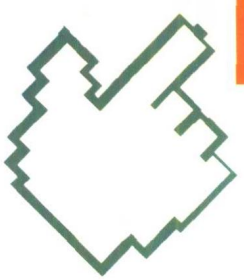
ISSUES IN CHINESE
INFORMATION PROCESSING

中文信息处理若干重要问题



徐 波 孙茂松 靳光瑾 主编

科学出版社
www.sciencep.com

中文信息处理若干重要问题

徐　波　孙茂松　靳光瑾　主编

科学出版社

北　京

内 容 简 介

受国家重点基础研究发展计划("973"计划)项目"图像、语音、自然语言理解与知识挖掘"(项目编号:G19980305)总体专家组的委托,编者邀请了国内近40位中文信息处理领域的专家学者,分别从中文信息处理的理论和方法、语义体系、语料库建设及规范制定、机器学习和语言处理、应用和技术等五个方面,对该领域的过去、现状以及未来的发展方向做了系统的阐述。各篇论文涉及不同的主题或侧面,论述全面而深入,是我国中文信息处理领域多年来研究和开发工作的结晶,对推动我国中文信息处理领域的发展具有重要意义。

本书可供从事中文信息处理研究人员及高等院校相关专业的师生参考。

图书在版编目(CIP)数据

中文信息处理若干重要问题/徐波,孙茂松,靳光瑾主编. —北京:科学出版社,2003
ISBN 7-03-012296-8

Ⅰ.中… Ⅱ.①徐…②孙…③靳… Ⅲ.汉字信息处理 Ⅳ.TP391.12

中国版本图书馆CIP数据核字(2003)第034458号

责任编辑:童安齐 沈 建/责任校对:包志虹
责任印制:刘士平/封面设计:耕者设计工作室

科学出版社 出版
北京东黄城根北街16号
邮政编码:100717
http://www.sciencep.com
新蕾印刷厂 印刷
科学出版社发行 各地新华书店经销

*

2003年11月第 一 版 开本:787×1092 1/16
2003年11月第一次印刷 印张:29
印数:1—2 000 字数:664 000

定价:58.00元

(如有印装质量问题,我社负责调换〈新欣〉)

序

——中文信息处理的研究应该走向高潮

许嘉璐

去年9月，国家重大基础研究计划“图像、语音、自然语言处理与知识挖掘”项目的专家们举行“自然语言处理若干重要问题”的研讨会，我因出访未能到会，失去了借此获得研究新进展的信息和了解专家们研究新思路的机会。时隔一年，那次会议上的成果已经结集，准备出版。参与操持其事的靳光瑾博士，代表编者要我写一篇序。我是乐于接受这项任务的。因为参加这次会议的许多学者是我多年的好友或研究工作的合作者，有些先生虽然未曾谋面，却都是我所钦佩的专家，能够在他们的论文集中表达我的敬意和祝愿，是我的荣幸，同时这也是我和他们的一次书面交流。

看了论文集中的部分稿件，十分高兴。一是专家们讨论的范围广，从自然语言处理的理论、方法、经验、教训，到研究成果的应用，都有所涉及；二是关于语义和语料库都有专门论述；三是在这些论文的背后，往往有着艰苦的实践、深层的思考、大胆的探索，文章所述多为实践的最近成果。可以说，这本论文集展示了当前从事中文信息处理研究的学者们所关注的方方面面；也可以说，上述三个方面都是今后突破中文信息处理学术和技术难关所必须解决的症结所在。

的确，现在我们遇到了难关。

中国正在频频地叩击信息化的大门，甚或可以说我们的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，但这却是一条充满荆棘坎坷的路。这荆棘与坎坷，是尚待完善的发展繁荣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能跟上，而其最大难点，或曰拦路虎，是我们对中文——一般指的是现代汉语——了解得不够；其中又以对语义——包括总结其规律——的认识并使之形式化还缺少能够全面解决的办法为突出。

汉语语义在组织句、段、篇，即形成口语或文本时有着特殊的功能，在这本论文集里，陆俭明、黄曾阳、陆汝占等先生的文章，分别从不同视角，运用不同的理论分析或介绍了对汉语语义的认识。他们三位的理论和方法已经初步显示出不同的风格，这是形成不同流派的开端。在一个研究领域里出现流派，这是该问题受到普遍重视、学科即将昌盛发达的标志。

汉语的语义问题太复杂了，但无论多么复杂的事物都有规律可循。惟其复杂，所以需

要更多的人，费很长时间（也许需要几代人），刻苦地研究，经过反复失败、反复探索，逐渐接近完善；又因其有规律可循，所以中国人一定可以自己解决这个拦路的难题。在这个相当漫长的过程中，我们既需要多种风格流派的竞争、互学互补，又需要自然地形成的发展的若干阶段，每个阶段都将以转化为应用技术为标志。

中国自古重视语义的解释和对其演变规律的探索，其成果集中体现在训诂学中。在训诂学对语义的解释中，语言环境和语义分类受到极大关注，这或许是训诂学区别于西方语义研究的特色。这一特色，是历代学者徜徉于语言事实的海洋中体味到的，换言之，是基于语言实际的。古人的这些经验（其中也有不少反面经验），是我们今天解开现代汉语语义之谜的重要参考。我们在陆俭明等先生的论述中已经明显地感到了传统学术的影响和对前人成果的关注。如何看待中国独有的训诂学，能否利用之，这或许又是解决中文信息处理应该注意的问题之一。

无论建立什么样的理论，使用什么样的方法，要研究口语和文本自动化处理问题，几乎都离不开大规模语料库作为基本工具。建立语料库当然不是目的，从一定意义上说，仅如飞机的试飞场，轮船下水后的试航区。中文信息处理的解决方案和技术需要在语料库中反复检验、逐步完善，才能达到比较理想的彼岸。因此建立、建设、开发语料库就是中文信息处理极为必要的工作环节了。这本论文集关于语料库的论文，是值得重视的。尤其让人高兴的是论文集中提出了“中文语言资源联盟”（ChineseLDC）的概念和会外建立这一联盟实践的启动。这一动向之所以重要，一是中文信息处理研究需要资源共享，二是多年来分散的相关资源缺乏整合，既浪费又不便。现在学者们不是依靠政府，而是自觉自愿地结成联盟，促进语言资源建设。我甚至觉得应该把这件事看做中文信息处理技术发展的一座里程碑：它标志着分散的研究力量开始了走向携手的阶段，虽然这种合作还是极初步的，还只限于资源互通一端，但是毕竟打破了相对封闭的局面。

中文信息自动化处理的终极目标之一，是机器翻译。几十年来我国的机器翻译研究始终没有停止过，近年来出现的许多可喜的动向反映了我国这一领域研究的新进展。一个是在研究文本翻译的同时，与语音识别与合成相结合，开始了口语翻译的探索，并取得了一定进展；一个是在研究汉语翻译的同时开始了少数民族语言的翻译研究；一个是有些成果已经在社会生活中试用。在这本论文集中这方面的论文并不多，但却反映了学术界最新的思考。

作为通向遥远目标途中必然遇到、也极为可能解决的诸多应用问题，一直为中文信息处理界的学者们所关心、所致力。这是非常可贵的。应用，的确是中文信息处理的“重要问题”。这不仅是因为中文信息处理本身就是应用学科，而且经济发展和社会发展对信息化的需求和渴望，也在吸引着学者们的目光，催促着他们的脚步。论文集中这方面的论著较多，尽管其中泛论和呼吁者占了相当比例，但一组文章的整体效应应该引起高度重视。确实，我们不能老在制作工具或写作论文中徘徊，总要用较成熟的理论、已实现的技术不断为社会提供可以实用的成果。这类成果将在使用中进一步检验已有理论和技术，为继续改进提供依据。

学者们孜孜矻矻多年，就是要实现中文信息的自动化处理，因为这对于我国的信息化事业太重要了。它关系到计算机的普及和使用效率，当然也就关系到全国各行各业信息化

的程度，各行各业现代化的水平。还有一点常常被一些业内人士忽略的，是中文信息处理对弘扬中华优秀传统文化的重要意义。

数字化，除了物质文化（或者可以把科学技术也暂时算在其中）首当其冲外，精神文化也将接受这一新的现实。中华民族丰富无比的传统文化，无论是文字记载的，还是以石、木、颜料等为介质的，靠动作形体表达的，都越来越靠数码技术记录、储存、传输。在传统文化的各种载体中，文字承载着大部分的信息。特别是中层文化（制度、法律、宗教、艺术等等）和底层文化（观念、意识、哲学等等），主要是靠文字传承。中文信息处理，应该包括对传统文化的处理。尽管现在我们还不能把更多的精力投入到这个领域里来，但是，任务已经明明白白地摆在那里了；中文信息处理技术进步的程度如何，决定着中华文化（包括传统的和现实的）继承、发展和弘扬的速度、范围和质量。甚至我们可以把这个话倒过来说：如果中华文化利用不了中文信息自动化处理技术，就可能在经济全球化的过程中衰落，被异质文化所淹没，而文化的萎缩，将是民族最大的灾难。

中文信息处理技术既然如此重要，就应该引起社会的更多的关心，国家应该逐渐加大这方面的投入。希望专家们一方面用自己的成果向社会展示这门技术对于国计民生、国家前途的重要意义，另一方面继续大声呼吁，争取国家更为有力的支持。

我还希望以这本论文集的问世为起点，中文信息处理界的朋友们能加强已有的团结和合作。未来的路有多长，没有人能说得出来，我们只知道科学技术是没有止境的，那么，中文信息处理技术的发展也将是没有止境的。我作为这一领域的蹩脚的“票友”、众多专家的朋友，想借此书一角表达这样的心情：祝各位事业一路蓬勃，祝中文信息处理早日出现突破性的进展。

2003 年 8 月 8 日于黑龙江

前　言

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。

众所周知，中国的汉文化是以在世界上独一无二的汉文字为基础的。而由于历史的原因，信息处理工具计算机的设计是以西方语言为基础，文字的编码与输入方式是汉语首先面临的难题。我国政府与科学家从上个世纪70年代就开始关注这个问题。继80年代我国解决汉字编码并颁布国家标准、北京大学王选教授解决出版系统中汉字字形的高倍率的压缩及复原后，各种输入方法(音码、形码及汉字识别)在90年代也取得了重要的研究成果，这些方面的研究成果和产业化为汉文化信息化奠定了基础，中国人至少可以使用计算机较方便地处理汉语信息。

然而，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，这类信息的利用率还不到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。在这近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，组织一次会议，讨论上述问题，以解决目前存在的问题。会议筹备组拟定了这次会议集中讨论的六个问题：

(1) 自然语言处理发展历程的经验与教训；

(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。

经过专家组与会议筹备组的努力，在2002年9月21～22日，本项目在中国科学院自动化研究所召开“自然语言处理若干重要问题研讨会”。国内大多数长期从事自然语言处

理的专家，代表他们领导的研究小组参加了这次会议。在这次会议上根据上述六个问题结合国内汉语处理研究中存在的问题，与会专家进行了认真、客观、无偏见的研讨，成功地举行了这次研讨会。本书就是以这次会议各位专家的发言记录为基础，同时另邀部分专家就上述六个问题发表意见后的汇总。

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。这些流派在汉语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。

“大工程实施”已被国际上证明是解决自然语言处理问题的有效方法。1992 年，在 DARPA 和 NSF 资助下，由美国宾夕法尼亚大学组织，搜集和发行语言资源用于语言信息处理领域的研究和开发，有 100 多所大学、公司和政府部门加盟，有各种语言资源 220 种，涉及英、德、法、西、汉等多种语言，称为语言数据联盟(Linguistic Data Consortium，简称 LDC)。这与我国的研究形成了鲜明的对照，我们目前手工作坊式的研究，说明我们的汉语处理研究至少在工作方式上落后 10 年。加快建立汉语数据联盟刻不容缓(徐波，孙茂松)。

制定规范与标准是与会者最为关心的问题之一，实现汉语数据资源共享，避免低水平重复是主要目的，同时，与国际接轨，以避免出现汉语数据资源(包括测试标准)的标准由国外制定的局面(俞士汶)。这需要解决大量组织问题与基础研究问题。现已存在的各种独立建立的汉语数据资源不能浪费，需要充分利用，这就需要在摒弃局部利益的条件下，尽快制定联盟规范，以便这些资源尽快纳入可以共享的统一汉语数据资源库之中。制定规范应该遵守三个原则：其一，在尽量保持一致性的条件下，充分利用已有资源。其二，汉语数据资源是国家民族的财富，规范必须可为社会共享。其三，规范不能成为发展自然语言处理新理论与新方法的障碍，必须为这些新理论与新方法的研究与开发保留充分大的空间(靳光瑾，宋柔)。

另外，测试与维护也是大工程实施中必须解决的一个重要的基础问题。传统的考虑，一般是针对系统，尽管这种测试思想还是重要的，并且测试的设计往往起着推动研究进展的作用(徐波)，但是，由于不同研究者使用不同的语言数据资源，而且由于自然语言的复杂性，这种测试的设计将非常困难(杨莹)，另外，这也不利于新的理论与新的方法的出现，也不利于解决局部关键问题的研究(董振东)。在汉语数据规范的基础上，建立推动汉语处理的测试体系，也是一个重要的基础研究任务，是实施大工程设想不可分割的一部分。

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(王珏)，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，人们将不会有耐心阅读它。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。

统计方法的优点是可以使语言现象数量化，这非常适合计算，也是这种方法可以大幅度提高系统质量的原因之一。目前的研究可以大致分为两类：其一，基于简单相关统计的方法，也可以称为语言资源性分析，其二，在统计意义下的建模方法，这涉及机器学习的算法研究。

统计方法最大的缺点是海量数据可能会淹没语言自身存在的某些性质，而这对深层次的自然语言处理来说可能是致命的(黄曾阳，陆汝占)。例如，对系统而言，其每个局部的正确率达到96%(事实上，这还不是统计方法可以普遍达到的指标)，系统的正确率也就只有70%，因此，使得统计方法在语言学意义下可解释，是揭示语言现象本质的重要途径，统计数据的语言学意义是需要重视的研究方向。

机器学习的方法最直接的应用是将语言数据的分析交给计算机自动完成，从而为某些自然语言处理的局部问题建立模型，以减轻人的枯燥的劳动。除此之外，还可以作为揭示统计数据的语言学含义的工具，一方面，可以形成规则，另一方面可以发现在不同规则粒度下的例外集合，这就是知识挖掘。

目前，自然语言处理逐步以统计方法占主导地位，语言规则方法则从与统计方法平分秋色到退居其次(黄泰翼)，在科学研究中，这是十分正常的。由此得出“统计方法是正确的，而其他方法是错误”的结论却是不科学的，而恰恰是机器学习这样一个大的框架正把统计方法和规则方法相融合，无疑这是一个非常重要的趋势。

目前，国内学者在国际上发表了大量的关于汉语处理的论文，但是，大多数是汉语数据资源的报道，因此，国际大公司与学术界非常了解国内汉语数据资源的开发情况。相对这类报道，语言学与计算方法的报道就少很多，这是不正常的。而国外，对这类数据资源的报道越来越少，而理论方法的研究结果却越来越多，这意味着，国外将在汉语处理中占有有利地位(李航)。改变这种状况刻不容缓。

最后，我们需要指出，通用机器翻译作为系统还存在着大量的基础问题需要解决，试图将这样的通用系统直接推到市场，似乎还欠成熟，但是，社会对汉语处理的需求不仅仅是通用机器翻译系统，更大多数的需求目前的技术就能够解决，例如，某些特定领域（旅游咨询、股票咨询）的汉语处理系统、文本分类、情报分析，以及有害信息的过滤与追踪等。将获得的阶段成果解决社会迫切需要解决的实际问题，开拓特别领域应用的市场，是汉语处理研究获得进一步支持的必要条件。

汉语处理的研究应该像其他学科一样，分为科学研究与技术研究，科学研究的目标是探索语言学的内在规律、计算方法与建设实际应用必不可少的基础设施；而技术研究则需要应用驱动，根据社会的实际需要，设计并研制经得起社会与市场考验的系统。科学研究与技术研究需要不同的检验标准。“上不顶天，下不着地”的研究必然导致低水平重复（张钹）。调查在目前技术水平条件下汉语处理可能应用的实际问题，列出清单，并逐一落实支持是需要尽快提到日程上来的问题。

目前，汉语数据联盟正在顺利形成，其主要目标是为汉语处理提供数据基础设施，以便政府部门、产业与包括语言学和计算机科学等研究单位可以方便地使用，但是，口语与语言研究密不可分，而口语又涉及语音特有的数据集合，因此，本项目专家组建议汉语数据联盟应该考虑将基于汉语的口语数据纳入这个联盟，进而以汉语文化信息化为目标，建立包括文字、语音与语言等与汉语信息处理有关的真正的汉语数据联盟。

最后，专家组对姚天顺教授在组织这次会议所做出的重要贡献表示感谢，对徐波、孙茂松与靳光瑾三位教授在建立汉语数据联盟与编辑这本书中付出的辛劳，以及对所有参加这次会议及投稿专家的真诚努力表示感谢。

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
2003 年 6 月 30 日

目　录

第五篇　应用和技术

第一篇　理论和方法

对统计语言模型的若干认识

孙茂松

(清华大学计算机系智能技术与系统国家重点实验室,北京 100084

Email: sms@s1000e.cs.tsinghua.edu.cn)

摘 要 本文首先讨论了自然语言处理研究中理性主义方法与经验主义方法的区别,指出在现实条件下理性主义与经验主义应达到的两个境界(并行不悖及兼收并蓄),之后结合"973"课题研究中的若干实际案例,从多个角度对统计语言模型进行了观察:适用的统计模型与统计模型的适用范围、统计量的繁与简、统计对象的升华和多统计量的组合。作者相信,随着对内容计算的呼声日益高涨,深层次的语义问题以及主要基于理性主义的大规模语义资源建设和语义计算研究将成为今后几年自然语言处理领域的研究前沿与重点,而统计语言模型在这个进程中会一如既往地扮演重要角色。

1 理性主义与经验主义

一说到统计语言模型,大家自然而然会想到理性主义与经验主义之间的区别。应该说,在方法论层面上,这两者中间确实存在一个截然的分水岭:理性主义的基本出发点是追求完美,企图以思辨去百分之百地解决问题;经验主义则追求一种残缺之美,在承认只能90%地解决问题的前提下,以统计手段来追求利益的最大化。

理性主义所追求的目标是不断的抽象。推至极致,势必会提出这样的问题:在自然语言系统中,存在不存在类似元素周期表的东西,存在不存在类似爱因斯坦 $E=MC^2$ 那样简洁的公式之类的东西?从语言认知或者纯粹的语言学理论研究的角度来看,这种可能性是有的(所谓"大道至简"),美国著名的语言学家诺姆·乔姆斯基(N. Chomsky)针对普遍语法(Universal Grammar)研究的演变历程——从20世纪60年代的"标准理论"(Standard Theory)、70年代的"扩展标准理论"(Extended Standard Theory)、80年代的"管辖与约束理论"(Government and Binding Theory)以至90年代的"最简方案"(Minimalist Program),便是逐步逼近语言"元素周期表"的生动写照[1,2](据《青年参考》2003年7月2日"前沿科技"版报道,来自汉堡-埃彭多夫大学医院的学者日前在《自然神经科学》杂志发表的最新研究成果初步验证了Chomsky的理论,即人类大脑中先天存在一种跨越不同语言的语法通则,语言获得的过程实际上是"普遍语法"向特定语言的语法转化的过程)。

然而,从语言计算及语言工程的角度来看,我认为并不存在这个周期表,这是由Chomsky所指出的语言能力(Linguistic Competence)与语言运用(Linguistic

Performance)的区别所决定的。语言体现了人类所独具的智能性,普遍语法研究的是与语言能力相对应的所谓的内化语言(I-Language),语言计算及语言工程则研究与语言运用相对应的所谓的外化语言(E-Language)。外化语言既有比较规律的一面,又有数不胜数的不规律的一面,变化无穷,个性飞扬,无法想像可以以一种简约的模型加以概括。只有在内化语言所揭示的一般原则的指导下,面对极其纷繁复杂的外化语言现象做艰苦、细致甚至繁琐的工作,才可能实现对语言的有效计算,而其他的终南捷径是断乎没有的。

在应用层面,迄今形成的基本态势是:面对真实环境下波诡云谲的外化语言,基于理性主义的方法往往顾此失彼、捉襟见肘,不禁使人产生"银样蜡枪头——中看不中用"之感;基于经验主义的方法(典型如基于字或词的 N-Gram 模型)虽然也是仓促上阵,但场面上大致还算应付得来,在某些应用中取得了长足的进步及有限的成功。换言之,一味地追求完美,却远达不到理想境界,退而求其次,追求残缺之美,效果却好得多。这个事实再一次提醒我们,经验主义也是一种客观规律,而非炼丹术,"上帝也会掷骰子"。

也许有人要问:你似乎认定经验主义的方法要优于理性主义的方法了?答曰:非也。两种方法各有所长,亦各有所短,不能一概而论。一般地,当研究任务企图覆盖外化语言的全部且分析深度比较浅时(典型如汉字识别的后处理,音字转换,语音识别的后处理,文语转换的前处理,文本检索、分类和过滤等),应主要诉诸经验主义的方法;而当研究任务仅关涉外化语言的一个侧面且分析深度需要比较深时(如智能型问答系统),理性主义的方法就成为适当的选择。

2 关于理性主义与经验主义的两个境界:并行不悖及兼收并蓄

第一个境界(即两者"并行不悖")应该不难达到。无论是理性主义的方法还是经验主义的方法,即使以孤立的眼光对待它们,也可比作挖隧道一样,从两个不同的角度向一个共同的目标掘进,谁也不可能包打天下。两种研究都应该充分得到鼓励,互相尊重,而不能以此轻彼、以此废彼。在自然语言处理的不同发展阶段,一个时期这种方法意气风发,而另一个时期那种方法又大行其道,这些都属于正常的现象,体现了科学研究的螺旋式上升。虽然近 10 年来主要基于词法单元(字、词等)的统计语言模型总体上占了上风,为自然语言处理研究与应用做出了显著贡献,但现如今其发展显然已接近饱和,正处于快速成长之后一个相对平稳的时期,面临新的瓶颈。而随着对内容计算的呼声日益高涨,深层次的语义问题逐渐被推到"风口浪尖"上(典型工作如 WordNet、Cyc、FrameNet、HowNet 及 HNC)。我认为,主要基于理性主义的大规模语义资源建设与语义计算研究将会成为今后几年自然语言处理领域的前沿与重点。那么,经验主义的方法是否就要淡出了呢?绝对不会的。我也相信,以人的抽象思辨构造的大规模语义资源与海量文本及其处理技术的进一步联手,将会为统计语言模型注入新的思想激情,促使其下一轮高潮的到来。

第二个境界(即两者"兼收并蓄")则又上升到方法论的高度。众所周知,历史上学术界曾多次在方法论上发生过重大争论。例如,20 世纪 80 年代人工智能界曾对符号主义和连接主义的地位问题展开过热烈讨论。讨论的焦点是基于神经元网络的连接主义方法会不会最终取代传统的符号主义方法。Minsky 在 1990 年的一篇著名的论文中对此做出了否

定性的回答。他认为人工智能不同于电路理论和电磁学。在这里不可能有像电路理论中的克希荷夫定律,或电磁学中的麦克斯韦方程那样神奇的统一理论(我在第1节中给出的断言"自然语言系统中,不存在类似元素周期表的东西",可谓与Minsky的这个观点一脉相承)。他主张人工智能必须利用各式各样的方法,包括各种不同的知识表示。他相信采用多元化的构件来建造复杂的人工智能系统的时候已经来到,这些构件中有的是连接主义的,有的是符号主义的,每个构件都有它自身存在的理由[3]。显然,Minsky"采用多元化的构件来建造复杂的人工智能系统"的思想同样适用于自然语言处理的研究。一个比较合理的格局是在研究与应用两个层面上打通理性主义(往往在系统中反映为规则形式)与经验主义(往往在系统中反映为统计形式)之间的"壁垒",实现两者的有机融合,形成一种"你中有我、我中有你"的状态。其实,规则为主的形式上可以附加统计信息(典型如概率型上下文无关文法),统计为主的形式也必须嵌进某种理性成分才可能更具价值(典型如词性自动标注,通常需要人工加工的语料库的支持,"人工"二字就意味着一定程度的"理性"的介入)。此外,还可以衍生出许多变化,如知识表示可以是基于规则的,但规则的获取算法是基于统计的(典型如E. Brill针对词性自动标注提出的基于错误驱动的转换策略),等等。这些已然是活生生的实践,只不过力度与规模尚远远不够罢了。

3 对统计语言模型的几点观察

本节结合本人指导的"973"课题研究中的若干具体案例,谈谈对统计语言模型的几点观察。

3.1 适用的统计模型与统计模型的适用范围

案例1:汉语自动分词中组合切分歧义的消解

首先,应根据待求解问题的特点,设计适用的统计语言模型。

"中将"是组合切分歧义的一个典型例子:

(例1) a. 1955年他被授予中将军衔。 (合)

b. 信息在社会发展中将起到关键作用。 (分)

以往的解决办法有两种:一种是由人去写一些规则,然而,即使是专家,也难以对"中将"的上下文作全面的把握,规则集的有效性、一致性均不易保证;另一种是利用基于词性N-Gram的隐Markov模型,以动态规划求一条分词及词性标注的最佳路径来排除歧义,但显然某些情形下"中将"的歧义消解(尤其是"中将"取合的情形)要取决于更大的上下文,而并非像词性标注那样仅仅严重依赖于邻接的BiGram或者TriGram。有鉴于此,我们将组合切分歧义的消解视作与词义排歧(Word sense disambiguation)同类的问题,并且很自然地将词义排歧常用的策略——向量空间模型(Vector space model)引入分析中。我们开辟了一个左、右上下文各为三个词的窗口,通过考察训练集中落进这个窗口的词集的相关统计信息做出排歧决策(图1):

图1 组合切分歧义消解中的上下文窗口