汉语名词短语和动词短语的自动识别方法研究

李 荣　曹建芳◎著

兵器工业出版社

汉语名词短语和动词短语的自动识别方法研究

李 荣 曹建芳 著

兵器工业出版社

内 容 简 介

短语识别是中文信息处理领域的一个重要组成部分。本书面向中文信息处理的实际需要，介绍了用规则方法识别汉语名词短语和动词短语的过程，然后介绍了用隐马尔可夫模型识别汉语名词短语，用支持向量机识别汉语动词短语的过程。在此基础上，探讨了解决计算机分析汉语短语结构碰到的各类歧义问题的途径。

本书可作为高等院校计算机专业高年级学生的教学参考书，也可供从事中文信息处理及人工智能研究的相关人员参考。

图书在版编目（CIP）数据

汉语名词短语和动词短语的自动识别方法研究/李荣，曹建芳著. —北京：兵器工业出版社，2008.5

ISBN 978-7-80248-031-5

Ⅰ.汉… Ⅱ.①李…②曹… Ⅲ.①汉语-名词-短语-自动识别-研究②汉语-动词-短语-自动识别-研究

Ⅳ.H146.3-39

中国版本图书馆 CIP 数据核字（2008）第 046960 号

出版发行：兵器工业出版社
发行电话：010-68962596，68962591
邮　　编：100089
社　　址：北京市海淀区车道沟 10 号
经　　销：各地新华书店
印　　刷：北京业和印务有限公司
版　　次：2008 年 6 月第 1 版第 1 次印刷

责任编辑：周宜今
封面设计：揽胜视觉
责任校对：郭　芳
责任印制：赵春云
开　　本：787×1092　1/16
印　　张：16.25
字　　数：235 千字
定　　价：38.00 元

前　言

中国正在频频地叩击信息化的大门，甚或可以说我们一只脚已经跨进了信息化的门槛。在我们面前伸展着一条有着无穷希望的路，但却是一条充满荆棘坎坷的路。这荆棘与坎坷，是尚待完善的发展信息事业的体制和机制，是信息技术科学研究的蹒跚步履。在信息技术中，中文信息处理是其关键之一。

中文信息自动化处理事业需要文、理、工三科的紧密结合。到20世纪90年代，我国在中文（口语和文本）信息的形式化，以及算法、编程、产业工程等环节上已经具备了相当实力，这为中文信息处理技术的发展和实用化准备了必要的条件。与之失衡的，是中文信息处理的基础研究和应用研究还没有跟上，而其最大难点，或曰拦路虎，是我们对中文了解得不够；其中又以对语义的认识并使之形式化还缺少能够全面解决的办法为突出。信息技术正在快速地渗入到我们工作与生活的方方面面，从而改变着我们的工作方式与生活方式。世界各国的传统观念与文化正面临着巨大的挑战，如何应对这种挑战，使之适应这种新的生活与工作方式，已成为全人类必须认真考虑的问题。

随着过去10年人们获得信息能力的指数增加，编码与输入方式的研究已不能满足人们在各个领域的需求，大量文本形式的信息使得各行各业的从业人员应接不暇。目前，这类信息的利用率不到1%。如何有效地利用信息，以提高生产率与生活质量，是今后10年全世界普遍关注的问题，也是国际竞争的主要焦点。由于我们获得与交流的信息有70%～80%是以语言文本形式出现的，因此，语言文本的处理成为关键问题之一。

我国开展汉语处理的研究是在70年代末，其原始目标是机器翻译。在最近30年中，经大批学者的艰苦努力，汉语处理的研究取得了重要的进展，然而，距实现这个目标还有相当长的路，还存

在着诸多理论与技术问题需要解决，特别是，组织的协调、研究方向的确立、理论与应用目标的结合等更是问题的关键。

为此，国家重大基础研究发展计划项目“图像、语音、自然语言理解与知识挖掘”专家组根据目前国内研究的现状与存在的问题，在2001年11月30日专家组会议决定，使用本项目2/3特别经费加强“汉语处理理论与建设可共享的汉语基础设施”的研究，并委托东北大学姚天顺教授、中国科学院自动化研究所徐波研究员与清华大学孙茂松教授组成筹备组，就上述问题，组织一次会议，以解决目前存在的问题。会议筹备组拟订了这次会议集中讨论的六个问题：

1. 自然语言处理发展里程的经验与教训；
2. 对统计自然语言处理的认识；
3. 语义及概念体系在自然语言处理中的作用；
4. 语料库建设及制定相关规范的可行性；
5. 机器学习与自然语言处理；
6. 自然语言处理研究阶段成果的可能应用领域。

目前，在国际上，自然语言处理有基于语言规则的流派、基于统计的流派，以及其他流派，如基于实例的流派。这些流派在汉字处理的研究中都有体现（姚天顺）。鉴于自然语言处理研究的不成熟性，没有一种现有理论可以独立概括这个研究的全部（冯志伟）。因此，在现阶段，多种流派并存是一件好事，“取长补短，百家争鸣”是当前必须采用的研究策略，试图以提倡一种理论，而压制其他理论的发展的思维，至少在当前的研究水平下是不可取的。另外，加强自然语言处理的新理论的发现是专家组一再强调的，同时，专家组也明确指出，无论提出何种新的理论与方法，必须以严谨的科学论证为基础，必须经得起语言学与计算机科学实践的锤炼。

在国内，汉语处理的研究现状是，汉语处理的基本方法已被越来越多的人所掌握，但是，缺少“大工程实施组织的魄力，绣花般的精雕细刻的耐心”（孙茂松）。这一方面说明，我国汉语处理的研究已取得了重要的进展，有些成果已可以走出实验室；另一方

面，对汉语处理研究的专业人士提出更高的要求，必须从大工程着眼，从“绣花”做起。

试图建立单一的能包打天下的语言计算处理理论的努力已经证明是不可行的（王钰），无论是语言规则流派、统计流派还是其他流派，都是建立语言的局部模型，进一步组装成完整模型。这样，“绣花般精雕细刻”的研究就凸现了重要性。“绣花般精雕细刻”研究的本质是提倡对自然语言局部现象进行深入独立的研究，这也是针对目前国内流行的以通用机器翻译系统为目标研究的批评。事实上，由于自然语言现象的复杂性与进化性，几乎不可能存在一种语言学理论与计算理论，可以概括自然语言的所有现象，因此，对自然语言处理的个别现象进行深入研究就是必要的，它是提高自然语言处理系统水平的必要条件之一。例如，在机器翻译的研究中，产生的英语文本，几乎没有系统可以准确地加上词尾上的“s”，这是一个需要研究的课题，而不仅仅是一个补丁式的措施（董振东）。尽管以系统为目标的研究是重要的，但是，如果没有这样的绣花式的研究，“大家能做的，我能做；大家不能做的，我也不能做”的局面将无法改变。另外，自然语言处理的困难在于人们对正确率的要求十分高，5% 的错误率，用户就可能不使用这个系统，这并不难理解，例如，一篇文章由一百个句子组成，其中有 5 个句子是不通顺的，人们将不会耐心阅读它。列出在汉语处理中迫切需要解决的“绣花般精雕细刻”问题的清单是需要进一步研究的问题。必须指出的是，这类研究也是产生新的自然语言处理理论与方法的源泉。

目前，面向人写的汉语语法书已经非常多了，面向计算机写的汉语语法书，尤其是现代汉语短语自动识别方面的书籍则还很少见。众所周知，计算机处理自然语言困难重重，最常被人们提及的恐怕莫过于计算机不懂得人类所用的自然语言这一问题。而要让计算机理解自然语言，处理中文信息，需要先让计算机学会自动句法分析，而短语识别是一种浅层句法，《汉语名词短语和动词短语的自动识别方法研究》可以看做是在前人已经开始的许多研究工作的基础上，为句法分析所迈出的新的一步。

本书面向中文信息处理的实际需要，先描述了用规则方法对真实文本中的名词短语和动词短语进行标注和划分的过程，然后分别介绍了用两种统计方法，即隐马尔可夫模型和支持向量机法识别汉语真实文本中的名词短语和动词短语的过程，在此基础上，探讨了解决计算机分析汉语短语结构碰到的各类歧义问题的途径。全书共分9章。

第1章是引论部分。主要是对论著主题进行解析，并阐述计算机自动短语识别的研究意义、研究难点和国内外研究现状。

第2章阐述了汉语的计算机理解。介绍了汉语的特点和汉语理解中的特殊问题。

第3章阐述了汉语短语的基本知识。介绍了汉语短语的标注体系和短语的组成定义，并确立了短语的句法功能分类框架。

第4章详细描述了基于规则的汉语短语识别过程。首先，对汉语名词短语和动词短语的结构进行了统计与分析；接着，分别确定了汉语名词短语和动词短语的定界规则；然后，尝试尽可能全面而细致地对现代汉语名词短语和动词短语的句法语义进行分析；最后，介绍了基于规则的汉语名词短语和动词短语的自动识别系统，其中，分别就识别算法、实验系统和实验数据分析进行了详细介绍。

第5章详细描述了基于隐马尔可夫模型的名词短语识别过程。首先对隐马尔可夫模型和层次分析法进行介绍；接着，描述了实验系统中的资源建设方法、资源组成和资源建设过程。然后，提出了隐马尔可夫模型的设计思想，其中，就隐马尔可夫模型的建立、隐马尔可夫模型的参数估计和汉语名词短语的识别过程进行了详细介绍。最后，给出了系统总体设计框图、总体算法流程图，介绍了主要模块的算法设计思想，并对整个模型生成的结果进行分析、评估和总结。

第6章详细描述了基于SVM的动词短语识别过程。本章的结构与第5章类似。首先对支持向量机技术作了概述；接着，描述了汉语动词短语的静态和动态特征提取；然后，介绍了基于支持向量机法的动词短语识别过程；最后，给出了实验模型的结果分析和

总结。

第7章细致分析了短语结构歧义类型与消解策略分析。现代汉语短语结构歧义是进行汉语句法分析的一大主要障碍。通过全面调查，本章详细分析了计算机处理汉语短语结构时面临的定界歧义和结构关系歧义问题，从不同角度区分了抽象的歧义格式的不同类型，分为包含终结符的歧义格式与不包含终结符的歧义格式、外显型歧义格式与内含型歧义格式以及真歧义格式、准歧义格式与伪歧义格式，并对这些类型的短语歧义格式进行了举例分析。针对一些典型的歧义格式，本章还对排歧策略方面进行了探讨，提出了一些短语结构歧义消解方法，并进行举例分析。

第8章介绍了关于短语识别的评测问题。具体描述了评测在软件开发中的位置和评测模型的定义，给出了短语识别的评测框架，并介绍了部分评测的实现过程。

第9章是结语部分。本章对本课题研究工作进行了全面总结，简要概括了本课题研究取得的主要成绩，讨论了本课题研究工作对中文信息处理研究的意义，通过本课题中规则法和统计法的短语识别实验，对规则方法和统计方法在中文信息处理领域中的应用进行了比较，并提出了进一步研究的计划和目标。

本书的研究工作是跨现代汉语语法和中文信息处理两个领域进行的。一方面，研究的具体结果对推进中文信息处理技术的发展有直接的应用和参考价值；另一方面，从中文信息处理的角度来审视现代汉语语法研究，可以为研究工作提供一个清晰的实用背景，可以注意到以往面向人的研究不容易注意到的一些问题。作者希望本书对从事汉语信息处理实际应用开发工作的科研人员、在计算机语言学这一交叉学科领域辛勤耕耘的研究人员，以及汉语语法研究工作者，都能起到一定的参考作用。

需要指出的是，本书得到了山西省忻州师范学院科研基金的资助（200623）。

书中内容在得到许多专家学者的指导和宝贵意见后经过若干次调整修正，但错误疏漏之处，恐仍难免。在请读者包涵谅解的同时，也恳请专家同行多批评指正。

目　录

第1章 引 论

1.1 课题的提出

本书的研究工作是尝试利用基于规则和基于统计的方法进行名词短语和动词短语识别研究，并对现代汉语短语结构歧义类型进行分析，以解决计算机的浅层句法分析问题。

有别于以往主要是面向人的语法研究，本课题的研究是面向计算机的。语法研究的应用对象由过去主要是面向人发展到现在还面向计算机，而且后一个面向显得越来越迫切和重要，这是计算机科学技术飞速发展以及信息社会对信息自动化处理的要求不断提高的必然结果。而目前的信息处理技术，越来越多地需要对自然语言进行深层分析，比如机器翻译、自动文摘等，就是如此。开发这类应用系统，要求计算机掌握尽可能多的有关自然语言的知识和非语言知识，前者又包括句法知识、语义知识乃至语用知识等。

衡量一个自然语言处理系统的水平，可以看它处理到语言单位中的哪个层级，同时也要看它对不同性质的语言知识掌握到了一个什么样的水平。无论是比较传统的基于规则的处理策略，还是20世纪90年代以来方兴未艾的基于统计的方法，在对语言知识的需求这一点上实际上都是共同的。所不同者，是走规则路线的研究者一般诉诸于专家的理性知识，由人来对语言知识进行抽象（比如以带有合一条件的规则形式给出），而走统计路线的研究者一般求助于计算机对大规模语料库的统计分析，由计算机来抽象出语言知识（比如以一定的数据结构记录的统计结果等）。两种路线孰优孰劣，不能笼统判断，只能跟具体的应用目标结合起来由实践结果来评价。统计方法已经在像语音识别、自动分词和词性标注这样相对浅层的自然语言处理中有不俗表现，但在深层分析方面（比如分析句子的树结构或者句

中成分间语义关系等）还没有显出特别的优势。于是又有学者提倡把两种方法结合起来使用（比如通过统计，给出带有概率值的规则）。在我们看来，无论采用哪种方法，首先都要求人自身先对自然语言有深入的了解。就规则方法来说，这一点是显然的；就统计方法来说，虽不那么明显，但道理也是一样的。现有的对自然语言深层知识的统计，一般是建立在经过标注的熟语料库基础上的。而从生语料库到熟语料库，就具体的加工方式而言，当然有人工方式，也有计算机自动加工方式或者人机互助的方式等，但加工什么内容，标注哪些信息，仍然取决于人对自然语言的认识。

具体到中文信息处理方面。如果以处理对象的单位大小为指标，宏观地看，中文信息处理技术已经走过了字处理阶段，分词和词性标注（词处理阶段）也有了基本可以实用的成果，目前可以认为是进入到句处理的前期阶段，即如何来对短语结构进行自动分析的阶段，包括划定短语边界、分析短语结构的内部句法关系、给出结构成分间的语义关系等不同深度的分析。这样定位，并不是无视目前也有研究者在开展更大单位（譬如篇章）上的中文信息处理研究，只是就这个领域的总体发展情况来看，目前以在句子一级上开展研究工作为主。另外本着解决问题由易到难，由简单到复杂的原则，把中文信息处理目前的发展定位在重点解决短语识别这么一个阶段，也是适宜的。

1.2 面向计算机的语言学研究工作的模式

面向计算机的语言学研究工作，涉及的范围相当广泛。研究模式根据工作内容或者目标的不同也有差别。这里不做全面讨论。下面概述跟发掘语言知识相关的研究工作的一般模式，目的是为本课题的研究工作勾勒工作方式上的背景。

面向计算机开展语言研究工作的语言学家实际上可以看作是处在跟计算机以及语料形成的一个三角关系中，如图 1.1 所示。

这是一个抽象环境尽量简化的示意。箭头表示谁“向”谁，或

谁“对”谁，比如语言学家向计算机提供语言知识，计算机对语料进行分析等。其中语料以不规则形状表示，是指其范围不确定。内部以不规则线分割，是表示语料可以按照不同标准大致区分为：包含各种复杂情况的真实语料与比较理想的“纯语料”；或者未经加工的生语料与所谓加工过的熟语料；或者合法的（或可接受的）语料与不合法的（或不可接受的）语料等不同情况。此外，这里的“计算机”是指由硬件加软件构成的一个信息处理系统，换言之，图 1.1 已经把计算机程序设计员这个也是由人充任的角色包含在“计算机”中了。

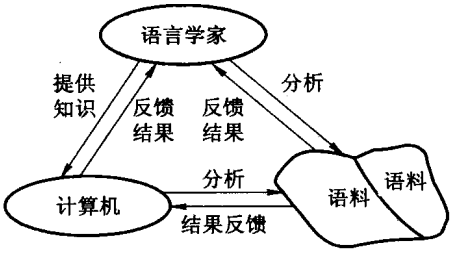

图 1.1　语言学家、计算机、语料构成的三角关系

前面提到过的两种主要的自然语言处理策略，基于规则和基于统计的方法，在图 1.1 中也都能得到反映。如果采用基于统计的方法，注重点就在图 1.1 等腰三角的底边上，即由计算机对语料（一般得是熟语料）进行统计得到语言知识（一般表现为参数），再利用得到的参数对语料进行分析，根据分析得到的反馈结果来调整已有的参数，从而提高分析能力。比如基于错误驱动的语法规则自动学习方法的研究，计算机可以用错误率为指标评价当前习得的语言知识的适用水平，进而做相应的调整。也就是说，语言学家可能提供一些初始的语言知识（譬如将生语料加工成熟语料的工作），而大部分归纳和调整语言知识的工作是由计算机来完成的。如果走基于规则的研究路线，注重的就是上图等腰三角的两个腰，即由语言学家提出一套语言知识（比如可以从对有限语料的分析中初步归纳出来）给计算机用，再根据计算机反馈的结果来改进原来的语言知识，形成一个循环处理系统。显然，在这个系统中，给出和调整语言知识这两步工作都主要是由语言学家完成的。

1.3 开展本课题研究工作的基础

开始研究工作之前，无疑应该先看看，对于解决本课题所关注的问题，前人在进行句法分析或短语识别方面，都做了哪些研究，采用的是基于规则的方法还是基于统计的方法，是统计方法中的哪一种技术，短语识别过程中采用什么样的思想，对这些方面的研究都是十分必要的。

1.3.1 研究意义

近年来，计算机的普及得以加速进行，国际互联网已进入千家万户，网上信息已经呈爆炸趋势。面对浩如烟海的网上信息，人们越来越需要搜索引擎、机器翻译、信息提取等技术的帮助，各大网络公司都在为改进和开发这方面的产品而努力。从另一方面说，由于以文本信息的智能化处理为主要对象的语言工程已成为国际上关注的热点，计算机必须从传统的对文本形式加工，发展到对文本内容进行加工处理，才能从 Internet 网上大量没有经过预先处理的非结构化的生语料中，提取有用的知识。而且，Internet 是一个以英语为主导语言的网络，极大地限制了我国在网上进行信息交换与实现资源共享的能力。这就迫切需要中文信息处理技术的发展必须适应当前形势的需求。而如何实现语言的计算机自动理解也正是当前计算语言学面临的一项难题。其实这项技术的关键是计算机的语言分析技术。汉语的计算机自动理解过程涉及分词、词性标注、短语标注、短语分析、语义理解等多级层次。每一层次的加工完成都需要形式化了的语言知识的介入。尽管国外计算语言学的发展为我们提供了多种语言知识形式的方法，但是，汉语知识本身的欠缺使在每个层次上汉语的自动理解都依然面临着重重困难。在分词和词性标注基础上，短语自动标注和短语分析研究，在国内也有了近十年的历史。近几年来，对汉语短语分析方法、依存关系标注、基本句型分析等方面的探索，为进行比较系统全面的短语分析积累了丰富的经验。自然语言处理经历了字处理、词处理阶段，在理论上和实