·博士论坛·

信息抽取研究综述

李保利　陈玉忠　俞士汶
(北京大学计算机科学与技术系计算语言学研究所,北京 100871)
E-mail:libl@pku.edu.cn

摘　要　信息抽取研究旨在为人们提供更有力的信息获取工具,以应对信息爆炸带来的严重挑战。与信息检索不同,信息抽取直接从自然语言文本中抽取事实信息。过去十多年来,信息抽取逐步发展成为自然语言处理领域的一个重要分支,其独特的发展轨迹——通过系统化、大规模的定量评测推动研究向前发展,以及某些成功启示,如部分分析技术的有效性、快速NLP系统开发的必要性,都极大地推动了自然语言处理研究的发展,促进了NLP研究与应用的紧密结合。回顾信息抽取研究的历史,总结信息抽取研究的现状,将有助于这方面研究工作向前发展。

关键词　自然语言处理　信息抽取　信息检索　命名实体识别

文章编号 1002-8331-(2003)10-0001-05　**文献标识码** A　**中图分类号** TP391

Research on Information Extraction: A Survey

Li Baoli　Chen Yuzhong　Yu Shiwen
(Department of Computer Science and Technology, Peking University, Beijing 100871)

Abstract: The research on Information Extraction aims at providing more powerful information access tools to help people overcome the problem of information overloading. Unlike Information Retrieval, Information Extraction Systems extract factual information directly from natural language texts. In the last decade, Information Extraction has become an important sub-field of Natural Language Processing. Its unique development track, i.e. accelerating research via systematical and large scale evaluation, and some successful experience, such as the effectiveness of partial-parsing techniques and the importance of fast development cycles, have made it a great and most important impetus to the research of NLP in the last decade. Moreover, Information Extraction has built a more effective connection between NLP researchers and NLP system developers. It will be helpful to review the history and investigate the state of the art of Information Extraction.

Keywords: Natural Language Processing, Information Extraction, Information Retrieval, Named Entity Recognization

1　引言

随着计算机的普及以及互联网(WWW)的迅猛发展,大量的信息以电子文档的形式出现在人们面前。为了应对信息爆炸带来的严重挑战,迫切需要一些自动化的工具帮助人们在海量信息源中迅速找到真正需要的信息。信息抽取(Information Extraction)研究正是在这种背景下产生的。

信息抽取系统的主要功能是从文本中抽取出特定的事实信息(factual information)。比如,从新闻报道中抽取出恐怖事件的详细情况:时间、地点、作案者、受害者、袭击目标、使用的武器等;从经济新闻中抽取出公司发布新产品的情况:公司名、产品名、发布时间、产品性能等;从病人的医疗记录中抽取出症状、诊断记录、检验结果、处方等等。通常,被抽取出来的信息以结构化的形式描述,可以直接存入数据库中,供用户查询以及进一步分析利用。

与信息抽取密切相关的一项研究是信息检索,但信息抽取与信息检索存在差异,主要表现在三个方面:

(1)功能不同。信息检索系统主要是从大量的文档集合中找到与用户需求相关的文档列表;而信息抽取系统则旨在从文本中直接获得用户感兴趣的事实信息。

(2)处理技术不同。信息检索系统通常利用统计及关键词匹配等技术,把文本看成词的集合(bags of words),不需要对文本进行深入分析理解;而信息抽取往往要借助自然语言处理技术,通过对文本中的句子以及篇章进行分析处理后才能完成。

(3)适用领域不同。由于采用的技术不同,信息检索系统通常是领域无关的,而信息抽取系统则是领域相关的,只能抽取系统预先设定好的有限种类的事实信息。

另一方面,信息检索与信息抽取又是互补的。为了处理海量文本,信息抽取系统通常以信息检索系统(如文本过滤)的输出作为输入;而信息抽取技术又可以用来提高信息检索系统的性能。二者的结合能够更好地服务于用户的信息处理需求。

信息抽取虽然需要对文本进行一定程度的理解,但与真正

基金项目:国家自然科学基金项目(编号:69973005);国家973重点基础研究发展规划项目(编号:G1998030507-4);北大985项目支持
作者简介:李保利(1971-),男,博士研究生,主要研究方向:中文信息处理。陈玉忠(1963-),男,副教授,博士研究生,主要研究方向:中文信息处理、机器翻译等。俞士汶(1938-),男,教授,博士生导师,主要研究方向:中文信息处理、计算语言学等。

的文本理解(Text Understanding)还是不同的。在信息抽取中,用户一般只关心有限的感兴趣的事实信息,而不关心文本意义的细微差别以及作者的写作意图等深层理解问题[1]。因此,信息抽取只能算是一种浅层的或者说简化的文本理解技术。

一般来说,信息抽取系统的处理对象是自然语言文本尤其是非结构化文本。但广义上讲,除了电子文本以外,信息抽取系统的处理对象还可以是语音、图像、视频等其他媒体类型的数据。在这里只讨论狭义上的信息抽取研究,即针对自然语言文本的信息抽取。

下面首先回顾了信息抽取研究发展的历史,然后介绍信息抽取系统的体系结构以及一些关键技术,最后对信息抽取研究的未来方向做了展望。

2 信息抽取研究的发展历史

从自然语言文本中获取结构化信息的研究最早开始于20世纪60年代中期,这被看作是信息抽取技术的初始研究,它以两个长期的、研究性的自然语言处理项目为代表[2]。

美国纽约大学开展的Linguistic String项目[3]开始于60年代中期并一直延续到80年代。该项目的主要研究内容是建立一个大规模的英语计算语法,与之相关的应用是从医疗领域的X光报告和医院出院记录中抽取信息格式(Information Formats),这种信息格式实际上就是现在作者所说的模板[1](Templates)。

另一个相关的长期项目是由耶鲁大学Roger Schank及其同事在20世纪70年代开展的有关故事理解的研究。由他的学生Gerald De Jong设计实现的FRUMP系统[4]是根据故事脚本理论建立的一个信息抽取系统。该系统从新闻报道中抽取信息,内容涉及地震、工人罢工等很多领域或场景。该系统采用了期望驱动(top-down,脚本)与数据驱动(bottom-up,输入文本)相结合的处理方法。这种方法被后来的许多信息抽取系统采用。

从20世纪80年代末开始,信息抽取研究蓬勃开展起来,这主要得益于消息理解系列会议(MUC,Message Understanding Conference)的召开。正是MUC系列会议使信息抽取发展成为自然语言处理领域一个重要分支,并一直推动这一领域的研究向前发展。

从1987年开始到1998年,MUC会议共举行了七届,它由美国国防高级研究计划委员会(DARPA,the Defense Advanced Research Projects Agency)资助。MUC的显著特点并不是会议本身,而在于对信息抽取系统的评测[5]。只有参加信息抽取系统评测的单位才被允许参加MUC会议。在每次MUC会议前,组织者首先向各参加者提供样例消息文本和有关抽取任务的说明,然后各参加者开发能够处理这种消息文本的信息抽取系统。在正式会议前,各参加者运行各自的系统处理给定的测试消息文本集合。由各个系统的输出结果与手工标注的标准结果相对照得到最终的评测结果。最后才是所谓的会议,由参与者交流思想和感受。后来,这种评测驱动的会议模式得到广泛推广,如1992年开始举行的文本检索会议TREC[2]等。

从历次MUC会议,可以清楚地看到信息抽取技术发展的历程。

1987年5月举行的首届MUC会议基本上是探索性的,没有明确的任务定义,也没有制定评测标准,总共有6个系统参加,所处理的文本是海军军事情报,每个系统的输出格式都不一样。

MUC-2于1989年5月举行,共有8个系统参加,处理的文本类型与MUC-1一样。MUC-2开始有了明确的任务定义,规定了模板以及槽的填充规则,抽取任务被明确为一个模板填充的过程。

MUC-3于1991年5月举行,共有15个系统参加,抽取任务是从新闻报告中抽取拉丁美洲恐怖事件的信息,定义的抽取模板由18个槽组成。从MUC-3开始引入正式的评测标准,其中借用了信息检索领域采用的一些概念,如召回率和准确率等。

MUC-4于1992年6月举行,共有17个系统参加,任务与MUC-3一样,仍然是从新闻报告中抽取恐怖事件信息。但抽取模板变得更复杂了,总共由24个槽组成。从这次会议开始MUC被纳入TIPSTER文本项目[3]。

MUC-5于1993年8月举行,共有17个系统参加:美国14个,英国、加拿大、日本各一个。此次会议设计了两个目标场景:金融领域中的公司合资情况、微电子技术领域中四种芯片制造处理技术的进展情况。除英语外,MUC-5还对日语信息抽取系统进行了测试。在本次会议上,组织者尝试采用平均填充错误率(ERR,Error Per Response Fill)作为主要评价指标。与以前相比,MUC-5抽取任务的复杂性更大,比如公司合资场景需要填充11种子模板总共47个槽,光任务描述文档就有40多页。MUC-5的模板和槽填充规范是MUC系列评测中最复杂的。

MUC-5的一个重要创新是引入了嵌套的模板结构。信息抽取模板不再是扁平结构(flat structure)的单个模板,而是借鉴面向对象和框架知识表示的思想,由多个子模板组成。模板中每个槽的取值除了可以是文本串(如公司名)、格式化串(如将日期、时间、金额等文本描述转化为某种规范形式)、有限集合中的元素(如组织类型可以分为公司、政府部门、研究机构等)外,还可以是指向另一个子模板的指针。

MUC-6于1995年9月举行,训练时的目标场景是劳动争议的协商情况,测试时的目标场景是公司管理人员的职务变动情况,共有16家单位参加了这次会议。MUC-6的评测更为细致,强调系统的可移植性以及对文本的深层理解能力。除了原有的场景模板(Scenario Templates)填充任务外,又引入三个新的评测任务:命名实体(Named Entity)识别、共指(Coreference)关系确定、模板元素(Template Element)填充等[5][6]。

---

[1] 遵循MUC(Message Understanding Conference)系列会议建立的术语,信息抽取最终的输出结果被称为模板(Template),模板中的域称为槽(Slot),而把信息抽取过程中使用的匹配规则称为模式(Pattern)。另外,要提取的特定事件或关系称为一个场景(Scenario),而领域(Domain)的概念要宽泛一些,通常一个领域可以包含多个场景。比如,在金融领域的新闻中,可能包含有建立合资公司、股票转让等很多个场景。

[2] http://trec.nist.gov/

[3] TIPSTER文本项目(http://www.itl.nist.gov/iaui/894.02/related_projects/tipster/)由美国国防高级研究计划委员会组织,1991年开始实施,1998年秋天终止。该项目致力于推动和促进提高文本处理技术水平,重点是文档检索(Document Detection)、信息抽取(Information Extraction)、自动文摘(Summarization)等技术,共分三个阶段实施。

命名实体识别任务主要是要识别出文本中出现的专有名称和有意义的数量短语并加以归类；共指关系确定任务是要识别出给定文本中的参照表达式，并确定这些表达式之间的共指关系；模板元素填充任务是要识别出特定类型的所有实体以及它们的属性特征。

最后一届 MUC 会议——MUC-7 于 1998 年 4 月举行。训练时的目标场景是飞机失事事件，测试时的目标场景是航天器（火箭/导弹）发射事件。除 MUC-6 已有的四项评测任务外，MUC-7 又增加了一项新任务——模板关系任务，它意在确定实体之间与特定领域无关的关系[6]。共有 18 家单位参加了MUC-7 评测。值得注意的是，在 MUC-6 和 MUC-7 中开发者只允许用四周的时间进行系统的移植，而在先前的评测中常常允许有 6~9 个月的移植时间。

在 MUC 中，衡量信息抽取系统的性能主要根据两个评价指标：召回率和准确率[7]。召回率等于系统正确抽取的结果占所有可能正确结果的比例；准确率等于系统正确抽取的结果占所有抽取结果的比例。为了综合评价系统的性能，通常还计算召回率（*REC*）和准确率（*PRE*）的加权几何平均值，即 *F* 指数，它的计算公式如下：

$$F\text{-}MEASURE=\frac{((beta)^2+1.0)*PRE*REC)}{((beta)^2*PRE)+REC}$$

其中，*beta* 是召回率和准确率的相对权重。*beta* 等于 1 时，二者同样重要；*beta* 大于 1 时，准确率更重要一些；*beta* 小于 1 时，召回率更重要一些。在 MUC 系列会议中，*beta* 取值一般为 1、1/2、2。表 1 给出了 MUC3-7 分任务最优评测结果[8]。

表 1　MUC3-7 分任务最优评测结果

子任务 / 评测	命名实体	共指	模板元素	模板关系	场景模板	多语言
MUC-3					R<50% P<70%	
MUC-4					F<56%	
MUC-5					EJV F<53% EME F<50%	JJV F<64% JME F<57%
MUC-6	E F<97% C F<85% J F<93% S F<94%	R<63% P<72%	F<80%		F<57%	
MUC-7	E F<94% C F<91% J F<87%	F<62%	F<87%	F<76%	F<51%	

说明：R-召回率　P-准确率　F-F 指数（相对权重取 1）　JV-合资
E-英语　C-汉语　J-日语　S-西班牙语　ME-微电子

MUC 系列会议对信息抽取这一研究方向的确立和发展起到了巨大的推动作用。MUC 定义的信息抽取任务的各种规范以及确立的评价体系已经成为信息抽取研究事实上的标准。

近几年，信息抽取技术的研究与应用更为活跃。在研究方面，主要侧重于以下几方面：利用机器学习技术增强系统的可移植能力、探索深层理解技术、篇章分析技术、多语言文本处理能力、WEB 信息抽取（Wrapper）以及对时间信息的处理等等。在应用方面，信息抽取应用的领域更加广泛，除自成系统以外，还往往与其他文档处理技术结合建立功能强大的信息服务系统。至今，已经有不少以信息抽取技术产品为主的公司出现，比较著名的有：Cymfony 公司[4]、Bhasha 公司[5]、Linguamatics 公司[6]、Revsolutions 公司[7]等。

目前，除强烈的应用需求外，正在推动信息抽取研究进一步发展的动力主要来自美国国家标准技术研究所（NIST）组织的自动内容抽取（ACE，Automatic Content Extraction）评测会议[8]。这项评测从 1999 年 7 月开始酝酿，2000 年 12 月正式开始启动，迄今已经举办过两次评测（2000 年 5 月、2002 年 2 月），最近正在进行第 3 次评测（2002 年 9 月）。这项评测旨在开发自动内容抽取技术以支持对三种不同来源（普通文本、由自动语音识别 ASR 得到的文本、由光学字符识别 OCR 得到的文本）的语言文本的自动处理，研究的主要内容是自动抽取新闻语料中出现的实体、关系、事件等内容，即对新闻语料中实体、关系、事件的识别与描述。最近一次评测（ACE Phase 2 summer evaluation）主要有两大任务：实体识别与跟踪（EDT，Entity Detection and Tracking）、关系识别与描述（RDC，Relation Detection and Characterization）[9]。

与 MUC 相比，目前的 ACE 评测不针对某个具体的领域或场景，采用基于漏报（标准答案中有而系统输出中没有）和误报（标准答案中没有而系统输出中有）为基础的一套评价体系，还对系统跨文档处理（Cross-document processing）能力进行评测。这一新的评测会议将把信息抽取技术研究引向新的高度。

中文信息抽取方面的研究起步较晚，主要的研究工作集中在对中文命名实体的识别方面，在设计实现完整的中文信息抽取系统方面还处在探索阶段。其中，国立台湾大学（National Taiwan University）和新加坡肯特岗数字实验室（Kent Ridge Digital Labs）参加了 MUC-7 中文命名实体识别任务的评测[10][11]。Intel 中国研究中心的 ZHANG Yi-Min 和 ZHOU Joe F 等人在 ACL-2000 上演示了他们开发的一个抽取中文命名实体以及这些实体间相互关系的信息抽取系统，该系统利用基于记忆的学习（MBL，Memory-Based Learning）算法获取规则用以抽取命名实体及它们之间的关系[12]。

3　信息抽取系统的体系结构

Hobbs 曾提出一个信息抽取系统的通用体系结构[13]，他将信息抽取系统抽象为“级联的转换器或模块集合，利用手工编制或自动获得的规则在每一步过滤掉不相关的信息，增加新的结构信息”。

Hobbs 认为典型的信息抽取系统应当由依次相连的十个模块组成：

（1）文本分块：将输入文本分割为不同的部分——块。

（2）预处理：将得到的文本块转换为句子序列，每个句子由词汇项（词或特定类型短语）及相关的属性（如词类）组成。

[4] http://www.cymfony.com/index.html

[5] http://www.bhasha.com/

[6] http://www.linguamatics.com/index.html

[7] http://www.revsolutions.com/index.shtml

[8] http://www.itl.nist.gov/iad/894.01/tests/ace/

(5)过滤:过滤掉不相关的句子。

(4)预分析:在词汇项(Lexical Items)序列中识别确定的小型结构,如名词短语、动词短语、并列结构等。

(5)分析:通过分析小型结构和词汇项的序列建立描述句子结构的完整分析树或分析树片段集合。

(6)片段组合:如果上一步没有得到完整的分析树,则需要将分析树片段集合或逻辑形式片段组合成整句的一棵分析树或其他逻辑表示形式。

(7)语义解释:从分析树或分析树片段集合生成语义结构、意义表示或其他逻辑形式。

(8)词汇消歧:消解上一模块中存在的歧义得到唯一的语义结构表示。

(9)共指消解或篇章处理:通过确定同一实体在文本不同部分中的不同描述将当前句的语义结构表示合并到先前的处理结果中。

(10)模板生成:由文本的语义结构表示生成最终的模板。

当然,并不是所有的信息抽取系统都明确包含所有这些模块,并且也未必完全遵循以上的处理顺序,比如(6)、(7)两个模块执行顺序可能就相反。但一个信息抽取系统应当包含以上模块中描述的功能。

图1给出了美国纽约大学Proteus信息抽取系统[14]的体系结构,具有一定的代表性。

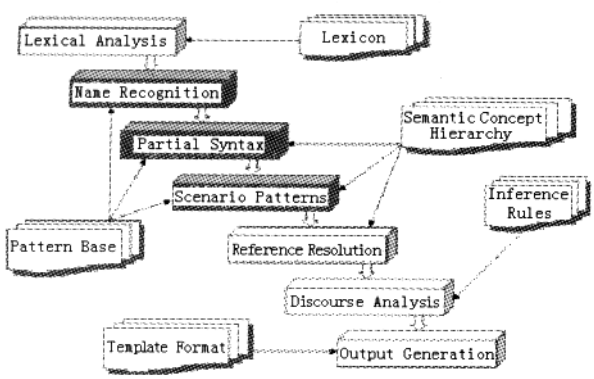

图1 纽约大学PROTEUS信息抽取系统体系结构

4 信息抽取中的关键技术

4.1 命名实体识别

命名实体是文本中基本的信息元素,是正确理解文本的基础。狭义地讲,命名实体是指现实世界中的具体的或抽象的实体,如人、组织、公司、地点等,通常用唯一的标志符(专有名称)表示,如人名、组织名、公司名、地名等。广义地讲,命名实体还可以包含时间、数量表达式等。至于命名实体的确切含义,只能根据具体应用来确定。比如,在具体应用中,可能需要把住址、电子信箱地址、电话号码、舰船编号、会议名称等作为命名实体。

命名实体识别就是要判断一个文本串是否代表一个命名实体,并确定它的类别。在信息抽取研究中,命名实体识别是目前最有实用价值的一项技术。根据MUC评测结果[8],英文命名实体识别任务的F-指数(召回率与准确率的加权几何平均值,权重取1)能达到90%以上。

命名实体识别的难点在于:

(1)在不同领域、场景下,命名实体的外延有差异;

(2)数量巨大,不能枚举,难以全部收录在词典中;

(3)某些类型的实体名称变化频繁,并且没有严格的规律可以遵循;

(4)表达形式多样;

(5)首次出现后往往采用缩写形式;

命名实体识别的方法主要分为:基于规则的方法和基于统计的方法。一般来说,基于规则的方法性能要优于基于统计的方法。但是这些规则往往依赖于具体语言、领域、文本格式,编制过程耗时且容易产生错误,并且需要富有经验的语言学家才能完成。相比而言,基于统计的方法利用人工标注的语料进行训练,标注语料时不需要广博的计算语言学知识,并且可以在较短时间内完成。因此,这类系统在移植到新的领域时可以不做或少做改动,只要利用新语料训练一遍即可。此外,基于统计的系统要移植到其他自然语言文本也相对容易一些。

4.2 句法分析

通过句法分析得到输入的某种结构表示,如完整的分析树或分析树片段集合,是计算机理解自然语言的基础。在信息抽取领域一个比较明显的趋势是越来越多的系统采用部分分析技术,这主要是由于以下三方面原因造成的[15]。

首先是信息抽取任务自身的特殊性,即需要抽取的信息通常只是某一领域中数量有限的事件或关系。这样,文本中可能只有一小部分与抽取任务有关。并且,对每一个句子,并不需要得到它的完整的结构表示,只要识别出部分片段间的某些特定关系就行了,得到的只是完整分析树的部分子图。

其次是部分分析技术在MUC系列评测中的成功。

SRI公司在其参加MUC-4评测的FASTUS系统[16]中开始采用层级的有限状态自动机(Cascaded Finite-State Automata)分析方法。该方法使FASTUS系统具有概念简单、运行速度快、开发周期短等优点,在多次MUC评测中都居于领先地位。

最后,部分分析方法盛行也是因为目前尚没有其他更好的选择。现在,完全分析技术的鲁棒性以及时空开销都难以满足信息抽取系统的需要。

但是,另一方面,也要清醒看到:部分分析技术只能使信息抽取系统的处理能力达到目前的水平(F-指数小于60%[17]),要想使其性能有大的飞跃,必须探索更有效的分析技术。

4.3 篇章分析与推理

一般说来,用户关心的事件和关系往往散布于文本的不同位置,其中涉及到的实体通常可以有多种不同的表达方式,并且还有许多事实信息隐含于文本之中。为了准确而没有遗漏地从文本中抽取相关信息,信息抽取系统必须能够识别文本中的共指现象,进行必要的推理,以合并描述同一事件或实体的信息片段。因此,篇章分析、推理能力对信息抽取系统来说是必不可少的。

初看起来,信息抽取中的篇章分析比故事理解中的篇章分析要简单得多。因为在信息抽取中只需要记录某些类型的实体和事件。但是,大多数信息抽取系统只识别和保存与需求相关的文本片段,从中抽取出零碎的信息。在这个过程中很可能把用以区分不同事件、不同实体的关键信息给遗漏了。在这种情况下要完成篇章分析是相当困难的。

除此之外,目前尚缺乏有效的篇章分析理论和方法可以借鉴。现有篇章分析理论大多是面向人、面向口语的,需要借助大量的常识,它们设想的目标文本也比真实文本要规范,并且理

论本身也没有在大规模语料上进行过测试。

信息抽取系统除了要解决文本内的共指问题外，还需要解决文本间的（跨文本的）共指问题。在文本来源比较广泛的情况下，很可能有多篇文本描述了同一个事件、同一个实体，不同文本间还会存在语义歧义，如相同的词有不同的含义、不同的词代表一个意思。为了避免信息的重复、冲突，信息抽取系统需要有识别、处理这些现象的能力。

由 MUC-6 和 MUC-7 对信息抽取系统部分篇章处理能力（即指称短语的共指消解）的评测结果看，篇章处理能力是目前信息抽取系统的弱项，是一个瓶颈，急需深入研究与改进。

4.4 知识获取

作为一种自然语言处理系统，信息抽取系统需要强大知识库的支撑。在不同的信息抽取系统中知识库的结构和内容是不同的，但一般来说，都要有：一部词典（Lexicon），存放通用词汇以及领域词汇的静态属性信息；一个抽取模式库（Extraction Patterns Base），每个模式可以有附加的（语义）操作，模式库通常也划分为通用部分和领域（场景）专用部分；一个概念层次模型（Ontology），通常是面向特定领域或场景的，是通用概念层次模型在局部的细化或泛化。除此之外，可能还有篇章分析和推理规则库、模板填充规则库等。

如前所述，信息抽取系统通常是面向特定应用领域或场景的。这种领域受限性决定了信息抽取系统中用到的主要知识是所谓的浅层知识。这种知识的抽象层次不高，通常只适用于特定应用领域，很难在其他领域复用。如果要把一个信息抽取系统移植到新的领域或场景，开发者必须要为系统重新编制大量的领域知识。一般说来，手工编制领域知识往往是枯燥的、费时的、易错的，费用较高，并且只有具有专门知识（应用领域知识、知识描述语言知识，熟悉系统的设计与实现）的人员才能胜任这种工作。另外，由于自然语言中存在的“长尾”综合效应（“long tail”syndrome）或称 Zipf 法则[9]，人工编制的知识库很难达到很高的语言覆盖面。因此，知识获取问题已经成为制约信息抽取技术广泛应用的一个主要障碍。它除了影响系统的可移植性外，也是影响系统性能的主要因素。正因为如此，近几年召开的多次专题学术研讨会都是以解决知识获取问题、建立具有自适应能力的信息抽取系统为主题的。

领域知识获取可以采用的策略通常有两种：手工+辅助工具（图形用户接口）；自动/半自动+人工校对。前者相对简单一些，人工工作仍然是主体，只是为移植者提供了一些图形化的辅助工具，以方便和加快领域知识获取过程。后者采用有指导的、无指导的或间接指导的机器学习技术从文本语料中自动或半自动获取领域知识，人工干预程度较低。实际上，这两种策略不是完全对立的，只是自动化程度高低不同而已。某种意义上讲，第一种策略仍然是一种人工编制知识库的过程，知识瓶颈问题只是得到某种程度的缓解。第二种策略才是解决信息抽取系统知识获取瓶颈问题的真正出路。近几年有不少研究者采用自扩展（Bootstrapping）技术从未经标注的语料中学习抽取模式[18]。

5 展望

信息抽取经过二十多年尤其是最近十多年的发展，已经成为自然语言处理领域一个重要的分支，其独特的发展轨迹——通过系统化、大规模的定量评测推动研究向前发展，以及某些成功启示，如部分分析技术的有效性、快速 NLP（Natural Language Processing）系统开发的必要性、知识工程研究以及软件工程技术的重要性等等[19]，都极大地推动了自然语言处理研究的发展，迫使 NLP 研究人员面向实际的应用重新考虑他们的研究重点，开始重视解决以前曾被忽视的一些深层问题，如语义特征标注、共指消解、篇章分析等等。

目前，影响信息抽取技术广泛应用的两个最主要的因素是：系统性能和系统可移植能力[18]。因此，今后信息抽取研究将紧紧围绕如何克服和解决这两个问题展开，重点解决知识获取、篇章分析、高效句法分析等问题，不断提高信息抽取系统的性能、增强其可移植能力。

未来的信息抽取系统将是动态（Dynamic）的、开放域（Open Domain）的[20]，前景光明。（收稿日期：2002 年 12 月）

参考文献

1.Applet D E,Israel D J.Introduction to Information Extraction Technology.A Tutorial for IJCAI-99,1999

2.Gaizauskas R,Wilks Y.Information Extraction:Beyond Document Retrieval[J].Journal of Documentation,1997

3.Sager N.Natural Language Information Processing.Reading,Massachusetts:Addison Wesley,1981

4.Dejong G.An Overview of the FRUMP System[C].In:LEHNERT W,RINGLE M h eds.Strategies for Natural Language Processing,Lawrence Erlbaum,1982:149~176

5.Grishman R,Sundheim B.Message Understanding Conference-6:A Brief History[C].In:Proceedings of the 16h International Conference on Computational Linguistics(COLING-96),1996-08

6.Chinchor N,Marsh E.MUC-7 Information Extraction Task Definition(ersion 5.1)[C].In:Proceedings of the Seventh Message Understanding Conference,1998

7.Douthat A.The Message Understanding Conference Scoring Software User's Manual[C].In:Proceedings of the Seventh Message Understanding Conference,1998

8.Chinchor N.Overview of MUC-7/MET-2[C].In:Proceedings of the Seventh Message Understanding Conference,1998

9.The ACE 2002 Evaluation Plan.ftp://jaguar.ncsl.nist.gov/ace/doc/ACE-EvalPlan-2002-v06.pdf,Site visited on August 30th,2002

10.Chen H H,Ding Y W,Tsai S C et al.Description of the NTU System Used for MET2[C].In:Proceedings of the Seventh Message Understanding Conference,1998

11.Yu S H,Bai S H,Wu P.Description of the Kent Ridge Digital Labs System Used for MUC-7[C].In:Proceedings of the Seventh Message Understanding Conference,1998

12.Zhang Y M,Zhou J F.A Trainable Method for Extracting Chinese Entity Names and Their Relations[C].In:Proceedings of the Second Chinese Language Processing Workshop,Hong Kong,2000-10

13.Hobbs J.The Generic Information Extraction System[C].In:Proceedings of the Fifth Message Understanding Conference(MUC-5),Morgan Kaufman,1993:87~91

14.Yangarher R,Grishman R.NYU:Description of the Proteus/PET System as Used for MUC-7[C].In:Proceedings of the Seventh Message Understanding Conference,1998

（下转 66 页）

[9]绝大多数事实采用经常出现的、非常少量的表达方式来描述，而剩余的事实却需要大量的、不经常出现的表达方式才能覆盖。

表 1　第一步特殊化时子空间上的搜索结果

搜索空间		最佳句节	覆盖的例子	加权信息赢取
Prolog 子空间		¬r (A)	9⊕/11⊖	1.8147
CLP 子空间	>/2,</2	−66X+82Y−101>0	9⊕/9⊖	3.1827
	=/2	X−4Y+38=0	2⊕/0⊖	1.2749
	≠/2	X+Y−16≠0	9⊕/11⊖	1.8147

第一步：由加权信息赢取，选择线性不等式约束 $-66X+82Y-101>0$ **作为整个搜索空间的最佳特殊化句节。当前子句为：** $t(A,B,X,Y)\leftarrow -66X+82Y-101>0$**，覆盖 9 个正例/9 个负例。**

表 2　第二步特殊化时子空间上的搜索结果

搜索空间		最佳句节	覆盖的例子	加权信息赢取
Prolog 子空间		$p(A)$	9⊕/6⊖	2.3673
CLP 子空间	>/2,</2	342X−259Y+1010>0	9⊕/7⊖	1.5293
	=/2	X−4Y+38=0	2⊕/0⊖	2.0000
	≠/2	X+Y−16≠0	9⊕/6⊖	2.3673

第二步：由加权信息赢取，选择 $p(A)$ **作为整个搜索空间的最佳特殊化句节。当前子句为：** $t(A,B,X,Y)\leftarrow -66X+82Y-101>0, p(A)$**，覆盖 9 个正例/6 个负例。**

表 3　第三步特殊化时子空间上的搜索结果

搜索空间		最佳句节	覆盖的例子	加权信息赢取
Prolog 子空间		$q(A,B)$	9⊕/3⊖	2.8974
CLP 子空间	>/2,</2	281X−344Y+2239>0	9⊕/4⊖	1.8581
	=/2	X−4Y+38=0	2⊕/0⊖	1.4739
	≠/2	X+Y−16≠0	9⊕/3⊖	2.8974

第三步：由加权信息赢取，选择 $q(A,B)$ **作为整个搜索空间的最佳特殊化句节。当前子句为：** $t(A,B,X,Y)\leftarrow -66X+82Y-101>0, p(A), q(A,B)$**。覆盖 9 个正例/3 个负例。**

表 4　第四步特殊化时子空间上的搜索结果

搜索空间		最佳句节	覆盖的例子	加权信息赢取
Prolog 子空间		$p(A)$	4⊕/1⊖	0.3724
CLP 子空间	>/2,</2	1239X−1406Y+8652>0	9⊕/1⊖	2.3673
	=/2	X−4Y+38=0	2⊕/0⊖	0.8301
	≠/2	X+Y−16≠0	9⊕/0⊖	3.7353

第四步：由加权信息赢取，选择 $X+Y-16\neq 0$ **作为整个搜索空间的最佳特殊化句节。当前子句为：** $t(A,B,X,Y)\leftarrow -66X+82Y-101>0, p(A), q(A,B), X+Y-16\neq 0$**。覆盖 9 个正例/0 个负例。**

当前定义子句不覆盖任何负例，即满足一致性，目标谓词的第一个定义子句学习完毕。当前已学习到的定义子句覆盖全部正例，即满足完备性，系统运行结束。算法学习到的目标谓词定义子句集包含一个子句，即 $t(A,B,X,Y)\leftarrow -66X+82Y-101>0, p(A), q(A,B), X+Y-16\neq 0$**。**

6　结论

论文提出了一种以自顶向下 ILP 系统为基础的 CILP 新方法。该方法中引入了模式识别领域中的 Fisher 法，克服了现有方法的一些不足，能够无诱导地导出不受变量个数限制的多种线性约束，从而由正负例和背景知识学习出含有约束的一阶谓词公式，且算法的实现不依赖约束求解器。

在 CILP 这个极具挑战性的领域中，未来的工作将主要致力于：研究如何导出多个变量的非线性约束；研究 CILP 的噪音处理方法；研究如何将 CILP 方法应用于 KDD 数据库知识发现。**（收稿日期：2002 年 11 月）**

参考文献

1.Nienhuys-Chen S-H,Wolf R de.Foundations of Inductive Logic Programming[C].In:Goos,Hartmanis eds.Lecture Notes in Computer Science 1228,Berlin:Springer-Verlag,1997:163~177

2.Quinlan J R.Learning logical definitions from relations[J].Machine Learning,1990;5:239~266

3.Muggleton S.Inverse entailment and Progol[J].New Generatin Computing,1995;13:245~286

4.刘椿年.约束逻辑程序设计 CLP—现状与未来[C].见：陆汝钤.世纪之交的知识工程与知识科学.北京：清华大学出版社,2001:251~279

5.Muggleton S,Page C D.Beyond first-order learning:Inductive logic programming with higher-order logic[R].Technical Report PRG-TR-13-94,Oxford University,Oxford,1994

6.Srinivasan A,Camacho R.Experiments in numeric reasoning with inductive logic programming[R].Technical Report PRG-TR-22-96,Oxford University,Oxford,1996

7.Sebag M,Rouveirol C.Constraint Inductive Logic Programming[C].In:de Raedt eds.Advances in ILP,IOS Press,1996:277~294

8.Anthony S,Frisch A.Generating numerical literals during refinement[C].In:Dzeroski,Lavrac eds.Proceedings of the 7th International Workshop on Inductive Logic Programming.Lecture Notes in Computer Science 1297,Berlin:Springer-Verlag,1997:61~76

9.边肇祺,张学工.模式识别[M].北京：清华大学出版社,1999:87~90

（止接 5 页）

15.Grishman R.Information Extraction:Techniques and Challenges[C].In:M-T Pazienza,editor,Information Extraction:a Multidisciplinary Approach to an Emerging Information Technology,Springer,Berlin,1997

16.Hobbs J,Appelt D,Bear J et al.FASTUS:A Cascaded Finite-State Transducer for Extracting Information from Natural-Language Text[C].In:Roche,Schabes eds.Finite State Devices for Natural Language Processing,MIT Press,Cambridge MA,1996

17.Appelt D E.Introduction to Information Extraction[J].AI COMMUNICATIONS,1999;12:()

18.Yangarber R.Scenario Customization for Information Extraction[D].Ph D Thesis.New York University,2001-01

19.Cowie J,Lehnert W.Information Extraction[J].Communications of the ACM,1996;39:()

20.Grishman R Adaptive information extraction and sublanguage analysis[C].In:Proceedings of IJCAI-2001 Workshop on Adaptive Text Extraction and Mining,2001