



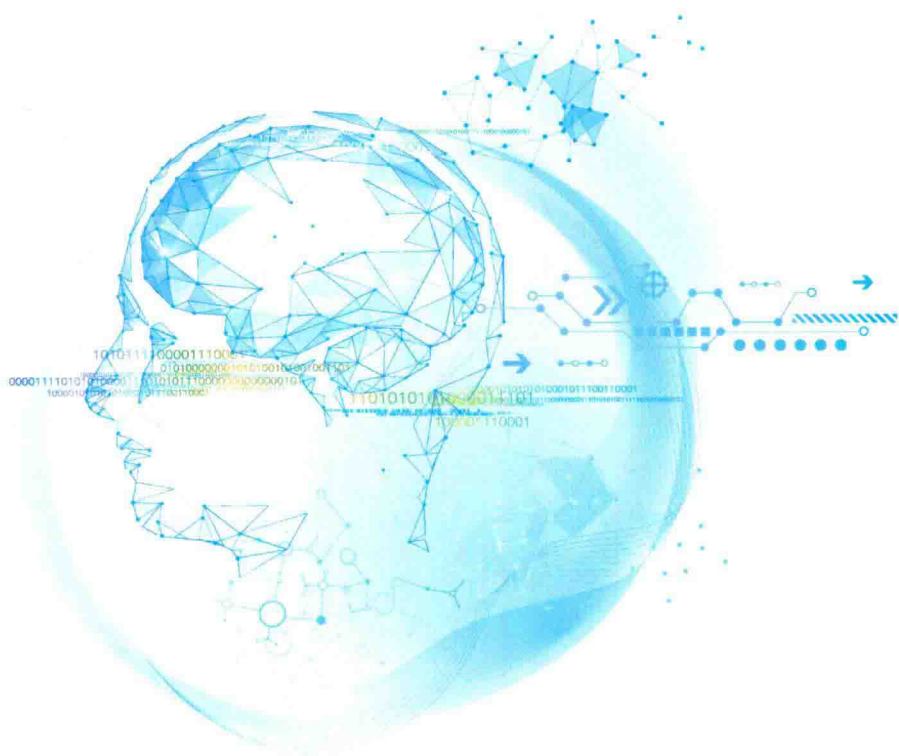
中国语言文学

一流学科建设文库



中文信息处理发展简史

刘云 肖辛格 著



科学出版社



中国语言文学

一流学科建设文库

中文信息处理发展简史

刘 云 肖辛格 著

科学出版社

北 京

内 容 简 介

本书以中文信息处理的学科发展脉络为线索,介绍了中文信息处理在不同历史时期中的发展情况与主要成果。本书首先简要介绍了中文信息处理学科的基本概况,然后以起步期、发展期、成熟期、繁荣期等不同的历史阶段,分章节对中文信息处理各个历史阶段的发展过程和成果进行了具体阐述。本书以时间为轴,将中文信息处理研究的方方面面介绍给读者,同时也将中文信息处理这一学科从无到有、从小到大、从弱到强的发展过程呈现出来。

本书适合高校语言学专业的教师、学生阅读,也适合中文信息处理相关工作人员参考。

图书在版编目(CIP)数据

中文信息处理发展简史 / 刘云, 肖辛格著. —北京: 科学出版社, 2019.12
ISBN 978-7-03-063843-4

I. ①中… II. ①刘… ②肖… III. ①汉字信息处理—技术史—中国 IV. ①TP391.12-092

中国版本图书馆 CIP 数据核字 (2019) 第 288887 号

责任编辑: 张 宁 赵 洁 / 责任校对: 贾伟娟

责任印制: 李 彤 / 封面设计: 蓝正设计

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>

北京虎彩文化传播有限公司印刷

科学出版社发行 各地新华书店经销

*

2019 年 12 月第 一 版 开本: 720 × 1000 B5

2019 年 12 月第一次印刷 印张: 13 3/4

字数: 280 000

定价: 98.00 元

(如有印装质量问题, 我社负责调换)

序

——计算机正改变着我们的语言生活

刘云把他与肖辛格合著的《中文信息处理发展简史》(以下简称《简史》)电子版发给我,希望我为《简史》写序。中文信息处理的发展史只有几十年,我还算了解,但当读起这部《简史》时,我仍然被它深深吸引了,觉得内容有趣,也有许多收益,特别是它引发了我的许多思考。

最主要的收益和思考,有如下三方面。

一、中文信息处理在解决一个个中文问题的过程中前进

科学技术都是由问题驱动而前的。中文信息处理是在处理汉字、词语和句子的一系列问题的过程中步步向前的。

(1) 汉字处理。中文信息处理一开始是处理书面语,遇到的第一个大问题便是汉字。在解决汉字信息处理问题时,它取得了诸多成就:大规模统计了汉字使用频率,总结出了“汉字效用递减率(周有光)”,制定了第一个汉字编码国家标准 GB 2312—80。汉字键盘输入的编码方案“万马(码)奔腾”,汉字自动识别技术、汉字存储技术及输出技术迅速发展。特别是王选教授研发的汉字激光照排系统,使汉字印刷告别了铅与火的时代。在汉字处理期,信息检索也有进步,主题词表、文献库、全文检索、自动标引、自动文摘等方面都有成绩可述。

(2) 词语处理。汉字进入计算机的难题逐步被攻克,词语处理成为亟待解决的问题,其基本任务是自动分词和词性标注。汉语书面语不像英语、法语、俄语、德语等语言以词为单位排列,而是以字为单位排列。信息处理是以词语为识别单位的,故而需要制定信息处理的分词规则,建立分词词表,妥善处理表外的“未登录词”(基本上是专有名词和专业术语)。汉语形态不发达,词类兼类现象较为普遍,需要制定适合信息处理的词类系统,建立词类标注规范,确立词性标记集。为支持词语处理,信息处理学界还建立了“语言工程”理念,建设了一些知识库、语料库,进行了大规模的词频统计。在词语处理期,汉语语音处理技术(特别是语音识别)、机器翻译、计算机辅助语言教学与测试等也都有进步。中文信息处理技术整体迈入实用化、商业化阶段。

(3) 句处理。词语处理进展到一定阶段,句处理的任务就紧迫起来了。句处

理需要进行句法分析和语义分析,还有更多的语用问题需要解决。句处理其实就是真实的自然语言处理,对于语言资源的需求也更为急切。为支持句处理工作,一批高质量、大容量的大型通用汉语语料库建立起来,如国家语言文字工作委员会(以下简称“国家语委”)、清华大学、北京大学、北京语言大学等都建立了自己的大型语料库。各类知识库也先后被开发出来,如汉语虚词知识库、汉语事件知识库等。在句处理期,机器翻译研究呈现新面貌,在语料库语言学的指导下,基于实例的翻译方法和统计翻译方法得到广泛运用,网上机器翻译系统发展迅速,口语机器翻译研究取得突破。互联网的进步和网络技术的发展,大大地促进了信息检索技术的发展,信息检索、文本分类、信息抽取、文本摘要等的研究价值也充分显现出来。计算机学界与语言学界合作紧密,中文信息处理呈现出蓬勃发展的态势。

(4) 数据驱动。2010年前后,中文信息处理扫除诸多中文处理的特殊困难,开始与国际自然语言处理更加同步。此时,“深度学习”“知识图谱”“大数据”等概念流行起来,中文信息处理走上了“数据驱动”之路。词向量技术带动着语义分析领域的进展,机器翻译、信息检索、人机对话、机器写作、虚拟主持人、社会计算等领域都有突破性进展,情感分析、机器语言行为等也受到关注。当前,社会对于包括中文信息处理在内的人工智能特别关注,国家制定了人工智能发展规划,一批人工智能、语言智能的研究院、学院相继建立起来。语言信息处理进入一个新的繁荣期。

二、中文信息处理带我们进入一个语言生活的新时代

中文信息处理开展的时间不长,机器的中文智能还比较低,但的确是“人工智能一小步,人类社会一大步”。回首过往,大家会发现我们的语言生活已经发生了巨大变化。

(1) 键盘时代。过去汉字要用笔书写,现在主要用键盘敲击。键盘时代使我们常常“提笔忘字”,很多人已是手拙字丑,只认识印刷体不认识手写体。键盘输入使得笔画、笔顺已经没有写字时代那么重要,那么,小学识字教学,是否还要特别强调“四、国”的第二笔有钩无钩?是否还要强调“万”的第二笔、“方”的第三笔是写一撇,还是写横折钩?从历史上看,汉字形体的发展变化都是由书写推进的,如篆书演变为隶书的“隶变”,隶书演变为楷书的“楷化”。那么,以后汉字的形体还会发生变化吗?键盘时代,带给人很多写字的忧虑,带给人很多关于汉字前景的思考与遐想。

(2) 融媒体时代。语言总是需要媒介、需要载体的。文字和印刷术的出现,使语言在声波的基础上又获得了光波这一新媒介,产生了书面语,有了图书、报

刊等平面媒体。广播、电视的出现,使语言又获得了电波媒介,人类有了有声媒体。互联网的发展为人类构筑了一个虚拟空间,产生了网络媒体。平面媒体、有声媒体、网络媒体的迭代融合,便形成了融合媒体。中文信息处理使我们在网络时代可以进入融媒体时代。

网络媒体和融媒体,正在改变着我们的阅读习惯。当前,碎片化阅读、行走中阅读、读标题、读图等所谓的“浅阅读”成为常态,专心阅读一篇长文已是“奢侈”之举、罕见之事。信息检索主要靠网络,靠知识库,纸媒图书已经退居知识的“桥栏杆”的地位。许多治学者也主要靠数据库查阅文献,平时的文献阅读、素材积累的功夫正在退化。图书、杂志等信息提供者已经不满足于向用户被动地提供书目单或内容提要,而探索着信息的二次加工和信息的精准投送。融媒体时代,知识的生产、储载、应用的模式都在悄悄地发生着变化。

(3) 虚拟语言生活。互联网,特别是移动互联网,为人类架构了一个“虚拟空间”。而正是因为有了中文信息处理的成果,国人方可在虚拟空间中过上虚拟语言生活。过去的语言交际方式是“人-人”交际,现在多数是“人-机”交际和“人-机-人”交际,在“人-机”交际和“人-机-人”交际的背后,其实还存在着“机-机”交际。在网络里,人们可以用虚拟身份进行交际,可以利用微信等聊天工具同时在多个群里出现,同时扮演不同的交际角色。人类的语言交际方式发生了重大变化。

人类在现实空间从事的各种活动,都被尝试迁移到虚拟空间中。在这种“空间大挪移”的“魔术”中,有些活动比在现实空间做得更好,有些则无法迁移,有些则可以在两个空间中合作进行。凡在虚拟空间做得更好的活动,就可能不在现实空间中进行了,比如信件逐渐被电子邮件代替,邮递员这个职业逐渐消失。虚拟语言生活改变着人类的活动方式,增加了许多新职业,也“取缔”了不少传统职业。

虚拟语言生活十分活跃,新词语及语言的新用法常在网络上孕育,再传播到现实语言生活中。就语言发展、语言风格等方面来看,现在是虚拟语言生活在引领现实语言生活。网络媒体成了新词语、新现象产生的温床,网民是语言发展十分重要的力量。我们对待网络语言的态度要与时俱进。

(4) 人与机器人的语言合作。随着计算机语言智能的提升,计算机可以做的语言工作越来越多,介入人类的语言生活也越来越广泛、越来越深入。语音输入、口语与书面语的自动转换、机器翻译、文献处理、人机对话、机器写作等,带来了人与机器在语言生活中的深度合作。特别是将来具有语言智能的机器人的出现,将使人类与机器人共处共事,共同进行语言生活。这不是童话故事,而是正在实现的未来。目前,网络上机器人的写作已经占到全部写作的15%左右。

在人与机器人的语言合作中,人类不仅驯化着机器,也在驯化机器中改变着

自己,改变着自己的书写习惯、阅读习惯、语言交际习惯甚至是语言思维习惯。这些习惯正在形成一种新文化,这种新文化需要通过教育传授给社会和未来人,正如文字产生之后要进行扫盲和学校语文教育一样。更为重要的是,语言智能会为社会各领域带来发展红利,也会促进社会劳动力的大转移,使其从被取缔的行业转移到新兴行业。就个人而言,人们需注意语言智能的前沿发展,为自己插上语言智能的翅膀,而不是被机器碾压。就国家来说,语言智能的发展已经严重影响到国家人力资源的分配与开发,国家需要及时调整学校的学科设置和课程设置,培养适合语言智能时代的人才,同时也要及时发出产业预警,并通过社会培训有计划地进行劳动力转移,防止出现大面积失业现象。

语言智能的发展在支撑着一个信息化时代,同时我们也须思考,语言智能(包括智能机器人)是否会用在危及人类的地方,机器人的智力“无限增长”是否会变得不可控制,而最终把人作为它的奴隶。这也许是杞人忧天,但也确实应为语言智能装上“牛缰绳”“马笼头”。这缰绳和笼头,包括从业者的自律、科学道德和法律的保障。2018年,北京语言大学开始设立“语言智能与社会发展高层论坛”,该论坛设立的初衷既是要促进语言智能的快速发展,保证社会获取人工智能的最大红利,也是要应对语言智能发展可能带来的伦理学、法学问题。

三、语言学与中文信息处理

(1) 数据与规则“双轮驱动”。数据驱动着中文信息处理驶入快车道,使其在许多领域已经走在世界自然语言处理的前列。数据驱动的能量还极其强大,要不失时机地释放数据的能量,推进语言智能的快速发展,推进语言智能的社会应用。但是,数据驱动缺陷也是比较明显的,比如数据所包含的知识并不全面,可能具有“偏颇性”;数据学习表现的是统计频率,而不是因果关系,不是客观规律;机器从数据中学到了什么,人们并不十分清楚,深度学习是一只“灰箱”。有学者预测,数据驱动的发展也会遇到“天花板”。而规则是人对事物的认识、对规律的把握,反映的是因果关系。语言智能的下一个学术发展周期,应当是数据与规则的“双轮驱动”。这规则可能来自多个学科,其中一个基础学科就是语言学,是语言学得到的关于语言及其运用的规则。

(2) 语言学的“规则之轮”。语言学与中文信息处理有过一段超长的“蜜月期”,只是到了语料库语言学时期,统计方法可以有效解决一些问题时,语言学的“规则”效力始被质疑。在深度学习的理论与方法流行之后,语言学的规则仿佛成了“无用之物”。语言学之“无用”源自三个方面。其一,数据效力遮蔽了语言学效力,其实语言智能关于语言的属性与概念、语言各层级各单位之间的关系、语言

与人类社会的关系的认识等，还都来自语言学的基本知识体系。其二，语言学的知识表述没有形式化，是供人看的知识，而不适合于机器阅读。形式化表述成了语言学知识到达语言智能车间的“最后一公里”。其三，语言学是以语言结构为学术基点的，其重点在语言结构的研究上，而中文信息处理在基本解决了字、词语的问题进入句处理阶段后，就开始了对真实话语的处理，但语言学对话语研究用力不够，积蓄不多。为了打造语言智能的语言学“规则之轮”，语言学必须实现“话语转向”，把学术基点转至“话语”。话语是语言的真实存在状态，本应成为语言学研究的重要对象。

在中文信息处理的发展过程中，语言学起到了重要的支撑作用，包括人才支撑和语言学知识体系的支撑。同时，中文信息处理也得到了一些新的数据，比如字频和词频；提出或强调了一些研究课题，比如词的识别和词性的识别、词语兼类、专有名词及其简称、数量结构、代词的指代关系、词语和句子的歧义问题等；建设了一批语言工程，如各种语料库、知识库等，这些资源支持着语言研究的现代化；问世了一批语言信息化产品，如电子词典、自动翻译机等，帮助语言学开展社会语言服务。这些新数据、研究课题、语言工程、语言信息化产品也在启发着语言学，装备着语言学，提升着语言学的研究能力，推进着语言学的现代化。

（3）国家语言规划。中文信息处理取得如此显著的成就，与国家的语言规划具有密切关系。1986年1月召开的全国语言文字工作会议就提出：汉语汉字的信息处理是一门新兴的边缘科学，有广阔的前景，加强这方面的研究，对经济、文化、科学技术的发展具有长远的意义。因此，当前语言文字工作的任务必须包括这项内容。之后，国家又进一步提出了语言文字的“三化”（规范化、标准化、信息化）和汉字整理工作的“四定”（定量、定形、定音、定序），并发布了信息处理用的多个规范标准，甚至还成立了“语言文字信息管理司”专司语言文字信息工作。1986年之后的国家语委一直把语言信息化作为重要工作，除了制定语言信息化的工作目标之外，所有工作都会考虑信息化的背景和对信息化的支持，许多语言文字规范，包括《通用规范汉字表》，都充分考虑到语言信息化的问题。如果没有国家语言政策的支持，我国的中文信息处理事业，乃至整个国家的信息化事业都不可能有今天的局面。在语言智能时代，需要什么样的国家语言规划，这是我们应该继续考虑的。多少有点遗憾的是，人们常常忽略国家语言政策、语言规划对信息处理发展做出的重要贡献。

（4）语言学与信息科学的“双栖教育”。学科交叉不仅是科学发展的要求，也是人才培养的要求。培养语言学与自然语言信息处理的交叉人才，是语言智能发展能够获得双轮驱动的必要保证，也是促进语言学能够充分利用现代信息技术、实现话语转向、乘借语言智能而发展自己的必要保证。在一些重要的文献中，学者已经理智地认识到人工智能的发展，要将其与神经科学、认知科学、量子科学、

心理学、数学、经济学、社会学等相关基础学科交叉融合，但是往往忽视了语言学，这是“语言意识薄弱”的一种表现，是不利于人工智能事业发展的。

关于交叉人才培养的重要性，本书的作者之一刘云教授就是一个典型例证。他是中文系出身，本科、研究生阶段受到的是人文科学训练，有着坚实的语言学基础。他在攻读博士学位期间，去北京大学计算语言学研究所进修。博士毕业后，再去俞士汶教授门下从事计算语言学的博士后研究。回到华中师范大学工作以后，刘云仍然做双栖型研究，发表了近 20 篇与中文信息处理相关的论文，出版了《汉语虚词知识库的建设》（华中师范大学出版社，2009 年），承担了多个与中文信息处理相关的国家社会科学基金、教育部人文社会科学基金等项目，他还是教育部教育信息技术工程研究中心兼职研究人员、国家数字化学习工程技术研究中心兼职研究人员，2011 年入选教育部“新世纪优秀人才支持计划”。刘云受到了一流的语言学培养和一流的计算语言学培养，“双一流”使他成为双栖型人才，并为语言学和中文信息处理做出了双重贡献。这部《中文信息处理发展简史》也是双栖型人才才能够写出的。目前，国家正在提倡“新文科”理念，把新技术融入人文社会科学中。其实，中国的语言学早就在这样做，是新文科的样板。

这个序写得有点长。这是因为这部书可读，脉络清晰，事实丰富，有回顾有展望。我们读后能够了解中文信息处理的情况，理清中文信息处理发展的来龙去脉，且有所启发。我曾经担任教育部语言文字信息管理司司长 12 年，近来又担任中国中文信息学会副理事长，比较了解教育部、国家语委的中文信息发展战略，比较了解中文信息处理业内发展的行情。书中提及的一位位人物，有许多都是我的师友；书中记述的一串串事件，有许多是我亲身经历的；书中的许多文献和数据，也都是我留意的、关心的；书中的人物、事件、文献、数据，都能唤起我的记忆、激起我的情感。

在此，我要感谢刘云和肖辛格用心地写出这么一部书，并请我作序，给了我忆往和抒情的凭借，给了我写读后感的机会。

李宇明

2019 年 8 月 10 日

于北京惧闲聊斋

目 录

序	
第一章 绪论	1
第一节 中文信息处理概述	1
一、什么是中文信息处理	1
二、研究中文信息处理的意义	3
第二节 中文信息处理的内容	5
一、中文信息处理的研究范围	5
二、中文信息处理的研究方法	6
三、常见的应用系统	8
第三节 中文信息处理研究的格局	10
一、中文信息处理与相关学科的关系	10
二、中文信息处理的研究取向	11
三、中文信息处理研究的基本历史分期	13
参考文献	14
第二章 中文信息处理的起步期	16
第一节 中文信息处理问题的诞生背景及萌芽	16
一、自然语言处理的理论准备	16
二、自然语言处理的发端	18
三、中文信息处理的萌芽	19
第二节 中文信息处理问题的兴起	22
一、沉寂后的复苏	22
二、汉字信息处理时代的到来	24
三、汉字精密照排系统的发展	27
第三节 汉字编码工作的开展	28
一、汉字编码的基本任务	28
二、字频统计工作的展开	30
三、万“码”奔腾时代的到来	32
四、汉字交换码标准的编制及发展	34
五、汉字编码的国际标准	37

第四节 汉字的输入及输出技术	38
一、键盘输入技术	38
二、汉字自动识别输入	44
三、汉语语音识别输入	47
四、汉字的存储及输出技术	50
第五节 起步期的应用研究及理论探索	55
一、从机器翻译到自然语言理解	55
二、自然语言理解与人机对话	59
三、信息检索及相关技术	61
四、学会组织的建立与发展	63
第六节 小结	65
参考文献	66
第三章 中文信息处理的发展期	70
第一节 “词”处理时代的到来	70
一、“词”处理的基本任务	70
二、词频统计与词表编制	72
第二节 语料库的发展与建设	74
一、统计方法的复苏与语料库的发展	74
二、我国的早期语料库建设	77
第三节 自动分词技术的发展	78
一、自动分词的困境	78
二、自动分词方法的探索	80
三、歧义与未登录词	85
四、自动分词系统的研制	89
五、分词规范的编订	91
六、自动分词技术的评测	93
第四节 词性标注技术的发展	95
一、词性标注与词类划分	95
二、词性标注中的兼类词难题	99
三、词性标注技术的发展	101
第五节 语音处理技术的突破	103
一、隐马尔可夫模型	103
二、语音识别技术的新进展	104
三、语音合成技术	105
第六节 应用研究的发展	109

一、“语言工程”概念的提出	109
二、机器翻译的新发展	110
三、计算机辅助语言教学与测试	112
第七节 小结	114
参考文献	115
第四章 中文信息处理的成熟期	119
第一节 成熟期的基本格局	119
一、统计方法的回归	119
二、从“词”处理到“句”处理	123
三、“句”处理的主要困难	125
第二节 句法分析与语义分析	127
一、句法、语义分析技术的发展背景	127
二、句法分析理论的两大体系	130
三、汉语句法分析方法的探索	134
四、语义资源及语义分析理论的发展	138
五、语义消歧与语义标注	142
第三节 基础资源库建设	147
一、语料库迅猛发展	147
二、大规模语言知识库建设	149
第四节 应用型技术及相关研究	158
一、机器翻译	158
二、网络技术下的新动向	161
第五节 小结	165
参考文献	167
第五章 中文信息处理的繁荣期	171
第一节 研究范式的彻底转向	171
一、从人工神经网络到深度学习	171
二、人工智能与大数据	173
三、语音识别技术的突飞猛进	176
第二节 新时期的中文信息处理	178
一、中文信息处理的新特点	178
二、基于深度学习的中文信息处理研究	180
三、新研究领域的拓展	184
四、应用型技术的百花齐放	186
第三节 深度学习语境下的新挑战	189

一、语言学知识有待回归	189
二、进一步拓展与深化新的应用领域	191
三、与认知科学及脑神经科学相结合	194
四、走向多模态信息处理	197
五、深入国际交流与合作	201
第四节 小结	203
参考文献	205
后记	207

第一章 绪 论

第一节 中文信息处理概述

一、什么是中文信息处理

自计算机诞生以来,能用自然语言与计算机进行自如的交流,便是人们一直以来追求的梦想。人们希望能够用自己最习惯、最擅长、最自然的语言来使用计算机,而无须再花费大量的时间和精力去学习各种计算机语言或操作;同时,人们也希望能够通过计算机来进一步了解人类自身的语言能力,探索人类自身的智能机制。为了实现这个梦想,关于如何对自然语言进行理解和处理的研究未曾中断过。在这种探索过程中,自然语言处理这样一门学科逐渐形成。

自然语言处理,又叫计算语言学^①,研究的是如何利用计算机来理解和处理自然语言。自然语言处理的对象是人类语言,而人类的语言现象极其复杂,不同的语言又各具特点,为了应对不同语言的特点,自然语言处理领域就产生了针对具体语言的研究分支。中文信息处理,便是自然语言处理这门学科下的一个分支。

对于“中文信息处理”这个术语,学术界一般有狭义和广义两种理解。狭义的中文信息处理,就是指利用计算机处理汉语信息,对汉语进行分析和理解。此时的“中文”,特指汉语,包括口语和书面语。所以狭义的中文信息处理也可以称为“汉语信息处理”。而广义的中文信息处理,其“中文”则表示“中国的语言文字”,除了汉语以外,它还囊括了中国各少数民族所使用的语言文字。本书所讨论的中文信息处理,在没有特别强调的情况下,一般取狭义的理解。

汉语与其他语言有一定的共同性,因此,中文信息处理与其他自然语言的信息处理有一定的共性。但由于汉语的一些特殊性,中文信息处理也还有很多自己的特点,概括起来主要有以下四个方面。

首先,汉字的特殊性。不同于其他拼音文字,汉字数量多,仅常用的汉字就

^① 通常人们未将“计算语言学”和“自然语言处理”严格区分开来,只是“计算语言学”主要用于说明理论,而“自然语言处理”侧重于工程,主要用于说明方法(刘颖,2002;江铭虎,2006)。不过,广义的“计算语言学”还包括了以获得语言中隐含的数量规律为目的,以计算机作为工具手段对自然语言进行的计量研究,而这部分内容一般属于“计量语言学”的研究范畴(俞士汶,2003)。为方便表述,本书所用的“自然语言处理”一词,在未做说明的情况下,一般等同于“计算语言学”。

有几千个；如果涉及古籍处理，则甚至要用到几万个汉字，因而汉字输入、汉字存储、汉字输出、汉字识别等问题的处理特别困难。

其次，书面汉语的特殊性。在书面汉语中，词与词之间没有明显的标记，自动分词成为书面汉语分析的第一道难关。

再次，汉语语音的特殊性。虽然汉语语音系统相对简单，但由于存在声调和变调问题，因此这对于语音识别和语音合成而言，也是一个难点。

最后，汉语语法的特殊性。汉语缺乏像一些西方语言那样的形态变化，依靠词序和虚词作为主要的语法手段，句法歧义问题复杂，这也使得汉语语句的自动分析比较困难（陈小荷，2002）。

面对汉语所具有的这些特殊性，汉语的信息处理任务就显得尤其艰巨。

用计算机处理汉语信息，大致可以划分为两个层次：一个是文字层次，即汉字信息处理；另一个则是语言层次，即对汉语的语音、词汇、语法等信息进行处理。具体内容包括对它们的输入、存储、传输、输出、识别、转换、压缩、检索、分析、理解和生成等。

中文信息处理是一门与计算机科学、语言学、逻辑学、数学、信息学、声学、认知科学等多个学科相关联的综合性交叉学科。针对不同层面的处理对象，或者不同的处理内容，中文信息处理往往会涉及不同学科、不同领域的知识，而其中与之直接相关联的，就是语言学和计算机科学。中文信息处理在语言学、计算机科学中所处的位置如图 1-1 所示。

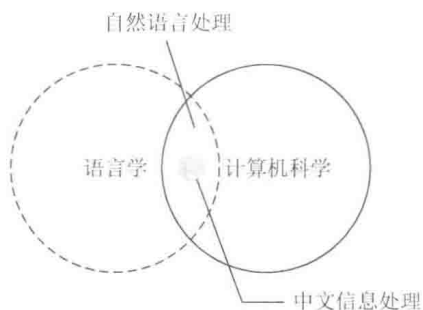


图 1-1 中文信息处理在语言学、计算机科学中所处的位置

作为语言学的一门分支学科，中文信息处理主要研究语言及语言处理与计算相关的方面，而不甚关心其在计算机上的具体实现（刘海涛，2005）。这时它一般被称为“计算语言学”。由于其具有实践性、应用性的特点，在语言学的学科范畴下，计算语言学一般也被归入应用语言学之中，属于应用语言学的子领域。

作为计算机科学的一个分支，中文信息处理一般被称为“自然语言处理”，是

人工智能(Artificial Intelligent, AI)的一个主要内容,是计算机模拟人类智能的一个重要方面(冯志伟,2012)。为了在计算机上实现自然语言能力,这一方向的研究与语言哲学(特别是逻辑学)、认知科学的关系也十分密切。它的主要研究目标是建立语言处理的形式化模型,进而通过计算机来验证这种模型(刘海涛,2005)。

总之,为了让计算机能够处理汉语,人们一方面需要研究如何以机械的、格式化的步骤或程序来处理汉语;另一方面需要研究如何将汉语的规律以精确的、形式化的、可计算的方式呈现出来(俞士汶,2003)。因而可以说,语言学知识和计算机科学知识是中文信息处理的两大支柱,缺一不可。

二、研究中文信息处理的意义

作为一门交叉学科,中文信息处理得到了来自不同领域的学者的重视与关注,这与其所具有的巨大价值密不可分。研究中文信息处理,既有重要的理论意义,又有明显的实际意义。^①

从理论研究的角度看,语言科学是人文科学与自然科学之间的桥梁,而自然语言信息处理正是这两大科学相结合的一个有代表性的交叉学科,自然语言信息处理研究可以带动多种学科和技术的发展,尤其是能为智能科学的突破性进展贡献力量。由于创造和使用自然语言是人类高度智能的表现,因此对自然语言理解的研究也有助于揭开人类智能的奥秘,深化我们对语言能力和思维本质的认识。在计算机上建立自然语言信息处理系统,可以为人类了解自身的语言活动提供一个观察的“窗口”。

从语言研究的角度看,中文信息处理研究可以促进语言研究。首先,中文信息处理研究的开展为汉语研究提供了新的手段和方法。例如,利用各种语料库人们可以自动检索、统计所需的语言现象,这样就可以缩短研究时间、提高研究效率、增加研究信度。其次,中文信息处理研究可以对汉语研究的相关成果加以检验。最后,中文信息处理研究对汉语研究提出了更高的要求:一是面向“人”的语言研究所得到的成果是关于人类自然语言的知识,通常是用自然语言来表述的,而面向“机器”的语言研究,强调研究在形式化的框架下进行,研究所得到的语言知识应以形式化的方式表示;二是面向“机器”的语言研究特别强调语言知识的系统性及对真实语料的覆盖能力,即语言知识要达到一定的规模。这两个要求对当前的汉语研究提出了更高的要求。

从语言发展的角度看,中文信息处理作为一种语言技术,一直是推动语言、

^① 关于中文信息处理的意义,参见邢福义主编:《现代汉语》,北京:高等教育出版社,2015年。

语言应用和语言学研究发展的重要力量。^①语言发展与语言技术的发展以及因语言技术而产生的不同交际方式的发展关系十分密切。此前报纸、广播、电视等大众传媒的产生与发展,已使人类的语言交际方式发生了巨大变化,语言交际变体异常丰富起来,不同方言、不同语言的接触更为频繁,更加深入,语言发展变化的速度加快、幅度加大。而现在,伴随着信息技术的发展,互联网为语言发展提供了更大的驱动力和更优越的实现条件。它将古今中外的语言数据汇集、“发酵”,将全社会的语言智慧网罗集中,使语言发展日新月异。在这样的背景下,网络媒体为语言研究展现了广阔视域——网络新词语现象、网络语言的传播特点、新媒体语言的丰富与规范、新媒体时代的语言舆情监测、网络语言政策研究等。而且,语言研究的手段也在不断进步。各种数据库技术、众包众筹研究模式、线上学术讨论与出版、虚拟研究与远程研究等,已是常见的或可以想见的研究手段。多学科共同讨论语言问题,协同解决语言问题,呈现出“大科学”和“大学科”的研究态势和研究旨趣,甚至在语言观念、语言研究旨趣、语言学派的形成方面,也都产生了广泛而深刻的影响。简言之,语言技术的每一次进步,都能带来语言、语言交际变体及语言功能的发展变化,这也就进一步促进语言学在研究材料、研究领域、研究手段、研究旨趣等诸多方面的进步与变化。此外,语言技术的发展和运用,还催生了一批语言产业,造就了一批新的语言职业,并促使传统语言产业现代化,语言技术、语言产业已成为当今经济发展的重要动力之一,其所得红利在现代社会中占有较大的经济份额。

从社会需求的角度看,信息化社会迫切需要自然语言信息处理研究取得实质性的突破。计算机和互联网的广泛应用昭示着信息时代的到来,信息技术(特别是网络技术)的进步推动了全球化的进程,全球化又反过来对信息技术提出更多新的要求,其中一个主要的要求是希望计算机能够尽量理解人类的自然语言;而要想让计算机领会人类自然语言,就必须开展自然语言信息处理的研究。自然语言信息处理研究必将对我们的生活产生深远影响,使人类社会加快信息化的步伐,对国民经济和社会的发展都会起到推动作用。纵观中文信息处理的研究历程我们不难看出,它极大地推动了中国计算机工业的发展,并先后催生了一批又一批实力强大的企业,如联想、方正、汉王、巨人网络、金山、搜狗等。而如今,互联网时代更是对汉语信息处理提出了新需求。譬如,互联网上的海量数据为信息检索、自动文摘等提出了新的课题;外文信息多到人工翻译根本不能满足需要,机器翻译的重要性被提到了空前高度;手机、掌上电脑等移动设备令信息处理需求变得无处不在,非标准键盘的汉字输入需要有新的解决方案;等等。

正因为中文信息处理具备这些重要意义,中文信息处理不仅是中国学者关注

^① 关于这部分内容,参见李宇明:《语言技术对语言生活及社会发展的影响》,《中国社会科学》2017年第2期。