，计算量大。1998 年，W. M. Soon 等人提出一种新的消解系统框架[8]，降低了分类实例特征向量的维数且减少了分类次数。他们将分类问题建立在指代语与一个候选先行语之上，使用决策树分类器确定该候选先行语是否可能是指代语的先行语。距离指代语最近的可能候选项即被确定为先行语。后来，Vicent Ng 改进了该算法，选择可能性最高的候选项为先行语，进一步提高了算法的性能[7]。

1999 年，Clair Cardie 和 Kiri Wagstaff 提出一种非监督的新方法进行同指消解[4]。他们将名词短语的同指消解问题作为一个聚类问题来解决。聚类算法将指向现实中同一实体对象的名词短语聚为一类。与用分类算法解决同指消解问题相比，使用聚类方法最大的优点是不依赖大量已标注的训练数据。同时，Clair Cardie 和 Kiri 等人还认识到，不同的特征在指代消解和同指消解中所处的地位不同，算法通过对每个特征赋权重，区分了不同类型的特征，且达到了将指代消解和同指消解区分处理的目的。但对特征之间的依赖性挖掘不足，排除性约束的使用忽略了语言知识获取的不准确性，降低了系统的容错能力。

考虑到语言知识获取的不准确性会影响消解算法的性能，2001 年 Sanda M. Harabagiu 等人提出一个基于知识挖掘的消解系统(COCKTAIL)。他们考虑到，消解关系具有传递性，一些需要复杂的语言知识才能获取的指代关系往往可以通过一些只需少量知识即可获取的指代关系的传递来得到。因此，该系统根据指代关系的传递性先将语料库中所标示的指代关系扩展，然后通过分析数据发现通过一致性、别称、同义词等语言知识的运用，扩展出的关系中约有 83%可以被正确消解，余下未消解的关系中约有 30%可以通过语义一致性知识消解。系统结合 WordNet 及 bootstraping 技术获取语义一致性信息，进一步提高了消解的查准率和查全率。

4 未来发展趋势分析

消解技术是随着对自然语言理解要求的加深不断向更深入、更智能的方向发展的。随着多文本自动摘要、文本分类的发展，跨文本的同指消解成为消解领域的新的挑战；随着信息抽取技术在各个语种中的发展，跨语言可移植性也成为消解算法的必然要求；而自动性的提高是消解算法一直以来追求的目标。

4.1 跨文档的同指消解

跨文档的同指消解建立在单文档的同指消解基础之上，它的主要任务是判断不同文档中的描述对象是否指向现实世界中的同一实体。进一步讲，它主要包括两个方面，即不同文档中不同的表达是否指向现实世界中的同一实体、不同文档中的相同的表达是否指向现实世界中的不同实体。跨文本的同指消解对多文本自动摘要和多文本的信息融合至关重要，已成为 TIPSTER 文本项目第三阶段的首要任务。由于不同文档来自不同的数据库，出自于不同的作者，写作的习惯与方法不同，甚至写作的语言不同，跨文档的同指消解困难重重。

ISOQuest 的 NetOwl 系统和 IBM 的 Textract 系统能够完成发现不同文档中指向同一实体的不同表达的任务，却无法判断不同文档中同样的表达是否指向不同的实体。1998 年，Bagga 和 Baldwin 提出一种基于向量空间模型的算法，能

够在较小数据集上较好地解决区分不同文档中相同表达的问题，但在大数据集上的效果差强人意[11]。2000 年开始的 ACE 评测提出跨文档的同指消解任务 EDT(Entity Detection and Tracking)促进了该领域的发展，但目前还没有令人满意的系统出现。

4.2 跨语言的同指消解

跨语言的同指消解要求针对某种语言开发的消解系统能够不加修改或稍加修改地移植到其它语言上。自 1998 年 R. Mitkov 提出同指消解算法的跨语言可移植性问题之后，这个问题开始受到广泛关注。由于各种语言的特点不同，在不同的语言进行指代消解过程中存在不同的难点。以零形式指代为例，汉语中这种指代是常见现象且是一个处理的难点，而在英语中零形式指代就很少出现。因此，很难将针对英语开发的指代系统不加修改地移植到汉语上来。但不同语种之间，除了存在差别之外，还存在一些共同的特点，例如指代语与先行语之间必须满足性别一致性、数的一致性、语义一致性的要求等。充分利用不同语言之间的共同特征以提高同指消解算法的跨语言可移植性已成为共识。

1998 年，R. Mitkov 提出一个可跨语言移植的系统，该系统起初是针对英语文档开发的，后经过少量修改移植到波兰语和阿拉伯语上，结果该系统在三种语言上都达到了 90%左右的成功率[1]。2004 年，IBM 提出一个独立于语言的统计模型，应用于中文、英文及阿拉伯语三种语言。模型在应用于不同语言时，只需引入少量的该语言的特殊特征即可[12]。

4.3 减少消解算法的人工干预

目前，影响消解算法的自动化程度的主要是以下几个方面：消解所需语言知识的获取；消解所需的特征的选取；消解过程中所涉及语言特征的相互关系的描述；大量已标注的训练语料的准备。完全人工来完成这几个方面的工作，不但费时费力，还会引入不可预测的错误。因此，针对以上几个方面减少消解算法中的人工干预也成为研究的热点。

1997 年，Balwin 提出一个只包含 7 条启发式规则的系统——COGNIAC，该消解系统能够得到高达 90%的消解查准率。随后，基于少量知识的消解系统不断出现(如 W. M. Soon 1998，S. M. Harabagiu 2001)，成为消解算法的一个新趋势。1997 年，R. Mitkov 曾对参与消解的语言特征的选择以及使用方法作过详细的讨论，但缺乏对特征的相互关系的讨论。特征之间的关系本身是错综复杂的，具体地描述特征之间的关系显得可行性不高，近年来，消解领域一些“判别式”机器学习方法(如决策树方法)的引入一定程度上缓解了这个问题。同时，基于 bootstrapping 和基于聚类的消解方法也不再完全依赖于大量的已标注的训练语料库。

结束语 消解问题是篇章理解的一个关键问题，也是一个难点问题。消解问题的解决对信息检索、文本自动摘要、文本自动分类等问题的研究都大有裨益。本文总结了消解算法中常用的语言知识，介绍了近 10 年来主流的消解模型，最后对篇章消解的发展趋势作了展望。自然语言理解的深度要求不断提高，要求篇章消解的查准率和查全率不断提高，以便更好地进行信息融合。同时，减少人工干预，提高消解算法的可移植性、自动性也成为算法研究的趋势。

参 考 文 献

1 Mitkov R. Amphora resolution: the state of the art. Working paper(Based on the(X)LING'98/ACL'98 tutorial on anaphora resolution), 1999

2 Harabagiu S M, Bunescu R C, Maiorano S J. Text and Knowledge Mining for Coreference Resolution. In: Proc. of NAACL2001, Carnegie Mellon University, USA, June 2001

3 王厚峰．指代消解的基本方法和实现技术．中文信息报，2001，16(6)

4 Claire C, Wagstaff K. Noun phrase coreference as clustering. In: Proc. of the 1999 Joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpora (ACL'99). University of Maryland, USA, 1999

5 钱伟，等．基于最大熵模型的英文名词短语指代消解．计算机研究与发展，2003，40(9)

6 Ng V. Learning Noun Phrase Anaphoricity to Improve Coreference Resolution: Issues in Representation and Optimization. In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL), Barcelona, Spain, July 2004

7 Ng V, Cardie C. Improving Machine Learning Approaches to Coreference Resolution. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002

8 Meng S W, Ng N T, Yong L C. Corpus-based Learning for Noun Phrase Coreference Resolution. In: Proc. of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-99)(285-291). College Park, Maryland, USA, 1998

9 Meng S W, Tou N H, Yong L D C. A Machine Learning Approach to Coreference Resolution of Noun Phrases. Computational Linguistics (Special Issue on Computational Anaphora Resolution), 2001, 27(4): 521～544

10 Mitkov R. Factors in anaphora resolution: They are not the only things that matter. A case study based on two different approaches. In: Proceedings of the ACL'97/EACL'97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution. Madrid, Spain, 1997

11 Gooi C H, Allan J. Cross-document Coreference on a Large Scale Corpus. In: Proc. Human Language Technology/North American Chapter of Association for Computational Linguistics Annual Meeting (HLT/NAACL). Boston, USA, May 2004

12 Florian R, et al. A Statistical Model for Multilingual Entity Detection and Tracking. In: HLT/NAACL 2004: Human Language Technology Conference / Annual meeting of the North American Chapter of the Association for Computational Linguistics, Boston, Mass, 2004

13 王厚峰．汉语篇章的指代消解浅论．语言文字应用，2004(4)：113～119

14 Baldwin B. CogNIAC: High precision coreference with limited knowledge and linguistic resources. In: Proceedings of the ACL Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts, 1997

15 Filippova K. A Memory-based Learning Approach to Pronominal Anaphora Resolution in German Newspaper Texts: [Master's Thesis in Computational Linguistics]. Oct. 2005

16 Niu Cheng, Li Wei, Srihari R K. Weakly Supervised Learning for Cross-document Person Name Disambiguation Supported by Information Extraction. In: Proceedings of ACL, 2004

17 McCarthy J F, Lehnert W. Using Decision Trees for Coreference Resolution. In: Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, 1995