篇章共指消解研究综述*

Intra-document Coreference Resolution: The state of the art

Jun Lang, Bing Qin, Ting Liu, Sheng Li

Information Retrieval Lab, Harbin Institute of Technology, Harbin, China
{bill_lang, qinb, tliu, lis}@ir.hit.edu.cn

Abstract

Intra-document coreference resolution, which is the process of partitioning intra-document mentions into different real world entities, includes personal pronoun resolution and noun phrase resolution, and is always one of the key problems of natural language processing. In recent years, there have been extensive studies and rapid progresses in coreference resolution. After distinguishing several related concepts and summarizing the basic coreference types, this paper analyses the past 30 years research line on coreference resolution. It is pointed out that the current key problems and research trends are global optimization, the use of deep linguistical and background knowledge, and the combination of linguistical models and statistical models. Finally, the related international evaluations, evaluation metrics, related corpora and tools are described in detail.

Key words

Intra-document coreference resolution; personal pronoun resolution; noun phrase resolution; natural language processing

摘　要

篇章共指消解就是将篇章内的所有表述划分为现实世界中不同实体等价描述的过程，主要包含人称代词消解和名词短语消解。该问题一直是自然语言处理中的核心问题，近年来得到的广泛关注和快速的发展。本文区分了几个容易混淆的相关概念，总结了共指的基本类型，分析了近三十年来共指消解的研究路线，认为全局优化、深层语言学知识以及背景知识的利用、语言学模型和统计模型的融合是当前的关键问题和研究趋势，并详细介绍了共指消解相关的国际评测、评测方法、语料资源和工具。

关键词

篇章共指消解;人称代词消解;名词短语消解;自然语言处理

近年来，随着互联网上信息的爆炸式增长，每天出现的新信息大大超过了人类的处理能力。在自然语言处理(Natural Language Processing)、信息检索(Information Retrieval)、数据库(Database)、计算机视觉(Computer Vision)等诸多领域中，现实世界

*本研究得到中国国家自然科学基金项目（60575042，60503072）资助

中的同一事物经常会有不同的名称、描述以及视角。将它们正确对应到具体的事物，对于数据的后续处理和深入理解是非常必要的。在计算机视觉领域，对物体的计数和跟踪之前必须解决物体的一致性判定。在数据库中，通过记录链接(Record Linkage)或者去重(De-duplication)创建一个干净的数据集可以使得后续挖掘更加精确。在自然语言处理中，对指向同一实体的名词、代词、以及普通名词短语进行消解，可以使后续的实体关系抽取更加完善。和含有结构化字符信息的数据库相比，非结构化的自然语言文本篇章中对同一实体的消解更加困难。本文主要讨论自然语言处理中的情形。

例如，在讨论中国、美国、日本等大国间贸易的文章中，开篇可能会写“中华人民共和国”，后面可能会说“中国”、“大中国”等，还会提到“这个国家”、“她”等。这些表述都是现实世界中“中华人民共和国”的不同体现。虽然人们可以毫无困难的区分文章中同一实体的不同体现，但对计算机而言，仍是非常困难的。所谓篇章共指消解(Intra-document Coreference Resolution)就是根据一篇文档中各个表述(Mention)的自身内容以及所在上下文来确定不同实体(Entity)的数量，以及确定各个实体分别包含哪些等价的表述。在某种意义上说，共指在自然语言中起到了超链接的作用。一方面，它使得作者在撰写文章时可以体现一定的风格并实现篇章的连贯性。另一方面，共指使得自然语言处理机制中增加了一种新的模糊成分。

共指消解是自然语言处理中最难的问题之一，因为共指消解不仅需要语言学方面的知识，例如浅层的词汇、句法知识，还需要较为宏观的语义和篇章知识。最为困难的是，很多时候共指消解需要丰富的背景知识才能完成，如图 1 所示。从这一点上来说，共指消解也是人工智能中的一个难题。

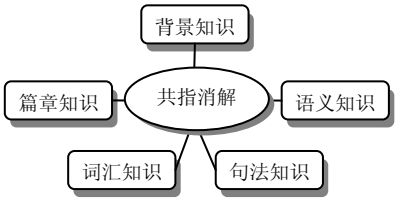

Fig.1 Knowledge requirement for coreference resolution

图 1 共指消解需要的知识

共指消解是传统的研究方向，见著于二十世纪三十年代，经过起初的蓬勃发展，于七十年代达到高潮，经历八十年代的低谷后，重新在九十年代初复兴(Mitkov, 2001)。近 20 年来，这方面的研究受到了格外的关注，ACL、COLING、EMNLP、EACL、NAACL 等重要的国际会议都召开过共指消解的专题会议，先后出现了 MUC(Message Understanding Conference)、ACE(Automatic Content Extraction)、ARE(Anaphora Resolution Exercise)等共指消解相关的国际评测。Computational Linguistics 期刊 2001 年出版了共指消解的专辑，随后几乎每年都有相关的文章发表。近两年来，越来越多的共指消解相关论文、会议以及国际评测，展现出了共指消解研究的繁荣景象。值得一提的是中文的共指消解研究开始于二十世纪末，中文共指消解的评测开始于 2003 年 10 月的 ACE Phase3。

共指消解研究早期的方法融合了大量的领域知识和语言学知识(Hobbs, 1978)，目前较新的方法基于更加强大的自动分析器和统计学习理论。本文概括了共指消解近三十年来的研究情况，分析了最新的三项趋势，概述了相关的国际评测，列举了相关系统及源代码。特别的，本文还针对中文上共指消解的研究进行了总结。

后续章节的内容安排如下：首先讨论共指消解的一些相关概念和研究意义；然后

概述国际上共指消解研究的现状以及最新的发展趋势；第 3 节总结了中文共指消解的研究情况；第 4 节总结了共指消解研究的三项主要的国际评测、四种重要的评测方法以及基于评测的语料库特征发现；第 5 节对共指消解相关的资源、系统等进行了列举；最后进行了结论和展望。

1 相关概念和研究意义

1.1 共指、指代、回指、预指、代词、名词短语六种消解的概念及关系

在各种自然语言处理相关的文献中经常会看到一些和共指消解相关的概念，主要有指代消解、回指消解、预指消解、人称代词消解、名词短语消解等。

下面先介绍两个基本概念。照应语(Anaphor)是消解过程中当前考察的用于指向的表述对象。先行语(Antecedent)是被指向的表述对象。以上几个概念主要的区别就在于照应语、先行语的类型以及二者之间的位置关系。

指代消解也叫参照消解(Reference Resolution)，就是确定参照表达式(Referring Expression)之间相互关系的过程。此时，照应语和先行语都统一叫做参照表达式。指代按照指示语和先行语的先后关系可以分为预指(Cataphora)和回指(Anaphora)。预指是照应语出现在先行语之前。例如，“想讨好[他]父亲的人，争先为[小张]开门。”回指是指照应语出现在先行语后面。例如，“[小张]很聪明。[他]总是能很快算出答案。”相对回指而言，预指比较少见。

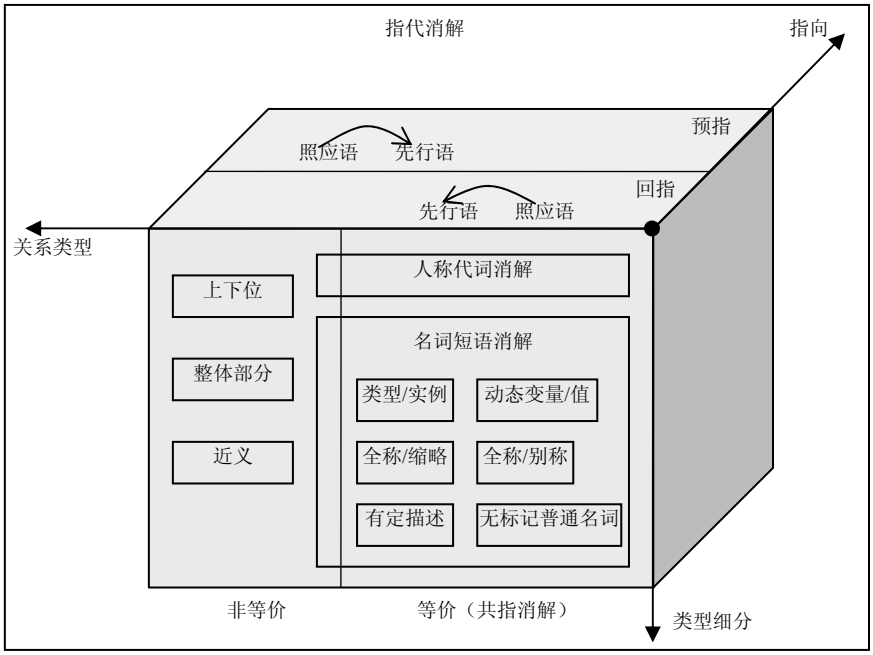

Fig.2 Three dimension concepts relationship related to coreference resolution

图 2 共指消解相关概念的三维关系

指代消解按照照应语和先行语之间的等价性可以分为等价指代和非等价指代。前者就是共指消解(Coreference Resolution)，后者主要包括照应语和先行语之间的上下位(Hyper-/Hypo-)、整体部分(Whole-part)、近义(Synonym)等关系。而共指消解又分为人

称代词消解(Personal Pronoun Resolution)和名词短语消解(Noun Phrase Resolution)。当然，单独提到人称代词消解和名词短语消解的时候也会考虑非等价关系。本文主要讨论共指消解，后面提到的人称代词消解、名词短语消解都只考虑等价关系类型的。共指消解的各种具体类型见下节。图 2 展示了各种概念之间的三维关系。三个坐标轴分别表示按照不同的分类方法来确定指代消解的研究范畴。在“指向”坐标轴上分为预指和回指两类，在“关系类型”上分为“非等价”和“等价”，等价关系就是本文主要讨论的共指消解问题。再按照类型细分，等价关系又可以分为两大基本类型，具体描述见下节。

1.2 共指的基本类型

总结共指消解，可以分为两大类：人称代词消解和名词短语消解。考察如下的几个例子。

a) [张国华]对人很热情，大家都叫[他$_1$][张三]。[张三]是[老师]，[他$_2$]讲课非常认真，同时，[他$_3$]也是[一个好爸爸]。
b) 花丛中跑出[一只小花猫]。['它]非常可爱。[这只猫]是老王家的。
c) [现在的气温]是[30 摄氏度]。
d) 随着重庆升级为直辖市，[河南]成为[中国人口最多的省]。
e) 两会闭幕后，今年下半年将召开[中国共产党第十七届全国代表大会]。目前，[十七大]的各项准备工作正有条不紊地展开。
f) [中国矿业集团有限公司]的领导大胆创新, 有效遏制住了经济滑坡，[公司]产值以平均每年 33 %的幅度递增。

人称代词一般都是在一个局部的上下文中对前面最近出现的事物进行替代。人称代词消解主要考察人称代词和近邻名词短语之间的共指关系。例如a)中的“张国华”/“他$_1$”，“张三”/“他$_2$”，“张三”/“他$_3$”；b)中的“一只小花猫”/“它”。

名词短语消解主要考查名词短语之间的共指关系，可以分为以下六种情况。

[1]. 类型/实例(Type/Instance)。这里的“实例”是指一种“类型”中的一个具体例子。例如 a)中的“老师”/“张三”，“一个好爸爸”/“张三”。
[2]. 动态变量/值(Dynamic Variable/Value)。这里的“值”是指在局部上下文限定的情况或者环境下“动态变量”能取得的唯一结果。例如 c)中“现在的气温”/“30 摄氏度”，d)中“中国人口最多的省”/“河南”。
[3]. 全称/缩略(Full Name/Acronym)。例如 e)中的“中国共产党第十七届全国代表大会”/“十七大”。英文的缩略一般是将短语中各个单词的首字母抽取出来组合而成，中文的缩略规律性不强。
[4]. 全称/别称(Full Name/Alias)。例如 a)中的“张国华”/“张三”。
[5]. 有定描述(Definite Description)(王厚峰, 2004)。有定描述有两种基本形式：其一，代词（主要是指示代词）引导的名词短语，如 b)中的“一只小花猫”/“这只猫”，英语中的这种描述常用定冠词“the”引导。其二，专有名词或者专有名词的一部分引导的名词短语，如“张三”/“张三老师”，“张三”/“张老师”。
[6]. 无标记的普通名词(Common Noun) (王厚峰, 2004)。普通名词或者名词短语也可以直接作为照应语，不带有任何引导信息或标记。如 f)中的“中国矿业集团有限公司”/“公司”。

1.3 研究意义

共指消解一直是自然语言理解中的核心问题，在机器翻译(Machine Translation)、信息抽取(Information Extraction)、自动文摘(Automatic Summarization)以及自动问答(Question Answering)等领域中都有重要应用。全自动的共指消解是计算机对自然语言理解的一项艰难任务。这方面的专门研究在国外已经进行数十年，但在国内才刚起步不久。

共指消解对于机器翻译的成功与否有关键作用，尤其是当把源语言翻译成有代词的性别标识的目标语时，解决好代词的共指先行语是最为重要的。刘礼进(2005)指出：二十世纪七八十年代开发的多数机器翻译系统，没有完全解决好源语言中共指先行语的识别和目标语言中共指成分的“对等先行语”的生成问题，以致机器翻译系统在篇章翻译方面很受限制。共指消解的引入，使得机器翻译软件工程的状况得以改观。对于一些共指消解中的专有名词短语的翻译也会存在类似的问题。例如篇章中出现的“十七大”等缩略型短语，如果不能在篇章中通过共指消解得知这是“中国共产党第十七届全国代表大会”的缩略，那么几乎不能翻译正确。

共指消解一直是信息抽取的子任务之一。在信息抽取中，共指消解可以合并或者链接篇章中同一实体的不同表述、实体之间的关系以及在不同篇章中描述的事件实体。在 MUC 的第 6 和第 7 届评测中，引入共指消解后信息抽取的效果有明显改善，产生了很大的影响力。ACE 的历次评测中都包含了共指消解。

自动文摘的研究人员对共指消解的研究兴趣越来越浓。目前主流的自动文摘技术是基于句子摘取的。常见的策略是通过不同的方法对篇章中的各个句子计算其重要度，然后通过考察重要度来选择需要的句子。如果在计算句子重要度之前完成共指消解，就能将篇章中同一实体的各种表述归一化，从而进行更加合理的句子重要度计算(Witte and Bergler, 2003; Witte, et al., 2006)。在将句子抓取出来形成文摘后，如果不进行任何处理，就可能会在篇章开始或者一个意义段开始处看到一个让读者感到不知所云的人称代词。如果之前进行了共指消解就能将这种情况下出现的人称代词替换为对应的实体名称，从而让人更容易理解生成的文摘。

还有研究表明，共指消解有助于自动问答。例如在 Morton(1999)提出的自动问答系统中，就是依靠建立问题中所述实体或者事件的共指链接来获取问题的答案。具体做法是将被检索文件的语句按共指关联排成等级顺序，级别最高的句子作为答案展示给用户。

2 研究现状和最新趋势

这个部分主要概述共指消解研究的现状以及最新趋势。共指消解研究的研究可以分为三个阶段：（1）1978 年～1995 年，以句法分析为基础的基于语言学方法的共指消解，代表方法是 Hobbs 算法以及中心理论；（2）1995 年～2002 年，这段时间主要是各种基于二元对的分类方法以及基于向量相似度的聚类方法；（3）2002 年至今，经过上一个阶段的发展，越来越多的研究人员意识到经典的二元分类框架存在各种问题，开始提出摆脱二元分类框架的各种篇章全局优化的方法，另一个趋势是考虑如何引入背景知识以及语义知识。下面分三个部分来分别介绍三个阶段的研究工作。

2.1 基于语言学方法的共指消解

这个部分总结和讨论了基于语言学的共指消解方法，包括 Hobbs 算法、中心理论(Centering Theory)以及一些基于中心理论的方法。

2.1.1 Hobbs 算法

Hobbs(1978)算法是最早的代词消解算法之一。该算法主要基于句法分析树进行相关搜索，包含两种算法：一种方法是完全基于句法知识的，也称朴素 Hobbs(Naïve Hobbs)算法，另一种既考虑句法知识又考虑语义知识。该算法不仅是一个具体的算法，同时更是一个理论模型，其中提到的约束方法对后来代词消解的研究起到了非常重要的指导意义。到现在为止，Hobbs 算法仍然是一个非常有效的算法。朴素 Hobbs 算法的流程及示例可以参见(Hobbs, 1978; 王厚峰, 2002)。

本质上，Hobbs 算法倾向于选择同一个句子中的实体，并且倾向于选择离代词近的实体。根据代词在句子中的位置，句子中不同实体的相关程度也不一样。在搜索前一个句子时，出现在主语位置的实体更加重要，因为搜索方式是从 S 节点开始在树中从左到右先广搜索。句法树中节点的深度在确定实体重要性时也是非常重要的因素。

诸多的 Hobbs 算法的研究都是在英文上进行，Converse(2006)首次将 Hobbs 算法应用在中文代词消解上，在 ACE2004 中 Chinese Pen Treebank 部分的共指关系标注语料的基础上，采用了三种模型来应用 Hobbs 算法。第一种方法，采用朴素 Hobbs，利用句法树来选择代词的先行语；第二种方法加入了性别和单复数等约束信息；第三种方法在候选先行语上加入了语义约束。

2.1.2 基于中心理论的代词消解

中心理论主要针对“在篇章结构中注意焦点、指代表达式选择、以及话语一致性等关系”提出的(Grosz, et al., 1995)。中心理论的一个主要目标就是在给定的句子中跟踪实体的焦点变化。Sidner(1981)中有这方面更加详细的早期工作介绍，其中详细分析了直接焦点以及应用直接焦点来消解人称代词和指示代词的算法和规则。

值得一提的是，中心理论的提出不是为了解决代词消解问题。它主要是提供了一个预测下一个句子焦点的模型。但由于代词指向的就是焦点实体，因而中心理论一直被应用于代词消解算法。扩展用于代词消解的中心理论还需要解决单复数识别、数量名词短语和其他不确定的语义约束信息。

Brennan, et al.(1987)提出了一种基于中心理论消解代词的 BFP 算法。这个想法能够用来寻找给定句子中代词指向的实体。BFP 算法中一个存在的问题是，必须考虑那些不是指代语的名词短语。这种不必考虑的情形会花费额外的语法信息和领域知识。

Left-Right Centering(LRC)是基于 BFP 算法和中心理论框架的算法(Tetreault, 1999)，主要解决了 BFP 中不能迭代式的消解代词以及过度构造候选对的问题。

Strube(1998)提出了一个不考虑回指中心的代词消解算法。算法中一直维护一个称为 S-List 的列表。代词消解时将列表中实体根据一些特定绑定约束、一致性检测等进行排序，排序最高的就是代词消解结果。这种方法允许增量式的进行代词消解，这更符合人们对代词的解释。

到目前为止，虽然有很多基于中心理论的代词/指代消解算法，原始的中心理论近来才被实验验证。Poesio, et al.(2004a)采用了一种多参数的方法来实际检验中心理论。他们指出，在进行回指中心代词消解优选性考察时，回指中心的唯一性约束是更加值得考虑的因素。在进行搜索时，参数空间是很大的。因为原始的中心理论在很多细节上没有明确，例如实体如何排序、什么是话语、如何计算上一个话语。这些细节都是和语言相关的因素，必须针对具体的语言来进行设定。

2.1.3 基于浅层自然语言处理的方法

前面介绍的 Hobbs 算法和中心理论方法都趋于理论性，二十世纪九十年代开始共

指消解的研究人员开始意识到共指消解的高度复杂性，开始了更加切合实际的研究。比如，经常把共指消解限定在一个单一的特定语境、语言知识或语域之内，获得了实际的应用、有效的方法和经验，涌现了一批标志性的成果。例如，Lappin and Leass(1994)提出的 RAP(Resolution of Anaphora Procedure)算法，Mitkov(1998)提出的“有限知识”的指代消解方法，王厚峰和梅铮(2005)提出的中文上的鲁棒性人称代词消解。

这些系统的一个共同特点都是借助性能日趋强大的自然语言底层处理工具，例如词性标注器、浅层句法分析器等针对代词进行消解。首先采用一些过滤和筛选规则，获得候选先行语，然后结合各种处理结果对各个候选先行语进行各种特征的加权。相关特征包含人称类型、性别、单复数、句法角色等。权值的设定根据具体的特征类型设定。最后根据各个候选先行语获得的加权得分来选取最好的候选先行语作为代词最后的消解结果。

2.2 基于机器学习方法的共指消解

在这一节中，展示了许多基于机器学习的共指消解方法。我们将那些通过学习算法在训练语料上获得相关知识的方法都称为基于机器学习的方法。机器学习方法应用到共指消解问题中兴起于 1995 年。下面主要介绍用于共指消解的分类以及聚类方法。

2.2.1 基于分类的方法

随着 McCarthy and Lehnert(1995)首次将共指消解问题视为二元分类并采用决策树(Decision Trees)C4.5 算法以来，共指消解开始在二元分类的框架下获得了长足的发展。总结相关系统和论文，基于二元分类的经典框架如图 3 所示。

图中①表示共指消解处理的对象。一般而言，共指消解系统的输入是预处理中获得的各种实体表述(Mention)。相关的预处理主要包括文本断句、词性标注、命名实体识别、嵌套名词短语识别等。针对中文等没有空格分隔的语言还需要在文本断句之后进行分词处理。这些前处理一般采用一些相关的模块来获得。共指消解的国际评测中，为了更加精准的评测共指消解算法的性能，组办方一般都会提供标注好 Mention 的语料。

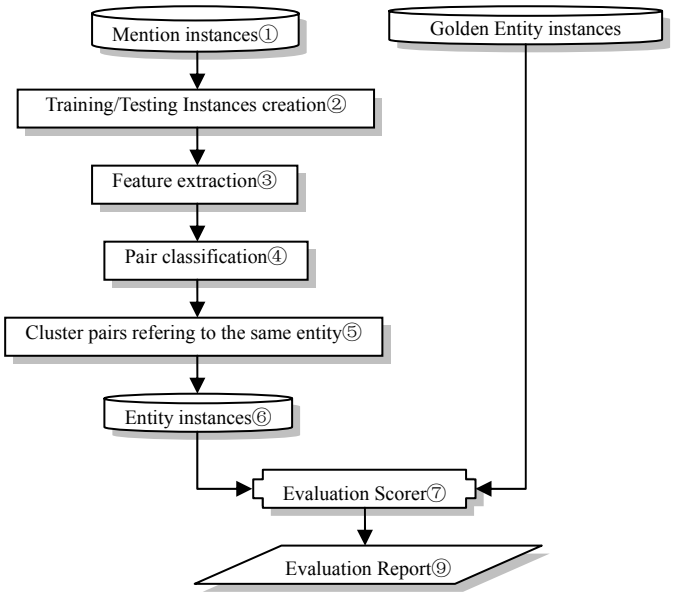

Fig.3 The classical framework of pair classification for coreference resolution
图 3 基于二元分类的共指消解经典框架

②表示从训练语料或者测试语料中构建用于分类器的输入实例。在训练语料上构建二元分类的训练实例时需要考虑如何构建正例和反例，到目前为止常见的构建训练实例的方法有三种。第一种是 McCarthy and Lehnert(1995)的方法，将篇章中任何两个不在同一共指链中的 Mention 构成反例，任何两个位于同一共指链的 Mention 构成正例。由于这种方法产生的训练实例数量巨大，而且反例数量远远大于正例数量，会严重影响机器学习算法的效率同时产生严重的不平衡训练，因此后来很少被采用。随后 Soon, et al.(2001)采用的方法产生的训练实例数量比第一种方法少得多。他们将篇章中各个共指链上任何两个最近的 Mention i、j 构成正例，其间的每个 Mention 都和 j 构成反例。这种构造训练实例的方法较多的考虑了共指的局部性。第三种是 Ng and Cardie(2002)采用的方法。和 Soon 方法不同在于创建正例的时候对于一个共指链中的任一 Mention j，如果 j 是代词则考虑同一共指链中 j 之前最近的 Mention 构成正例；如果 j 不是代词则考虑同一共指链中 j 之前最近的不是代词的 Mention 构成正例。创建反例的方法和 Soon 的一样。比较而言，第三种方法创建的实例最为合理。

在测试阶段创建测试实例时，一般是将篇章中任何两个 Mention 构成实例用于测试。在测试过程中可能会设置一些优先条件进行选择性的过滤。

③表示特征抽取。事实上，在二元分类框架下，如何设计需要选定的特征，对于最终的共指消解性能具有决定性的影响。自然语言处理是一个强不适定问题(Stringly Ill-posed Problems)，只有通过提供大丰富的“约束”（包括知识，经验等），才能使之成为适定性的、可解的问题(张钹, 2007)。共指消解也需要采用大量的约束才能解决，而对于二元分类等具体的框架，添加约束的方法就是采用更多合理的特征。共指消解需要考虑的特征主要分为以下几类：词汇、距离、一致性、语法、语义等。词汇特征主要考虑两个 Mention 的字符串的匹配程度，一般而言字符串相同程度越高的 Mention 共指概率越大。距离特征主要考察两个 Mention 的句子距离，这个主要依据是共指事实上也是一种局部性的替代关系，越是临近的 Mention 之间共指概率越大。一般而言，两个 Mention 相隔超过三个句子，共指的可能性就会很小了。一致性特征详细可以分为性别、单复数、语义类别等是否一致。这组特征主要起到筛选的作用。语法关系用来判断两个 Mention 的语法角色之间的关系，由于对句子深层的语法分析还很难办到，这里主要采用的是一些基于特定模板的方法，例如判断两个 Mention 之间是否被逗号格开或者相邻等来决定是否具有同位关系。语义特征主要是考察两个 Mention 在语义类别不一致时是否满足上下位或者同义、近义关系。这种判断主要依赖于具体的语言学词典，例如英文上的 WordNet(Fellbaum, 1998)、中文上的 HowNet(董振东,董强, 2001)等。

各种特征的使用经历了一个由少到多，再由多到精的过程。最早的 McCarthy and Lehnert(1995)的方法中采用了 8 个最为简单的特征。Soon, et al.(2001)采用了 12 个特征，主要都是一些在 Mention 的字符串本身就能处理获得的。Ng and Cardie(2002a)在 Soon, et al.(2001)的基础上又加入了 41 个更为复杂的语法和语义特征，但是特征的增加并没有大幅度的提高系统的性能。事实上，当系统的语料规模受限时，并不是选用的特征越多得到的效果越好。对于机器学习方法这个较为显然。因为语料受限决定了训练实例受限，这时如果特征越多特征空间中的各种相关参数训练就越不充分，从而会导致出现数据稀疏并最终导致实验结果中封闭测试性能较好，但表示机器学习模型泛化能力的开放测试性能较差。事实上，Hoste and Daelemans(2005)证明了特征选择

对于共指消解是有用的，选取出最为有效的特征并完成训练会在训练速度以及模型泛化能力上得到明显的改观。

图中④表示二元分类的机器学习算法。到目前为止，用于共指消解二元分类的机器学习方法主要有贝叶斯(Naïve Bayes)(Ge, et al., 1998)、决策树(McCarthy and Lehnert, 1995; Soon, et al., 2001; Ng and Cardie, 2002)、最大熵(Maximum Entropy)(钱伟,等,2003; Luo, 2004)、条件随机域(CRF, Conditional Random Field)(McCallum and Wellner, 2004)、遗传算法(GA, Genetic Algorithm)(Byron and Allen, 1999; 杨佳,罗振声,2005)、互训练(Co-Training)(Müller, et al., 2002; Ng and Cardie, 2003)等。这些方法的一个共同点是都在各种相关特征构成的特征向量的基础上训练得到各种特征的权值或者优选性（主要是决策树能得到优选性），只是各自在使用时采用的相关特征不尽相同。

图中⑤表示 Mention 二元分类的结果合并为 Entity。合并方法主要分为三类：最近合并是指在当前指代语前面符合共指条件的候选先行词中选取最近的一个作为先行语(Soon, et al., 2001; Strube, 2002)；最优合并是将当前指代语前面满足共指约束的候选先行语列表中选取共指概率最大的一个作为先行语(Ng and Cardie, 2002; Iida, et al., 2003)；最大化合并就是将指代语前面满足共指条件的所有候选先行语合并起来一起作为先行语(McCarthy and Lehnert, 1995)。经过合并后就得到了最终用于共指消解评测的实体（图中⑥所示）。

随后进行的就是共指消解结果的评价（图中⑦所示）以及得到最终的实验结果（图中⑧所示）。具体的评价方法见后面的第 4 节。

2.2.2 基于聚类的方法

分类的经典框架中由于采用了有指导的机器学习方法，不可避免的需要人标记好的训练语料。但是在共指消解领域，标注语料的工作相对于其他底层的自然语言处理任务（例如分词、词性标注、命名实体等）困难得多。有人采用了不需要训练语料的无指导方法来进行共指消解研究。

Cardie and Wagstaf(1999)采用特征向量来表示各个名词短语，然后在各个特征向量上采用聚类算法来实现名词短语的共指消解。聚类过程中采用凝聚式方法，每次选择两个最适合合并的类来进行合并。这种方法可以很好的避免类似于“Mr. Powell”被放入已经存在“She”的类中，从而避免不一致问题。但是这种方法并不是完全无指导的，因为其中的距离函数以及加权方法都由启发式方法确定。

Wagstaff(2002)针对名词短语的共指消解提出了一种修正版的聚类方法，称为约束聚类。在算法中规定了一些“不能链接”和“必须链接”的约束。“不能链接”规定了一些名词短语对不能位于同一组中，“必须链接”规定了一些名词短语对必须位于同一组中。在他们的实验中绝大多数都是“不能链接”类型的，主要实现了一些语言学约束，例如性别、单复数、语义类别一致、篇章等。需要指出的是，不是所有的约束都非常有效。例如，单复数约束采用的是非常简单的启发式规则，比如考察是否是“-s”或者“-es”结尾。同时，在这些约束中并没有清晰的指出约束之间的涵盖关系，所以采用了一种层次模型来实现配置各种约束之间的优先级。

Finley and Joachims(2005)描述了一种有指导的聚类方法。这种算法先学习一种用于聚类的相似度函数。和二元分类类似，相似度函数主要用于判断两个名词短语是否具有共指关系。训练数据中生成的正反例实例对非常不平衡（得到的具有共指关系的二元对只有 1.6%）。但是，Ng and Cardie(2002b)对一个给定的名词短语只使用其最近的共指短语作为正例，最近共指短语和当前名词短语之间的短语和当前名词短语都构成反例。更进一步，关系聚类的方法还能考虑传递性的依赖关系(Bansal, et al., 2004)，

和普通聚类方法的主要区别就是目标函数从最大化距离变成了关系聚类中采用的那种目标函数。Finley and Joachims(2005)存在的一个问题是这种形式下约束的数量会随着待聚类实体的数量增加呈现比指数递增还快的复杂度。适当的优化这种目标函数是一个“NP-完全问题”，因此大都采用近似算法。一个优点是这种方法能够处理传递性依赖，但是如果语料中传递依赖不是很多的话，这种方法只是能够和经典的二元分类算法差不多。

2.3 共指消解研究当前的发展趋势

共指消解的研究对于自然语言理解具有重要的作用。以往的研究框架主要有两种。一种是基于语言学处理工具的基于规则的方法，一种是抽取指代语和候选先行语上下文特征后采用各种机器学习模型来选取最优先行语。经过分析对比目前的各种共指消解方法和系统，我们发现共指消解的研究主要存在以下几种发展趋势。

2.3.1 篇章全局优化技术

目前主流的共指消解算法采用的框架多数都是分为两步的。首先采用经典的二元分类模型计算文档中实体两两之间的共指概率，然后采用不同的方法来完成共指链的合并。多数的共指消解系统都集中在判断篇章中两个实体表述是否指向同一实体，将这个问题看成是二元对共指关系的分类问题，从而在实体描述对上采用判别模型(Discriminative Model)来综合利用各种相关特征，例如距离、浅层句法分析等(Soon, et al., 2001; Ng and Cardie, 2002a)。

虽然这些方法取得了很大的成功，但是他们存在两个主要的问题。

第一，指代语的识别潜在的融合于二元共指决策中。如果二元分类决策输出的分值高于设定阈值，就判定这个二元对具有指向关系，其中的实体描述也就具有指代语功能，反之如果低于阈值就判定不具有指向关系，其中的实体描述就不具有指代语功能。这样就会产生两个问题：（1）系统错误的将一个先行语指向了一个非指代语的实体描述；（2）系统不对一个指代语的实体描述进行共指消解。为了解决这个问题，Ng and Cardie(2002b)、Vincent(2004)、Versley(2007)以及 Luo(2007)等在共指消解决策之前都采用了一个单独的指代语过滤分类器，用于判断一个实体描述是否应该将一些实体描述作为其先行语。这种方法提高了共指消解系统的性能。但是由于采用了两个级联分类器，系统中需要对指代语过滤分类器的阈值进行仔细的设定，才能避免过滤掉过多的实体描述。

第二，共指消解本质上是一个聚类或者实例划分为不同等价类的问题。共指关系是一个等价关系，满足传递性，自然会出现如果 *a* 和 *b* 具有共指关系，*b* 和 *c* 具有共指关系，那么 *a* 和 *c* 也具有共指关系。因此二元分类方法很容易产生不一致的共指等价类，因为一种经典的合并方式是采用贪心算法在篇章中从左到右的进行实体描述对合并。例如可能会在二元分类时将“Mr. Powell”和“Powell”，“Powell”和“She”分别判断为具有共指关系，但是合并等价类时就出现了“Mr. Powell”和“She”在一起的矛盾现象。

Yang, et al.(2003)提出的基于竞争的双候选模型(Twin-candidate Model)可以看作是一种对纯粹的局部上下文的脱离。对比经典的二元分类考察指代语和一个候选先行语是否具有共指关系，他们采用的是指代语和两个候选先行语，一个和指代语具有共指关系（即正例），另外一个和指代语不具有共指关系（即反例）。想法的来由是单独的候选先行语不能够有效的用于学习问题，双候选模型能够帮助学习针对一个指代语正例和反例的区别。但是，他们没有清晰的指出为什么设计成正例－反例双候选对，

以及为什么采用决策树模型。

C. Nicolae and G. Nicolae(2006)的 BestCut 方法和 Denis and Baldridge(2007)的整数线性规划方法以及 Luo(2004)的 Bell 树模型都可以看成是全局优化的方法。三种方法在进行全局处理之前都是在采用最大熵模型来计算任何两个实体之间的共指概率。在后处理阶段，BestCut 方法采用类似于图理论中的 Min-Cut 方法来完成篇章内实体的分组，一个采用 Min-Cut 方法之前的初始图如图 4 所示。整数规划方法将共指消解结果的划分考虑成整数规划的优化模型，然后求解。Bell 树模型通过在构建的树型结构上搜索来完成最终的实体分组任务，一个搜索树如图 5 所示。将篇章中的实体按照先后顺序从左到右排列，从第一个实体开始扫描。初始状态就是第一个实体构成一个等价类，然后依次扫描后续实体，经过判断后将实体放入已经存在的某个等价类中，或者新建一个等价类放置这个实体。这种方法对全部实体形成所有可能的等价类划分。可能的划分数量被称为 Bell 数。Bell 数随着实体数量的增大呈现指数增长的趋势，所以很多实际应用 Bell 树的系统都会采用一些减少搜索空间的方法。Luo(2004)中也有类似处理。

McCallum and Wellner(2004)采用基于条件随机域的图分割方法将篇章中的各个实体表述合并到不存在矛盾的等价类中，其中考虑了共指消解的传递性，其第三个模型中对可能出现的不一致三角情况进行了约束。Ng(2005)采用各种二元分类器来对篇章中的实体描述产生多种候选等价类划分，然后根据一个聚类模型的输出来将各种等价划分进行排序，从而得到最佳的结果。

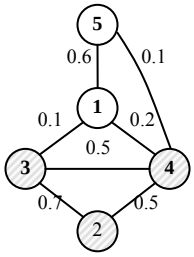

Fig.4 The initial graph for BestCut
图 4 BestCut 方法中的初始图

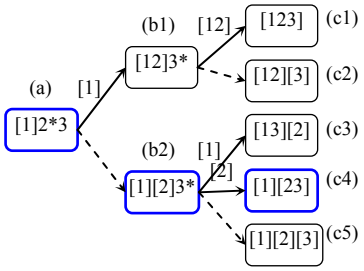

Fig.5 Bell tree for three mentions
图 5 三个 Mention 的 Bell 树

尽管上面这些方法都是基于篇章全局优化策略的，但它们都是基于二元分类模型来获取 Mention 之间的共指概率。因为基于判别模型的等价划分在进行等价划分的时候很难进行其他方式的融合。这种方式的全局优化仅仅采用二元分类模型得到的分值来进行后续处理肯定会存在一些偏差。我们认为需要采用脱离这种框架的全局优化方法，例如可以在全局处理时不但考虑两两之间的共指概率，还需要考虑每个 Mention 下具体的特征属性值。

目前基于全局优化的方法正在成为共指消解研究的一种趋势，越来越多的研究人员意识到局部优化的弊端。众多新近出现的论文向我们展示了一个共指消解研究的宏观方法。

2.3.2 利用深层语言学知识以及背景知识

现在主流的各种共指消解系统都在不断的探索新的数学模型，和发明新的计算模型同等重要的是针对共指消解发掘新的特征。但是，自从二十世纪九十年代中期以来，

共指消解的研究一直采用的是“知识匮乏”的方法。因为之前的研究都是基于深层句法分析和语义分析的理论模型，获取深层的语言学知识在当时比较困难。虽然结合着一些浅层的词法分析模块，一些“知识匮乏”的系统获得了较大的实用性能(Mitkov 2001)，但 Kehler, et al.(2004)强调为了使共指消解的系统性能提升到一个新的水平必须要利用深层的语言学知识。

这一点其实并不奇怪，因为一些共指关系不可能通过字符串匹配和句法分析技术来进行确定。例如，需要语义知识才能确定两个不相似的字符串是否具有共指关系，例如“美国首都”和“华盛顿”。在篇章中没有相关信息时，需要借助一些背景知识才能消解“美国现任总统”和“小布什”。

获取深层的语言学知识可以从三种途径产生。第一，基于常规知识库。这种方法常见的策略是借助一些人类已经编撰好的知识词典，例如英文上有 WordNet (Fellbaum, 1998)，中文上有 Hownet (董振东,董强, 2001) 、《同义词词林》(梅家驹,等,1996)、《现代汉语分类词典》(董大年,1998)等。第二种途径是从大规模语料库中挖掘模式信息。这种方法主要是启发式的总结一些槽模板，然后在大规模语料库（可以经过一些相关加工）中，统计各种匹配信息。例如，Bergsma(2005)在一个经过 Minipar 依存分析的语料库上获取了大量的指代信息，实现了英文名词短语性别和单复数信息的模板化提取，Vincent(2007)在语料库上通过一些模板获取了多种名词短语语义类信息，增强了共指消解的性能。Yang and Su(2007)利用语料库中发现的模板信息来增强共指消解。第三种方法是将整个互联网当成一个巨大的语料库，利用搜索引擎显示的各个查询得到的返回数来计算各种相关信息，例如通过计算互信息来考察两个短语的关联程度(Poesio, et al., 2004b; Markert, et al., 2003)。

在共指消解中结合背景语义知识的常见方法都是在基于特征向量的机器学习方法中引入一些语义相似度的特征，例如基于 Wikipedia 的语义相似度(Strube and Ponzetto, 2006; Ponzetto and Strube, 2006)，性别相似度(Bergsma, 2005)等。目前基于特征向量的机器学习方法很难更加直接的引入背景语义知识。我们正在尝试一种新的共指消解方法：基于归纳逻辑程序(ILP, Inductive logic programming)技术的共指消解。ILP 建立在一阶谓词逻辑(First Order Logic)的基础上，同时将正例、反例、背景知识一起编码，能够很好的融入背景知识(Lavrac and Dzeroski, 1994)，而且 ILP 对实例的属性采用逻辑表达式，可以避免特征向量方法中的属性缺失的问题。目前已经有诸如词性标注、句法分析、信息抽取等任务采用 ILP 取得了很好的效果(Dzeroski, et al., 2000)。最近，Specia, et al.(2007)采用 ILP 很好的融合了背景知识来完成词义消歧研究。和词义消歧类似，我们认为共指消解对于背景语义知识的应用也可以采用 ILP 的技术来实现，但是到目前为止还没有见到相关的研究成果。

另外一种不采用基于特征向量的方法来融合背景知识的途径是 Bean and Riloff(2007)等采用的 Dempster-Shafer 概率模型。他们的系统中采用 AutoSlog 系统从源语料中挖掘一些模式(Pattern)，来表示上下文的角色关系，还使用一些简单的特征作为背景知识源。最后将多种知识源采用 Dempster-Shafer 概率模型融合起来。这种概率模型可以融入每一个背景知识源得出的置信度，依次对置信度进行重置，在一个统一的框架下找到最佳的假设。但是这种模型不是很容易理解，而且在经过第一次置信度重置后集合的分散程度已经不能很好发挥 Dempster-Shafer 概率模型的优势。当然这种框架也是非常值得深入拓展的。

2.3.3 语言学模型和统计学习模型的融合

我们认为可以结合语言学和机器学习的方法，利用语言学思路来构建更加丰富的

机器学习模型。以往很多研究都是在机器学习的分类、聚类框架中选取相关特征时加入一些语言学知识，例如加入句法角色、单复数、性别等特征。这种语言学融合到机器学习中的方法还比较初级。

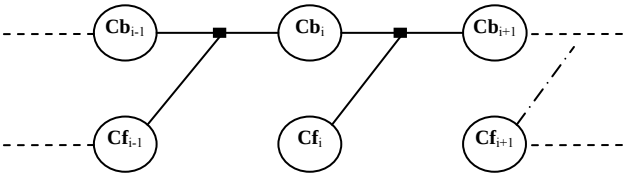

Fig.6 Coreference resolution model based on centering theory and CRF

图 6 基于中心理论和 CRF 的共指消解模型

作为开始，Elango(2006)提出了一种初始化的建议：结合中心理论和条件随机域模型(CRF)来实现人称代词消解。图 6 展示了模型的基本框架，可以综合考虑各种相关特征因素，例如给定子句的回指中心、前一个回指中心、前一个预指中心列表之间的关系。这些因素对中心转移的各种优选性建立了模型。基于 CRF 模型的灵活性，依赖于上下文的传递优选性能被很好的融入到模型中。正如 Poesio, et al.(2004)中讨论的那样，子句作为话语单元是一个较为合理的假设。从而，篇章可以表示成一系列子句的集合，进而可以被表示为一系列预指中心集合的特征空间。这个预指中心列表构成的特征空间可以融合一些相关特征，例如语法角色、性别、单复数等。类似的序列 CRF 模型上的推理和估计，可以采用 Sutton and McCallum(2006)中讨论的技术。

3 中文共指消解

英文属于印欧语系，中文属于汉藏语系。相对于英文，中文上的共指消解研究更加困难。因为中文上能够在词汇层获取的额外信息比英文上要少得多。例如，英文能够通过词形和词性标注结果来判断单复数信息，而中文上没有简单规则来直接判断一个中文词汇是单数还是复数。英文上专有名词和缩略语都体现在英文单词的首字母组合上，而中文上专有名词和缩略语大多没有采用这种方式。另外一个问题是英文词汇用空格分开，而中文上没有词语的天然边界，一个句子往往会有能表示多种意思的分词结果，人们不能通过逐字处理来直接理解一个句子，因此分词对于中文共指消解也是一个严重的问题。中文上共指消解的人工标注语料相对较少。据我们了解，目前只有 ACE 评测上有公开的中文共指消解的手工标注语料，但这些语料都不能轻易的免费获得。构建一个可行的大规模共指消解语料事实上是一件极其耗费人力的事情。

王厚峰(2004)简要分析了汉语指代消解中存在的三个问题：照应语的识别，尤其是零形式照应语和无标记的普通名词或名词短语作为照应语的辨识缺乏标记；有些照应语对先行语的属性分析和构成形式的判断没有指导作用，从而导致了潜在先行语识别的困难；最重要的是，指代消解所需要的语言知识在目前分析技术下不容易得到。因此，完全的汉语指代消解仍然是困难的。结合汉语的特点，王厚峰和梅铮(2005)提出了一种弱化语言知识的鲁棒性人称代词消解方法，仅仅用到了单复数特征、性别特征和语法角色特征。该方法主要分为两步。首先，利用这三种特征的简单约束关系，过滤与人称代词特征不一致的词，并形成可能的先行语候选集；然后，使用一个权值算法，计算候选的权值，并将最高权值的候选作为代词最终的先行语。权值算法并不是枚举式的计算每个候选的权值，而会通过动态评测机制，在合适的条件下自动终止计算，因而有效地控制了计算复杂度。此外，该方法不需要对文本进行深层的分析处理，实现起来也很容易。更具特色的是，王厚峰和何婷婷(2001)结合 HNC 理论提出了

汉语人称代词的消解方法，主要结合句类基本知识，根据人称代词所在语义块中的语义角色和人称代词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，同时，也从句法的角度，结合局部焦点法（类似于中心理论）给出了优选性规则。

王德亮(2004; 2006)仔细研究了中心理论，并且结合中文的特点进行了中文零型回指的研究，还提出了一个进行中文指代消解的算法框架。这是对中文零型回指的一项实证研究。

王智强(2006)针对中文共指消解中需要首先处理的中文基本名词短语识别问题进行了深入研究，采用了多种模型来实现中文基本名词短语的识别，随后提出了三种中文人称代词消解算法和一种中文基本名词短语的共指消解模型。中文人称代词消解分别是基于规则、基于机器学习、规则和统计相结合的方法。实验效果显示统计处理之后应用规则的方法取得了最好的效果。实验中详细对比了决策树(王智强,等,2006)、最大熵和 CRF 模型，结果显示 CRF 能够达到最好效果。中文基本名词短语的共指消解只是提出了模型框架，并没有具体实现。

周俊生等(2007)开始考虑进行全局优化的策略。他们的方法先采用基于规则方法获得任意两个 Mention 之间的共指概率，然后采用基于模块度的自动确定类别数量的聚类方法，从而实现最终的无指导的共指消解。

Wang and Ngai(2006)结合中文的特点，提出了一种中文共指消解的聚类方法。和 2.2.2 中研究的不同之处是，他们采用的很多特征都考虑了中文的特点。还有 2.1.1 节提到的 Converse(2006)首次将 Hobbs 算法应用在中文代词消解上。

4 共指消解评测

4.1 国际评测

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正在成为一种大力推动相关研究进展的方式。共指消解的研究也不例外。到目前为止，共指消解的相关评测有如下三种。

最早开始共指消解评测的是消息理解系列会议MUC[1]。MUC主要包括信息抽取的评测和召开相关的讨论会议，但是其显著特点在于对各种信息抽取系统的相关评测。正是MUC系列会议使得信息抽取发展成为自然语言处理领域的一个重要分支，并一直推动这一领域向前发展。

从 1987 年开始到 1998 年，MUC 会议一共举办了 7 次，它由美国国防高级研究计划委员会(DARPA)资助举行。共指消解的评测出现在 1995 年 9 月举行的 MUC6 和 1998 年 4 月举行的 MUC7 中，而且都是进行英语上的共指消解评测。

目前正在推动信息抽取研究进一步发展的动力来自美国国家标准技术研究所(NIST)组织的自动内容抽取ACE评测会议[2]。这项评测从 1999 年 7 月开始酝酿，2000 年 12 月正式开始启动。迄今已经举办过七次评测(ACE Pilot：2000 年 5 月，ACE Phase1：2002 年 2 月，ACE Phrase2：2002 年 9 月，ACE2003：2003 年 10 月，ACE Phase3：2004 年 8 月，ACE2005：2005 年 11 月，ACE2007：2007 年 1 月)。目前ACE评测的主要标注任务之一是实体检测与识别(EDR, Entity Detection and Recognition)。该任务将篇章中出现的各种表述(Mention)指向对应的实体(Entity)，从而给出一个实体全面的描述。这项任务中首先需要识别出各种表述，然后将描述同一实体的表述合并，该合

[1] http://www-nlpir.nist.gov/related_projects/muc/
[2] http://www.nist.gov/speech/tests/ace/index.htm

并过程就是共指消解的过程。可见，共指消解是ACE评测中的一项重要任务。

值得一提的是，从 2003 年开始 ACE 中开始包含中文的相关评测，至今已经开展四次评测。其中的共指消解也是迄今为止唯一的中文共指消解国际评测。

2006 年 11 月到 2007 年 3 月，英国伍尔佛汉普敦大学发起了一个名为指代消解练习(ARE)的共指消解评测[3]。这项评测是在英文上进行的迄今为止最全面的共指消解评测，包含四项评测任务:

1. 预标注文档上的人称代词消解：文档内的名词短语都被识别出来，而且需要消解的代词也被标注出来。参加系统需要对每个人称代词在一个不包含人称代词的名词短语列表中找到正确的先行语。

2. 预标注文档上的共指消解：文档内所有的名词短语都被识别出来，参加系统需要将文档内的所有共指链识别出来。

3. 生语料上的人称代词消解：和第一项任务不同的是，评测文档没有经过任何标注，需要参加系统自行识别相关信息。

4. 生语料上的的共指消解：和第二项任务不同的是，评测文档没有经过任何标注，需要参加系统自行识别相关信息。

前面两项任务在预标注文档上进行，主要是为了给参加系统提供公平的输入内容，主要评测各自系统的算法。后两项评测在完成相关的名词短语识别后就能像前两项任务那样展开，主要是为了考察名词短语识别对最终的人称代词消解以及共指消解产生的影响。

4.2 评价方法

评测一个系统的得分对于共指消解研究具有非常重要的意义。首先，通过得分可以判断一个系统的好坏；其次，可以从中发现系统继续改善的方向。不管是什么评测方法，都需要一个标注好正确共指消解结果的样本库，这样才能在不同的系统之间进行可以比较的评测。

正如 1.1 中分析的那样，共指关系是一种等价关系，具有共指关系的实体属于同一个等价类。共指消解就是将文本中所有名词短语组成的全集划分为一系列互不相交的等价类子集的过程。评价之前需要有标准答案的划分(Key)，待评价系统的输出划分(Response)。评价过程就是比较两个划分：Key 和 Response。

例如下面的对话(根据 Passonneau(1997)修改)：

M：Okay. We need to ship a boxcar of oranges to Bath by 8 AM today.

S：Okay.

M：Okay. So I guess I would suggest that we use [engine E1]$_1$ and have [it]$_1$ pick up [a boxcar]$_2$ at Dansville. How long'll [it]$_1$ take

S: That'll take hours to get to Dansville and get [the boxcar]$_2$.

M: Okay. And then, how long to go on to Corning with [the boxcar]$_2$ coupled to [E1]$_1$?

S: Another hour.

M: Ok. So that's okay. And then if we loaded [the oranges]$_3$ at Corning and sent [E1]$_1$ on to Bath with [the oranges]$_3$

S: We'd get there at 7.

这段对话的共指消解结果如表 1 所示。表中CA_1为标准标注，CA_2 为待测系统标注，都是对NP集的不同划分。NP集合{A,B,C,D,E,F,G,H,I,J}被CA_1划分为{A,B,D,G,I},{C,E,F}和{H,J}，被CA_2划分为{A,B,G,I},{C,D,E,F}和{H,J}。

[3] http://clg.wlv.ac.uk/events/ARE/

Table 1 Coreference Resolution result of a dialogue
表 1 一段对话的共指消解结果

Token	String	CA_1	CA_2
A	engine E1	1	1
B	it	1	1
C	a boxcar	2	2
D	it	1	2
E	the boxcar	2	2
F	the boxcar	2	2
G	E1	1	1
H	the oranges	3	3
I	E1	1	1
J	the oranges	3	3

除了集合划分外，共指消解的结果还有一种链式表示方式。链式表示中每个 NP 只和距离最近的先行共指 NP 链接在一起，如图 7 所示。本质上，集合划分和共指链式是等价的，通过在链式表达上求传递闭包就能获得。

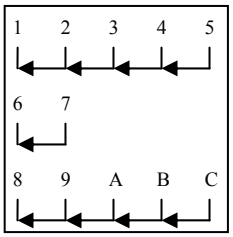

Fig.7 Sketch of coreference chain
图 7 共指链示意图

4.2.1 Valain 评测方法

多数共指消解系统都在 MUC6 和 MUC7 数据集上进行评测。MUC6 的评测方法是在 Valain et al.(1995)工作的基础上进行的，这种方法是专门用于共指消解评测的。

Recall 和 *Precision* 都是通过计算共指链中丢失链接的数量得到的。丢失链接数量是基于共指链上生成的共指划分计算的，有点类似于划分方法中计算精确率时采用的错误链接。直观的观察，影响 *Recall* 和 *Precision* 的因素主要是 Response 未能正确包含目标等价类成员，导致召回率降低；Response 中包含了不属于目标等价类的成员导致精确率降低。由此，计算召回率时需要统计向 Response 中至少添加多少 Link 才能将所有 NP 都归于 Key 中的一个等价类。类似的，可以对称的计算 *Precision*。

将共指消解结果链式表示中的一个等价类，按照两个节点是否链接可以采用最小生成树来表示。其中 Link 为边数，如果其中包含 N 个节点，那么 Link 数为 N-1。设目标等价类集合为 C，那么 C 中的 Link 数为$|C|$-1。设 $p(C)$为 Response 对 C 的一个划分，则对应于 C，Response 缺少的 Link 数为$|p(C)|$-1。注意：若目标集合内的 NP 未能正确识别，则默认为将该 NP 划分为单独一个等价类。例如：Key:{A,B,C,D,E}, Response:{A,B,C},则 p(C): {A,B,C}, {D}, {E}。

Valain 定义的等价类 C 的召回率为公式(1)；对于一个完整结果的评价如公式(2)所示。

$$Recall_C = \frac{(|C|-1)-(|p(C)|-1)}{|C|-1} = \frac{|C|-|p(C)|}{|C|-1} \quad (1)$$

$$Recall = \frac{\sum_i(|C_i| - |p(C_i)|)}{\sum_i(|C_i| - 1)} \tag{2}$$

例如表 1 中得到的两个划分： Key:{A,B,D,G,I}, {C,E,F},{H,J}， Response: {A,B,G,I},{C,D,E,F},{H,J}。令C = {A,B,D,G,I},则$|C|$ = 5，$|p(C)|$ = 2 则$Recall_C$ = (|C|-|p(C)|)/(|C|-1) = 3/4 。对于整体的召回率 *Recall* = [(5-2)+(3-1)+(2-1)]/[(5-1)+(3-1)+(2-1)] = 6/7。对称的，将Key和Response交换，采用同样的计算方法可以得到*Precision* = [(4-1)+(4-2)+(2-1)]/[(4-1)+(4-1)+(2-1)] = 6/7。

4.2.2 B-CUBED 评测方法

虽然 Valain 的方法是专门用于共指消解评测的，但它将各种共指消解错误都平等的对待了。例如，图 8 中对于 Response A：*Recall* =9/10，*Precision*=9/10；对于 Response B：*Recall*=9/10，*Precision*=9/10。

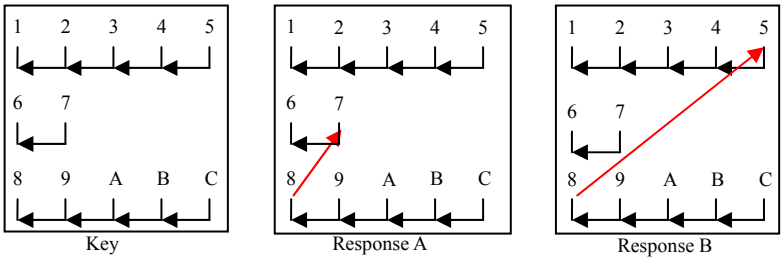

Fig.8 Sketch of the problem in Valin method

图 8 Valain 方法存在问题示意图

考虑一种极端情况，如果 Response 中只有一个等价类，即将篇章中所有的实体都认为是等价的，那么此时的 *Precision* 如公式(3)所示。

$$Precision = \frac{N - m}{N - 1} \tag{3}$$

公式中 N 为篇章中实体的个数，m 为 Key 对应的共指链的个数。此时，如果 $N>>m$，那么 $Recall \to 1$。这显然是不合理的。

事实上，图 8 中 Response B 将两个 Key 中的大类合并到一起了，比起 Response A 算是出现了较大的错误。两个大组之间的一个错误链接会比两个小组之间的错误链接产生更大的破坏性。为此，Bagga and Baldwin(1998)提出了 B-CUBED 评测方法。

对于篇章中的每个实体E_i，精确率和召回率定义如公式(4)、(5)所示。

$$Precision_{E_i} = \frac{|Right(Response(E_i))|}{|Response(E_i)|} \tag{4}$$

$$Recall_{E_i} = \frac{|Right(Response(E_i))|}{|Key(E_i)|} \tag{5}$$

公式中$Response(E_i)$表示包含E_i实体的输出共指等价类，$Key(E_i)$表示包含E_i实体的目标共指等价类，$Right(Response(E_i))$表示包含E_i实体的输出等价类中被正确划分的实体构成的集合。最终的*Precision*和*Recall*的计算公式如(6)、(7)所示。

$$Precision = \sum_{i=1}^{N} w_i \times Precision_{E_i} \tag{6}$$

$$Recall = \sum_{i=1}^{N} w_i \times Recall_{E_i} \tag{7}$$

式中w_i一般依赖于具体的算法和应用，也可以采用相等权值另w_i=1/N。

例如，对于图 8 中的两个 Response 的 *Precision* 分别计算如公式(8)、(9)所示。由公式(5)得知，两个 Response 的 *Recall* 都是 100%。

$$Precision_A = \frac{1}{12} \times (\frac{2}{5} \times 5 + \frac{2}{7} \times 7 + \frac{5}{7} \times 5) = \frac{16}{21}(76\%) \tag{8}$$

$$Precision_B = \frac{1}{12} \times (\frac{5}{10} \times 5 + \frac{2}{2} \times 2 + \frac{5}{10} \times 5) = \frac{7}{12}(58\%) \tag{9}$$

4.2.3 ACE 评测方法

ACE任务之一的实体检测与识别(EDR)需要识别出文章中的实体及其类型，并且将表示现实世界同一实体的不同描述合并到一起[4]。评价方法中考虑了系统输出的丢失率和错误率。总体评价指标主要考虑两个层面：

- 实体类型：考察实体的类型正确与否。主要有七大类实体，包括设施名(FAC，Facility)、行政区划(GPE，Geo-Political Entity)、地名(LOC，Location)、机构名(ORG，Organization)、人名(PER，Person)、交通工具名(VEH，Vehicle)、武器名(WEA，Weapon)。
- 实体内部：主要考察实体的名称、名词性共指、代词性共指等。

计算得分时将两种类型的分值相乘，公式如(10)所示。

$$Values_{sys\_entity} = Entity\_Values(sys\_entity) \cdot Mentions\_Values(\{sys\_mentions\}) \tag{10}$$

式中 *Entity* 表示等价类，即现实世界中的实体；*Mention* 表示等价类中的元素，即实体在篇章中的不同表述。

ACE 的这种评测方法严格算来不是专门针对共指消解进行评测的，因为其中涉及到了很多属性信息的检查。事实上，很少有论文采用这种方法进行共指消解评测。

4.2.4 CEAF 评测方法

虽然 B-CUBED 方法避免了 Valain 方法的一些不足之处，但是由于本质上 B-CUBED 方法和 Valain 方法都是基于集合交叉的，容易导致一个 Entity 被多次计算。例如在图 8 的基础上考虑图 9 中的两个 Response。计算得到的 R、P、F 分别如表 2 中 B-CUBED 列所示。Response C 将全部 Mention 合并为一个 Entity，得到的召回率为 1.0；Response D 将各个 Mention 分别看成单独的 Entity，得到的精确率为 1.0。事实上，这种结果是违背人的直觉的。

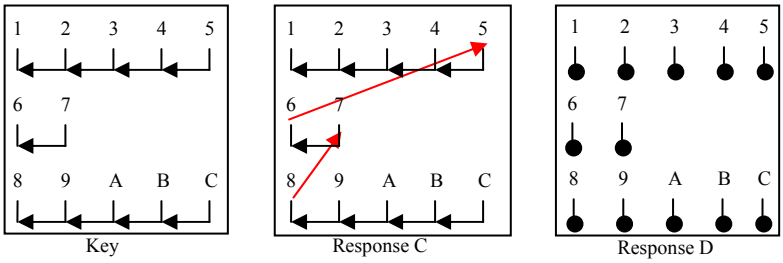

Fig.9 Sketch of the problem in B-CUBED method

图 9 B-CUBED 方法存在问题示意图

Table 2 Comparision of coreference evaluation metrics between B-CUBED and CEAF

[4] http://www.nist.gov/speech/tests/ace/ace07/doc/ace07-evalplan.v1.3a.pdf

表 2　B-CUBED 和 CEAF 评价结果对比

System response	B-CUBED			CEAF					
	R	P	F	ϕ_3-R	ϕ_3-P	ϕ_3-F	ϕ_4-R	ϕ_4-P	ϕ_4-F
Response C	1.0	0.375	0.545	0.417	0.417	0.417	0.196	0.588	0.294
Response D	0.25	1.0	0.400	0.250	0.250	0.250	0.444	0.111	0.178

为了更加合理的计算共指消解的实验效果，Luo(2005)提出了 CEAF(Constrained Entity-Aligned F-Measure)评价方法。其主要思想是将 Key 的各个共指链和 Response 的各个共指链之间建立一一映射，然后采用 Kuhn-Munkres 算法来求取最优的带权二分图匹配，最后将匹配结果转化为 R、P、F。两个共指链 K 和 R 的匹配程度采用共享的 Mention 个数来进行度量，例如公式(11)、(12)的两种度量：

$$\phi_3(K,R)=|K\cap R| \tag{11}$$

$$\phi_4(K,R)=\frac{2|K\cap R|}{|K|+|R|} \tag{12}$$

计算 F 值的算法如下：

算法 1. Computing the F-measure
Input：Key entities: K, response entities: R
Output：Optimal alignment g*; F-measure
1. Initial: g*=∅; Φ(g*)=0.
2. **for i=1 to** |K|
3. **for j=1 to** |R|
4. Compute $\phi(K_i, S_j)$.
5. [g*,Φ(g*)]=**Kuhn-Munkres** ({$\phi(K, S)$: R∈ℝ, S∈S}).
6. $\Phi(K)=\Sigma_{K\in\mathbb{R}}\phi(K, K)$; $\Phi(R)=\Sigma_{R\in\mathbb{R}}\phi(R, R)$.
7. r=Φ(g*)/Φ(K); p=Φ(g*)/Φ(R); F=2pr/(p+r).
8. **return** g* and F.

对图 9 中 Response C、D 计算得到的 CEAF 的 R、P、F 值如表 2 中 CEAF 列所示。事实上，CEAF 评价方法越来越得到研究人员的认可，新近出现的很多共指消解相关文献都采用这种评价方法(Vincent, 2007; Luo and Zitouni, 2005; Eisenstein and Davis, 2006)。

4.3 基于评测的语料库特征发现

Bagga(1998)提出了一种共指类型分类的方法，能够帮助分析一个给定的共指消解系统的优势和劣势所在。主要的几个类别是同位语、句法等价、专有名词、人称代词、引用话语代词、指示性描述等。在华尔街日报（WSJ）语料上统计发现，专有名词和人称代词两个类别的比例分别是 27.8%和 21.0%，但是需要额外背景知识才能消解的共指类型只占 5.9%的比例。

Harabagiu, et al.(2001)将数据挖掘的技术应用到共指消解上。他们在 MUC6 和 MUC7 的数据集上标注指代链然后生成更多的共指数据。有趣是，在他们的实验中统计发现<指代语，专有名词>链接的比例达到 29.1%，<普通名词短语，普通名词短语>链接的比例达到了 10%。更进一步，MUC6 语料上接近 83%的共指链可以被简单规则或者简单特征正确消解，例如重复、别名、公共开头字符串等。随后多种知识规则组合起来形成规则集，采用熵来计算每条规则的置信度。给定一组名词短语，通过最大

化关系聚类中采用的目标函数得到最佳划分。

5 相关资源和工具

像其他自然语言处理研究一样，共指消解的研究也需要一些相关的辅助资源和工具。但是由于共指消解在篇章自然语言处理中处于相对上层，需要的底层处理工具和辅助资料就会相对多一些。对于初涉该领域的研究者了解并掌握这些资料对于快速进入深入的研究具有重要意义。下面列出我们已经了解到的一些语料资源，相关工具、源码和集成了共指消解模块的的系统。

共指消解研究的语料，在众多论文中提及的主要有如表 3 所示的一些。目前网上能够找到源代码的共指消解系统如表 4 所示。目前共指消解系统大量存在于一些集成系统中，共指消解模块在其中发挥着重要的作用，相关的系统如表 5 所示。

Table 3 Corpora on coreference resolution research
表 3 共指消解研究常用语料库

Item Name	Release Institiute	Scale	Language(s)	Release Year
Message Understanding Conference (MUC) 6	Linguistic Data Consortium (LDC2003T13)	318 annotated Wall Street Journal articles	English	2003
Message Understanding Conference (MUC) 7	Linguistic Data Consortium (LDC2001T02)	107 annotated newswire articles	English	2001
ACE 2004 Multilingual Training Corpus	Linguistic Data Consortium (LDC2005T09)	Annotated articles of Arabic(689)，Chinese(646)，English(451)	Arabic, Chinese, English	2005
ACE 2005 Multilingual Training Corpus	Linguistic Data Consortium (LDC2006T06)	Annotated articles of Arabic(433)，Chinese(633)，English(599)	Arabic, Chinese, English	2006
Anaphora Resolution Exercise	Research Group in Computational Linguistics，University of Wolverhampton， UK	74 annotated newswire articles on four levels	English	2007
KNACK-2002	CNTS，University of Antwerp, Belgium	267 annotated newspaper texts	Dutch	2002

Table 4 Famous open source coreference resolutio systems
表 4 著名的共指消解系统源代码

System Name	Current Version	Release Institiute	Main Function	Preprocess Module	Processing Language(s)	Programming Language
GuiTAR (Poesio and Kabadjov, 2004)	3.0.3[5]	Language and Computation Group, University of Essex, UK	It can process original text with syntax analysis, mention detection, anaphora resolution, and can be used for system	Charniak's full parser	English	Java

5 http://privatewww.essex.ac.uk/~malexa/GuiTAR/

			evaluation.			
JavaRAP (Qiu, et al., 2004)	1.11[6]	Department of Computer Science, National University of Singapore	A freely-available JAVA anaphora resolution implementation of the classic Lappin and Leass (1994) paper	Charniak's full parser	English	Java
MARS (Mitkov, et al., 2002)	Re-implemented in 2002[7]	Research Group in Computational Linguistics，University of Wolverhampton，UK	On going internal project which develops a knowledge-poor anaphora resolution for English.	Conexor's FDG Parser	English	Java
ROSANA	5.1[8]	Dr. Roland Stuckardt personal software	An algorithm for anaphora resolution that focus on robustness against information deficiency in the parsed output.		English, German	LISP

Table 5 Famous systems integrated with coreference resolutio
表 5 集成有共指消解模块的系统

System Name	Current Version	Release Institute	Main Function	Processing Language	Programming Language
LingPipe	3.1.1[9]	Alias-I company, USA	Determining entity coreference within documents	English	Java
OpenNLP	1.3.0[10]	Open source project in Sourceforge.NET	An open source package of natural language processing components written in pure Java.	English	Java
GATE	4.0[11]	NLP group, University of Sheffield, UK	General Architecture for Text Engineering, Natural Language Processing system	English	Java
LTP	2.0[12]	Information Retrieval Laboratory, Harbin Institute of Technology	XML-based open Chinese processing platform for web application.	Chinese	C++

6 结论和展望

篇章共指消解的研究已经发展了三十多年，已经涌现出了大量的研究成果。在经

6 http://www.comp.nus.edu.sg/~qiul/NLPTools/JavaRAP.html

7 http://clg.wlv.ac.uk/MARS/

8 http://www.stuckardt.de/rosana.htm

9 http://www.alias-i.com/lingpipe/index.html

10 http://opennlp.sourceforge.net/index.html

11 http://gate.ac.uk/

12 http://ir.hit.edu.cn/demo/ltp

历了早起的基于语言学规则的发展后，伴随着整个自然语言处理领域中统计方法的兴盛，共指消解也进入了统计时代。随着研究者们对共指消解问题本质的不断深入理解，开始更加的集中在篇章全局优化、深层语言学知识和背景知识的利用，以及语言学模型和统计模型的融合等三大趋势上。共指消解是 AI 完全问题，需要解决很多的相关问题，随着方法、工具、语料、源码、系统的大量涌现，共指消解的研究正在进入一个崭新的阶段。

对于中文共指消解的研究来说，还需要积累很多的相关技术以及语料资源。在不断借鉴国外最新研究成果、研究路线的基础上，我们应该进一步结合中文自身的特点来推动中文共指消解的研究。随着现在中文自然语言处理研究的不断推进，相信中文共指消解的研究会走上更加顺利的发展道路。

References:

A. Bagga and B. Baldwin. 1998. Algorithms for scoring coreference chains. In: Proc. of the Linguistic Coreference Workshop at The First International Conference on Language Resources and Evaluation. GRANADA, SPAIN, 563-566.

A. Bagga. 1998. Evaluation of coreferences and coreference resolution systems. In: Proc. of the First Language Resource and Evaluation Conference. Granada, Spain, 563-566.

A. Kehler, et al. 2004. The (non)utility of predicate-argument frequencies for pronoun interpretation. In: S. Dumais, D. Marcu, and S. Roukos eds. Proc. of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. Boston, Massachusetts, USA: Association for Computational Linguistics, 289-296.

A. McCallum and B. Wellner. 2004. Conditional models of identity uncertainty with application to noun coreference. In: L. Saul, Y. Weiss, and L. Bottou eds. Proc. of Neural Information Processing Systems. Vancouver, British Columbia, Canada: MIT Press, 905-912.

B. Grosz, A. Joshi, and S. Weinstein. 1995. Centering: A framework for modelling the local coherence of discourse. Computational Linguistics, 21(2):203-225.

C. Cardie and K. Wagstaf. 1999. Noun phrase coreference as clustering. In: P. Fung and J. Zhou eds. Proc. of the 1999 Joint SIGDAT conference on Empirical Methods in NLP and Very Large Corpora. College Park, MD, USA: Association for Computational Linguistics, 82-89.

C. Fellbaum (editor). 1998. Wordnet. An electronic lexical database: MIT Press

C. Müller, S. Rapp and M. Strube. 2002. Applying co-training to reference resolution. In: P. Isabelle ed. Proc. of the 40th Annual Meeting on Association for Computational Linguistics. Philadelphia, Pennsylvania: Association for Computational Linguistics, 352-359.

C. Nicolae and G. Nicolae. 2006. Bestcut: A graph algorithm for coreference resolution. In: D. Jurafsky and E. Gaussier eds. Proc. of the 2006 Conference on Empirical Methods in Natural Language Processing. Sydney, Australia: Association for Computational Linguistics, 275-283.

C. Sidner. 1981. Focusing for interpretation of pronouns. Computational Linguistics,7(4):217-231.

C. Sutton and A. McCallum. 2006. An introduction to conditional random fields for relational learning, In: L. Getoor and B. Taskar, eds. Introduction to statistical relational learning: MIT Press.

C. Wang and G. Ngai. 2006. A clustering approach for unsupervised Chinese coreference resolution. In: H. Ng and O. Kwong eds. Proc. of the Fifth SIGHAN Workshop on Chinese Language Processing. Sydney: Association for Computational Linguistics, 40-47.

D. Bean and E. Riloff. 2007. Unsupervised learning of contextual role knowledge for coreference resolution. In: S. Dumais, D. Marcu, and S. Roukos eds. Proc. of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. Boston, Massachusetts, USA: Association for Computational Linguistics, 297-304.

D. Byron and J. Allen. 1999. Applying genetic algorithms to pronoun resolution. In: J. Hendler, et al. eds. Proc. of the sixteenth national conference on Artificial intelligence and the eleventh Innovative applications of artificial intelligence conference innovative applications of artificial intelligence. Orlando, Florida, United States, 957.

J. Eisenstein and R. Davis. 2006. Gesture improves coreference resolution. In: R. Moore, et al. eds. Proc. of the North American Chapter of the Association for Computational Linguistics. New York City, USA: Association for Computational Linguistics, 37-40.

J. Hobbs. 1978. Resolving pronoun references, In: B. Grosz, K. Sparck-Jones, and B. Webber, eds. Readings in natural language processing. Lingua, 44:339-352.

J. McCarthy and W. Lehnert. 1995. Using decision trees for coreference resolution. In: C.R. Perrault ed. Proc. of the Fourteenth International Joint Conference on Artificial Intelligence. Québec, Canada: Springer, 1050-1055.

J. Tetreault. 1999. Analysis of syntax-based pronoun resolution methods. In: R. Dale and K. Church eds. Proc. of the 37th Annual meeting of the Association for Computational Linguistics on Computational Linguistics. College Park, Maryland, USA: Association for Computational Linguistics, 602-605.

K. Markert, M. Nissim, and N. Modjeska. 2003. Using the web for nominal anaphora resolution. In: R. Dale, K. Deemter, and R. Mitkov eds. EACL 2003 Workshop on the Computational Treatment of Anaphora. Budapest, Hungary, 39-46.

K. Wagstaff. 2002. Intelligent clustering with instance-level constraints Ph.D. Thesis]. Cornell University.

L. Qiu, M. Kan, and T. Chua. 2004. A public reference implementation of the rap anaphora resolution algorithm. In: Proc. of the 4th International Conference on Language Resources and Evaluation. Lisbon, Portugal, 291-294.

L. Specia, M. Stevenson, and V.N. Maria das Graças. 2007. Learning expressive models for word sense disambiguation. In: J. Carroll, A. Bosch, and A. Zaenen eds. Proc. of the 45th Annual Meeting of the Association of Computational Linguistics. Prague, Czech Republic: Association for Computational Linguistics, 41-48.

M. Poesio and M. Kabadjov. 2004. A general-purpose, off-the-shelf anaphora resolution module: Implementation and preliminary evaluation. In: Proc. of the 4th International Conference on Language Resources and Evaluation. Lisbon, Portugal.

M. Poesio, et al. 2004a. Centering: A parametric theory and its instantiations. Computational Linguistics, 30(3):309-363.

M. Poesio, et al. 2004b. Learning to resolve bridging references. In: D. Scott ed. Proc. of the 42th annual meeting on Association for Computational Linguistics. Barcelona, Spain: Association for Computational Linguistics, 143-150.

M. Strube and S. Ponzetto. 2006. Wikirelate! Computing semantic relatedness using wikipedia. In: Y. Gil and R. Mooney eds. Proc. of the Twenty-First National Conference on Artificial Intelligence. Boston, Massachusetts, 1419–1424.

M. Strube, S. Rapp, and C. Müller. 2002. The influence of minimum edit distance on

reference resolution. In: J. Haji and Y. Matsumoto eds. Proc. of the 2002 Conference on Empirical Methods in Natural Language Processing. Philadelphia, USA: Association for Computational Linguistics, 312-319.

M. Strube. 1998. Never look back: An alternative to centering. In: C. Boitet and P. Whitelock eds. Proc. of the 17th international conference on Computational linguistics. Morristown, NJ, USA: Association for Computational Linguistics, 1251-1257.

M. Vilain, et al. 1995. A model-theoretic coreference scoring scheme. In: Proc. of the Sixth Message Understanding Conference. Columbia, Maryland: Morgan Kaufmann Publishers, 45-52.

N. Bansal, A. Blum, and S. Chawla, 2004. Correlation clustering. Machine Learning, 56(1-3):89-113.

N. Ge, J. Hale, and E. Charniak. 1998. A statistical approach to anaphora resolution. In: E. Charniak ed. Proc. of the Sixth Workshop on Very Large Corpora. Montreal, Canada: Association for Computational Linguistics, 161-170.

N. Lavrac and S. Dzeroski. 1994. Inductive logic programming: Techniques and applications. New York: Ellis Horwood, 1994.

N. Vincent. 2004. Learning noun phrase anaphoricity to improve conference resolution: Issues in representation and optimization. In: D. Scott ed. Proc. of the 42th annual meeting on Association for Computational Linguistics. Barcelona, Spain: Association for Computational Linguistics, 151-158.

N. Vincent. 2007. Shallow semantics for coreference resolution. In: M.M. Veloso ed. Proc. of International Joint Conferences on Artificial Intelligence. Hyderabad, India: AAAI Press, 1689-1694.

P. Denis and J. Baldridge. 2007. Joint determination of anaphoricity and coreference resolution using integer programming. In: C. Sidner, et al. eds. Proc. of Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics. Rochester, New York, USA: Association for Computational Linguistics, 236-243.

P. Elango. 2006. Coreference resolution: A survey. Project report of the course "Advanced natural language processing" In computer science departments, university of Wisconsin Madison.

R. Iida, et al. 2003. Incorporating contextual cues in trainable models for coreference resolution. In: R. Dale, K. Deemter, and R. Mitkov eds. EACL 2003 Workshop on the Computational Treatment of Anaphora. Budapest, Hungary, 23-30.

R. Mitkov, R. Evans, and C. Orasan. 2002. A new fully automatic version of mitkov's knowledge-poor pronoun resolution method. In: A. Gelbukh ed. Third International Conference on Computational Linguistics and Intelligent Text Processing. Mexico City, Mexico: Springer, 168-186.

R. Mitkov, S. Lappin, and B. Boguraev. 2001. Introduction to the special issue on computational anaphora resolution. Computational Linguistics, 27(4):473-477.

R. Mitkov. 1998. Robust pronoun resolution with limited knowledge. In: C. Boitet and P. Whitelock eds. Proc. of 36th Annual Meeting of the Association for Computational Linguistics and 17th International. Conference on Computational Linguistics. Montreal, Quebec, Canada: Association for Computational Linguistics, 1998.869-875.

R. Passonneau. 1997. Applying reliability metrics to co-reference annotation. Technical Report CUCS-017-97. Department of Computer Science, Columbia University, New York.

R. Witte and S. Bergler. 2003. Fuzzy coreference resolution for summarization. In: S. Harabagiu and R. Delmonte eds. Proc. of 2003 International Symposium on Reference

Resolution and Its Applications to Question Answering and Summarization. Venice, Italy, 43-50.
R. Witte, R. Krestel, and S. Bergler. 2006. Context-based multi-document summarization using fuzzy coreference cluster graphs. In: Proc. of Document Understanding Workshop at HLT-NAACL 2006. Brooklyn, New York, USA.
S. Bergsma. 2005. Automatic acquisition of gender information for anaphora resolution. In: B. Kégl and G. Lapalme eds. Canadian Conference on AI. Victoria, Canada: Springer-Verlag, 342-353.
S. Brennan, M.W. Friedman, and C.J. Pollard. 1987. A centering approach to pronouns. In: Proc. of the 25th annual meeting on Association for Computational Linguistics. Stanford, California: Association for Computational Linguistics, 155-162.
S. Converse. 2006. Pronominal anaphora resolution in Chinese Ph.D. Thesis]. University of Pennsylvania.
S. Dzeroski, J. Cussens, and S. Manandhar. 2000. An introduction to inductive logic programming and learning language in logic, In: J. Cussens and S. Deroski, eds. Learning language in logic: Springer-Verlag, 3–35.
S. Harabagiu, R. Bunescu, and S. Maiorano. 2001. Text and knowledge mining for coreference resolution. In: L. Levin ed. Second Meeting of the North American Chapter of the Association for Computational Linguistics. Pittsburgh, Pennsylvania: Association for Computational Linguistics, 1-8.
S. Lappin and H. Leass. 1994. An algorithm for pronominal anaphora resolution. Computational Linguistics, 20(4):535-561.
S. Ponzetto and M. Strube. 2006. Exploiting semantic role labeling, wordnet and wikipedia for coreference resolution. In: R. Moore, et al. eds. Proc. of Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics. New York City, USA: Association for Computational Linguistics, 192-199.
T. Finley and T. Joachims. 2005. Supervised clustering with support vector machines. In: S. Dzeroski, L.D. Raedt, and S. Wrobel eds. Proc. of the 22nd international conference on Machine learning. New York, NY, USA: ACM Press, 217-224.
T. Morton. 1999. Using coreference for question answering. In: A. Bagga, B. Baldwin, and S. Shelton eds. ACL 1999: Workshop on Coreference and Its Applications. College Park, Maryland, USA: Association for Computational Linguistics, .85-89.
V. Hoste and W. Daelemans. 2005. Comparing learning approaches to coreference resolution. There is more to it than 'bias'. In: C. Giraud-Carrier, R. Vilalta, and P. Brazdil eds. Proc. of the Workshop on Meta-Learning (held in conjunction with ICML-2005). Bonn, Germany: MIT Press, 20-27.
V. Ng and C. Cardie. 2002a. Improving machine learning approaches to coreference resolution. In: P. Isabelle ed. Proc. of the 40th Annual Meeting on Association for Computational Linguistics. Philadelphia, Pennsylvania: Association for Computational Linguistics, 104-111.
V. Ng and C. Cardie. 2002b Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution. In: S. Tseng, T. Chen, and Y. Liu eds. Proc. of the 19th international conference on Computational linguistics. Taipei, Taiwan: Association for Computational Linguistics, 1-7.
V. Ng and C. Cardie. 2003. Bootstrapping coreference classifiers with multiple machine learning algorithms. In: M. Collins and M. Steedman eds. Proc. of the 2003 Conference on Empirical Methods in Natural Language Processing. Sapporo, Japan: Association for Computational Linguistics, 113-120.

V. Ng. 2005. Machine learning for coreference resolution: From local classification to global ranking. In: K. Knight ed. Proc. of the 43rd Annual Meeting on Association for Computational Linguistics. Ann Arbor, Michigan: Association for Computational Linguistics, 157-164.

W. Soon, H. Ng and D. Lim. 2001. A machine learning approach to coreference resolution of noun phrases. Computational Linguistics, 27(4):521-544.

X. Luo and I. Zitouni. 2005. Multi-lingual coreference resolution with syntactic features. In: R. Mooney ed. Proc. of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. Vancouver, British Columbia, Canada: Association for Computational Linguistics, 660-667.

X. Luo, et al. 2004. A mention-synchronous coreference resolution algorithm based on the bell tree. In: D. Scott ed. Proc. of the 42th annual meeting on Association for Computational Linguistics. Barcelona, Spain: Association for Computational Linguistics, 135-142.

X. Luo. 2005. On coreference resolution performance metrics. In: R. Mooney ed. Proc. of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. Vancouver, British Columbia, Canada: Association for Computational Linguistics, 25-32.

X. Luo. 2007. Coreference or not: A twin model for coreference resolution. In: C. Sidner, et al. eds. Proc. of Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics. Rochester, New York, USA: Association for Computational Linguistics, 73-80.

X. Yang and J. Su. 2007. Coreference resolution using semantic relatedness information from automatically discovered patterns. In: J. Carroll, A. Bosch, and A. Zaenen eds. Proc. of the 45th Annual Meeting of the Association of Computational Linguistics}. Prague, Czech Republic: Association for Computational Linguistics, 528-535.

X. Yang, et al. 2003. Coreference resolution using competition learning approach. In: E. Hinrichs and D. Roth eds. Proc. of the 41st Annual Meeting of the Association for Computational Linguistics. Sapporo, Japan: Association for Computational Linguistics, 2003.176-183.

Y. Versley. 2007. Antecedent selection techniques for high-recall coreference resolution. In: J. Eisner ed. Proc. of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Prague, Czech Republic: Association for Computational Linguistics, 496-505.

董大年. 1998. 现代汉语分类词典.上海:汉语大词典出版社.

董振东,董强. 2001. 知网和汉语研究.当代语言学, 3(1):33-44.

刘礼进. 2005. 自然语言理解中的回指解析研究概述.外语教学与研究, 37(6):439-445.

梅家驹,等. 1996. 同义词词林.第二版.上海:上海辞书出版社.

钱伟,等. 2003. 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, 40(9):1337-1342.

王德亮. 2004. 汉语零形回指解析--基于向心理论的研究.现代外语, 27(4):350-359.

王德亮. 2006. 基于向心理论的汉语回指消解研究博士学位论文].北京师范大学.

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

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

王厚峰. 2002. 指代消解的基本方法和实现技术.中文信息学报,16(6):9-17.

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

王智强,李蕾,王枞. 2006. 基于决策树的汉语代词共指消解.北京邮电大学学报,

29(4):1-5.
王智强. 2006. 汉语指代消解及相关技术研究博士学位论文].北京邮电大学.
杨佳,罗振声. 2005. 使用遗传算法的人称代词消解. In: S. Maosong, L. Teng ed. Proc. of International Conference on Chinese Computing. Singapore.
张钹. 2007. 自然语言处理的计算模型.中文信息学报,21(3):3-7.
周俊生,等. 2007. 一种基于图划分的无监督汉语指代消解算法.中文信息学报, 21(2):77-82.