文章编号：1003-0077(2011)06-0098-13

开放式文本信息抽取

赵 军，刘 康，周光有，蔡 黎

（中国科学院 自动化研究所 模式识别国家重点实验室，北京 100190）

摘 要：信息抽取研究已经从传统的限定类别、限定领域信息抽取任务发展到开放类别、开放领域信息抽取。技术手段也从基于人工标注语料库的统计方法发展为有效地挖掘和集成多源异构网络知识并与统计方法结合进行开放式信息抽取。该文在回顾文本信息抽取研究历史的基础上，重点介绍开放式实体抽取、实体消歧和关系抽取的任务、难点、方法、评测、技术水平和存在问题，并结合课题组的研究积累，对文本信息抽取的发展方向以及在网络知识工程、问答系统中的应用进行分析讨论。

关键词：开放式信息抽取；知识工程；文本理解

中图分类号：TP391　　**文献标识码**：A

Open Information Extraction

ZHAO Jun, LIU Kang, ZHOU Guangyou, CAI Li

(National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing 100190, China)

Abstract: The research on information extraction is being developed into open information extraction, i. e. extracting open categories of entities, relations and events from open domain text resources. The methods used are also transferred from pure statistical machine learning model based on human annotated corpora into statistical learning model incorporated with knowledge bases mined from large-scaled and heterogeneous Web resources. This paper firstly reviews the history of the researches on information extraction, then detailedly introduces the task definitions, difficulties, typical methods, evaluations, performances and the challenges of three main open domain information extraction tasks, i. e. entity extraction, entity disambiguation and relation extraction. Finally, based on our researches on this field, we analyze and discuss the development directions of open information extraction research and its applications in large-scaled knowledge engineering, question answering, etc.

Key words: open information extraction; knowledge engineering; text understanding

1 引言

文本信息抽取(Text Information Extraction)指的是从自然语言文本中抽取指定类型的实体(Entity)、关系(Relation)、事件(Event)等事实信息，并形成结构化数据输出的文本处理技术[1]。例如从有线新闻和广播电视的文本中抽取恐怖事件相关情况：时间、地点、作案者、受害者、袭击目标等信息。从 20 世纪 80 年代开始，在 Message Understanding Conference (MUC)[2]、Automatic Content Extraction (ACE)[3] 以及 Text Analysis Conference (TAC)[4]等评测会议的大力推动下，文本信息抽取技术的研究得到蓬勃发展。MUC 从 1987 年到 1997 年总共进行了七届，其五大评测任务是命名实体识别、同指关系(Co-reference)消解、模板元素(Template element)填充(类似于实体属性抽取)、模板关系(Template relation)确定(类似于实体关

收稿日期：2011-09-22　**定稿日期**：2011-10-10

基金项目：国家自然科学基金资助项目(60875041,61070106)

作者简介：赵军(1966—)，男，研究员，博士生导师；刘康(1981—)，男，博士，助理研究员；周光有(1983—)，男，博士生；研究方向皆为自然语言处理、信息抽取和问答系统。

系抽取）和场景模板（Scenario Template）填充（类似于事件抽取）。数据来源是限定领域语料，例如海军军事情报、恐怖袭击、人事职位变动等；ACE 从 1999 年到 2008 年总共进行了九届，涉及实体检测与跟踪（Entity Detection and Tracking，EDT）、数值检测与识别（Value Detection and Recognition，VDR）、时间识别和规范化（Time Expression Recognition and Normalization，TERN）、关系检测与描述（Relation Detection and Characterization，RDC）、事件检测与描述（Event Detection and Characterization，EDC）、实体翻译（Entity Translation，ET）等评测任务。数据来源主要是书面新闻语料。TAC-KBP 从 2009 年开始到目前共进行了三届，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。

纵观信息抽取技术的发展历程，传统信息抽取评测任务是面向限定领域文本的、限定类别实体、关系和事件等的抽取，这大大制约了文本信息抽取技术的发展和应用，例如问答系统所需要的信息抽取技术远远超越我们通常研究的人名、地名、机构名、时间、日期等有限实体类别；上下位（Hypernym-hyponym）、部分整体（Part-whole）、地理位置（Located/Near）等有限关系类别；毁坏（Destruction/Damage）、创造（Creation/Improvement）、所有权转移（Transfer of Possession or Control）等有限事件类别，甚至所需要的类别是未知的、不断变化的。这种应用需求为信息抽取技术的研究提出了新的挑战。另一方面，从信息抽取的技术手段来讲，由于网络文本具有不规范性、开放性以及海量性的特点，使得传统的依赖于训练语料的统计机器学习方法遇到严重挑战。

为了适应互联网实际应用的需求，越来越多的研究者开始研究开放式信息抽取技术，目标是从海量、冗余、异构、不规范、含有大量噪声的网页中大规模地抽取开放类别的实体、关系、事件等多层次语义单元信息，并形成结构化数据格式输出。其特点在于：①文本领域开放：处理的文本领域不再限定于规范的新闻文本或者某一领域文本，而是不限定领域的网络文本；②语义单元类型开放：所抽取的语义单元不限定类型，而是自动地从网络中挖掘语义单元的类型，例如实体类型、关系类型和事件类型等；③以“抽取”替代“识别”：相对于传统信息抽取，开放式文本信息抽取不再拘泥于从文本中精确识别目标信息的每次出现，而是充分利用网络数据海量、冗余的特性，以抽取的方式构建面向实际应用的多层次语义单元集合。在这一过程中，不仅需要考虑文本特征，同时需要综合考虑网页结构特征、用户行为特征等。

本文以开放式文本信息抽取为主题，在回顾文本信息抽取研究历史的基础上，重点介绍开放式实体抽取、关系抽取和实体消歧的任务、难点、方法、评测、技术水平和存在问题，并结合课题组的研究积累，对文本信息抽取的发展方向以及在网络知识工程中的应用进行分析讨论。由于篇幅限制，面向开放式的其他信息抽取技术，例如事件抽取[5-7]、观点信息抽取[8]等不在本文论述的范围。

2　开放式实体抽取

传统的命名实体识别任务就是识别出待处理文本中三大类（实体类、时间类和数字类）、七小类（人名、机构名、地名、时间、日期、货币和百分比）命名实体[2,9]，也有一些研究针对一些特定领域的特定类型的命名实体（例如：产品名称、基因名称等）进行研究[10]。开放式实体抽取的任务是在给出特定语义类的若干实体（称为“种子”）的情况下，找出该语义类包含的其他实体，其中特定语义类的标签可能显式给出，也可能隐式给出。比如给出“中国、美国、俄罗斯”这三个实体，要求找出“国家”这个语义类的其他实体诸如“德国、法国、日本……”。从方式上，传统意义上的实体识别关注的是从文本中识别出实体字符串位置以及所属类别（比如人名、地名、组织机构名等），而开放式实体抽取关注的是从海量、冗余、不规范的网络数据源上抽取出符合某个语义类的实体列表。传统方法更侧重于识别，而开放式实体抽取更侧重于抽取。相对而言，实体抽取比实体识别在任务上更加底层，实体抽取的结果可以作为列表支撑实体的识别。在互联网应用领域，开放式实体抽取技术对于知识库构建、网络内容管理、语义搜索、问答系统等都具有重要应用价值。

（1）开放式实体抽取的难点

开放式实体抽取目标是根据用户输入的种子词从网络中抽取同类型的实体，在这一过程中需要自动判别用户输入种子词的类别信息或者根据用户输入的类别进行类别词扩展。具体难点如下：

① 初始信息少：实体抽取通常采用半监督或无监督的方法，已知信息一般有以下三种：种子实例、语义类别标签以及预先定义的信息。其中给出

的种子通常少于 5 个，语义类别标签有时会给出有时不会给出，而预先定义的信息通常是若干模板，可以利用的已知信息非常少。

② 语义类别难以确定：在没有给定语义类别标签的情况下，种子实体可能会同时属于多个语义类，使得目标语义类别的确定非常困难。比如给出“中国、美国、俄罗斯”三个种子实体，这三个种子实体都可归为“国家”类别，但同时又都可归为“联合国安理会常任理事会成员”类别，或者归为“有核武器的国家”类别。

③ 缺乏公认的评测：实体抽取缺乏公认的评测，研究者在各自构造的实例集上进行研究，评价指标也各有不同，造成不同方法之间横向可比性差。而且由于测试语义类别很少，算法的可推广性不足。

(2) 现有方法

开放式实体抽取的基本假设是：“同类实体在网络上具有相似的网页结构或者相似的上下文特征”。因此抽取过程就是首先找到这样的网页或者文本，然后从中抽取未知的同类型实体。例如在图 1 中，“奥迪”、“宝马”、“保时捷”等具有相同的网页结构，如果已知“奥迪”、“宝马”为汽车品牌名，那么该网页中其他具有相同上下文特征的字符串也很可能是汽车品牌类型实体。

图 1 开放式实体抽取示例

图 2 开放式实体抽取基本流程

开放式实体抽取过程通常包括两个步骤：①候选实体获取；②候选实体置信度计算和排序。其主要方法是：从种子实体出发，通过分析种子实体在语料中的上下文特征得到模板，根据模板得到更多候选实体，选取置信度高的候选实体作为新种子进行迭代，满足一定条件后停止迭代，返回历次置信度高的候选实体作为结果输出。其基本抽取过程如图 2 所示。

目前绝大多数方法都基于上述思路，具体区别在于目标语料来源不同，例如从网页中进行实体抽取，从具有特殊性质的文本(查询日志、网页表格、维基百科)中进行实体抽取等等，以下分别介绍。

使用网页语料：网络上存在大量含有同类实体列表的网页，可以利用这类网页的结构信息辅助类别实例抽取。Wang 等人[11-13]首先通过搜索引擎返回包含全部种子实体且排名靠前的前 100 个网页作为语料；然后从这些语料中学习模板，进而获取候选；最后使用网页、模板和候选以及它们彼此的链接关系构造图，使用随机步算法为候选打分。Whitelaw 等人[14]首先根据种子实体在网页文本中的出现情况及上下文获得高质量的训练数据，然后选用有效的特征训练分类器，利用分类器判定候选是否是给定类别的实体。

使用某种具有特殊性质的文本作为语料(查询日志、网页表格、维基百科等)：Pasca[15]利用查询日志进行实体抽取，首先利用种子实体在查询日志中的上下文特征抽取出特定模板；然后通过模板获取候选实例；最后对种子和候选实例分别构建上下文向量，通过计算相似度来打分。He 等人[16]利用出现在同一个网页中同一个表格中的文本串，很有可能是同类实体这一假设，将文本串和表格作为两类不同的节点构建二分图，通过图算法对文本串进行打分并排序。

综合使用多种资源：Pennacchiotti 等人[17]认

为对于不同数据源应该根据其特性采用不同方法进行处理，最后将融合结果。他们针对不同数据源设计不同的抽取器来抽取实体，同时从不同数据源中抽取特征，构建排序函数，对于不同数据源抽取得到的实体进行融合和排序，输出最终结果。这种方法有效地利用了多源数据的冗余特性，利用大规模数据中的统计特征对于目标实体进行抽取。实验结果表明准确率得到有效提升。

另外，种子的质量对于实体抽取的结果具有重要的影响。Vyas 等人[18]通过定义种子的典型度、歧义度和覆盖度来衡量种子的质量，进而选择更好的种子实体。而为了减少人工校对扩展结果的工作量，Vyas 等人在先前的工作基础上提出了一种多次迭代，每次迭代由人工指定一个错误候选之后重新打分的提纯方法[19]。

(3) 系统评测和技术水平

实体抽取目前还没有举办过公开的评测，研究工作的数据来源也不统一。通常使用平均准确率(Average Precision，AP)或者 P@N 作为评价指标。表 1 是 Wang 等人[20]对中英文各 12 种语义类别进行实体抽取的结果。其中，E1、E2 是两种不同的模板获取方法，E1 表示取 3 个种子实体的所有命名性指称项的公共上下文作为模板；E2 表示取 3 个种子实体中每个种子的至少 1 次命名性指称项的公共上下文作为模板。EF、GW 是两种不同的打分排序方法，EF(extracted frequency)表示按照抽取出的候选出现的次数进行排序，GW(graph walk)表示按照图漫步方法的输出的结果进行排序。

表 1　中英文各 12 种语义类别进行实体抽取的结果

(a) 英文结果

		Max. 100 results			Max. 200	Max. 300
English	**Google Sets**	**E1+EF**	**E2+EF**	**E2+GW**	**E2+GW**	**E2+GW**
classic-disney	37.62%	79.36%	74.45%	84.42%	88.20%	89.39%
cmu-buildings	0.00%	87.85%	87.75%	87.83%	87.83%	87.83%
Common-diseases	1.12%	17.94%	52.84%	57.46%	75.79%	76.87%
constellations	10.45%	89.61%	99.97%	100.00%	100.00%	100.00%
countries	14.24%	95.95%	97.86%	98.17%	98.67%	98.53%
mlb-teams	70.06%	98.61%	99.50%	99.80%	99.84%	99.81%
nba-teams	90.73%	100.00%	100.00%	100.00%	100.00%	100.00%
nfl-temas	94.26%	99.22%	99.98%	100.00%	100.00%	100.00%
periodic-comets	0.22%	69.24%	79.04%	84.78%	84.77%	84.77%
popular-car-makers	73.61%	79.18%	88.23%	95.16%	96.23%	96.95%
us-presidents	56.77%	91.64%	97.07%	99.99%	100.00%	100.00%
us-states	76.00%	99.96%	93.55%	100.00%	100.00%	100.00%
Average	**43.76%**	**84.05%**	**89.19%**	**92.30%**	**94.28%**	**94.51%**

(b) 中文结果

	Max. 100 results			Max. 200	Max. 300
中　文	**E1+EF**	**E2+EF**	**E2+GW**	**E2+GW**	**E2+GW**
china-dynasties	25.45%	33.86%	65.20%	64.62%	65.22%
china-provinces	94.97%	99.19%	99.21%	99.34%	99.35%
class-disney	80.73%	91.17%	91.68%	91.68%	91.68%
constellations	92.00%	96.25%	99.99%	99.99%	99.99%
countries	94.79%	95.39%	96.94%	97.76%	97.72%
mlb-teams	94.42%	84.05%	99.98%	99.96%	99.96%

续表

	Max. 100 results			Max. 200	Max. 300
中 文	E1+EF	E2+EF	E2+GW	E2+GW	E2+GW
nba-teams	90.29%	95.04%	99.90%	100.00%	100.00%
nfl-teams	68.08%	88.43%	95.75%	95.75%	95.75%
popular-car-makers	71.44%	83.29%	94.36%	94.47%	94.55%
taiwan-cities	95.26%	98.04%	100.00%	100.00%	100.00%
us-presidents	62.84%	82.61%	93.03%	94.24%	94.24%
us-states	98.47%	97.08%	99.48%	99.48%	99.48%
Average	**80.73%**	**87.03%**	**94.63%**	**94.77%**	**94.83%**

从上表来看，似乎实体抽取问题已经得到很好解决，但实际上并非如此。现有方法对不同类别实体抽取的效果差别很大，有些语义类别比较容易处理，比如"国家"这一类别，主要原因是这些语义类别的相关语料较多（比如在网络上出现的次数多），或者该类别实体的集中程度更好（比如经常在同一个网页中，甚至经常以列表的形式出现）。但是对于一些小的语义类别，由于数据的稀疏性，语义的歧义性，使得抽取结果中噪声严重，影响应用效果。

(4) 实体抽取存在的问题

尽管目前存在着各种不同的实体抽取方法，有些方法的实验性能也达到了较高水平，但是实体抽取还存在着很多问题，其中最突出的问题是：

• 算法的可扩展性问题

由于缺少相关评测，目前用于测试方法的数据皆由研究者自行构造，不同方法在不同数据上得到的结果难以比较。由于实验中采用的数据类别很少，使得算法的可扩展性差，无法满足面向互联网大规模真实应用的需求。

• 模板的获取问题

目前的方法主要依靠模板来获取候选实体，而模板主要包括自定义的语义模板（比如"such as、kinds of"）以及简单统计得到的上下文模板。这类模板对语义类别的描述能力有限，而且与特定的数据格式和上下文密切相关，如何挖掘和抽取有效的模板是今后研究的重点。

• 目标数据源的置信度问题

目前实体抽取的数据源有普通网页、查询日志、维基百科等，这些数据源的质量层次不齐，严重影响了实体抽取的性能，如何过滤掉低质量的数据源是下一步的重要研究课题。

• 开放式中文实体抽取

开放式中文实体抽取，尤其是当不存在网页结构特征的情况下，抽取任务变得更加困难。其中一个重要原因是分词问题，未知实体往往在分词过程中被分开。针对纯文本环境下开放式中文实体抽取的任务，本课题组[21]利用启发式规则来判别目标实体被错分的边界，然后利用上下文特征判别目标是否为实体以及实体类别，在搜狗语料上测试，能够达到70%的准确率，有效地改善了中文开放式实体抽取的性能，但是这一结果还远远不能达到实用程度，还需进行进一步深入研究。

3 实体消歧

实体歧义指的是一个实体指称项可对应到多个真实世界实体（或称实体概念）的问题。例如，给定如下的三个实体指称项"华盛顿"：

美国开国元勋华盛顿。

美国首都华盛顿特区。

华盛顿州，位于美国西北部。

它们分别指向"美国的第一任总统"、"美国首府"及"美国的华盛顿州"三个真实世界实体。在许多任务中，需要确定一个实体指称项所指向的真实世界实体，这就是实体消歧。

(1) 实体消歧的难点

实体消歧任务与普通的词义消歧（Word sense disambiguation）任务有很多相似之处[22]，但是有其自身的难点。

① 实体消歧目标不明确：传统的词义消歧任务是在具体上下文环境中确定多义词的确切词义，其词义候选来源于专家编撰的词典，目标明确。而实体消歧任务中，往往不能提供实体概念列表，或者提供的实体概念列表不完整，实体消歧难以完成。

② 指称项的多样性(Name variation)：指一个实体概念可以用多种命名性指称项指称，例如全称、别称、简称、拼写错误、多语言名称等。例如：NBA篮球明星 Michael Jeffrey Jordan 在文本中可以用 Michael Jordan、MJ、Jordan 指称。

③ 指称项的歧义性(Name ambiguity)：指一个命名性指称项在不同上下文中可以指称不同的实体概念。例如："迈克尔·乔丹获得今年 NBA 的 MVP"中有三个歧义实体："迈克尔·乔丹"可能是篮球明星 Michael Jeffrey Jordan，也可能是 University of California, Berkeley 的教授 Michael I. Jordan；NBA 可能是"National Basketball Association"，也可能是"National Bicycle Association"；MVP 可能是 Most Valuable Player，也可能是 MVP：Health Care。

(2) 现有方法

目前命名实体消歧任务分为两种类型：实体聚类消歧和实体链接消歧，主要解决单语言实体消歧问题，多语言实体消歧有其特有的方法[23-24]，由于篇幅限制，本文不再介绍。

• 实体聚类消歧

实体聚类消歧任务为：给定一个包含某个歧义实体的网页集合，按照网页中实体指称项所指向的实体概念对网页进行聚类，并抽取一个网页中关于某个实体的特定属性来辅助进行实体消歧。目前，实体聚类聚类消歧一般采用如下步骤：①对每一个实体指称项，抽取其上下文特征(包括词、实体等)，并将其表示成特征向量；②计算实体指称项之间的相似度；③基于指称项之间的相似度，采用一定聚类算法将其聚类，将每个类看作是一个实体概念。核心是如何计算实体指称项之间的相似度。

传统方法主要利用上下文的词信息建立 Bag-of-words 模型(BOW)，从而进行实体指称项相似度计算[25-29]。针对人名消歧，基于图算法[30-32]，充分利用社会化关系的传递性而考虑隐藏的实体关系知识，在某些情况下(特别是结构化数据，如论文记录、电影记录等)能取得更为准确的实体指称项相似度计算结果。但是，基于社会化网络的相似度度量的缺点在于它只用到上下文中的实体指称项本身的信息，不能利用实体指称项的其他上下文信息，因此通常不能在文本实体消歧领域取得有竞争力的性能。

为了克服基于表层特征的实体消歧方法的缺陷，一些研究者开始使用知识资源来提升实体消歧的效果，所使用的知识资源包括：Wikipedia[33-34]、Web 上的链接信息[35-36]、命名实体的同现信息[37]、领域特定语料库[38]等。Bunescu and Pasca[39] 将 Wikipedia 中的类别信息用于 Wikipedia 中的实体消歧；Cucerzan[40] 同时利用 BOW 和 Wikipedia 类别信息对 Wikipedia 中以及普通网页上的实体名进行消歧；利用 Wikipedia 条目信息对于目标实体的上下文进行语义表示的优点在于可以更加精确地捕捉目标实体的语义关联度，而缺点在于这种表示具有稀疏性。针对这一问题，本课题组[33] 利用 Wikipedia 中的知识链接信息计算实体指称项之间的相似度，其中采用概念对齐策略来捕捉不同百科条目之间的语义关联度，使得实体消歧的性能得到改善。但是单一使用 Wikipedia 知识库进行语义表示仍然具有语义稀疏性，针对这一问题，我们综合利用 WordNet、Wikipedia、网页信息等多种知识源挖掘实体指称项的上下文语义信息，并提出了基于图的知识表示模型，将异构语义信息融合在统一的基于图的知识表示框架下，以此为基础挖掘概念之间的潜在语义关联，从而同时集成来自于不同知识源的语义知识[34]。与基于单一知识源的方法相比，该方法显著提升了实体消歧的性能。

• 实体链接消歧

基于聚类的实体消歧方法尽管可以将不同语义的实体指称项区分开，但是不能显式地给出实体的语义信息。针对这一问题，现在越来越多的研究者转向实体链接 Entity Linking(也称 Entity Resolution，Record Linkage 和 Entity Disambiguation)研究。实体链接消歧任务为：给定一个实体指称项，将其链接到知识库中的实体概念上。例如：将"Michael Jordan has published over 300 research articles on topics in computer science, statistics, electrical engineering, molecular biology and cognitive science."中的实体指称项"Michael Jordan"链接到知识库中的实体概念"UC Berkeley 大学教授 Michael Jordan"上，而不是链接到实体概念"NBA 球星 Michael Jordan"上。

实体链接消歧主要有两个步骤：①候选实体的发现：给定实体指称项，链接系统根据知识、规则等信息找到实体指称项的候选实体。例如：对"Michael Jordan is a former NBA player, active businessman and majority owner of the Charlotte Bobcats."中的 Michael Jordan 进行实体消歧，首先要找出 Michael Jordan 可能指向的真实世界实体 Michael Jordan (basketball player)、Michael Jordan

(mycologist)、Michael Jordan (footballer)、Michael B. Jordan、Michael H. Jordan、Michael-Hakim Jordan、Michael Jordan (Irish polotician)等等；②候选实体的链接：链接系统根据指称项和候选实体之间的相似度等特征，选择实体指称项的目标实体。

候选实体发现目前有两种方法，一种是通过挖掘 Wikipedia 等网络百科得到，我们可以利用 Wikipedia 中锚文本的超链接关系、消歧页面(Disambiguation page)以及重定向页面(Redirection page)获得候选实体。另一种是通过挖掘待消歧实体指称项的上下文文本得到，这种方法主要用于发现缩略语的候选实体。缩略语在实体指称项中十分常见，据统计，KBP2009 测试数据的 3 904 个实体指称项中有 827 个为缩略语[41]，缩略语指称项具有很强的歧义性，但它的全称往往是没有歧义的。Zhang 等人[41]利用规则方法从上下文中获取缩略语候选实体，取得不错的效果。

实体链接的核心仍然是计算实体指称项和候选实体的相似度，选择相似度最大的候选实体作为链接的目标实体。从相似度计算的方式上，可以分成单一实体链接和协同实体链接，以下分别介绍。

单一实体链接：该方法仅仅考虑实体指称项与目标实体间的语义相似度。Honnibal 等人[42]和 Bikel 等人[43]将实体指称项的上下文与候选实体的上下文分别表示成 BOW 向量形式，通过计算向量间的余弦值确定指称项与候选实体的相似度，系统选择相似度最大的候选实体进行链接。Bunescu 等人[39]考虑到候选实体的文本内容可能太短，会导致相似度计算不准确，加入指称项文本中的词语与候选实体类别的共现特征。Han[44]认为实体链接与三个因素相关：①实体指称项与目标实体之间的关联度；②目标实体在上下文中的语义一致性；③目标实体在语料中的流行度。基于这三个考虑给出了一种产生式模型，充分融入了候选实体的背景知识与先验信息，显著提升了实体链接的性能。

协同实体链接：传统的单一实体链接只是孤立的单个实体的消歧问题，但是在现实文本存在大量的歧义实体，如果把每个歧义实体看作是一个孤立点，就忽略了实体之间的语义关联。而协同实体链接的目的就是利用协同式策略综合考虑多个实体间的语义关联，建立全局语义约束，从而更好地对于文本内的多个实体进行消歧。Cucerzan 等人[40]考虑不同实体的类别信息，利用实体类别重合度计算目标实体的语义相似度。Kulkarni 等人[45]采用 pairwise 策略，将多个目标指称项分解为多个目标对，计算每个对之间的语义关联度，然后累加起来作为文本内部多个实体之间的语义一致性度量。这种方法尽管考虑了目标实体之间的语义一致性，但是 pair-wise 策略仍然是一种局部寻优方法，在寻优过程中考虑的仅仅是局部语义一致性。因此，本课题组在充分分析问题的基础上，给出了一种基于图的方法，利用图上的计算，充分考虑文本内部目标实体之间的全局语义一致性、指称项与目标实体之间的关联度[46]。相对于传统单一消歧方法以及 pairwise 方法能够有效地提高消歧的精度。

(3) 系统评测和技术水平

目前主流的命名实体消歧评测平台主要有两个：一个是 WePS(Web Person Search Clustering Task)评测[47-48]，主要针对基于聚类的命名实体消歧系统进行评测；第二个是 TAC KBP 的 Entity Linking 评测[49]，主要针对基于实体链接的命名实体消歧系统进行评测。

WePS 主要针对 Web 人名搜索结果的消歧技术进行评测，其任务是通过对人名搜索结果进行聚类来消除歧义。目前 WePS 评测已经开展了两届，正在进行的是第三届：其中第一届评测作为 SemEval 2007 的子任务进行，共有 15 家单位参加；第二届评测作为 WWW 2009 的子任务进行，共有 17 家单位参加。目前 WePS 评测共包含三个数据集，分别为第一届的开发集(WePS1_Training)、第一届的测试集(WePS1_Test)和第二届的测试集(WePS2_Test)。这些数据集共包含 109 个待消歧人名，其中每个人名下大约有 100 个网页(第二届为 150 个)。

与 WePS 不同，TAC KBP 评测对实体链接(Entity Linking)任务进行评测。目前，TAC 实体链接任务的目标实体知识库使用 2008 年 10 月版本的 Wikipedia 构建，共包含了约 82 万个实体，其中有人物实体 11 万，占 14%；组织实体 5.5 万，占 6.8%；地理实体 11 万，占 14.2%；其他类别的实体 53 万，占 65%。目标知识库的总大小约为 2.6Gb。图 3 是 TAC KBP 2010 评测的结果。从图 3 中可以看出，各个系统的平均水平在 70%，还无法满足真实的应用需求，因此仍然需要深入研究和探索。

(4) 实体消歧存在的问题

• 空目标实体问题(NIL Entity Problem)

实体链接的一个未解决的问题是空实体问题(实体知识库中不包含某指称项的目标实体)，现有的框架使用基于相似性阈值的处理方法，不能很好

地建模和解决这个问题。我们正在尝试在语言模型框架下，用一个伪实体语言模型来建模这个问题，从而为有效地解决空实体问题提供一种思路。

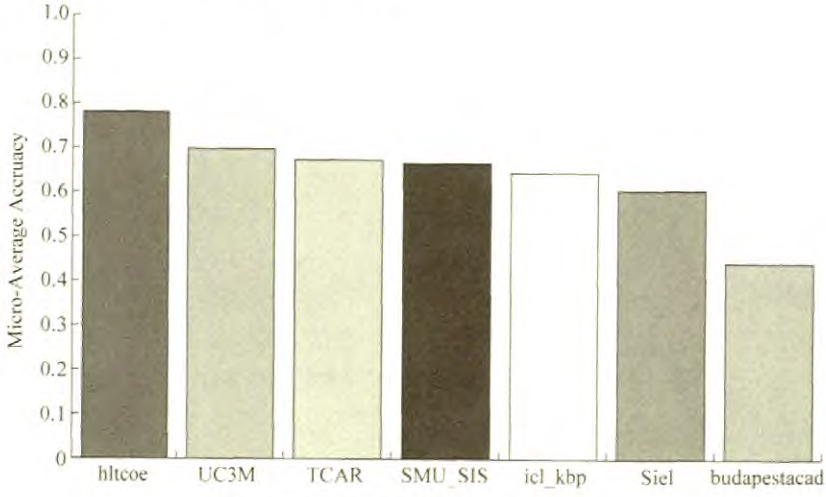


图 3　TAC KBP 2010 结果

• 知识库的覆盖度问题

目前的研究表明，基于知识资源设计更精确的实体指称项相似度计算方法可以在某种程度上提升实体消歧的性能，但常常面临知识覆盖度问题。例如，仅仅使用社会化网络并不能对所有特征关联进行建模，如概念之间的语义关联、词汇之间的语义关联。对多源异构网络知识资源进行有效挖掘和集成是解决上述问题的一种途径。

• 知识不确切的问题

互联网上的知识源通常面临着不准确的问题，甚至包含错误的知识。通常有两个方面的原因：①知识本身的不可靠：网络百科（如维基百科、百度百科等）本身存在错误；②由于抽取技术不可靠带来的知识不可靠：利用信息抽取、网页抽取等技术从社会化网络以及 Web 中抽取出来的知识很可能存在错误。因此需要研究能够容错的知识集成和推理技术。

• 知识库使用的问题

利用知识库进行实体消歧时，对于知识库的使用目前所有方法都集中于使用单文档特征，例如：对实体概念的描述仅仅使用其 Wikipedia 页面。但是，单文档特征常常面临数据稀疏问题，不足以描述实体概念。另外，仅仅使用单文档特征也忽略了其他知识，如语料库中存在的聚类结构和网页链接结构、概念或实体的层级结构等。因此，有必要在语言模型框架下提出新的可以有效使用这些知识的方法。

4　开放式实体关系抽取

实体关系抽取指的是确定实体之间是否存在关系并确定其关系类别的任务。例如，给定“国家财政部部长项怀诚发表了重要讲话”这个句子，实体关系抽取需要识别其中的实体“国家财政部”和“项怀诚”之间存在“Employee_of”类别的关系。传统的实体关系抽取大都是给定关系类别，要求在限定语料中判别两个实体之间是否存在给定关系，可以看作是一个模板填充或者槽填充的过程。例如在 MUC-6[2] 中，其机构模板中包含 LOCATE 和 COUNTRY 两个填充槽，分别表示该机构所处的位置和所在国家。MUC-7[50] 把命名实体之间潜在的关系从实体的属性值中分离出来，正式引入了模板关系（TR，Template Relation）任务，它要求识别实体之间的三种相互关系（即 location_of、employee_of 和 product_of 等）。在 TAC KBP Slot Filling 任务中，针对不同类型实体，定义了不同的实体关系（是各种属性关系），要求系统从大规模文本中找到指定实体的属性值。

在面对海量网络文本资源时，不同的实体类型具有不同关系（或属性）。传统实体关系抽取研究受到人工定义关系类型的限定以及训练语料的限制，很难适应网络文本快速增长、变化的需求。因此，开放式实体关系抽取的目标就是突破封闭的关系类型限定以及训练语料的约束，从海量的网络文本中抽

取实体关系三元组(Arg_1, $Pred$, Arg_2),这里 Arg_1 表示实体,Arg_2 表示实体关系值,通常也为实体,$Pred$ 表示关系名称,通常为动词、名词或者名词短语。例如对于下面这句话:

"McCain fought hard against Obama, but finally lost the election"

从中,我们可以抽取出如下两组三元组(McCain, fought, Obama)和(McCain, lost, election)。

(1) 开放式关系抽取的难点

开放式实体关系抽取包含两个子任务:①实体关系类型抽取;②实体关系值抽取。

实体关系类型抽取:面对开放领域,如何针对每一领域内实体类别确定其关系类别,是开放式关系抽取的首要难点,这种关系不仅仅包含概念之间的上下位关系、部分整体关系、属主关系等通用关系,也包含不同类别实体概念所特有的语义关系,例如"篮球运动员"的以下属性关系:身高、臂展、命中率、篮板等。Web 上存在着大量结构化知识源,其中蕴含着大量易于获取的实体语义关系类别(如维基百科的 Infobox),挖掘和利用 Web 知识源中的语义知识,并充分利用数据冗余性进行知识验证是可行的解决方案。

实体关系值抽取:基于给定类别体系,如何在网络文本中挖掘其关系值是传统关系抽取任务的主要研究点。以往方法依赖于训练语料,通过上下文特征进行关系值抽取。然而,面对开放领域,针对每一个领域构建相应的训练语料不具有可行性。那么面对开放的网络资源,如何利用结构化网络知识与非结构化网络知识的冗余性,自动构建训练语料,同时建立自适应的关系抽取算法,是开放式关系抽取的另一个难点问题。

(2) 现有方法

在开放式实体关系抽取方面,Washington 大学的人工智能研究组在这方面做了大量代表性的工作,并且开发了一系列原型系统:TextRunner[51]、WOE[52]、ReVerb[53] 等。对于关系名称的抽取,TextRunner[51]把动词作为关系名称,抽取过程类似于语义角色标注,通过动词链接两个论元,从而挖掘论元之间的关系。WOE[52]是以 Wikipedia 为目标,从中抽取实体关系类型,从而构建实体的属性描述框架。在 Wikipedia 中,在每个概念条目中,通常都会有人工标注的 Infobox 信息,其中包含了大量实体关系类别,如图 4 所示。依据 Infobox 中蕴含的大量实体关系对,WOE 对于概念条目正文进行回标,以此来自动产生关系值抽取的训练语料,从而解决了开放式关系抽取训练语料不足的问题。

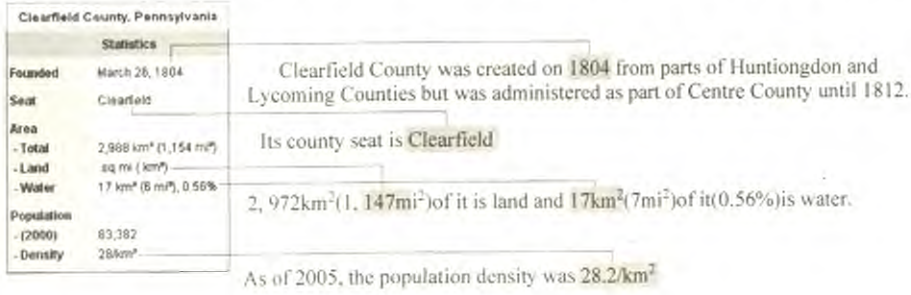


图 4 从 Wikipedia 中抽取关系

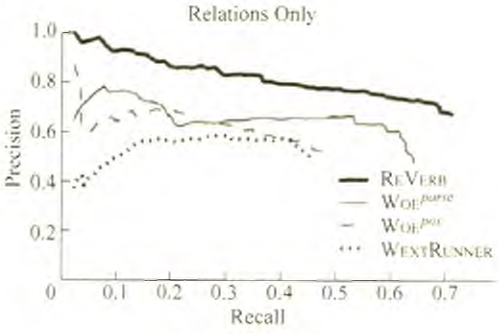
除了从纯文本以及半结构化网页中进行关系类别抽取外,Pasca 等人[15]以用户日志为数据源,利用其中实体、属性和关系词的共现信息获取目标实体类别的属性类别列表。比如输入目标类别"电脑厂商"和种子实体"联想、苹果、戴尔",输出排序后的目标类别的属性类别列表为"笔记本、售后、CEO,…"。实验表明,这种方法在前 50 个结果中平均可以达到 76%的准确率。

在关系值抽取方面,TextRunner 直接从网页的纯文本中抽取实体关系,在这一过程中只考虑文本中词与词之间的关系特征,而不考虑网页内部的结构特征。TextRunner 首先利用简单的启发式规则,在宾州树库上产生训练语料,提取一些浅层句法特征,训练一个分类器,用来判断两个实体间是否存在语义关系;然后在海量网络数据上,找到候选句子,提取浅层句法特征,利用分类器判断所抽取的关系对是否可信;最后利用网络数据的冗余信息,对初步认定可信的关系进行评估。但是,TextRunner 的问题在于往往从文本中抽取出无信息量的三元组(Un-informative Extractions)和错误的三元组(Incoherent Extractions),其中无信息量三元组在抽取结果中占 7%的比例,错误三元组占 13%的比例。针对这一问题,Etzioni 等人[53]开发了 ReVerb 系

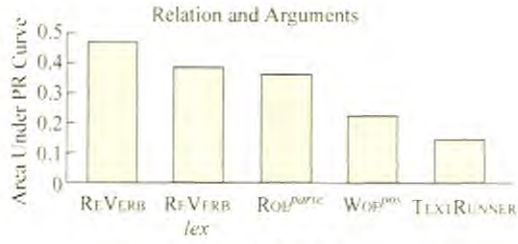
统，提出了利用句法和词汇信息对抽取过程进行约束，实验证明这种方法可以较大幅度地提升关系值抽取的准确率和召回率。

(3) 系统评测和技术水平

开放式关系抽取目前还没有举办过公开评测，研究工作的数据来源也不统一。目前，评价指标仍然和传统信息抽取评价指标一样，采用正确率(Precision)、召回率(Recall)以及 F 值作为评价指标。Wu[52]给出了几个开放式关系抽取系统的实验比较，如图 5 所示：



(a) 关系名抽取



(b) 关系名与关系值抽取

图 5　开放式关系抽取系统比较

从图 5 我们可以看出，对于关系名抽取，目前 F 值可以达到 70%左右的，而综合考虑关系值的抽取，性能下降很多。从面向互联网的真实应用需要来看，未来还需要深入研究。

(4) 需要解决的问题

从传统给定类别的关系抽取到开放式的关系抽取，是关系抽取研究思路上的一个转变，目前开放式抽取系统还存在的不足是：

• 针对真实网络数据的关系抽取问题

目前的关系抽取研究大多数是在干净的文本上进行的，而网络数据格式不规范，噪声大，质量层次不齐，如何针对真实网络数据研究鲁棒的关系抽取方法是需要重点研究的问题之一。

• 单纯利用 Infobox 抽取关系名覆盖率不高的问题

Wu[52]利用 Infobox 信息进行回标产生训练集，这种方法对于中文百科页面仍然具有局限性。在中文百科页面中(百度百科、互动百科等)并不是所有的类别条目下都有 Infobox 信息，这使得 Wu[52]的方法具有很大局限性。同时，Infobox 中往往是一些同类型条目共有的信息，而大部分条目特有的属性信息散落在百科条目的文本中，以半结构化或者纯文本形式出现。开放式关系抽取不能忽略这一部分信息。

4　结束语

信息抽取技术的研究从上世纪 80 年代开始至今走过了 20 多年的历程，研究内容和技术手段随着互联网的发展而发展。在研究内容上，已经从面向限定领域、限定类型的信息抽取任务逐渐发展为开放领域、开放类别的信息抽取任务。在技术手段上，从早期基于人工模板的方法，到基于语料库的统计方法，再到目前 Web2.0 时代从大规模用户生成内容(User Generated Content，例如网络百科、社区问答等)进行知识挖掘，进而融合知识和统计方法进行开放式信息抽取，技术手段越来越有效。在以上进展过程中，信息抽取技术乃至自然语言处理技术的研究越来越面向互联网应用，而互联网也为信息抽取技术和自然语言处理技术的研究提供了越来越多的宝贵资源和技术创新的源泉。近年来，研究人员利用网络上丰富的数据资源开展了一系列的研究工作，比如利用网络海量数据提升句法分析的性能[54-55]；利用网络上积累的大量问答对开展社区问答方面的研究[56-59]，等等。在这种交叉融合的趋势下，信息抽取技术和自然语言处理技术的研究和应用必将得到加速发展。

作为开放式信息抽取技术的应用，大规模知识库的自动构建是一个典型代表。很多互联网应用任务都需要背景知识库的支撑，这个知识库不仅包含 WordNet[60]、HowNet[61]等常识知识库中的通用语义知识，而且包含百科全书、领域知识库中的领域语义知识。如果能把多源知识集成为一个大的知识系统，将可能提高很多互联网应用系统的性能，并开创语义网时代的很多应用。现有的知识库如 WordNet[60]、HowNet[61]和 CYC[62]等大多数依靠专家人工编撰。随着互联网的发展，知识呈爆炸式增长，人工构建知识库特别是领域知识库遇到了很大困难[63]：不仅费时费力，而且知识覆盖率低，数据稀疏，更新缓慢。另一方面，机器自动构建知识库的方

法目前仍旧只能完成简单粗浅的任务[64]，无法达到构建高质量知识库的要求。开放式信息抽取技术研究的不断深入以及 Wikipedia、Freebase、百度百科、互动百科等大规模网络知识库的大量出现，为大规模知识工程的构建提供了新的契机。信息抽取和知识工程领域的研究人员在这方面做出了积极有效的探索。YAGO[65]从 Wikipedia 的 category pages 中提取出实体实例和关系实例候选，并与 WordNet 进行衔接，准确率达到 97%。这样，YAGO 既具有 WordNet 干净的概念层级结构，又拥有 Wikipedia 的海量实例。目前，YAGO 有 100 万实体及其 500 万事实。本研究组利用在信息抽取方面的技术积累，以《中国大百科全书》知识体系作为目标知识库的结构，从网络知识库中抽取概念实例并综合利用网络百科网页中蕴含的丰富的语义标签、半结构化信息和非结构化信息进行概念实例挂载，从而将百科知识库从 8 万条目扩展为目前的百万条目级别，在此基础上进行概念属性抽取，为下一步研发面向开放式的自动问答系统提供了知识资源的支撑[66]。

综上所述，信息抽取在互联网应用中具有非常重要的应用前景。面对互联网的实际需求以及网络文本的特点，传统信息抽取技术已经遇到技术瓶颈，无法得到广泛应用，迫切需要更加系统深入的研究。本文重点介绍了面向互联网应用的开放式信息抽取技术，以实体为核心，重点分析介绍实体识别与抽取、实体消歧和实体关系抽取等三个开放式信息任务的研究现状、存在的问题和值得深入研究的方向。从研究方法上来看，研究人员已经开始突破传统的依赖人工标注语料库的统计学习方法，有效地挖掘和集成多源异构的网络知识并与统计方法结合进行开放式信息抽取。因此，研究领域知识的表示、挖掘、集成和推理机制，探索构建高性能、大规模知识系统的方法，为克服传统方法在面向开放式信息抽取时的推导和泛化能力不足的问题提供解决方案，具有重要的学术意义。

致谢 感谢研究生们对本文的贡献，特别是韩先培和张涛(实体消歧)，杨帆(多语言实体消歧)，齐振宇(实体抽取)，刘芳(属性抽取)，徐立恒、刘洋和来斯惟(网络知识工程)等。

参考文献

[1] Ralph Grishman. 1997. Information Extraction: Techniques and Challenges[R]. New York: New York University, 1997.

[2] Ralph Grishman, Beth Sundheim. Message Understanding Conference-6: A Brief History[C]//Proceedings of COLING, 1996.

[3] http://www.itl.nist.gov/iad/mig/tests/ace/[OL].

[4] http://www.nist.gov/tac/[OL].

[5] Martina Naughton, N. Kushmerichand J. Carthy. Event Extraction from Hetergeneous News Sources [C]//Proceedings of AAAI, 2006.

[6] D. McClosky, M. Surdeanu, C. D. Manning. Event Extraction as Dependency Parsing[C]//Proceedings of ACL-HLT, 2011.

[7] Yu Hong, Jianfeng Zhang, Bin Ma, Jianmin Yao, Guodong Zhou, Qiaoming Zhu. Using Cross-Entity Inference to Improve Event Extraction[C]//Proceedings of ACL-HLT, 2011.

[8] 刘康. 文本倾向性分析技术研究[D]. 中国科学院自动化研究所博士学位论文, 2010.

[9] 赵军. 命名实体识别、排歧和多语言关联[J]. 中文信息学报, 2009, 23(2): 3-17.

[10] Jun Zhao, Feifan Liu. Product Named Entity Recognition in Chinese Texts[J]. International Journal of Language Resource and Evaluation. 2008, 42(2): 132-152.

[11] Richard C. Wang, William Cohen. Automatic Set Instance Extraction using the Web[C]//Proceedings of ACL-IJCNLP, 2009.

[12] Richard C. Wang, William Cohen. Iterative Set Expansion of Named Entities using the Web[C]//Proceedings of ICDM, 2008.

[13] Richard C. Wang, Nico Schlaefer, William Cohen, Eric Nyberg. Automatic Set Expansion for List Question Answering[C]//Proceedings of EMNLP, 2008.

[14] Casey Whitelaw, Alex Kehlenbeck, Nemanja Petrovic. Web-Scale Named Entity Recognition[C]//Proceedings of CIKM, 2008.

[15] Marius Pasca: Organizing and searching the world wide web of facts-step two: harnessing the wisdom of the crowds[C]//Proceedings of WWW, 2007.

[16] Yeye He, Dong Xin. SEISA: Set Expansion by Iterative Similarity Aggregation [C]//Proceedings of WWW, 2011.

[17] Marco Pennacchiotti, Patrick Pantel. Entity Extraction via Ensemble Semantics [C]//Proceedings of EMNLP, 2009.

[18] Vishnu Vyas, Patrick Pantel, Eric Crestan. Helping Editors Choose Better Seed Sets for Entity Set Expansion[C]//Proceedings of CIKM, 2009.

[19] Vishnu Vyas, Patrick Pantel. Semi-Automatic Entity Set Refinement[C]//Proceedings of NAACL, 2009.

[20] Richard C. Wang, William Cohen. Language-Independent Set Expansion of Named Entities using the Web[C]//Proceedings of ICDM, 2007.

[21] 齐振宇，赵军，杨帆. 一种开放式中文命名实体识别的新方法[C]//第五届全国信息检索学术会议，上海，2009 年.

[22] Philip Edmonds. SENSEVAL: The Evaluation of Word Sense Disambiguation Systems[R]//ELRA Newsletter, October, 2002.

[23] Fan Yang, Jun Zhao, Bo Zou, Kang Liu. Chinese-English Backward Translation Assisted with Mining Monolingual Web Pages[C]//Proceedings of ACL, 2008.

[24] Fan Yang, Jun Zhao, Kang Liu. A Chinese-English Organization Name Translation System Using Heuristic Web Mining and Asymmetric Alignment[C]//Proceedings of ACL, 2009.

[25] Bagga, Baldwin. Entity-Based Cross-Document Coreferencing Using the Vector Space Model[C]//Proceedings of HLT/ACL, 2008.

[26] Gideon S. Mann, David Yarowsky. Unsupervised Personal Name Disambiguation[C]//Proceedings of CONIL, 2003.

[27] Cheng Niu, Wei Li, Rohini K. Srihari. Weakly Supervised Learning for Cross-document Person Name Disambiguation Supported by Information Extraction [C]//Proceedings of ACL, 2004.

[28] Ted Pedersen, Amruta Purandare, Anagha Kulkarni. Name Discrimination by Clustering Similar Contexts [C]//Proceedings of CICLing, 2005.

[29] Ying Chen, James Martin. Towards Robust Unsupervised Personal Name Disambiguation [C]//Proceedings of EMNLP, 2007.

[30] Bradley Malin. Unsupervised Name Disambiguation via Social Network Similarity[C]//Proceedings of SIAM, 2005.

[31] Bradley Malin, Edoardo Airoldi. A Network Analysis Model for Disambiguation of Names in Lists[J]. Computational & Mathematical Organization Theory, 2005, 11: 119-139.

[32] Kai-Hsiang Yang, Kun-Yan Chiou, Hahn-Ming Lee, Jan-Ming Ho. Web Appearance Disambiguation of Personal Names Based on Network Motif[C]//Proceedings of WI, 2006.

[33] Xianpei Han, Jun Zhao. Named Entity Disambiguation by Leveraging Wikipedia semantic knowledge [C]//Proceedings of CIKM, 2009.

[34] Xianpei Han, Jun Zhao. Structural Semantic Relatedness: A Knowledge-Based Method to Named Entity Disambiguation[C]//Proceedings of ACL, 2011.

[35] Joseph Hassell, Boanerges Aleman-Meza, I. BudakArpinar. Ontology-Driven Automatic Entity Disambiguation in Unstructured Text[C]//Proceedings of ISWC, 2006.

[36] Ron Bekkerman, Andrew McCallum. Disambiguating Web Appearances of People in a Social Network[C]//Proceedings of WWW, 2005.

[37] Dmitri V. Kalashnikov, Rabia Nuray-Turan, Sharad Mehrotra. Towards Breaking the Quality Curse. A Web-Querying Approach to Web People Search[C]//Proceedings of SIGIR, 2008.

[38] Yiming Lu, Zaiqing Nie, Taoyuan Cheng, Ying Gao, Ji-Rong Wen. Name Disambiguation Using Web Connection[C]//Proceedings of AAAI, 2007.

[39] Razvan Bunescu, Marius Pasca. Using Encyclopedic Knowledge for Named Entity Disambiguation[C]//Proceedings of EACL, 2006.

[40] Silviu Cucerzan. Large-Scale Named Entity Disambiguation Based on Wikipedia Data[C]//Proceedings of EMNLP, 2007.

[41] Wei Zhang, Yan Chuan Sim, Jian Su, Chew Lim Tan. Entity Linking with Effective Acronym Expansion, Instance Selection and Topic Modeling[C]//Proceedings of IJCAI, 2011.

[42] Matthew Honnibal, Robert Dale. DAMSEL: The DSTO/Macquarie System for Entity-Linking[C]//Proceeding of TAC, 2009.

[43] Dan Bikel, Vittorio Castelli, Radu Florian, Ding-Jung Han. Entity Linking and Slot Filling through Statistical Processing and Inference Rules[C]//Proceedings of TAC, 2009.

[44] Xianpei Han, Le Sun. A Generative Entity-Mention Model for Linking Entities with Knowledge Base [C]//Proceedings of ACL, 2011.

[45] Sayali Kulkarni, Amit Singh, Ganesh Ramakrishnan, Soumen Chakrabarti. Collective annotation of Wikipedia entities in web text[C]//Proceedings of KDD, 2009.

[46] Xianpei Han, Le Sun, Jun Zhao. Collective Entity Linking in Web Text: A Graph-Based Method[C]//Proceedings of SIGIR, 2011.

[47] Javier Artiles, Julio Gonzalo, Satoshi Sekine. The SemEval-2007 WePS Evaluation: Establishing a benchmark for the Web People Search Task[C]//Proceedings SemEval, 2007.

[48] Javier Artiles, Julio Gonzalo, Satoshi Sekine. WePS2 Evaluation Campaign: Overview of the Web People Search Clustering Task[C]//Proceedings of WWW Workshop of WePS2, 2009.

[49] Paul McNamee, Hoa Dang. Overview of the TAC 2009 Knowledge Base Population Track[C]//Proceedings of Text Analysis Conference (TAC-2009),

2009.

[50] http://www-nlpir. nist. gov/related _ projects/muc/proceedings/muc _ 7 _ proceedings/overview. html [OL].

[51] Michele Banko, Michael J Cafarella. Stephen Soderland, Matt Broadhead and Oren Etzioni. Open Information Extraction from the Web[C]//Proceedings of IJCAI, 2007.

[52] Fei Wu, Daniel S. Weld. Autonomously Semantifying Wikipedia[C]//Proceedings of CIKM, 2007.

[53] Oren Etzioni, Anthony Fader, Janara Christensen, Stephen Soderland, Mausam. Open Information Extraction: the Second Generation[C]//Proceedings of IJCAI, 2011.

[54] Mohit Bansal, Dan Klein. Web-Scale Features for Full-Scale Parsing[C]//Proceedings of ACL-HLT, 2011.

[55] Guangyou Zhou, Jun Zhao, Kang Liu, Li Cai. Exploiting Web-Derived Selectional Preference to Improve Statistical Dependency Parsing[C]//Proceedings of ACL-HLT, 2011.

[56] Xiaobin Xue, Jiwoon Jeon, W. Bruce Croft. Retrieval Models for Question and Answer Archives[C]//Proceedings of SIGIR, 2008.

[57] Guangyou Zhou, Li Cai, Jun Zhao, Kang Liu. Phrase-Based Translation Model for Question Retrieval in Community Question Answer Archives [C]//Proceedings of ACL-HLT, 2011.

[58] Li Cai, Guangyou Zhou, Kang Liu, Jun Zhao. Learning the Latent Topics for Community QA[C]//Proceedings of IJCNLP, 2011.

[59] Li Cai, Guangyou Zhou, Kang Liu, Jun Zhao. Learning to Classify Large-Scale Questions in Community QA by Leveraging Wikipedia Semantic Knowledge[C]//Proceedings of CIKM, 2011.

[60] George A. Miller, WordNet: A Lexical Database for English[J]. Communication of the ACM, 38(11): 39-41.

[61] HowNet: http://www. keenage. com/[DB/OL].

[62] Douglas B. Lenat. CYC: A Large-Scale Investment in Knowledge Infrastructure[J]. Communications of the ACM 1995,38(11): 33-38.

[63] Alexander Madche and Steffen Staab. Ontology Learning for the Semantic Web[J]. IEEE Intelligent Systems, 2001, 16(2): 72-79.

[64] L. Brainbridge. Ironies of automation[J]. Automatica, 1983, 19: 775-779.

[65] Fabian M. Suchanek, Gjergji Kasneci and Gerhard Weikum. YAGO: A Core of Semantic Knowledge Unifying WordNet and Wikipedia[C]//Proceedings of WWW, 2007.

[66] 徐立恒,刘洋,来斯惟,等. 基于多特征表示的本体概念挂载研究[C]//全国第十一届计算语言学学术会议,洛阳,2011.

(上接第89页)

[41] Junguo Zhu, Muyun Yang. All in Strings: a Powerful String-based Automatic MT Evaluation Metric with Multiple Granularities[C]//The 23rd International Conference on Computational Linguistics. 2010.

[42] 王博. 机器翻译系统的自动评价及诊断方法研究[D]. 哈尔滨工业大学博士学位论文, 2009.

[43] Dequan Zheng, Hao Yu, Tiejun Zhao et al. Research on a Chinese Language Model Based on Ontology and Statistical Method[J]. Journal of Chinese Language and Computing, 2004, 14(4): 305-315.

[44] Honglei Zhu, Dequan Zheng, Tiejun Zhao. Research on query translation for clir based on a combination of statistical method and web information[J]. Journal of Computational Information Systems, 2009, 5(3): 1115-1122.