中文信息处理技术发展简史*

张华平©

(中国科学院计算技术研究所软件实验室 北京 100080)

E-mail: zhanghp@software. ict.ac.cn

Homepage: www.nlp.org.cn; http://pipy_world.y365.com

摘要：真正意义上的中文信息处理迄今已经有 20 余年的历史了，随着计算机的普及和 Internet 的蓬勃发展，中文信息处理技术实实在在的改变了人们的生活。本文根据目前所能收集的文献资料，集中整理了中文信息处理技术发展的简史，并从自身专业的角度，针对重大的研究工作做了粗浅的评述，旨在提供一个中文信息处理技术发展的脉络，达到“以史为鉴”或者“温故而知新”的效果。

关 键 词： 中文信息处理技术，简史

1、引言

在我国，中文信息处理已经不是什么新鲜事物了，随着科学技术的发展，中文信息处理技术已经深入到了社会生活的各个方面。所谓“中文信息处理”，指的是用计算机对汉语（包括口语和书面语）进行转换、传输、存贮、分析等加工的科学。它是一门与语言学、计算机科学、心理学、数学、控制论、信息论、声学、自动化技术等多种学科相联系的边缘交叉性学科，是自然语言信息处理的一个分支，需要以大量的语言知识、背景知识为依据，对中文信息的人脑处理过程进行模拟。其中，“中文”是指中国通用的所有语言种类，包括汉语及其他少数民族的语言；但一般都是指汉语。“信息”是指能通过视觉、听觉、嗅觉、味觉、触觉等器官或仪器获取，并有一定交际功能的东西，“信息”是不确定性的减少，是负熵。所谓“处理”，是指用计算机对信息进行各种加工，主要的是图像信息和语言信息的识别、模拟、分析、转换和传输。严格意义上讲，“汉语计算机自动分析”比“中文信息处理”更加确切，为表述的习惯，在这里，我们依然沿袭这一称呼。

2002年9月，笔者有幸参加了在台北市举行的第十九届国际计算语言学学术会议(The 19th International Conference on Computational Linguistics) SIGHAN（Special Interest Group on HAN）研究兴趣组关于“十年后的中文处理”的讨论，台北“中研院”的黄居仁教授详尽的回顾了中文信息处理在台北的二十年发展史。实际上，祖国大陆的中文信息处理历史更加悠久、而且取得了许多实实在在的、改变了人们生活的成就，然而境外的研究群体、我们国家非中文信息处理领域的人员、乃至从事这一方向研究的人员也知之甚少。因此，笔者依据目前所能收集的文献资料，整理出中文信息处理二十年的科学发展史，并从自己专业的角度出发，予以评述，希望能对投身这一领域的研究人员或者工程技术人员提供一些历史参考资料，并盼望有心人能够“以史为鉴”或者“温故而知新”。

本文的第一部分将综述中文信息处理的难点，第二部分按照发展的各个阶段，阐述中文信息处理的发展史，最后探讨目前中文信息处理的问题及应对方案。

2、中文信息处理的难点

汉语在世界上属于汉藏语系，是一种孤立语。汉语在历史上先后吸收和同化了匈奴、鲜卑、突厥、契丹、满、蒙古、梵语等语言里面的许多成分[1]。其主要特点有：

*基金项目： 国家重点基础研究项目(G1998030507-4；G1998030510)

©作者简介：**张华平**(1978.2-)：男，江西波阳人，硕博连读生，主要研究领域为计算语言学与中文信息处理。

(1) 汉语的独一无二的特色是：完全使用由象形文字演化而来的方块汉字；

(2) 词语没有形态标记；

汉语是以字为基本单位，词之间没有明显的标记，词本身也没有明显的形态标志。所以中文信息处理的基础课题和特有的问题就是中文分词，分词本身的也有一定的错误率[2]，这无疑降低了后续处理的实际效果。

(3) 结构松散，比如：我上街买菜，看见一个人，穿着一件军大衣，打了卖菜的一巴掌，脸都肿了。

(4) 语法灵活，即缺乏狭义的形态，汉语句子中各个成分之间的关系一靠词序，二靠“意合”，三靠虚词。[3]

(5) 语义灵活，一方面语法的灵活主要来源于语义的灵活；另一方面同一结构可以表达不同的意思，同一意思可以用不同结构表达。[3]

另外.现有的自然语言处理理论和技术大多都是以英语为研究对象语言发展起来的.而汉语无论在语音、文字表示，还是在词汇，语法，语义及其语用等各个层面上都与之存在着很大的差异.这使得无法直接套用西方已成熟的理论和技术，汉语无疑是计算模型比较不发达的语言。这对从事中文信息处理的研究者来说是一个巨大的挑战和压力。

3、中文信息处理发展史

从我国早在1956年的开始了俄汉机译研究，并于1959年取得成功，至今差不多有50年的历史， 但当时的技术主要是词与词翻译和模式匹配，缺乏句法和语义分析[4]，几乎谈不上真正的中文信息处理。下面笔者依据时间顺序，根据当时的主流研究方法和研究的主要问题，将中文信息处理技术的发展史分为如下6个阶段进行阐述。

3.1 学习和理论探索的萌芽阶段

这一阶段以介绍国外计算语言学领域的理论方法为主。

对国外相关领域的介绍，理论内容相对较少，主要偏重在各种上机实现的系统方面。范继淹[5]、徐志敏[5]、李家治[6]、陈永明[6]、冯志伟[7]等人的介绍及其所研制的实验系统报告，是这方面的代表。早期将国外的理论方法进行全面系统汉化的主要刊物有：86 年底创刊的《中文信息学报》，语言学界的《国外语言学》和《语言文字应用》。

学者们在介绍国外先进的理论和方法同时，也有不少人结合汉语自身的特点，对这些理论和方法做了深入一步的探索，极少数人对自然语言理解做了深层次的带有哲学色彩的思考，如：80 年代中期宁春岩发表的《自然语言理解中的几个根本问题》[8]，以及他译介的美国哲学家休伯特.德雷福斯（Hubert L.Dreyfus）的专著《计算机不能做什么－－人工智能的极限》[9]，语言学界袁毓林 1993 年发表了《自然语言理解的语言学假设》[10]。

这些早期的的研究和探索对确立中文信息处理的宏观格局起到了决定性的作用[11]、奠定了中文信息处理后期的理论基础。

3.2 汉字信息处理为主的早期阶段

1974 年周恩来总理亲自批准了“七四八”工程，它标志着计算机中文信息处理技术受到了国家高度重视并且进入了他的第一个发展阶段——汉字信息处理时代。[12] 。在新技术面前，完全使用由象形文字演化而来的方块汉字不能直接进入电脑，因而受到了变革的冲击。

1880 年，丹麦人编制了汉字电报码本，用于电报传输汉字；1956 年，我国钱文浩提出了“码化理论”，他认为把汉字编为 4 位数字的电码，又把数字换成点和划的系统，这两个过程都是码化过程，汉字被码化后就可以作为信息来传输和处理了。从那时到现在，研究汉字信息处理的有识之士，克服种种困难，已经创造出近 1000 个汉字输入编码方案了，1986

年 3 月，国家有关部门举办了全国汉字编码方案评测，有 33 个方案参评，评出大众码、五十字元码、部形编码、笔形编码等 11 个 A 类方案。1987 年 10 月，中国中文信息学会等组织的“中华杯”汉字录入赛，操作员在规定字比赛中最高输速达 70 字/分； 1990 年，在海峡两岸中文电脑表演赛上，专业操作员单字输入达 147.8 字/分，词语输入达 203.3 字/分。在经历了所谓万“码”奔腾的汉字编码战国时代之后，这方面的问题已经基本解决。从键盘到 OCR 到手写识别到语音输入，汉字的输入方式已经是多种多样，能够满足多种需要了。

跟汉字的输出密切相关的是汉字字库的信息压缩技术。享有“当代毕升”美誉的北京大学教授王选与其同事一道研制成功的汉字折线段压缩技术，很好地解决了这个难题。从而划时代地使汉字文献的印刷出版告别铅与火，进入电子时代。

3.3 字、词等表层处理为特征的初级阶段

汉字信息处理成功解决之后，接着面对的是更为复杂的词法分析问题。在这一阶段主要研究和解决的问题就是字、词等表层问题。[12]其中重要的史实有：

1.北京大学开发的华光排版系统被评为 1985 年中国十大科技成就之一，并荣获中国发明协会发明奖。

2.“六五”期间(1981-1985)，北京航空学院主持，中国人民大学等十几个院校，研究机构参加的“现代汉语词频统计”工程是这一阶段代表性的重大科研成果，这是国内首次使用计算机进行大规模语料（2000 万字）的词频统计研究的大型语言工程。专家们把这次词频统计工程称之为经国大业，不朽盛事。

3.第一个汉语自动分词系统——CDWS，建立了一个有 13 万余词条的计算机词典，研制了一个有 52 个属性的汉字信息库。

4.“七五”期间(1986-1990)，建立了功能完备、实用有效的“汉字属性系统”，编纂并出版了汉字属性字典。

5.1988 年初，北京航空航天大学在承担国家“七五”科技攻关项目《信息处理用规范现代汉语词库》的同时，提出并经过了三年的努力，汲取了语言界和计算机界数百名专家的宝贵建议和意见，最终制定了《信息处理用规范现代汉语分词规范》，从计算机工程应用的需求出发，解决了语言学界争论了几十年而未解决的汉语的词的定义问题。为我国从汉字处理进入词语、语句处理打下了基础。

3.4 句法和语义等深层处理为代表的中期阶段

“八五”期间，中文信息处理技术的研究开发重点逐步由字、词的表层处理转向了以句法、语义分析为核心的深层处理。电子部计算机与微电子发展研究中心(CCID)联合国内从事中文信息处理的主要单位，从信息处理用汉语语法、语义体系的应用研究着手，以中文信息处理产品的智能化为目标，组织实施了并形成了一个完整的中文信息处理应用平台工程。

从80 年代开始，在借鉴国外的自然语言语义理论的基础之上，先后提出了一系列符合汉语特点的语义分析方法和语义表示理论。如汉语格语法理论，汉语的各种信息在语义网络中的表示方法等。 在构造语义规则时，基本上采用上下文无关文法（CFG）。与语法规则不同的是表示非终止符和终止符的内容是与语义有关的概念知识而不是VP(动词短语)或N(名词)等语法术语。

3.5 语料库统计方法兴起的近期阶段

语言学的研究必须以语言事实作为根据，必须详尽地、大量地占有材料，才有可能在理论上得出比较可靠的结论。在这种工作中逐渐创造了一整套完整的理论和方法，形成了一门新的学科 -- 语料库语言学（corpus linguistics），并成为了自然语言处理的一个分支学科。[13]

其中有影响力的中文生语料库、词语语料库、句法语料库有：

(1) 1979 年，武汉大学建设的汉语现代文学作品语料库，共计 527 万字，是我国最早的机器

可读语料库。

(2) 《人民日报》收集了 48 年的全部文字和图像内容，公开发行。

(3) 北京大学计算语言学研究所与富士通公司（Fujitsu）合作，加工 2700 万字的《人民日报》语料库，加工项目包括词语切分、词性标注、专有名词（专有名词短语）标注。还要对多音词注音。他们还建立了一个小型汉语树库：与新加坡国立大学计算机系合作，内容为新加坡中学语文教材（1995 年），所有的句子都分析为树形图。北大语料库的特点有：规模大、加工深、覆盖面广、正确率高、无著作权纠纷。

(4) 1998 年，清华大学建立了 1 亿汉字的语料库，着重研究歧义切分问题。现在生语料库已达 7-8 亿字。

(5) 北京邮电大学在美国 LDC 的汉语句法树库的基础上进行自动获取语法规则的研究。LDC 的树库包含新华社 1994 到 1998 年的 325 篇文章，包含 4185 颗树，10 万个词。

(6) 香港城市理工大学语言资讯科学研究中心建立了 LIVAC(Linguistic variety in Chinese communities)语料库，其宗旨在于研究使用中文的各个地区使用语言的异同。总字数为 15,234,551 字,经过自动切词和人工校对之后总词数约为 8,869,900 词。

(7) 台湾建立了平衡语料库（Sinica Corpus，中央研究院）和树图语料库（Sinica Treebank，中央研究院）。两个都是标记语料库，有一定加工深度。语料库规模约 500 万字。

口语语料库主要是中国社会科学院语言所、中国科学院自动化所建设的；

用来翻译和研究各种不同语言对比的语料库有：北大、哈工大、东北大学建立的英汉双语语料库；北京外国语大学的北京日本学研究中心建立了 2000 万字的汉语和日语并行语料库；山东海洋大学的《蝴蝶》（王蒙小说）德汉对照语料库；复旦大学计算机系建立了容量为 1GB 汉日英分类熟语料库，包含数千个类别，数十万篇文章。

同时，我国少数民族语料库有：新疆师范大学 200 万词的维吾尔语语料库；中国社会科学院民族研究所 500 万藏语字符的藏语语料库；内蒙古大学的蒙古语语料库，并进行了初步的切分和标注。

3.6 以Internet为主要应用对象、大规模真实文本、智能信息访问的现阶段

近年来，Internet 迅猛发展，根据中国互联网络信息中心发布的报告，截止到 2002 年 6 月 30 日，中国上网计算机总数 1613 万，上网用户总数 4580 万[14]。人们在享用 Internet 带来的各种便利的同时，却又被如何从浩如烟海的网上资源中，如何快速、高效的查找自己的信息所困扰，典型的主要需求有信息分类、信息提取、自动问答、基于内容的快速信息检索、基于个性的信息推送，数字化图书馆和信息网格等。因此中文信息处理技术必须解决网络环境下的、大规模的、信息（文本或语音）智能访问、加工处理、自动分析理解。

现阶段，中文信息处理的特征主要表现为：统计方法与规则方法相结合、基础理论研究与实用系统并重、面向 Internet 的大规模真实文本的智能信息访问。

1、统计与规则结合

现在人们已经不再做更多的“经验主义”和“理性主义”的争论，更多的是汲取两家之长，以实用的智能化系统为目标，以大规模语料测试为评价目标。近年来，国家863计划智能计算机专家组，曾对语音识别、汉字（印刷体和手写体）识别、文本自动分词、词性自动标注、自动文摘和机器翻译译文质量等课题进行过多次有统一测试数据和统一计分方法的全国性评测[15]。最近刚刚结束的973专家组第二次汉英机器翻译评测系统评测了国内主要的汉语词法分析系统，获得最好成绩的中科院计算所汉语词法分析系统ICTCLAS[16]就是采用了统计方法与规则相结合的手段。清华大学的黄昌宁教授等人就成功地结合语料库统计与规则的优点.设计了一个统计与规则并举的汉语句法分析模型CRSP, 在这个模型中.语料库用来支持各类知识和统计数据的获取.并检验句法分析的结果，规则主要用于邻接短语的合并和依存的关系网的剪枝， 他们的实验取得了令人满意的结果。

2、基础理论研究与实用系统并重

当前重大的基础理论研究成果有：

1）董振东教授的知网；知网是一个以汉语和英语的词语所代表的概念为描述对象，以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库，它为语言信息处理的研发提供了丰富的知识资源。[17]

2）黄曾阳先生的 HNC 理论; HNC 理论是"Hierarchical Network of Concepts（概念层次网络）"的简称，是关于自然语言理解处理的一个理论体系。它以概念化、层次化、网络化的语义表达为基础，把人脑认知结构分为局部和全局两类联想脉络，认为对联想脉络的表达是语言深层（即语言的语义层面）的根本问题。中心目标是建立自然语言的表述和处理模式，使计算机能够模拟人脑的语言感知功能。该理论使自然语言理解获得了突破性的进展，它所蕴涵的精深丰富的思想对人工智能、语言学、计算机科学和认知科学等都具有重要的理论和应用价值，对中文信息处理和汉语研究尤其具有实际意义。

3）北京大学计算语言所的《现代汉语语法信息词典》；它是以朱德熙先生提出的词组本位语法体系作为设置各项语法范畴的理论基础。首先是选取一些具体的功能标准确定了汉语的词语分类系统，并对照一个词语的句法功能表现按义项把它归入某个词类；然后是以功能理念指导词语语法属性项目的设置，并根据一个词语的实际用法情况标记它的属性值。[11]

另外，北京大学计算语言所已经启动一项建设中文类 WordNet 的重大基础知识工程，它的建成，也将从根本上促进中文语义理解、句法分析等深层次的核心理解问题的解决。

3、面向 Internet 的大规模真实文本的智能信息访问

主要的方向和系统有： 1）基于内容的搜索引擎，代表性的系统有北京大学天网、计算所的“天罗”、百度、慧聪等公司的搜索引擎； 2）信息自动分类、自动摘要、信息过滤等文本级应用，如上海交通大学纳讯公司的自动摘要、复旦大学的文本分类，计算所基于聚类粒度原理 VSM 的智多星中文文本分类器；3）信息自动抽取，即将 Internet 上大量的非结构化的信息，抽取出格式化的数据，以备进一步的搜索应用。目前是研究热点，至今还没有实用的系统； 4）自动问答、机器翻译等需要更多自然语言处理和理解的应用。

4、中文信息处理技术发展的问题与应对

二十余年来，经过中国语言学家和计算机专家的艰辛努力，中文信息处理技术取得了非常惊人的成绩。但是，相对于日益发展的 Internet，相对于快速膨胀的中文信息、相对于十几亿中文语种用户来说，现代中文信息处理技术依然滞后，很多技术和系统依然是实验室的原型，离实际的应用还有较大差距。主要问题[3]体现在：

1、 汉语言学家没有为中文信息处理作好语言分析的准备，长期以来，对汉语的研究方法基本上是例举性的，而非穷尽的；材料和对象基本上是书面的，而非口语的。

2、 中文信息处理研究力量分散而且存在着低层次重复、缺乏统一规范和标准的问题。

3、 现代汉语研究领域和计算机领域的隔绝状态没有出现根本性的改变。

笔者认为，应对的措施关键在于：（1）联合汉语研究专家和计算机专家，培养精通语言学和计算机技术的“两栖”人才，紧密合作，集体攻关。(2)改变目前研究单位封闭、大而不细，全而不精、低水平重复、小作坊式的研究方式；将国家支持的研究成果开放，供广大的研究人员自由共享，实现合作、互补、共赢。9 月，中国科学院计算所向社会免费发布 15 项研究成果的做法很值得中文信息处理的研究机构学习。同时，我们很欣喜地看到，在计算语言学专家白硕研究员、刘群副研究员的倡导下，中国科学院计算所自然语言处理组搭建了中文自然语言处理开放平台[18] (www.nlp.org.cn)，并将他们多年的研究成果（包括所有相关的论文、源代码、文档等）无偿的在平台发布，并采取开放自由源码的方式，为广大的中文自然语言处理感兴趣者、研究者、业界提供了一个可以共同建设的自由社区。自由社区里，

大家是建设者，同时也是共享的受益者，最终形成一种良性循环。这种合作机制能从根本上解决中文信息处理技术封闭、低水平重复的弊病。我们呼吁更多的中文信息处理领域的研究人员、工程技术人员加入到这个社区，真正的将我国的中文信息处理事业推向新的高度，造福广大的中文语种社区，并让中文真正走向世界！

附记 本文对中文信息信息处理领域 20 多年的历史和评述，是极为简略和粗浅的。一方面受篇幅限制，一方面也因为作者知识水平和认识的局限，有很多重要的研究成果文中没有提及，而评述不当也势必存在，疏漏有误之处恳请专家学者指正。

致谢 感谢刘群副研究员、张浩学友提供的资料，感谢孙健博士、骆卫华硕士、邹纲学友严格而有善意的讨论；感谢 Azalea 朋友的建议。

参考文献

[1] 白硕 . 计算语言学教程 . 2001.6:4-5

[2] 张华平,刘群 . 基于 N-最短路径的中文词语粗分模型 . 中文信息学报 . 2002. 16(5): 1-7

[3] 许嘉璐 . 现状和设想——试论中文信息处理与现代汉语研究 . 中国语文 . 2000. 6

[4] 郭艳华,周昌乐 . 自然语言理解研究综述 . 杭州电子工业学院学报 . 2000.2. 20(1)

[5] 范继淹,徐志敏 . RJD-80 型汉语人机对话系统的语法分析 . 中国语文 . 1982(3)

[6] 李家治,陈永明 . 机器理解汉语——实验 I . 心理学报 . 1982(1)

[7] 冯志伟 . 国外自然语言理解系统简介 . 计算机科学 . 1984 年第 2 期

[8] 宁春岩 . 自然语言理解中的几个根本问题 . 语言研究 . 1985(2)

[9] 休伯特·德雷福斯(Hubert L.Dreyfus). 计算机不能做什么——人工智能的极限 . 宁春岩译,马希文校 . 三联书店 . 1986

[10] 袁毓林. 自然语言理解的语言学假设 . 中国社会科学 . 1993(1)

[11] 詹卫东 . 80 年代以来汉语信息处理研究述评 . 当代语言学 . 2000.2 (1)

[12] 中国中文信息学会 . 我国中文信息处理的发展与展望 . 中国科学技术协会"科学技术面向新世纪"学术年会 . 1998.9 . 137-140

[13] 冯志伟. 中国语料库研究的历史与现状. 国际中文电脑会议 ICCC2001 论文集(新加坡) . 2001.11 . 1-15

[14] 中国互联网络信息中心. 中国互联网络发展状况统计报告(2002/7). . 2002.7 . 5

[15] 黄昌宁. 统计语言模型能做什么？. 语言文字应用 . 2002,2002(1): 77-84

[16] Kevin Zhang (Zhang Hua-Ping), Qun Liu, etc. Automatic Recognition of Chinese Unknown Words based on Roles Tagging. SIGHAN, COLING2002 . 2002.9 .

[17] 杜飞龙. 知网辟蹊径共享新天地—董振东先生谈知网与知识共享. 微电脑世界. 2000.9 .

[18] 刘群,张浩,白硕. 中文信息处理开放平台的设计.第一届学生计算语言学研讨会论文集. 2002.8 . 339-345

Brief History of Development in Chinese Information Processing

ZHANG Hua-Ping

(*Software division*, *Institute of Computing Technology*, *The Chinese Academy of Sciences*, *Beijing* 100080, *China*)

E-mail: zhanghp@software. ict.ac.cn

Abstract: Strictly speaking, Chinese information processing has a history of over 20 years until now. With the popularization of computer and development of Internet, Chinese information processing really change our life. This paper focuses on coordinating the brief history of development in Chinese information processing based on the documents collected. Upon some important research projects, the author give his own comments from the view of his knowledge. It aims to provide a clear development map and help to achieve “Learning experience from the history” or “Knowing new knowledge after going over history”.

Keywords: Chinese Information Processing, Brief History