实体名称规范的相关理论、方法及工具调研报告

刘建华、邹益民、曲云鹏、岳婷

实体名称是指在文本中出现的指称如人物、机构、地点等实体的名称，这些实体名称承载着文本中的重要信息，他们对于文本内容挖掘、非结构化的知识管理、科研机构评价等工作具有非常重要的作用。按照 Automatic Content Extraction(ACE)评测计划的定义，实体概念在文本中的引用(entity mention，也可称为指称项)可以有三种形式：命名性指称、名词性指称和代词性指称，如在下面一段话中：“[President] [Barack Obama] hosts the second White House Science Fair celebrating the student winners of a broad range of science, technology, engineering and math (STEM) competitions from across the country. [He] talked....”，“President”、“Barack Obama”和[He]分别代表了同一个人物的名词性指称、命名性指称和代词性指称，这些形式的存在为实体名称的识别与规范带来了很大的挑战。事实上，在实际的科技文献中，命名性指称具有更多变的表达形式，包括全称、简写、缩写、别称等等，如“Steven A. Morris”与“S.A. Morris”“Professor Morris”,“Mr Morris”等；“America”与“USA”、“United States of America”、“U.S.A”等。面对这些现象，需要解决的问题就包括以下两个方面：其一，同一个人物名称出现在多篇科技文献中，而这些相同的人物名称在事实上并不一定指代了同一个人物实体，如“Steven A. Morris”可能是美国维吉尼亚大学医学院的一名泌尿学研究人员，也可能是美国田纳西州大学商学院的一名计算信息系统研究人员，还可能是美国俄克拉荷马州大学从事科学计量的一名研究人员，如何甄别其真实指代含义。其二，对一个人物名称、机构团体或国家，不同的科技文献源（如不同来源的专利库）或同一篇文献中可能采用不同的表达方式，如何确定其是否相关，如何选用最合适的表达作为该实体的规范名称。本研究重点围绕这两个方面进行相关理论、方法和工具的探讨，由于研究的精力有限，本研究中的实体名称更关注于命名性的指代方面。本报告主要围绕目前在实体名称规范方面所开展的相关研究理论、方法和工具进行总结，以期会实际的解决方法提供更多有力的参考。

1 实体名称规范的理论方法研究

1.1 实体名称规范的概念、思路、方法和趋势研究

本章着重从什么是实体名称规范、针对命名实体规范主要方法的演化与发展，以及未来的发展趋势等几个方面总体介绍命名实体规范的相关情况。

1.1.1 实体名称规范的概念

在现实世界中，人们对同一个事物经常会给予不同的名称、描述或视角。随着信息科技的不断发展，网络资源越来越多，这种事物的名称也越来越多样化，这为实现计算机的自动理解和计算带来了很大的挑战。为了计算机的后续自动处理和深入分析，如机器翻译、信息检索、数据挖掘等，将这些名称、描述与其对应的事物对应起来，并从中选择一种规

范的表达作为不同名称或描述之间的核心关联非常有必要，由此产生了实体名称规范这样一个概念。简而言之，名称实体规范即解决两个方面的问题：确定两个不同的命名是否指代同一个实体，确定多个不同的命名中最为规范的名称。这其中，尤以第一个问题为重。从任务角度而言，第一个问题又包括了如下两个子问题：（1）一个对象有多种名称称为命名实体的共指问题；（2）一个名称可能指代不同的实体为实体歧义问题1。 实体名称的共指消解是比较复杂的一个问题，包含的内容比较多，既包括代词的消解，如“he，she”等人称代词实际指称对象的查找，也包括名词性称呼的消解。另外，实体由于一个词义的表达方法（从含义的有限集合枚举到基于规则的新含义的产生）、含义列表的细粒度（从细微的区别到反义词）、面向领域的与非严格定义的自然文本等原因，往往会出现一个实体名称可以对应到多个命名实体概念上的问题，比如，“Washington”既可能指称华盛顿州，也可能指代美国的第一任总统，因此就需要明确从两个实体名称是否为同一个概念，具体是什么概念几个方面入手。针对这两个问题，研究者们为了区分任务的专指性，又特意单独提炼出缩略语的识别这样一个专属任务。在科技文献中存在着大量实体名称缩写的情况，而且还充斥着大量不同实体全称缩写相同的情况，如“Automatic Content Extraction”、“American Collegiate English”与“ACE”，因此在进行文本分析时，就需要判定出这些缩略语所对应的全称。基于此，实体名称的缩略语识别研究即着重关注实体名称的缩略语与全称的匹配，目前在该方面的研究多数集中于术语的缩略语识别，且多数集中于特定的领域，主要为医学、生物学等。

总而言之，实体名称规范事实上主要是一个以计算的方式自动辨析词语在上下文中的真实含义的过程2，与实体名称规范这一概念密切相关的研究主题主要有缩略语的识别、实体名称共指消解和实体名称的消歧，这些研究在英文中被称为“Named Entity Disambiguation， Abbreviation Reorganization， Co-reference Resolution， Named Entity Normalization、Word Sense Disambiguation”等。

1.1.2 实体名称规范的主要思路与方法

从上述的概念中可以看出，实体名称规范中核心任务是以计算的方式自动辨析词语在上下文中真实含义的过程。要完成这样的过程，其中需要涉及到很多知识，不仅仅需要语言学方面的常用知识，如浅层的词汇、语法、句法等的分析，还需要用到很多语义及其背景知识信息。与此相关的研究正是遵循着这样由浅入深的知识利用，不断在思路与方法上进行着新的探索。哈尔滨工业大学的郎君、秦兵老师对相关的研究做过比较翔实的综述，他们将共指消解的研究划分为三个阶段：（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，引入背景知识以及语义知识，同时采用一些全局考虑篇章信息的方法来实现最优化的篇章共指消解[3]。他们的划分依据主要是底层的算法，为了更进一步地区分，我们结合算法及算法中采用的知

识，从两条两条主线来探讨实体名称识别的思路：（1）算法框架上的不断进化与调整；（2）算法中涉及的知识特征上的不断进化与调整。

一、算法框架上的不断进化与调整

Hobbs 算法是最早的共指消解算法之一，该算法主要基于句法分析树进行相关的搜索。而中心理论在更侧重于在给定的句子中跟踪实体的变化焦点。这两类算法都是比较经典的理论算法，使用可操作性并不是特别强，关于这两类算法的概述可以参见郎君老师的研究[3]。在 Hobbs 算法之后，基于二元对的分类方法以及基于向量相似度的聚类方法逐渐开始大规模地出现在研究中。这两类方法结合机器学习方法，在共指消解方面取得了很多研究进展。首先将二元分类思想引入到共指消解问题中的 McCarthy 和 Lehnert，采用了决策树 C4.5 算法，他们将篇章中任何两个不在同一共指链中的提及描述构成反例，反之，位于同一共指链上的则构成正例，通过大量的训练实例，构建反例库和正例库，再通过分类方法对测试实例进行归类，从而实现共指的判定[4]。在这一方法的基础上，很多的研究者进行了更深入的改进调整。郎君老师将这一类方法的基本框架进行了归纳，形成了如图 1 所示的经典框架。

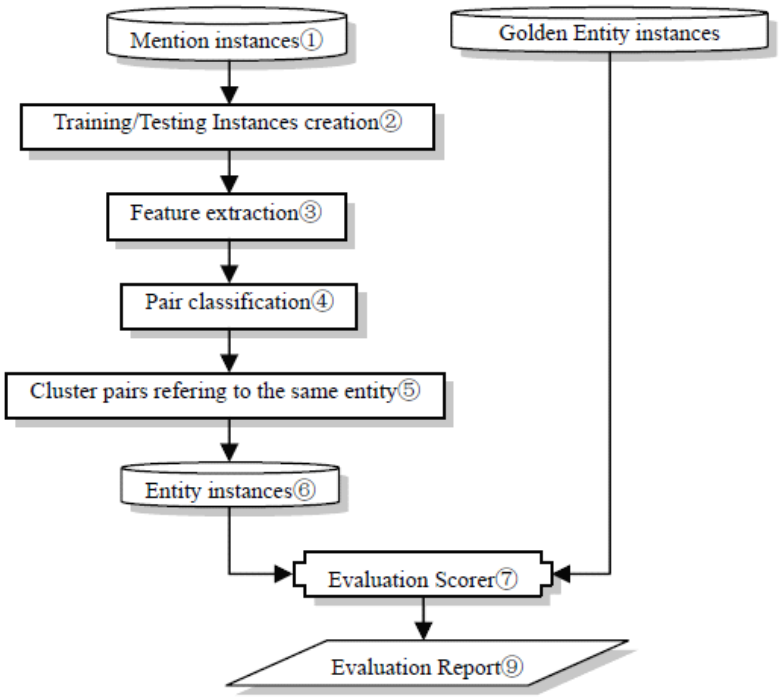

图 1 基于二元分类的共指消解经典框架[3]

在这一框架中，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)。相关预处理主要包括断句、词性标注、命名实体识别、嵌套名词短语识别等。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。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。针对训练和测试分别采用不同的实例构建方法。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。本文的各种背景语义特征在共指消解上的应用就主要体现在这个环节。图中④表示二元分类的机器学习算法[3]。

为了避免分类算法中所需要消耗的大量人工标记语料的准备工作，也有不少研究者在不需要训练语料的自动聚类方法方面做了不少工作。具体内容可以参见 Cardie and Wagstaf5、Finley and Joachims6等人的研究。

在单独的缩略语识别任务中，主要包含的工作一是确定某个词是否为缩略语，二是查找该缩略语对应的全称概念。因此，在确定缩略语过程中，一般选用模板匹配的方式，通过一定的语言学特征，如全大写字母、大小写混合、音节、词的前缀后缀、词根等方法，构建相应的规则进行识别。确定出缩略语之后，一种思路是从利用构建缩略语的规则方式将候选的全称 进行缩略语的构建，从而匹配构建出的缩略语与待判定缩略语之间的相似度，另一种思路则是利用上下文规则模板（如全称与缩略语出现的顺序、全称与缩略语的类别等）或动态的全局规划对齐，判定模板的匹配度。在这一方面主要可以参照 Yun Xu 和 ZhiHao Wang7，Naoaki Okazaki8等人的研究。

二、算法中涉及的知识特征上的不断进化与调整

上文中已经提到，要辨析词语的真实含义需要涉及很多知识，这些知识作为词语在文本中对外表征出来的特征集合，在辨析过程中发挥着重要的作用。传统的方法体系中，基于语言学特征是主体，在 Hobbs 算法中，句法分析结果即是其主要利用特征知识。随着研究的不断深入，研究者们在共指消解方面引入的知识特征越来越多。二十世纪九十年代时，一些研究者尝试着将共指消解的任务限定在单一的特定语境或语言知识内，同时这些研究也逐渐开始借助日渐强大的自然语言处理，如词性标注、浅层句法分析等设定一些过滤与筛选规则，然后对各种特征进行加权，入选的特征也逐渐开始包括人称类型、性别、单复数形式、句法角色、实体间的距离、字符串的匹配相似度等方面。这些研究主要可以 Mitkov9、王厚峰10等人提出的相关研究。随着各种语料资源库的不断开发和网络技术的不断发展，以 WordNet、WikiPedia、DBPedia、Yago、各领域本体、社会网络等为代表的语义背景知识资源库被纳入到了共指消解的研究中，研究者们利用这些语料库进行深层次的实体特征标注，获取实体同义词、上下位类词、共现背景词等特征，构建更为丰富的共指消解特征向量，从而提高实际的应用效果。具体而言，主要的几类研究包括：利用已经存在的知识体系（如 ontology、Wikipedia 等），将实体名称对应到其对应的知识体系概念下，利用知识体系中已经存在的属性、关系等信息进行统分类统计计算，判定实体名称具体的指称对象；利用实体名称的上下文信息，将其上下文中出现的词及该实体名称构建成向量空间模型或图模型，基于这些特征模型实施聚类分析，从而明确实体名称的具体指称；利用文本相似度、关注领域相似度、共同作者等信息，基于半结构化的文本（如引文条目）等进行人物名称的消歧判别。

1.1.3 实体名称规范的趋势

经过上述的研究，我们认为在共指消解的研究主要存在以下几种发展趋势。

一、算法方面：

在算法方面目前的研究中逐渐趋向于多模型融合的方法。在过去的研究中，基于语言学

特征的统计学方法和机器学习方法主流是分开思考的，很多研究都是在机器学习的分类或聚类中选择特征是再考虑加入一些语言学特征，这种融合方式对提高共指识别的效率比较有限。目前的研究中，研究者们逐渐开始考虑利用语言学思路来构建更加丰富的机器学习模型。Elango 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。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[11]。Poesio[12]等人将子句作为话语单元，将篇章可以表示成一系列子句的集合，进而将篇章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，还可以采用 Sutton and McCallum 讨论的技术[13]。

二、特征模型方面：

从当前发表的研究论文集中的研究主题上看，研究者们越来越重视在共指消解识别中引入越来越多的特征，单纯从算法上进行改进而实施基于“知识匮乏”的研究方法越来越不被主流研究所看重。而不断涌现出的各种语料资源库也为这些深层的语言学知识获取提供了非常好的途径。这些知识主要可以从以下三种途径获取：（1）常规的知识库。如 WordNet、HowNet、WikiPedia、DBPedia、Yago 等；（2）利用大规模的语料库挖掘模式信息。如 Hearst 等通过构建了“is-a”等模板，用于从文本中发现同义词14；Bergsma15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；Vincent16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 Su17利用语料库中发现的模板信息来增强共指消解 。（3）充分利用互联网这一语料库，利用搜索引擎显示的各个产寻得到的返回数来计算各种相关信息。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 Poesio 等人通过计算互信息来考察两个短语的关联程度。

1.2 国内外主要的实体名称规范项目

1.2.1 英国国家档案馆 TNA-Search 项目[18]

英国国家档案馆 TNA（the National Archives）中存储了大量政府信息，政府投入了大量经费资助在政府网站以公开可获取的格式发布越来越多的材料，以便民众更快更便捷地获取所需信息。但因为检索工具非常基本，仅支持基于关键词的检索。公众依然很难获取所需的信息，TNA-search 作为 Govenment Web Archive Project 中的一部分，主旨在于如何用简单直观的机制，提高 TNA 中与政府网站相关的记录（记录回溯到 1997 年，包含了大概 7 亿的网页）的开放利用度。TNA 通过使用关联数据原则和上百亿的事实，以简单可操作的形式，在一个可扩展的语义知识库中导入、存储并索引与 web 存档相关的结构化数据，在 web 存档文档与结构化数据间建立关联。在该项目中，大规模的语义标注是支撑整个项目的重要关键，在语义标注过程中，实体名称的规范又是关键问题之一。为了解决项目中的实体名称规

范问题，TNA-Search 项目主要利用 GATE，联合了 FactForge[1]和 SKB（Semantic Knowledge Base）Ontology[2]，构建了大规模的语义仓储库，通过仓储库所提供的详细的对象描述等背景信息，计算实现实体名称的规范。

具体而言，在语义标注过程中，该项目基于 LKB 直接将文档中的实体与各种不同的本体建立关联，或者通过其中的实例，或者通过概念。LKB 使用了一系列 SPARQL 查询集合的配置文件到 SKB 中检索。标注的实体与 SKB 中的实例关联是通过两个互补的途径完成的: 通过 LKB 词典找到一个匹配时，SKB 中类与实例信息被添加到文本中的相关实体上；文本中的实体与 SKB 中的类或实体没有直接关联时，通过共指的方式实现关联。即如果文本中某段提及在上述过程中已经与 SKB 建立关联时，该实体所有共指提及均可通过 TNA Instance Generator 自动获得相同类和实例信息。在进行规范标注时，项目将一篇文档中同一个实体的不同表达关联在一起，同时还添加通过 semantic tagger 发现的标注间的特征关系。该过程仅在同一类型的标注内开展，同时一篇文档中同一类型标注不超过 500 个实体时才进行。TNA Instance Generator：基于共指信息将 SKB 类与实例的 URIs 分配给相应的标注。例如：三个 Person 提及：“David Cameron” “D Cameron” and “Mr. Cameron”，并确定这三者为共指，其中“D Cameron” 已经由 LKB 创建一个标注。 TNA Instance Generator 则会将“D Cameron”的实例、类、URIs 信息 copy 给其它二者。三个 Person 提及:“David Cameron” “D Cameron” and “Mr. Cameron”，并确定这三者为共指，尚无确定的标注。 TNA Instance Generator 自动选择一个最长的表达产生一个 URIs。但 TNA Instance Generator 不会为识别出的有歧义的实体或仅由一个单词组成的名字创建新的 URLs。

通过这种规范标注方式，TNA-Search 实现了人物、地理名称、机构、时间等 11 种命名实体的自动标注与规范。

1.2.2 OKKAM[19]

OKKAM 是由欧盟委员会资助的第七框架项目(FP7)下的一个大规模集成项目,其基本理念是,根据 14 世纪的“奥卡姆剃刀(Occam’s razor)”原则,提倡如果没有必要则不增加实体的标识符。OKKAM 为内容创建者、编辑和开发人员等提供一个全球性的基础设施,称为实体命名系统(entity name system,简称 ENS),用于帮助人们便捷地查找相关实体的公用标识符.其中,实体命名系统包含了一种基于特征的实例匹配方法 FBEM，该匹配方法通过集成两个实例标识符的多种不同特征属性及其属性值之间的相似度，识别出可能的对象共指.例如,FBEM 使用了基于 Levenstein 编辑距离的方法来比较实例标识符的本地名。

[1] OntoText 开发的知识库，该知识库包含了超过 22 亿声明和来源于多个源的数据集

[2] OntoText 开发，基于 CGO（Central Government Ontology）与 UK 政府的官员职位、8138 个官员名字以及无歧义的 UK 政府机构名称

1.2.3 Sig.ma[20]

Sig.ma 主要由爱尔兰 DERI 研究所设计开发，提供了一个在线的语义 Web 数据聚合(mash-up)服务，其原始数据主要来源于 RDF 数据和网页中的 RDFa 或 Microformat 格式的数据.Sig.ma 以关键词作为输入，输出为聚合后的语义 Web 数据。严格意义上说，Sig.ma 并不完全是对象共指的消解系统，但是它提供的数据浏览能力使用户能够查询对象共指。技术实现上，Sig.ma 主要使用了 owl:sameAs 和反函数型属性以及查询 OKKAM 系统来发现对象共指。另外，Sig.ma 系统提供了精心设计的用户界面，允许用户通过交互反馈来过滤错误或不一致的 RDF 数据和数据源,以提高共指消解的准确度。在该项目中，用户的手动消解冲突是非常重要的特征之一。

1.2.4 国内典型的共指消解项目

共指消解是文本处理中非常重要的任务之一，它对于提高信息检索的效率、深度的文本挖掘有着非常重要的作用，国内目前在此方面也有不少相关的研究项目在开展。本项目主要选取了清华大学的 RiMoM[21]和南京大学的 ObjectCoref 两个作为代表进行介绍。

RiMOM 是清华大学研发的一种集成了多种本体匹配方法的多策略本体匹配系统，其中也包含了多种实例匹配方法。针对实例匹配，RiMOM 将每个实例所含信息分为 6 类：URL、元信息、名称、字符串类型信息、非字符串类型信息和邻居信息。通过基于编辑距离的方法和向量空间模型，计算实例所含各种信息之间的相似度，并使用元信息和非字符串类型信息进一步过滤,最后通过多种策略将各种相似度集成起来用于发现对象共指。

与 RiMoM 不同，南京大学的 ObjectCoref 基于语义 Web 搜索系统 Falcons 提供的数据集，目前已经包含 7300 多万个实例标识符。ObjectCoref [22]首先利用语义等价推理，包括 owl:sameAs、函数型或反函数型属性以及基数或最大基数限制，构建出一个初始训练集; 随后，基于这个训练集不断学习,自举式地识别对象共指，其中的关键技术是从训练集中学习出最适合识别对象共指关系的属性及属性值。该系统还考虑了频繁属性组合，同时使用两个属性识别对象共指(例如经度和纬度、 姓和名),进一步提高消解的准确度。另外，还基于语义等价关系是否可以解引以及实例标识符在不同 RDF 文档中的出现次数等，对共同指称同一对象的实例标识符进行排序。ObjectCoref 提出了一种新的语义等价推理与相似度计算相集成的体系结构，能够较为全面地识别对象共指，但是训练集中的错误共指关系可能会导致学习过程中的错误积累。使得识别的准确性降低。

1.3 国内外主要的实体名称规范评测会议

为了促进共指消解研究的不断发展，国际上有不少与之相关的评测会议，这些评测会议通过细化评测任务，提供相应的语料集合，提供交流的平台，推动者相关研究的不断发展。

本项目中筛选了几个比较典型的评测会议进行了介绍，以期为其它研究提供一些参考。

1.3.1 自动信息抽取-Automatic Context Extraction(ACE)

ACE 会议是从 1999 年 7 月开始酝酿，2000 年 12 月正式启动，由美国国家安全局（NSA），美国国家标准和技术学会（NIST），以及中央情报局（CIA）共同主管，到今年为止已经举办过 8 届23。测评中需要的大量训练集和测试集均由语言资源联盟（Linguistic Data Consortium, 简称 LDC）24提供。ACE 主要关注 6 个领域的信息：网络上的专线新闻（Newswire）、通过 ASR（自动语音识别）得到的广播新闻（Broadcast Conversations）、及通过 OCR（光学字符识别）得到的报纸新闻（Newspaper）、新闻组（Usenet）以及对话性的电话谈话（Conversational Telephone Speech）和网络日志（Weblog）。其测评任务定义为：实体探测与识别(Entity Detection and Recognition ，简称 EDR)、价值探测与识别(Value Detection and Recognition ，简称 VAL)、时间表达识别与标准化(Time Expression Recognition and Normalization ，简称 TERN)、关系探测与识别(Relation Detection and Recognition ，简称 RDR)以及事件探测与识别(Event Detection and Recognition，简称 VDR)25。测评系统要求按照用户制定的内容种类进行抽取，采用了基于漏报（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）为基础的一套方法。ACE 的目标是发展包括自动识别和标识文本在内的自动内容抽取技术，以支持对语料库的自动处理。

在该评测会议中，共指消解的评测任务主要蕴含于实体探测与识别中，即 EDR 中。该任务将篇章中出现的各种提及表述指向对应的实体，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，该合并过程就是共指消解的过程。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 5 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。

在 2008 年之后，ACE 会议被 Text Analysis Conference（TAC）[26]会议所取代，在该评测会议中，TAC-KBP 从 2009 年开始到目前共进行了四届，该评测任务包括实体链接(Entity Linking)和实体属性值抽取(Slot Filling)，数据来源是新闻和网络数据。

1.3.2 web 环境中人名消歧任务评测会议-Web People Search Evaluation（WePS）

WePS 是针对英文网页中人名消歧任务进行评测的一个专门会议，有 Julio Gonzalo 和 Satoshi Sekine 主要负责组织，至今为止共组织过三次[27]。该任务集中于在 web 检索场景中人名的消歧。参加测试的系统将在接收到一个以人名为检索式的 web 检索后，确定有多少个不同的涉及的人员在检索结果中，并将特定的指称分配给相应的文档。从总体上来说，这个任务是个聚类问题。对给定的一组文档，按照文档中出现的某个指定的人名所指向的人进行聚类。最后，在每个类中，所有指定的人名都必须是指向现实生活中的同一个人。从 WePS3

发布的评测任务看，在该评测中，需要重点从人物的属性角度出发，包括人员的生日、出生地、别名、工作、所属机构、获得奖项、学校、学位、专业、民族、电话等多个方面年代信息。受该项目启发，李文捷等人也于2010年组织发起了专门针对中文人名消歧的评测任务[28]。

1.3.3 指代消解练习（ARE）[29]

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。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:

- 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。
- 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。
- 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。
- 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。

除上述的三种不限于领域的评测外，还有一些领域特定的共指消解任务评测，如生物医药领域的生物医药领域的自然语言处理及应用联合工作组 JNLPBA(Joint Workshop on Natural Language Processing in Biomedicine and Its App lications) 和以及生物学领域信息抽取的关键评价 BioCreAtIve(Citical Assessment of Information Extraction Systems in Biology)。

1.4 结语

本小节围绕着实体名称规范的概念、传统的思路演化、当前主流的研究方法、当前的发展趋势等几个方面进行了概括性地探索分析，理清了实体名称规范的任务内涵，主要的一些思路方法。为了进一步了解国内外的发展现状，项目组还调研了目前国际上相关的项目与评测会议，从而为进一步的研究提供了基础。

2 实体名称规范的关键技术路线研究

2.1 基于 Web 对象属性信息的实体名称规范研究

Web 页面中往往嵌入了各种各样的对象，如人、产品、组织机构等。从 Web 页中抽取并集成这些对象，可以实现功能强大的对象层内容揭示。

Zaiqing Nie[3031]等认为，Web 对象是描述某一 Web 信息的数据单元，通常可以看作是与应用领域相关的概念。一个 Web 对象可以通过一系列的属性表示，如 A={a1, a2,……, am}。对象的属性集可根据领域的需要预先设置。具体识别过程主要包括：（1）Web 对象的抽取，从多个来源的数据中抽取出对象及其属性；（2）Web 对象的标识和集成，将每个抽取出的对象实例映射到现实世界中并存储到数据仓储中，需要对同一对象进行识别，共用一个对象标识。基于此，Zaiqing Nie 等提出分两步实现 Web 对象抽取的方法，如图 2 所示。

第一步，实现对象记录级别（record-level）的抽取。Zaiqing Nie 等将 Web 上一系列有一定结构的相同条目（如产品列表、服务列表等）称为数据记录。抽取 Web 对象时，先从数据源中抽取出与领域相关的数据记录，形成对象记录级别的标识。每一个对象被表示成一系列被抽取出来与对象相关的数据记录，在此级别中不识别对象的具体属性。

第二步，实现对象属性级别（attribute-level）的抽取。这一过程中，需对上一步抽取出的数据记录进行分析，将数据记录中的不同部分标识成为不同的属性，并且从多个来源的记录中，实现同一对象不同属性值的获取。

基于这种思路，Zaiqing Nie 等人进行了 Windows Live Product Search 项目实验。该项目从 Web 数据源中，自动抽取出大规模产品对象（record-level）。用户查询某一特定产品时，系统返回的并不是相关的 Web 页面，而是一系列与所查询内容相关的产品，及产品的标题、图像、价格及特性等属性信息。

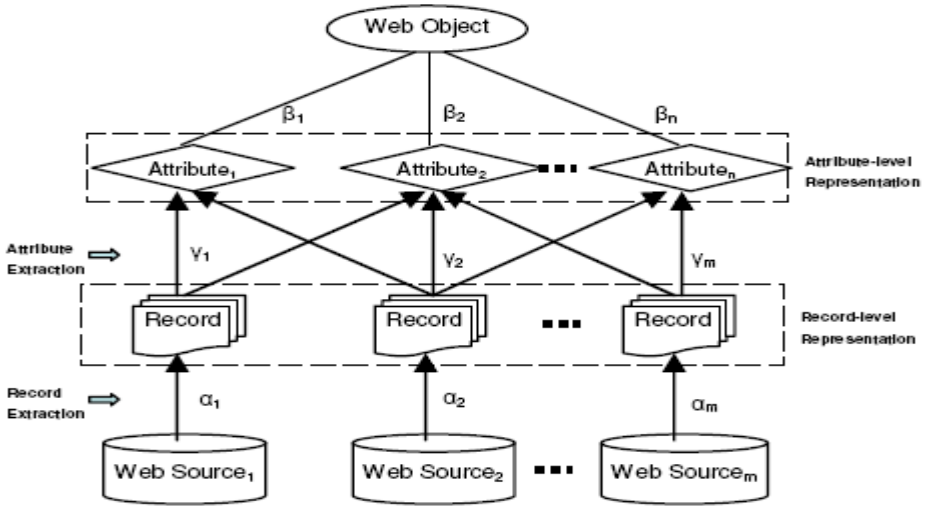

图 2 Web 对象抽取技术框架

此类方法的优势在于其来源数据的特殊性。这些来源于 web 网页的资源在获取其属性方面具有很大的便利性，从而为基于属性模板的共指消解提供了很大的便利条件。

2.2 基于 Wikipedia 的实体名称规范研究

命名实体消歧的关键问题是测度实体名称出现的相似度，传统的测度方法是利用 BOW 模型，但它忽略了语义关系，为弥补以往方法的不足，不少学者提出了利用 Wikipedia 作为共指消歧的背景知识，因为它覆盖很多概念，每篇文章中都包含了一个实体或一个概念的信

息，具有丰富的语义信息且内容时时更新。

Xianpei Han 通过扩充 Wikipedia 的语义知识，纳入各种语义关系提出了可准确测度实体名称之间相似度的新方法32。Silviu Cucerzan 提出了基于 Wikipedia 数据的命名实体识别和语义识别的大规模系统，该系统是通过最大化描述（maximizing the agreement）从 Wikipedia 抽取的语境信息和文献中的语境以及与候选实体相关的类标签来提高消歧精度33；Anthony Fader 等介绍了 GROUNDER 系统，通过利用 Wikipedia 上用户贡献的信息和新的消歧模型，有效利用先验信息，组合先验信息和语境信息以提高消歧精度34；Hien T 等人将文本中提到的实体映射到 Wikipedia 中正确的实体，在基于候选实体统计秩序模型基础上，证明 Wikipedia 和文本的功能组合是消歧的最好选择35；Danica Damljanovic 则认为 Linked Data 是扩充已可用语境的有效资源，并将先进的命名实体工具与基于 Linked Data 相似度测度方法进行结合，证明该方法能提高 Wikipedia 消歧精度36；Danuta Ploch 等人将实体名称消歧看做是将文本中的实体提及与预定义在知识库中的指称词相关联的任务，他们在研究中通过挖掘共现的实体间在 Wikipedia 里的关联关系，通过实体共现与歧义形式的关系推导出可用于分类候选实体的功能范围，并将消歧功能进行组合，利用 SVM 分类器得到了有效的结果37。Anna Lisa Gentile 等人提出利用语义关系图中得到的语义关系分值来进行命名实体消歧，该方法只需考虑文本中的命名实体，无需建立 BOW 也能得到精确的结果38。Razvan Bunescu 等人通过训练一种消歧 SVM kernel 方法，以开发在线 encyclopedia 编码的高覆盖和丰富结构的知识库39。Hien T. Nguyen 等人利用共现实体关系，结合级别加权和共指方法来进行命名实体消歧，此方法不限制实体级别，不要求结构化的文本[1]。

2.3 基于本体的实体名称规范研究

本体作为经过人工加工过的一种包含了丰富的语义信息的语料库，可以为名称实体规范的相关研究提供非常有价值的语义信息。因此，基于本体的实体名称的规范研究也成为当前的关键技术之一。

Horacio Saggion 等基于欧盟的 MUSING3平台，在跨数据源的知识单元获取与集成任务方面做出了一定探索。整个过程分为两个部分，一是基于本体的信息抽取，二是基于本体的跨数据源对象集成。其中，由领域专家构建的商业本体是系统的首要特征。在具体实验中，其主要思路如下（图 3）：

（1）定期收集 Yahoo! Finance、100 多个公司站点及该项目合作公司提供的报告或新闻报道等数据源，存储于 MUSING 文档库中。基于 Proton4设计出适用于商业领域的 Ontology，该 Ontology 包含商业领域的类层次结构、关系和属性；其定义的对象主要包括：公司名、公司雇员数目、公司地址、网址、电话、传真和盈利状况等。

（2）基于 GATE 提供的自然语言处理平台，根据定义的对象实现规则等方面的扩展，

[3] MUlti-industry, Semantic-based next generation business INtelliGence，基于语义的下一代多产业商业情报
[4] SEKT 项目的成果之一，是一个用于知识管理的本体，已经在 OntoText 的 KIM 项目中得到应用。

生成标注系统。在扩展时，需要注意该系统与本体的兼容，同时提供与领域专家的交互接口，以便专家在标注过程中及时调整标注结果。

（3）针对每篇文档标注的结果，获取各标注对象所在的文档和描述内容部分，计算其相似度，实现多数据源中同一个标识对象的聚类。通过本体映射模块，将自动标注出的结果分别映射到预先定义的本体类和属性中，生成本体。

（4）获取的有效数据最终以结构化方式存入知识库中，提供查询和推理。根据这些语义数据，可以生成公司的商业报告，亦可提供区域内特定经济指标下的公司排名等。

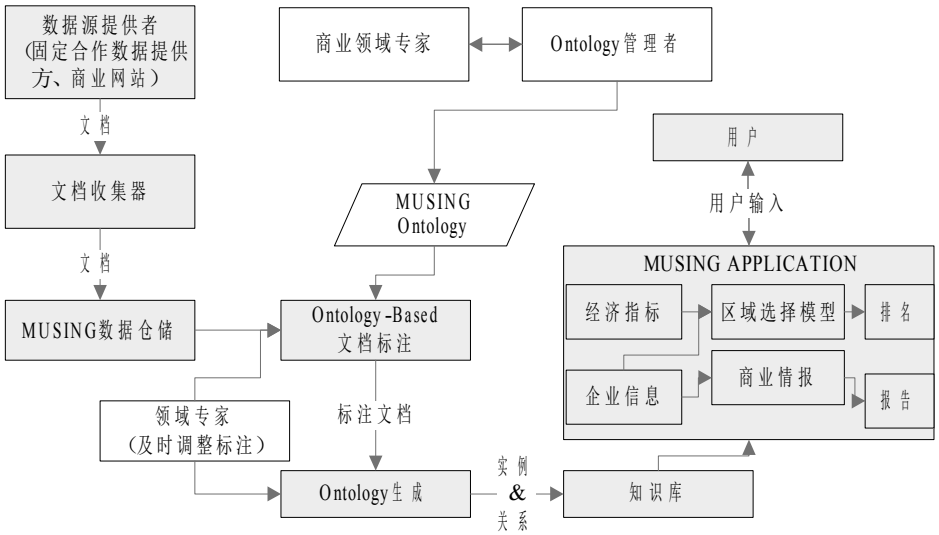

图 3 MUSING 系统框架

清华大学研发的 RiMOM 也集成了多种本体匹配方法，其中也包含了多种实例匹配方法。针对实例匹配，RiMOM 将每个实例所含信息分为 6 类：URL、元信息、名称、字符串类型信息、非字符串类型信息和邻居信息。通过基于编辑距离的方法和向量空间模型，计算实例所含各种信息之间的相似度，并使用元信息和非字符串类型信息进一步过滤,最后通过多种策略将各种相似度集成起来用于发现对象共指。

英国 Open 大学开发的语义数据融合系统 KnoFuss[40]主要基于本体标注的语义数据集的融合或互联解决数据集成问题，其中的关键位问题即重点考虑对象共指的消解和知识库的更新问题。在实际的共指消解过程过程中，KnoFuss 特别关注本体类和属性的匹配与对象共指消解之间的相互促进。具体而言，KnoFuss 通过对象共指来发现类和属性之间的匹配，再利用这些类和属性的匹配来进一步提高对象共指消解的准确度以及扩大其覆盖面。其中，对象共指的识别用到了多种 OWL 语言定义的原语和信念传播(belief propagation)算法。

此外，WordNet 和其扩展作为本体的一种，在共指消解中被广泛地用于获取同义词、上下位类、背景相关词的语义扩展等，其在基于本体的实体名称消解研究中也占据着非常重要的席位。

2.4 基于社会网络的实体名称规范研究

随着搜索引擎和社会网络挖掘技术的不断发展，利用人物社会关系关联构建社会网络，进而实现相应的实体消解方法也逐渐成为目前的关键思路之一。

RonBekkerman[41]等人提出了一种非监督的框架来解决检索某个特定人物时返回大量无关人员页面的问题。其中两个关键内容包括网页间的链接关系与 Agglomerative 重复聚类。在该方法中，网页间的链接关系即主要用于构建人物的社会网络。

郎君[42]等人依据同名的不同人物具有不同的社会网络的思想，利用检索结果中共现的人名发现并拓展检索人物相关的潜在社会网络，结合图的谱分割算法和模块度指标进行社会网络的自动聚类，在此基础上实现人名检索结果的重名消解．在人工标注的中文人名语料上进行实验，整体性能达到较好水平，图聚类算法能帮助连通社会网络的进一步划分，从而提高消解效果。图 4 是典型的基于社会网路的共指消解框架。

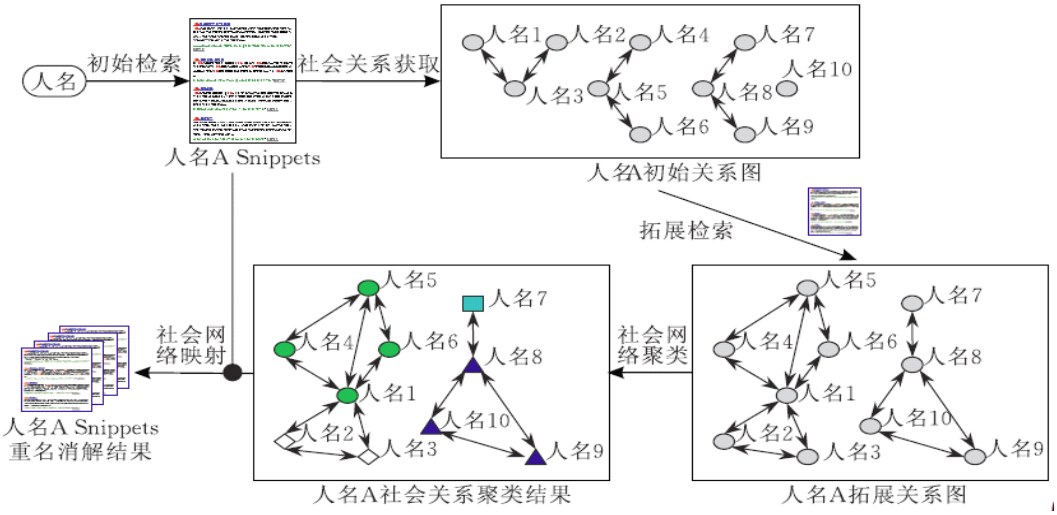

图 4 典型的基于社会网路的共指消解框架

陈晨等人先使用谱聚类对社会网络中的人名聚类,然后根据不同社会网络边权值和不同图划分准则对人名消歧效果的影响,引入了模块度阈值作为社会网络划分的停止条件，在共指消解方面取得了较好的效果43。

2.5 结语

本章在第一章对实体名称规范的概念、理论和趋势调研的基础上，进一步围绕目前的一些关键技术方法进行了深入探讨，希望能从中对本项目的实施获取到相应的启发，从而推动本项目的开展。

3 实体名称规范中涉及的实体特征研究

通过上述的调研分析，我们可以发现，在实体名称规范相关的研究中，非常重要的内容之一是实体特征集合的筛选。对这一部分进行深入的探讨将为本项目在设计相应的特征模式时提供很大的帮助。综观当前实体名称规范相关的研究中理论方法的探索，目前主要采用的指标主要包括三个方面的特征：词汇特征、语法特征（这两者又可以归纳为语言学特征）、语义特征、语用特征三类。不同的研究中对这几类特征的子特征归纳又有所不同，本章将主要从这四个方面结合其获取途径来探讨实体名称规范中涉及的相关实体特征。

3.1 语言学特征

语言学特征是实体名称规范中使用时间最长的特征，从早期的句法树到后来的浅层文本处理，语言学特征在该任务的发展中占据着非常重要的地位。具体而言，该类特征又可以划分为词汇特征和语法特征两类。

3.1.1 词汇特征

对词汇特征而言，主要是指字符串的匹配与别称的匹配两种。

在字符串匹配方面，主要是指待消解的两个共指元 A 与 B 在去掉其中包含的各种非实意词（包括冠词、介词、助词、虚词）等之后，同时将 A 与 B 进行相应词形上的处理（主要指单复数的处理），如果剩余的次能实现超过一定阈值的匹配，则认为 A 与 B 是共指的。这种特征事实上在实际计算中往往会导致大量的无关共指对的出现，阈值的设定该特征中一个关键之处。这类特征主要通过分词来获得，在处理过程中可以配以相应的停用词表来辅助预处理工作的完成。

除字符串匹配外，基于别称词典的匹配也是词汇特征之一。这种特征方式主要是通过构建一些别称词典，直接在词典中匹配 A 与 B，若完全吻合，则实施共指。这类特征往往受词典的规模的限制，一方面，读取词典是一个很消耗资源的问题，另一方面，词典的编制、整理也是一个非常繁复的过程。更为重要的是，有可能会出现不同对象有相同别称的情况，因此，这类特征在使用中需要配合其他特征进行筛选。这类特征主要通过人工编撰相应的词典工作来获取。

3.1.2 语法特征

语法特征主要包括句法角色功能、句法关联关系两类。句法角色功能主要指其实体在句子承担的语法功能，包括主谓宾三种。这类的特征通过句法分析即可以获取，通过此类特征，

仅仅能确定共指实体元的角色功能，并不能有效单独地确定其是否共指。而句法关联关系主要指共指实体元之间相隔的距离，如果间隔大于一定的阈值，则可以忽略该共指对。

3.2 语义特征

语义特征是近年来在共指消解领域备受关注的一个特征类型。在该类特征下，主要的子特征类型包括：语义标注信息、属性信息、上下位类、同类词信息、关联关系信息、上下文背景信息等几类。

3.2.1 语义标注信息

语义标注信息即实体的语义类别，如机构、人物、会议、地理名称等详细的语义标注，这类信息往往需要借助于信息抽取来完成。这类信息的存在可以将共指元对现定于同一种语义类别的集合中。

3.2.2 属性信息

在语义标注过程中，往往还需要完成的一项很重要的工作即实体属性的标识。在实际的文本中，往往在实体名称出现的不同位置上提供了很多描述性信息，这些描述性信息可以有效地完成对该实体的属性描述，如通过称呼“Mr”、“Miss”等关键词，可以区分出的性别属性，通过“Present of”等职称属性可以区分出人物的职业属性，这类信息往往也是在语义标注的过程中进行完成。

3.2.3 语义相关性信息

这里的语义相关性信息主要包括了上下位类、同类词信息几类。在实际的标注过程中，有可能我们无法从实际的文本中抽取出足够的属性描述信息，那这里就需要借助于获取实体的上下位类、同类词等信息来进行综合的判定。这类信息往往是通过 WordNet、HowNet 甚至领域本体来获取。借助于一些语义语料库，我们可以获取好的语义相关性集合信息。

3.2.4 上下文背景信息

这里的背景信息主要是指与共指元在文本中一定范围内共同出现的实意词的信息。如“the National Association for Patient Participation promotes and supports patient participation in primary care”一句中，与该机构同时出现的往往有 patient、primary 一类的词，若在“NAPP”出现的页面正文中也经常出现这一类的词，就可以认为这两个名称的关系非常密切，很可能指代同一个实体。共现背景词特征表示如下：

A_contextWord：取缩略语周围的几个词（设定为前后各 4 个窗口词）作为上下文，这里需要去除连词、介词、形容词、冠词等虚词（项目组收集整理了相应的统一停用词表用于这类词的筛选），仅仅保留实义词。在其可能匹配的全称页面上，判定是否含有这些词中的一个或多个。

F_contextWord：取候选的全称周围的几个词（取词方式与 A_contextWord）相同，然后在其可能的缩略语页面上，判定是否含有这些词中的一个或多个。

3.2.5 其它

除上述的一些语义信息外，基于关系挖掘的共指消解研究中还提出了通过共现的实体关系对来获取相应的语义知识的方法。这主要是指通过关系抽取，识别出与共指关系元分别共现的其它实体名称，从而构建共指实体的特征集合，通过计算共指实体的重合度来判定共指的可能性。

3.3 词用特征

词语在不同的使用语境中往往会呈现不同的含义，这也是人工智能研究中最难以处理任务之一。目前在语用方面主要的研究还集中于领域的相关性等方面。为了简化研究任务，研究者们在开展共指消解研究时往往会将研究限定于某一个领域中。因此目前在这一方面的研究并不多见，在实际使用中，主要的使用途径是结合术语抽取的工作，将与实体共现的篇章中的术语构建成相应的领域特征集合，再进一步地匹配领域特征集合的相似度。

3.4 其它

除上述说总结的几类特征外，事实上在实际的文本中还存在着一些共指模板的特征，这一类的共指模板可能是通过位置、特殊标记构成，也可能通过一些词组模板链接。对位置而言，即指两个共指元在篇章中是否是否有前后相关联的位置、结构，特殊标记则指的是是否有-、()这一类的特殊标记。通过词组的模板往往是指“also called”、“named”等一类的表达。这一类的模板需要借助于人工进行很好的整理收集，通过规则的方式在实际的共指任务中发挥作用。

3.5 结语

本章在前面几章调研的基础上，对实体名称规范中涉及的实体特征进行了更为深入的调研分析，从语言学、语义、语用几个方面分别讨论了其所包含的子类特征的具体含义、获取方式，可能存在的缺陷等问题，以便为本项目中特征模型的构建提供有力的参考。

4 实体名称规范的相关资源研究

与实体名称规范相关的研究至今已经开展了几十年，在这几十年的发展中，逐步形成了一些较为成熟的工具与语料库，利用这些工具与语料资源进行二次开发或特征获取，将为相关的应用研究提供很大的便利，从而节约大量的时间、人力和物力。本章即从国内外典型

的实体名称规范工具与语料库两个方面进行了相应的调研归纳。

4.1 国内外典型实体名称规范的工具分析

4.1.1 GATE[44]

一、GATE 概况介绍

经过 10 多年发展，英国谢菲尔德大学研究开发的 GATE(General Architecture for Text Engineering, 文本工程通用框架)不断发布新版本的同时，凭借优秀的组织架构和开源的优势，在科研、教育、商业等领域获得广泛应用。为简化语言工程系统开发流程，GATE 构建了“算法＋数据＋图形用户界面＝应用”基本结构。按照此结构，GATE 选用面向对象的编程语言和基于 JavaBean 组件的软件开发方式，开发出一个核心库和一系列可重用语言工程组件（CREOLE，a Collection of Reusable Objects for Language Engineering）。每个 CREOLE 组件包括语言资源（LRs）、处理资源（PRs）和可视化资源（VRs）三类资源，资源参数存储于 creole.xml 文件中。用户可根据应用快速灵活定制、修改、扩展各组件。图 5 中 GATE 详细架构清楚展示了 GATE 中各类资源所属层级。

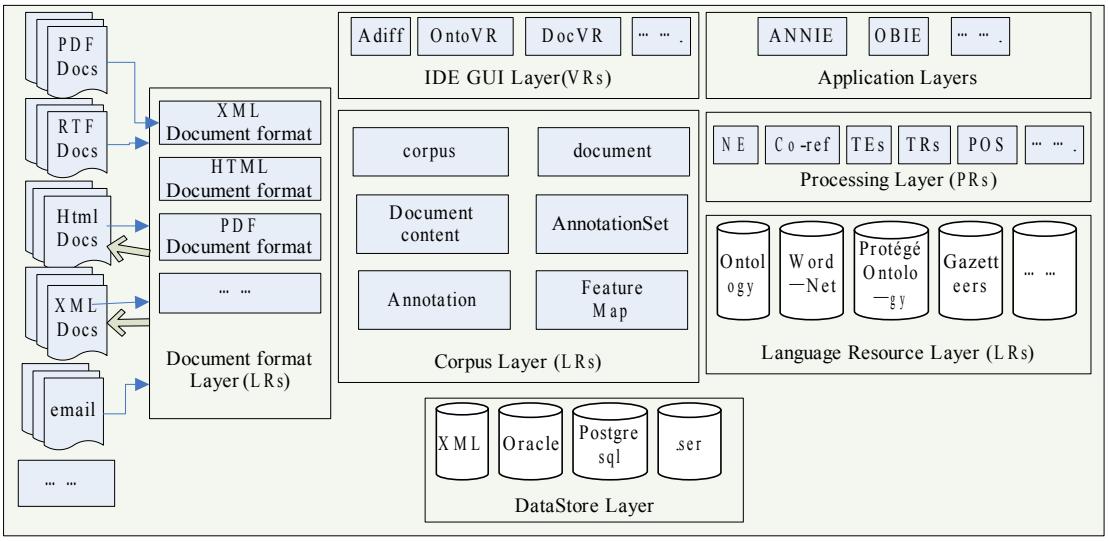

图 5 GATE 详细架构

经过十多年发展，GATE 逐步完善。目前发布的 7.1 版本提供 30 多个组件，并集成 Lucene、Sesame、Minipar 等众多相关领域组件接口，可处理包括 XML、HTML、email、纯文本等 7 种格式文本，对标注结果提供关系数据库、java 序列化、xml 存储几种存储方式。

凭借优秀的组织架构和开源的优势，GATE 已在众多领域发挥作用。如 KIM（Knowledge and Information Management，知识和信息管理）系统中，GATE 作为底层的实体标注工具，为构建知识库实现语义检索等提供支持45。Horacio Saggion 等人利用 GATE 的实体标注和

Ontology API 构建 MUSING5，自动查找、收集、分析商业报道，为商业决策服务46。英国国家档案馆 TNA-search 项目中也将 GATE 作为很重要的语义标引工具，为后续的语义索引和语义检索提供支撑。

二、GATE 中实体共指消解的组件与解决思路

在上文中已经说过，GATE 工具各模块的组织方式采用了组件的方式，在 GATE 中提供实体名称共指消解的模块位于其核心组件 ANNIE（a Nearly-New Information Extraction System）中，该核心组件提供了分词、分句、词性标注、词典匹配、规则匹配等系列浅层信息处理功能，用户可以按照自己的需求，进一步对其提供的词典、规则进行更为深入灵活的修改。其中，共指消解模块名称为 othomatcher，该模块中的资源由一系列的词典组成，用户可以将一些确定的无歧义的特定实体名称表达共指对设置在相应的词典中，从而在处理过程中进行相应的匹配。在实际处理是，由于存在更多无法通过词典直接匹配的信息，GATE 的解决思路如下：

GATE 在进行实体标注时，尽可能多地标注了一些基本命名实体（主要是人物、时间、地理位置）的属性信息，如人的职称、性别等，在处理过程中，GATE 首先从词典里进行完全匹配，若无匹配项，则利用属性信息进行综合的匹配处理。

4.1.2 Stanford Deterministic Coreference Resolution System[47]

Stanford Deterministic Coreference Resolution System 是由斯坦福大学自然语言处理小组开发的一款共指消解系统，该系统中集成了多种共指消解模型，其中包括了语言学、语法、语义等信息。所有这些模型中都通过在篇章全局模式下共享提及属性（如性别、数字等）来实现。其中语言学信息主要包括了距离、单复数、相关字符串等方面的信息，在语义方面主要借助了 WordNet、Wikipedia infoboxes 和 Freebase 记录来获取相应的背景信息。在该系统中，可以自由加入其它类型的知识库。这为其它系统的灵活扩展提供了很大的便利。

4.1.3 Illinois Coreference Package[48]

Illinois Coreference Package 是由 Illinois 大学 E. Bengtso 研究开发的一个共指消解工具包，该工具包中包含了一个共指消解解决器和一个共指消解关联的特征集合。在该工具包中主要用到的特征包括性别、数字匹配，在语义方面，该工具包也选择从 WordNet 中获取同义词、上下位类词等信息，在该工具包中解决的共指实体类别主要参照了 ACE 的语义类表。在实际实现是，该工具包通过机器学习的方式训练学习出一个消解分类器，从而应用于测试集上进行共指对的判定。下图展示了该工具包中对各种消解特征的分类、特征获取来源方式与详细说明。

---

[5] MUlti-industry, Semantic-based next generation business INtelliGence，基于语义的下一代多产业商业情报

Category	Feature	Source
Mention Types	Mention Type Pair	Annotation and tokens
String Relations	Head Match	Tokens
	Extent Match	Tokens
	Substring	Tokens
	Modifiers Match	Tokens
	Alias	Tokens and lists
Semantic	Gender Match	WordNet and lists
	Number Match	WordNet and lists
	Synonyms	WordNet
	Antonyms	WordNet
	Hypernyms	WordNet
	Both Speak	Context
Relative Location	Apposition	Positions and context
	Relative Pronoun	Positions and tokens
	Distances	Positions
Learned	Anaphoricity	Learned
	Name Modifiers Predicted Match	Learned
Aligned Modifiers	Aligned Modifiers Relation	WordNet and lists
Memorization	Last Words	Tokens
Predicted Entity Types	Entity Types Match	Annotation and tokens
	Entity Type Pair	WordNet and tokens

图 6 Illinois Coreference Package 各种消解特征的分类、特征获取来源方式与详细说明

4.1.4 CREAP[49]

CREAP 与 GATE 类似，也是一个以组件形式组织的共指消解平台，目前发布的版本为 1.0.0，该平台目前通过配置和开发相应的插件，实现添加新的特征和消解器等，已经可以支持所有的共指算法和策略等。该平台支持数据源为语料和纯文本，并且定义了语料的标准格式，为了便于用户将自己的语料转换成该平台所支持的标注格式，该平台还实现了几个常用的语料转换器，比如针对 ACE 语料的转换器，还有部分文本标注格式的转换器。

4.1.5 其它

除上面几种比较常见的共指消解工具或共指消解包外，还有如 BART ARKref、ARKref、MARS（Mitkov's Anaphora Resolution System）、Reconcile 等，这些工具在推动共指消解研究中发挥了很重要的作用。

4.2 国内外典型实体名称规范语料库分析

在实体名称规范识别中，知识是任务中的重要组件之一，知识资源为词语语义的关联提供了必要的数据。它们的范围非常广泛，覆盖了从文本集合（无论是未标注的还是标注过词语含义的）到机器可读的词典、叙词表、词汇表、本体等各种类型。本项目主要按照资源的结构化程度从非结构化与结构化两个方面对相应的一些语料资源进行了介绍，以期为其它项目提供一些参考。

4.2.1 结构化资源

一、叙词表。叙词表中提供了诸如同义词、反义词或者更多的词语间关联关系信息，比如（carn 是 motorcarn 的同义词、uglya 是 beautifula 的反义词）。在 WSD 领域中应用最为广泛的叙词表是 Roget’s International Thesaurus。该词表最新的版本中包含了 250，000 个词条，分别以 6 个大类、1000 个类目进行组织。当然也有研究人员在研究中选用 Macquarie Thesaurus，这个词表中包含了逾 200,000 个同义词。

二、机器可读的字典。从 1980 年第一部可以以电子形式获取的字典面世开始，这类字典就成为自然语言处理中一类流行的知识资源。在这些字典中，我们可以选取 Collins English Dictionary、the Oxford Dictionary of English 以及 the Longman Dictionary of Contemporary English (LDOCE)。在 WordNet 流行之前，LDOCE 一直是 NLP 研究团体中使用最为普遍的机器可读式字典之一。之后，WordNet50成为英语的词语含义消歧中使用最多的资源，WordNet 是由普林斯顿大学创建的一个基于心理语言学规则的英语计算辞典，它常常被认为是超出一般机器可读字典的一站式资源，因为它包含了概念间丰富的语义网络信息。正因为此，它常常被定义为计算辞典。利用 WordNet 词典，用户常常可以计算获取词语的相似度、上下位类、同义词等众多信息。目前最新的版本为 WordNet3.0，但该版本对 windows 平台适应性不行，在 windows 平台上广泛使用的仍然以 WordNet2.1 版本为主。其最新的版本 WordNet3.0 中包含了大约 155,000 个单词，这些单词以超 117,000 的 synset 进行组织。例如，automobile 这一概念由以下的 synset 方式表达（上标和下标分别代表了单词的语义标识符和词性标注）

$$\{car_{n}^{1}, auto_{n}^{1}, automobile_{n}^{1}, machine_{n}^{4}, motorcar_{n}^{1}\}.$$

我们可以把一个synset看做是表达同一个意思的词语意义集合，依据在2.1中介绍的标记，下述的公式将每一个词性标注过的词语Wp与其WordNet的含义集合进行了关联。

$$Senses_{WN} : L \times POS \rightarrow 2^{Synsets}$$

这里Synsets代表了wordnet中的synsets全集。例如：

$$Senses_{WN}(car_{n}) = \{\{car_{n}^{1}, auto_{n}^{1}, automobile_{n}^{1}, machine_{n}^{4}, motorcar_{n}^{1}\},$$
$$\{car_{n}^{2}, railcar_{n}^{1}, railwaycar_{n}^{1}, railroadcar_{n}^{1}\}, \{cablecar_{n}^{1}, car_{n}^{3}\}, \{car_{n}^{4}, gondola_{n}^{3}\},$$
$$\{car_{n}^{5}, elevatorcar_{n}^{1}\}\}$$

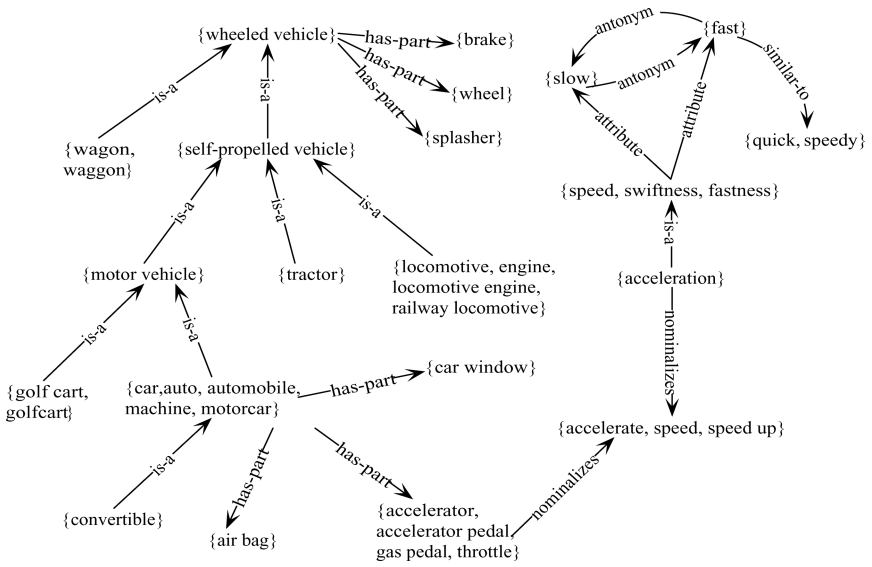

图7 WordNet语义网络片段

我们注意到每一个词语含义单一地标识了一个synset。例如，与 car_n^1 对应的synset $\{car_n^1, auto_n^1, automobile_n^1, machine_n^4, motorcar_n^1\}$ 是唯一确定的。在图3中，我们从WordNet的语义网络中摘录了一段包含 car_n^1 synset的内容。对于每一个synset，WordNet提供了以下相关信息。

——注释，也就是该synset的文本解释，可能还包含一系列的使用示例。（比如，car_n^1 的注释就是“a 4-wheeled motor vehicle; usually propelled by an internal combustion engine; ‘he needs a car to get to work’”）。

——辞典和语义的关联。通过这两个关联分别关联起词语的含义和synsets，其中语义关系用于全面的synsets，而辞典关系则用于连接包括各个synsets在内的词语含义。针对后者，我们有如下的解释：

- 反义词组：如果X表达了与Y相反的概念，则X是Y的反义词（比如 $good_a^1$ 是 bad_a^1 的反义词）。反义词组对所有的词性都有用。
- 包含词组：一个形容词X可以被定义为一个名词Y（或者极少数情况下另一个形容词）的一个部分或附属于Y（比如 $dental_a^1$ 附属于 $tooth_n^1$ ）。
- 名词化处理：名词X是动词Y的名词化形式（如 $service_n^2$ 是动词 $serve_v^4$ 的名词化）。

针对语义关联，我们有如下的解释：

- 上位关系Hypernymy（或者称之为kind-of或is-a）：如果每一个X都是Y的一种，那Y是X

的上位类（比如 $motorvehicle^{1}_{n}$ 是 car^{1}_{n} 的一个上位。）。上位关系存在于一对儿名词或动词synsets之间。

- 动词的下位关系（troponymy）或名词的下位关系（hyponymy），这两个关系分别是动词和名词synsets的上位关系的逆转。
- 部分关系Meronymy（也称为part-of）:如果Y是X的一部分，则Y与X是部分关系（比如 $flesh^{3}_{n}$ 是 $fruit^{1}_{n}$ 的部分）。部分关系仅存在于名词synsets中。
- 整体关系Holonymy。如果X是Y的一部分，则Y与X之间是整体关系（也就是Meronymy的逆转）。
- 蕴涵关系Entailment：如果做X必须先做Y，则动词Y就蕴涵于东西X中（例如，$sleep^{1}_{v}$ 蕴涵于 $snore^{1}_{v}$）。
- 相似关系。形容词X与形容词Y相近（例如，$beautiful^{1}_{a}$ 与 $pretty^{1}_{a}$ 相近）。
- 属性关系。一个名词X是形容词Y所传达的含义的属性表达。（例如，hot^{1}_{a} 是 $temperature^{1}_{n}$ 的一个值。）
- 又见。这是形容词之间关联的一种关系。（比如，$beautiful^{1}_{a}$ 和 $attrictive^{1}_{a}$ 通过又见的关系发生关联。）

Magnini 和 Cavaglia 针对 WordNet synsets 开发了一个领域标签的数据集。WordNet synsets 已经被半自动地赋予了一个或多个领域标签，这些标签来源于杜威十进制分类法中预定义的 200 多个标签（如，Food，Architecture，Sport 等），如果在杜威十进制分类中没有领域标签信息，则采用了通用的标签（ Factotum）。这些标签以层级式的方式组织，比如（Psychoanalysis 是 Psychology 领域的一类），图 8 展示了这种领域分类体系的一个片段。

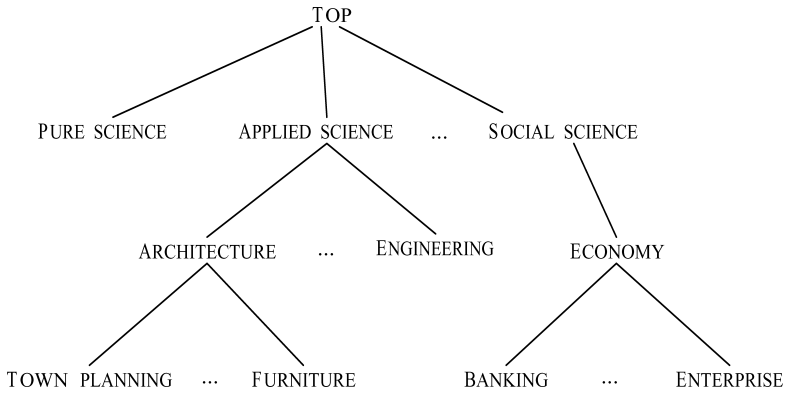

图 8 WordNet 领域标注分类体系的片段摘要

鉴于其在研究团体中的广泛接受度，WordNet 可以看做是一个事实上的标准。伴随着

其应用的成功，多语种的 wordnets 已经被开发出来，并与原始的 Princeton WordNet 进行了关联。在这方面的第一个尝试就发生在 EuroWordNet 项目中，该项目提供了国家 wordnets 间的一个中间语言对齐功能。现在，针对不同语种的 wordnets 的创建、丰富和维护工作还在继续，如 MultiWordNethe BalkaNet.此外，还成立了一个名为 Global WordNet 的联盟，其主要负责将世界上所有语种的 wordnets 实现共享和关联。这些成功都为其它语种的实体含义计算提供了有力的支撑。

三、本体。本体是某个特定领域中概念的详细说明，通常包括了分类体系和一系列的语义关系。从这个角度上，WordNet 和其扩展可以看做是一种本体，同样的情况还有 Omega Ontology（对 WordNet 重组织并将其概念化的词典）、SUMO 上层本体等。在面向领域方面比较出色的有医学统一叙词表（Unified Medical Language System ，UMLS），它提供了医学概念类目间的语义网络。

四、开放关联数据（Linked Open Data，LOD）。开放关联数据是这几年新兴的一种结构化数据，采用 RDF 数据模型，利用 URI（统一资源标识符）命名数据实体，来发布和部署实例数据和类数据，从而可以通过 HTTP 协议揭示并获取这些数据，同时强调数据的相互关联、相互联系以及有益于人机理解的语境信息。由于其组织方式的特殊性，在利用机器自动处理获取相应的语义背景知识时非常简单便捷，如有研究者从文本文档中抽取出语义关系，然后将这些关系并入一个语义图集中，之后文章在这个语义图集合上应用频繁子图发现算法来生成常用的模式。最终，利用关联数据来标注这些发掘出来的模式。因此，LOD 目前也逐渐被研究者们所关注。目前 LOD 云图中覆盖了两百多个数据集合，主要包括政府、生命科学、新闻媒体、地理、用户产生等领域的数据，基本上涵盖了目前主流的命名实体识别中会涉及的实体类别。该方法可以应用于恐怖主义网络分析和生物学网络分析这样的领域。目前比较典型的社会 LOD 主要有：Yago[51]、DBPedia[52]、FOAF[53]等。其中发布的 Yago2 数据主要来源于 Wikipedia, GeoNames, and WordNet，包含了经过人工确认过的 98 个实体的 800 万事实描述。。而 DBPedia 作为关联数据化的 wikipedia,由于 wikipedia 本身即是由网络用户共现的知识库，因此 DBPedia 没有经过专家进行更为精细的确认与筛选。其开放数据量更为庞大，目前已经包含了 47 亿条数据，涉及地理、人物、公司、电影、音乐、基因、药品、书籍以及科学出版物，

4.2.2 非结构化资源

一、语料集。语料集就是用于学习语言模型的文本集合，它可以是语义标注过的，也可以是未标注过的。这两种资源在都有应用，分别在可监督的方法和无监督的方法中发挥作用。

- 未标注的语料集。这类语料集有 the Brown Corpus （包含了 1961 年在美国出版的一百万单词的文本集合）、the British National Corpus [54](BNC)-一个包含了 10 亿英语写作或口语中的词语，其常常用于统计词频以及确定词语间的语法关系、the Wall Street Journal (WSJ)—包含了从华尔街期刊上来的大概三千万词语、the American National Corpus—包

含了美式英语的写作和口语中大概两千两百万个词语、the Gigaword Corpus—包含了新闻文本中大概两百万个词语。

- 语义标注过的语料集。这类语料集中有：SemCor[55]—最大也是用的最多的语义标注语料集，其中包括标注了大概234,000个语义的352篇文章；MultiSemCor—一个利用WordNet的英语版和意大利语版标注了语义的英语—意大利语双语语料集；the linehard-serve corpus—包含了对三类词语（名词、形容词、动词）标注的4000个标注样本；the interest corpus—含有对名词进行语义标注的2369个样本；the DSO 语料集—由新加坡国防科学组织研制的，包括了从Brown和Wall Street Journal语料集中来191个词语的192,800个语义标注分词；the Open Mind Expert data set---该句子语料集中包含了由web用于协作标注的288个名词实例语料标注；the Senseval and Semeval data sets—从四个评价活动中来的语义标注语料集。除the interest corpus采用LDOCE语义标注以及the Senseval-1 corpus 采用HECTOR语义列表（一个来源于joint Oxford University Press/Digital project的词典和语料集）进行标注的外，所有这些语料集在标注中都采用了WordNet语义列表的不同版本。

二、集合资源。

该类资源中记录了词语与其它词语经常共现的趋势，这一类资源示例中有：theWord Sketch Engine, JustTheWord, The British National Corpus collocations, the Collins Cobuild Corpus Concordance等。近来，一个名为Web1T Corpus的文本共现海量数据集发布了，其中包含了从web上获得的一百万兆文本中多达5个词串的共现频率，该集合在相关的研究中已经引起了非常大的反响。

三、其它资源。其它非结构化的知识资源还有如词语频次列表、停用词表（一些没有实在意义的单词，如a、an、the等等）、领域标签等。

4.3 结语

本章通过梳理目前与实体名称规范研究相关的工具与语料，旨在为本项目的进一步实验实施提供进一步的工具筛选参考，也是希望能为相关的研究提供一些参考。

5 总结

本调研报告围绕国内外与实体名称规范相关的理论、方法与工具进行了深入广泛深入的调研，调研首先从实体名称规范的概念、思路、方法和趋势研究入手，明确了实体名称规范的概念、任务划分，然后从传统的思路方法和当前主流的思路与方法两个方面，初步探讨了国内外在实体名称规范方面的发展演化，并进一步探索了实体名称规范的未来研究趋势。在此基础上，项目还从目前国内外典型的几个实体名称规范项目和典型的实体名称规范评测会

议入手，分析这些项目和会议的研究内容、研究思路，并将其进行了对比，从而进一步深入了解实体名称规范的主要内容。在概况调研的基础上，项目集中针对实体名称规范的几类关键技术路线和实体名称规范中涉及的实体特征开展了详实而具体的研究，主要围绕着当前主流的基于语义、背景知识等关键方法展开。为了充分利用现有的相关资源，项目还重点筛选了当前几种比较典型的实体名称规范工具及相关的语料资源库进行了分析介绍，一方面为本项目的实验系统开发提供资源储备，另一方面也希望为其他项目提供一些资源上的参考。

参考文献

[1] Hien T. Nguyen1, Tru H. Cao. A Knowledge-Based Approach to Named Entity Disambiguation in News Articles. AI 2007, LNAI 4830, pp. 619 – 624, 2007

2 ROBERTO NAVIGLI. Word Sense Disambiguation：A Survey.ACM Computing Surveys, Vol. 41, No. 2: 10-69

[3] Jun Lang，Bing Qin, Ting Liu, Sheng Li. Intra-document Coreference Resoulution: the state of the art. Journal of Chinese Language and Computing 17（4）：227-253

[4] J. McCarthy, W. Lehnert. Using decision trees for coreference resolution. In: C.R. Perrault ed. Proc. of the fourteenth International Joint Conference on Artificial Intelligence. Quebec, Canada: Spring, 1050-1055

[5] C. Cardie and K. Wagstaf. 1999. Noun Phrase coreference as clustering. In: P. Fung and J zhou eds. Proc. Of the 1999 Jonit SIGDAT conference on Empirical Methods in NLP and Very Large Corpora. College Park, MD, USA:Association for Computational Linguistics, 82-89

[6] T. Finley, T. Joachims. 2005. Supervised clustering with support vector machines. In: S Dezroski, L.D Raedt, and S. Wrobel eds. Proc. of the 22nd international conference on Machine learing. New York, NY, USA: ACM Press, 217-224

[7]Yun Xu, ZhiHao Wang. A new alignment algorithm to identify definitions corresponding to abbreviations in biomedical text. In:2008 Workshop on Knowledge Discovery and Data Mining

[8] Naoaki Okazaki,Sophia Ananiadou.A Discriminative Alignment Model for Abbreviation Recognition. In: Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pages 657–664

[9] R.Mitkov Robust pronoun resolution with limited knowledge. In:Proc. COLING '98 Proceedings of the 17th international conference on Computational linguistics - Volume2,Pages 869-875

10 王厚峰，梅铮.鲁棒性的汉语人称代词消解.软件学报，2005，16 (05): 700-707

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

12 M. Poesio，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

13 C. Sutton，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

14 M.A. Hearst. Automatic acquisition of hyponyms from large text corpora.[C].In:Proceedings of the 14th International Conference on Computational Linguistics, 1992. [2011-06-01] http://acl.ldc.upenn.edu/C/C92/C92-2082.pdf

15 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
16 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
17 X. Yang，J. Su. 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
18 Diana Maynard, Mark A. Greenwood. Large Scale Semantic Annotation, Indexing and Search at The National Archives
19 OKKAM. http://whois.gwebtools.cn/okkam.info
20 Sig.ma EE- Semantic Information Mashup Enterprise Edition.http://sig.ma/
21 Li JZ, Tang J, Li Y, Luo Q. RiMOM: A dynamic multistrategy ontology alignment framework. IEEE Trans. on Knowledge and Data Engineering, 2009,21(8):1218−1232
22 ObjectCoref.http://ws.nju.edu.cn/objectcoref/
23 Automatic Content Extraction (ACE) Evaluation.http://www.itl.nist.gov/iad/mig/tests/ace/
24 Linguistic Data Consortium[EB/OL]. [2012-05-28].http://projects.ldc.upenn.edu/ace/
25 ACE08 Evaluation Plan v1.2d. [EB/OL]. [2011-05-28]. http://www.itl.nist.gov/iad/mig/tests/ace/2008/doc/ace08-evalplan.v1.2d.pdf
26 Text Analysis Conference.http://www.nist.gov/tac/
27Web People Search Task, WePS, information retrieval, clustering, information extraction. http://nlp.uned.es/weps/index.php
28 中文人名消歧.http://www.cipsc.org.cn/clp2010/task3_ch.htm
29 ARE - Anaphora Resolution Exercise. http://clg.wlv.ac.uk/events/ARE/
30 Nie, Z., et al., Web object retrieval.In: Proceedings of the 16th international conference on World Wide Web, 2007: 81-90.
31 Nie, Z., et al., Object-level ranking: bringing order to Web objects.In: Proceedings of the 14th international conference on World Wide Web, 2005: 567-574.
32 Xianpei Han. Named Entity Disambiguation by Leveraging Wikipedia Semantic Knowledge. In:Proceedings of the 18th ACM conference on Information and knowledge management,2009:215-224
33 Silviu Cucerzan. Large-Scale Named Entity Disambiguation Based on Wikipedia Data, In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague, Czech Republic, pp. 708–716 (2007)
34 Fader, A., Soderland, S., Etzioni, O.: Scaling Wikipedia-based named entity disambiguation to arbitrary web text. In: Proceedings of the IJCAI Workshop on User-contributed Knowledge and Artificial Intelligence: An Evolving Synergy, Pasadena, CA, USA, pp. 21-26 (2009)
35 Hien T. Nguyen, Tru H. Cao Exploring Wikipedia and Text Features for Named Entity Disambiguation.INTELLIGENT INFORMATION AND DATABASE SYSTEMS Lecture Notes in Computer Science, 2010, Volume 5991/2010: 11-20
36 Danica Damljanovic, Kalina Bontcheva.Named Entity Disambiguation using Linked Data.9th Extended Semantic Web Conference (ESWC2012)
37 Danuta Ploch. Exploring Entity Relations for Named Entity Disambiguation.In:Proceedings of the ACL 2011 Student Session,Portland, OR, USA,2011
38 Anna Lisa Gentile,Ziqi Zhang,Lei Xia,Jos´e Iria. Graph-based Semantic Relatedness for Named Entity Disambiguation.In 1st International Conference on Software, Services and Semantic Technologies (S3T) (2009)
39Razvan Bunescu,Marius Pasca.Using Encyclopedic Knowledge for Named Entity Disambiguation.In: Proceedings of 11st Conference of the European Chapter of the Association for Computational Linguistics, April 3-7, 2006, Trento, Italy
40 Nikolov A, Uren V, Motta E, de Roeck A. Overcoming schema heterogeneity between linked semantic repositories to improve coreference resolution. In: Gomez-Perez A, Yu Y, Ding Y, eds. Proc. of the 4th Asian Semantic Web Conf. LNCS 5926,Heidelberg: Springer-Verlag, 2009. 332−346.
41Ron Bekkerman, McCallum Andrew. Disambiguating web appearance of people in a social network. In WWW '05 Proceedings of the 14th international conference on World Wide Web，2005:463-470
42 郎君，秦兵等.基于社会网络的人名检索结果重名消解.计算机学报,2009(7)：1-10
43 陈晨.王厚峰. 基于社会网络的跨文本同名消歧.中文信息学报 2011(5)

[44] The GATE User Guide.[EB/OL]. [2011-02-12]. http://gate.ac.uk/sale/tao/tao.pdf

[45] Dimitar Manov, Borislav Popov. Massive Automatic Annotation[EB/OL]. [2012-02-12]. http://www.sekt-project.com/internal/deliverables/folder.2005-08-22.7344914457/D2.6.1%20-%20Massive%20Automatic%20Annotation.pdf

[46] Horacio Saggion, Adam Funk, Diana Maynard, Kalina Bontcheva. Ontology-based Information Extraction for Business Intelligence[J]. The Semantic Web, pp. 843-856, 2008

47 Heeyoung Lee, Yves Peirsman. Stanford's Multi-Pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task.

[48] Illinois Coreference Package.http://cogcomp.cs.illinois.edu/page/software_view/18

[49] Creap.http://code.google.com/p/creap/

[50] What is WordNet? [EB/OL].[2011-02-12]. http://wordnet.princeton.edu/

[51] YAGO2：A Spatially and Temporally Enhanced Knowledge Base from Wikipedia. [EB/OL].[2012-02-12]. http://www.mpi-inf.mpg.de/yago-naga/yago/downloads.html

[52] DBpedia.http://dbpedia.org/About

[53] The Friend of a Friend (FOAF) project.http://www.foaf-project.org/

[54] CLEAR, J. 1993. The British National Corpus. In The Digital Word: Text-Based Computing in the Humanities, P. Delany and G. P. Landow, Eds. MIT Press, Cambridge, MA. 163–187.

[55] SemCor. Corpushttp://www.gabormelli.com/RKB/SemCor_Corpus