第十三章

计算语言学研究 70 年

第一节　引言

计算语言学是利用计算机技术，从计算的角度出发，寻找自然语言的规律，以使得计算机系统能够像人类那样理解和生成自然语言的研究。计算语言学是一门交叉学科，它涉及语言学、数学以及计算机科学等。在计算机领域，被称为自然语言处理。当处理的对象为中文时，称中文信息处理。

1950 年提出的图灵测试一般被认为是自然语言处理思想的发端。时至今日，图灵测试的场景依然是计算语言学的主要研究对象，而研究方法则几经变革。纵观计算语言学研究方法的演化，我们可以将研究方法归纳为规则、统计，以及深度学习三类。

表 13—1　　研究方法分类

	理性主义	经验主义
符号主义	规则方法	统计方法
联结主义		深度学习方法

这种划分本质上反映的是人工智能从符号主义（Symbolicism）方法向联结主义（Connectionism）方法演进的过程。符号主义方法认为人工智能来源于数理逻辑。它主张使用离散的符号表示知识，将知识系统地归纳成公理体系，采用某种形式化的语言来描述推理过程。而联结主义方法则认为人工智能来源于对人脑机制的模拟。它认为知识不存在于特定的地点，而是分布在神经网络内相互联系的神经元中。当对这些神经元的刺激超过了某个阈值，这些神经元将被激活，神经元的整体活动构成了人类的认知。

与此同时，计算语言学也经历了从理性主义方法向经验主义方法的演进，知识获取的途径由语言学家通过内省获得，发展到由机器自动地从语料库里学习和提取。

本章对新中国成立 70 年以来我国计算语言学的研究成果进行梳理。在国际计算语言学研究方法演化的大背景下，我国计算语言学的发展历程，可以大致划分为四个时期：汉字信息处理时期、规则时期、统计时期，以及深度学习时期。关于这样的划分，有几点说明。

（1）我国最早开展的计算语言学研究是机器翻译，它甚至早于汉字信息处理时期的研究，可视为我国计算语言学研究的萌芽。

（2）汉字信息处理时期解决的主要问题是如何在计算机中使用汉字，考虑汉字本身，而不考虑其承载的语言学意义。后面三个时期主要研究汉语各种语言单位的计算及其应用问题，包括词法分析（分词、词性标注）、句法分析（依存句法分析、短语结构句法分析）、语义分析（语义角色标注）、篇章分析（篇章结构分析、指代消解）、知识图谱、信息抽取（命名实体识别、关系抽取、事件抽取）、信息检索、机器翻译、文本分类与聚类、情感分析、问答、推荐、社会计算、多模态信息处理等多个领域[①]。

---

① 中国中文信息学会：《中文信息处理发展报告（2016）》2016 年 12 月，http：//www. cipsc. org. cn/download. php? file = cips2016. pdf。

（3）统计方法在我国计算语言学研究的各个时期均有所贡献。早在 1959 年，刘涌泉在《我国机器翻译工作的进展》中就曾提及统计词尾和词的频率。这种朴素的统计方法，通过统计语料库，从定量的角度对语言事实进行刻画，有时进而对语法理论进行检验。而“统计时期”使用的“统计”方法，主要是指使用传统机器学习方法，利用统计数据直接参与语言运算的研究。

（4）在计算语言学内的各个领域，创新的节奏不尽相同，而且诸多研究采取多种研究方法相融合的思路。因此，上述按照研究方法划分的几个时期在时间上有时存在交叠。

鉴于计算语言学研究领域繁多，囿于篇幅，本章选取机器翻译和中文分词两项任务作为梳理的重点。中文分词作为我国独具特色的计算语言学任务，目前技术已臻成熟。而机器翻译作为图灵测试的一部分，不论在国际上还是国内都是最早开展的计算语言学研究，研究历程最长。而且其研究内容涉及词法、句法、语义等多个领域，几乎计算语言学的研究成果都可以直接或间接地用于机器翻译。因此，本章从这两项具有代表性的任务出发，对我国计算语言学的研究脉络进行梳理，以求管中窥豹。

第二节　汉字信息处理时期

在我国推广和普及计算机，首当其冲遇到的问题就是如何在计算机中处理汉字。1974年，国家计委批准了“关于研制汉字信息处理系统工程”（“748工程”）的建议。自20世纪70年代中期至90年代初期，汉字信息处理技术成为研究热点。

汉字信息处理技术以汉字为处理对象，包括汉字字符集的确定、编码、字形设计、存储、输入、输出、编辑、操作系统、排版等

①。这个时期的研究带有浓厚的工程技术色彩。

一　编码

汉字编码技术是整个中文信息处理的基础，总体上经历了从无到有，从基础到完善，从国家标准（GB）向国际标准（ISO）发展的历程。在最初的机器翻译研究中，曾使用四角号码做汉字编码②。1978 年，我国自主完成的第一个汉字编码输入系统面世③。1980 年，中国国家标准总局颁布《信息交换用汉字编码字符集·基本集》（GB2312—80），包括 6763 个汉字。这是我国第一个汉字信息技术标准。1993 年颁布了与 ISO/IEC10646 相对应的国家标准 GB13000《信息技术通用多八位编码字符集（UCS）第一部分：体系结构与基本多文种平面》，包括 20902 个汉字，字符集涵盖简体及繁体汉字。2000 年发布了 GB18030—2000《信息技术信息交换用汉字编码字符集基本集的扩充》的国家强制性标准，包括 27533 个汉字，该标准同时收录了藏族、蒙古族、维吾尔族等主要的少数民族文字，为推进少数民族语言信息化奠定了基础。

二　输入

与汉字编码问题密切相关的是输入法。1979 年，中国科学院计算所倪光南等在《计算机辅助汉字输入的一种方法》中首次提出了汉字输入的联想功能。1983 年，王永民五笔字型汉字编码方案通过鉴定，该方案具有速度快、重码少的特点，因此在“万码奔腾”的

① 俞士汶：《民族特点的文化要求——汉字汉语民族语言进入信息系统》，罗沛霖主编：《信息电子技术知识全书》，北京理工大学出版社 2006 年版，第 298—311 页。

② 刘涌泉：《机器翻译和文字改革（下）》，《文字改革》1963 年第 3 期，第 6—9 页。

③ 支秉彝、钱锋：《浅谈“见字识码”》，《自然杂志》1978 年第 6 期。

时代脱颖而出，成为专业录入人员普遍使用的输入法。

三　输出

在汉字输出方面，1968 年，中国科学院计算所研制的 717 机显示器能显示 256 个汉字，是我国最早的汉字显示器之一。

四　操作系统

1983 年，电子部六所研制的 CCDOS 是我国第一套与 IBM PC-DOS 兼容的汉语操作系统。1983 年，中国科学院计算所研制的 GF20/11A 系统通过鉴定，是我国第一台在操作系统核心部分进行改造的汉字操作系统。1985 年，第一台具备完整中文信息处理能力的国产微机长城 0520CH 研制成功。同年，联想汉卡通过软件和硬件相结合的方式，进一步提升了汉字处理能力。1986 年，四通集团推出的四通 MS2400 中外文文字处理机面世，在全国掀起了从旧式的机械打字机向具有编辑功能的新式电脑打字机过渡的热潮。1988 年，希望公司发布的 UCDOS 中文操作系统、金山公司推出的 WPS 中文办公软件，以及基于 Windows 3.0 开发的外挂式中文平台中文之星 1.0 版等，一度成为当时广为普及的汉字信息处理产品。

五　排版

汉字精密照排是“748 工程”的项目之一。1981 年，第一台汉字激光照排系统“原理性样机”通过鉴定。1987 年，《经济日报》为我国第一家采用方正汉字激光照排系统的单位，开创了全部废除铅字排版的先河。到 1993 年，国产激光照排系统已覆盖国内 99% 的报纸和 90% 以上的黑白书刊的出版印刷，在中国掀起了“告别铅与火，迎来光与电”的印刷技术革命[①]。

---

① 中国工程院编，常平主编：《20 世纪我国重大工程技术成就》，暨南大学出版社 2002 年版，第 29—41 页。

六　其他

汉字信息处理时期还有多项汉字统计成果问世。70 年代末期，冯志伟的《汉字的熵》率先对汉字信息熵展开研究，通过对语料库进行手工统计，测定汉字的信息熵为 9.65 比特，为汉字双字节编码奠定了理论基础。1980 年，“748 工程”组织研究者手工统计了 2100 余万字的语料，编印了《现代汉字综合使用频度表》。

汉字信息处理问题的解决，对汉语计算语言学研究起到了重要的奠基作用，标志着研究主战场从“字处理”转战“词处理”。

第三节　规则时期

早期计算语言学研究普遍采用规则方法。规则方法具备符号主义和理性主义的特点。一方面，语言单位以离散符号表示，通过符号逻辑进行推理；另一方面，由语言学家通过内省，手工编写规则和词典。我国计算语言学研究的规则时期从 20 世纪 50 年代开始至 90 年代末期。

一　机器翻译

1950 年代中期至 1980 年代初期，汉语计算语言学研究主要围绕着机器翻译方向展开。

我国是世界上继美、苏、英之后，第四个开展机器翻译研究的国家。1956 年，《机器翻译、自然语言翻译规则的建立和自然语言的数学理论》的课题被纳入《1956—1967 年科学技术发展远景规划》。

1957 年，中国科学院语言研究所（现为中国社会科学院语言研究所）与计算技术研究所合作开展俄汉机器翻译研究，开创了我国

机器翻译研究的先河。1959 年，俄汉机器翻译试验[①②]翻译了九种不同类型较复杂的句子。该系统由俄语词典（2370 条）和语法分析规则、汉语词典和调整词序规则等部分组成，其中，语法规则系统由 29 个线路图表组成。系统采用穿孔纸带输入，输出的不是汉字而是代码。1958 年底至 1960 年初，又研制了一套英汉机器翻译规则系统。1964 年，整理出版了《机器翻译浅说》一书，初步总结了当时我国机器翻译研究进展，具有里程碑式的意义。

研究者意识到，通过查词典的方法实现的词对词机器翻译，难以克服译文可读性低的问题。因此，他们从最初就注重句法分析，着重研究了俄汉机器翻译中的词序调整问题。刘涌泉等发表的《机器翻译的发展概况和我国机器翻译的特点》，根据外汉机器翻译的特点提出了“中介成分理论”。中介成分介于原语和译语之间。确定中介成分需要对原语进行语法和语义分析，同时考虑到译语转换，遵循逻辑语义原则、结构层次原则以及对比差异原则。中介成分不仅能表示语言成分的功能意义（如主谓宾定状补）和语言成分的分布关系（如直接、间接成分），而且还能反映两种语言的对比差异（如前置、后置成分）。中介成分理论曾在早期的中国机器翻译研究中产生了重要影响。

1966—1975 年，由于国内“文化大革命”和国际机器翻译进入低潮的双重原因，机器翻译研究工作陷于停顿。

1975 年后，国内多家单位纷纷开展机器翻译研究。

1981 年，冯志伟根据依存语法和配价语法理论，研究汉语到外语的机器翻译，将 20 多篇中文科技文章翻译成俄英法德日等五种语言[③]。在此基础上，1983 年，冯志伟的《汉语语句的多标记多叉树形图分析

---

① 高祖舜：《俄汉机器翻译初步试验成功》，《中国语文》，1959 年第 11 期。

② 中国科学院计算技术研究所三室五组：《俄汉机器翻译试验》，《电子计算机动态》，1960 年第 4 期，第 1—14 页。

③ Feng, Z., 1982, “Memoire pour une tentative de traduction automatique multilangue de chinois en francais, anglais, japonais, russe et allemande”, *Proceedings of COLING'82*, Prague.

法》根据从属关系语法（即依存语法）[①] 理论，提出了多标记多叉树形图分析法。针对单标记短语结构语法存在的生成能力过强的问题，他提出了“多标记”概念，主张采用多标记函数代替短语结构语法的单标记函数。同时，他认为语句具有非线性的多层次的树形结构。在机器翻译的过程中，首先生成原语树形图，将其转换为译语树形图。然后取译语树形图中各叶子节点的词，进行形态变化，即可得到译文。这种树形图转换的思想与基于句法的统计机器翻译中树到树模型的思路比较类似。

随着转换生成语法的兴起，人们普遍认为要实现机器翻译必须对语言理解进行形式化。然而，也存在不同的声音。宁春岩等[②]将机器翻译视为利用计算机将一个符号序列向另外一个符号序列进行转换的过程，可以通过机械地进行等值复写、调位、增项、减项等四种类型的运算完成。因此，机器翻译无须建立在对原语及译语理解的基础之上。这种思想对机器翻译在形式上进行了高度抽象。不难看出，这种思想与深度学习中序列到序列的机器翻译模型存在相似之处。

1987 年之后，市场上涌现出一大批商品化的机器翻译系统[③]。

在外汉机器翻译方面，1987 年，中国人民解放军军事科学院董振东等研制的“科译 I 号”英汉机器翻译系统[④]通过了鉴定。次年，该系统经由中软公司的二次开发，成为我国第一个商品化英汉机器翻译系统，命名为“译星 I 号”。

“科译 I 号”系统以董振东在《“科译 I 号”机译系统的语言学理论基础》中提出的逻辑语义结构为理论基础。他认为，语言之所

---

① 冯志伟：《特思尼耶尔的从属关系语法》，《国外语言学》1983 年第 1 期。

② 宁春岩、邱润才：《HT83 英汉机器翻译系统研制报告》，《情报科学》1984 年第 3 期，第 68—74 页。

③ 冯志伟：《迈向实用化和商品化的机器翻译研究》，《语文建设》1994 年第 8 期，第 36—39 页。

④ 董振东、张德玲：《科译 I 号英汉机译系统概要》，《现代图书情报技术》，1986 年第 3 期，第 18—25 页。

以能够成为交际工具是由于它的表义性。因此，机器翻译应该以原语和译语共同的逻辑语义为基础。对一个句子或短语进行逻辑语义分析，就要求解其中各个逻辑语义核的逻辑语义以及它们所处的层次。这里逻辑语义核指在逻辑语义结构中承担了逻辑语义的词或词组。翻译时，从根节点逐层展开，形成译语的线性结构，得到相应的译文。概括地说，多标记多叉树形图分析与逻辑语义结构都生成语言的层级结构。多标记多叉树形分析法生成的是句法树；而逻辑语义结构生成的是语义树，并且不进行原语树到译语树的转换。

1986 年 3 月启动的国家高技术研究发展计划（863 计划）旨在提高我国的自主创新能力。在 863 计划的支持下，中国科学院计算所陈肇雄等研制了《智能化英汉机译系统 IMT/EC》。该系统有英语基本词 35000 条，汉语词 25000 条，通用规则 1500 条，以及大量的特殊规则和成语规则。

20 世纪 90 年代，陈肇雄提出的面向机器翻译《SC 文法功能体系》，“是一种基于传统的上下文无关文法、语义文法，以及超前与反馈分析技术和格框架约束分析等基础上的上下文相关处理文法”。它将语法及语义分析结合起来，在语法规则中引入了上下文相关条件测试，简化了分析与转换的操作过程。

此外，高立英汉及日汉机器翻译系统、国防科技大学史晓东研制的 Matrix 英汉系统、通译英汉—汉英系统、LIGHT 英汉系统、雅信英汉系统、铁道部研制的“TECM 英语机译系统”、信通“英汉机器翻译系统 Marco Polo”等也是具有代表性的外汉机器翻译系统。

在汉外机器翻译方面，1989 年，哈尔滨工业大学周明等[①]研制的汉英机器翻译系统 CEMT－I 成为我国第一个通过技术鉴定的汉英机器翻译系统。该系统能处理八类汉语简单陈述句和规范的题录。

---

① 周明、李生、胡铭曾、石淼：《交互式汉英机器翻译系统 CEMT－Ⅱ》，《情报学报》1990 年第 2 期，第 151—154 页。

1993 年，由中软公司吴蔚天等开发的 SinoTrans 汉英—汉日机译系统[①]通过了电子工业部的部级鉴定，翻译速度达到每小时 20000 个汉字。SinoTrans 从黎锦熙的句本位学说出发，提出了汉语完全语法树的概念，用于表达汉语陈述句型。

事实上，虽然这些商品化机器翻译系统也有一定销量，但译文质量普遍不高，用户实际使用的情况并不理想。译文质量可以通过评测反映出来。国家 863 计划计算机软硬件主题（原智能计算机主题）专家组从 1991 年开始，多次举办中文信息处理与智能人机接口技术评测，机器翻译的评测结果并不乐观。

俞士汶等研制的 MTE（机器翻译译文质量自动评估系统）是世界上较早提出的机器翻译自动评测系统之一。MTE 借鉴了语言测试中分离式测试的方法，每个句子着重测试一个语言点。可实现分项评测，如单独测试系统的词汇能力或者语法能力，亦可进行整体评测，缺点是需要专家编写试题，成本高，题库难以扩充。

综上所述，基于规则的机器翻译系统主要使用双语词典进行词对词翻译，并使用包括源语言分析规则、语言转换规则和目标语言生成规则等的翻译规则调整词序。在方法论上，机器翻译的语法研究与传统汉语语法研究存在区别。传统的汉语语法研究通常着眼于从某种理论出发，刻画典型的语言规律，采用枚举例句或创造例句的方法。而机器翻译研究者面临的则是真实场景下的典型和非典型语言现象，需要语法规则具备良好的完备性以及一致性，尽可能覆盖较广的语言现象，并要尽量减少规则之间的冲突。

除了方法论不同，语法研究的侧重点也存在差异。从最初的以传统汉语语法范畴为基础的翻译规则，到以短语结构语法为基础的翻译规则，研究者逐渐认识到基于单一标记的短语结构规则无法独立完成句法分析，并且有限数目的短语结构规则也无法覆盖大规模

① 吴蔚天：《汉外机器翻译与汉语分析器》，《语言文字应用》1997 年第 3 期，第 85—90 页。

语料中的语法现象[①]。进一步地，研究者发现机器翻译必须保持原语和译语在语义上的一致性，因此，语义分析越来越受到重视。有鉴于句法分析的局限性以及语义分析的必要性，机器翻译研究普遍引入了复杂特征集、合一语法以及词汇主义等思想，典型地如刘倬等提出的《基于词专家的机器翻译系统——系统研制的语言学基础》、冯志伟提出的多标记概念，以及陈肇雄提出的 SC 文法等。这些研究通过向语法系统中加入细化的约束，避免粗粒度规则描述带来的生成能力过强、约束能力弱等问题。

二　分词

长期以来，汉语计算语言学研究的首要预处理步骤就是分词。它也是汉语计算语言学研究中独具特色的一项任务。

孙茂松、邹嘉彦在《汉语自动分词研究中的若干理论问题》中指出汉语分词研究面临三个困难，即汉语中词的概念缺乏清晰的定义，切分歧义处理，以及未登录词识别。由于一般采用统计的方法识别未登录词，因此与未登录词相关的讨论放在第四节。

（一）词的概念

邢福义等编写的汉语语法教科书《现代汉语》对“词”的定义是“具有一定语音形式的、能独立运用的、最小的语言单位”。然而，汉语中词的概念具有模糊性。孙茂松、邹嘉彦在《汉语自动分词研究评述》中指出，词的概念涉及汉语语言学理论研究中长期争论的语素、词、短语之间的界限以及词类等一些经典问题。在日常使用中，即使在汉语母语者之间中文词语的平均认同率也只有 0.76 左右[②]。

---

① 黄昌宁、张小凤：《自然语言处理技术的三个里程碑》，《外语教学与研究》2002 年第 3 期，第 180—187 +239 页。

② Sproat, R., Shih, C., Gale, W. et al, 1996, “A stochastic finite-state word segmentation algorithm for Chinese”, *Computational Linguistics*, 22 (3): 377 – 404.

1982—1986 年，由北京航空学院主持的《汉语处理的基础工程——现代汉语词频统计》工程是我国首次使用计算机进行的大规模语料词频统计研究，并实现了首个汉语自动分词系统 CDWS。

1993 年，经过计算机界和语言学界研究者的共同努力，我国颁布国家标准 GB－13715《信息处理用现代汉语分词规范》。该规范尝试解决语言学界争论了几十年的汉语的词的定义问题，以期满足计算机工程应用的需求。

为与语言学中更严格的“词”的概念以示区别，《规范》称文本中的词语为“分词单位”。分词单位具有确定的语义或语法功能，包括语言学中词的全部，以及满足条件的某些词组，例如“由此可见”。这样的处理方式，在一定程度上避免了词表规模急剧膨胀，也避免了词语被切碎后违背其原有组合的意义，或影响后续处理。

《规范》中多处使用“结合紧密、使用稳定”作为判断分词单位的条件。但是，孙茂松、邹嘉彦在《汉语自动分词研究中的若干理论问题》中指出，“紧密”和“稳定”的判断依然是主观而抽象的，具体操作时难以把握其度。例如《规范》认为动宾结构“吃饭”是分词单位，而“喝水”则不是。

孙茂松的《谈谈汉语分词语料库的一致性问题》主张分词语料库以切成“心理词”为宜，例如“大海”“唱歌”“吸烟”“打断”“缩短”“等于”等。即便如此，心理词依然具有模糊性，无法保证严格意义上的切分一致。

（二）切分歧义

1. 切分歧义类型[①]

梁南元《书面汉语自动分词系统—CDWS》最早从结构的角度对切分歧义进行划分，定义了交集型和多义组合型两种切分歧义类型。针对交集型切分歧义，还定义了链长。刘挺、王开铸的《关于

① 孙茂松、邹嘉彦：《汉语自动分词研究评述》，《当代语言学》2001 年第 1 期，第 22—32、77 页。

歧义字段切分的思考与实验》统计发现交集型切分歧义与多义组合型切分歧义的出现比例约为 1∶22。孙茂松、黄昌宁等在《利用汉字二元语法关系解决汉语自动分词中的交集型歧义》中建议称多义组合型切分歧义为包孕型或者覆盖型。

董振东的《汉语分词研究漫谈》从歧义消解的角度对切分歧义进行划分，分别称交集型和多义组合型切分歧义为偶发歧义和固有歧义。类似地，还有孙茂松、左正平的《汉语真实文本中的交集型切分歧义》，提出区分真切分歧义和伪切分歧义，其中伪切分歧义的消解与上下文无关。

2. 切分歧义检测

孙茂松、邹嘉彦的《汉语自动分词研究评述》将切分歧义处理在逻辑上分成检测和消解两个相对独立的步骤。

最早出现也是最基本的汉语自动分词方法是最大匹配法。它将切分歧义检测与消解这两个过程合二为一，给出唯一候选解。1963 年，刘涌泉的《机器翻译和文字改革（下）》介绍过这种方法。1986 年，刘源、梁南元在《汉语处理的基础工程——现代汉语词频统计》中首次将最大匹配法应用到中文分词系统 CDWS。在词典完备的情况下，大约每 169 至 245 字发生一次切分错误[①]。揭春雨等《论汉语自动分词方法》分析了最大匹配法的结构及时间效率。最大匹配法具有简单、快速的优点；缺点是由于长词一定覆盖短词，无法检测出包孕型歧义，而且严重依赖词典，无法处理未登录词。

依将句子与词典中的词进行匹配的方向不同，最大匹配法又分为正向最大匹配 FMM 和逆向最大匹配 RMM。总体来说，逆向匹配比正向匹配更有效。M. Sun 和 B. K. T'sou[②] 对双向最大匹配法的有效性进

---

① 梁南元：《书面汉语自动分词系统—CDWS》，《中文信息学报》1987 年第 2 期，第 44—52 页。

② Sun, M., and T'sou, B. K., 1995, "Ambiguity resolution in Chinese word segmentation", *Proceedings of the* 10*th Asia Conferenceon Language, Information and Computation.*

行了考察，发现只有不到 1% 的句子，双向的切分都是错误的。

王晓龙等《最少分词问题及其解法》提出的最少分词法，将分词建模为求解有向图两点间最短路径问题。这种方法不仅分词精度较正向和逆向最大匹配有所提高，而且改变了当时主要以字或固定词为单位的拼音输入模式，以短语或句子为单位进行输入，提高了输入效率。

3. 切分歧义消解

与机器翻译研究类似，切分歧义消解也经历了一个对语言分析的由浅及深、由简单到复杂的过程。

研究者普遍认为，95%左右的切分歧义可以借重句法以下的语言分析解决①②。早期的研究主要使用语素③、词频等浅层信息进行消歧。随着研究的深入，出现了各种类型的分词规则知识库，典型的研究包括李国臣等《汉语自动分词及歧义组合结构的处理》的联想—回溯法、黄祥喜《书面汉语自动分词的"生成—测试"方法》的扩充转移网络、徐辉等《书面汉语自动分词专家系统的实现》的专家系统，以及韩世欣等《基于短语结构文法的分词研究》短语结构文法等。

然而，随着语言分析的层次加深，切分精度提升不显著，效果有限。

三　基础资源建设

在规则时期，随着汉语计算语言学研究对大规模语言知识本体的需求日益迫切，一批有影响的词汇知识库、语法语义词典等作为汉语语言知识表示理论研究的代表性成果陆续问世。这些研究以各具特色的汉语理论为背景，以知识工程为方法，对汉语语法和语义知识从各

① 梁南元：《书面汉语自动分词系统—CDWS》，《中文信息学报》1987 年第 2 期，第 44—52 页。

② 徐辉、何克抗、孙波：《书面汉语自动分词专家系统的实现》，《中文信息学报》1991 年第 3 期，第 38—47 页。

③ 梁南元：《书面汉语自动分词系统—CDWS》，《中文信息学报》1987 年第 2 期，第 44—52 页。

个层面进行了刻画，构成了汉语计算语言学研究的基础资源。

（一）现代汉语语法信息词典[①②③]

北京大学计算语言学研究所与中文系合作开展语言资源建设工作，1998 年出版了《现代汉语语法信息词典》。这部词典是面向汉语信息处理的语言知识库。它以朱德熙、陆俭明、俞士汶等倡导的词组本位语法体系作为设置各项语法范畴的理论基础，根据语法—义项相结合、语法功能分布等原则，建立词语分类体系。

该机器词典由包含全部 5 万多词语的总库，以及 23 个各类词库组成。除此之外，某些类词库下面又设分库。属性字段用于标记一个词语的语法属性，例如能否重叠等。每个词在实际话语中的使用分布情况，大致可以根据其在属性字段上的取值反映出来。

詹卫东在《80 年代以来汉语信息处理研究述评——作为现代汉语语法研究的应用背景之一》中认为《现代汉语语法信息词典》是利用复杂特征集方法对汉语词语的语法知识进行形式化描述的一次大规模实践。

（二）现代汉语动词大词典

1994 年，中国人民大学语言文字研究所林杏光等主编的《人机通用现代汉语动词大词典》出版。该词典将汉语动词与名词性成分的语义搭配概括为一个由 22 个格组成的格系统，对 1000 多个汉语常用动词的语义格关系按义项进行了描写。

（三）知网

2000 年，董振东建立的知网（HowNet）1.0 Beta 版公开发布。知网（http：//www. keenage. com）是“一个以汉语和英语

---

① 俞士汶、朱学锋、王惠等：《现代汉语语法信息词典详解》，清华大学出版社 1998 年版。

② 俞士汶、朱学锋、王惠等：《现代汉语语法信息词典规格说明书》，《中文信息学报》1996 年第 2 期，第 1—22 页。

③ 俞士汶、朱学锋、王惠：《〈现代汉语语法信息词典〉的新进展》，《中文信息学报》2001 年第 1 期，第 59—65 页。

词汇所代表的概念为描述对象，以揭示概念与概念之间，以及概念的属性与属性之间的关系为基本内容的常识知识库”。“它着力要反映的是概念的共性和个性。”《知网导论》认为，“义原指最基本的、不易于再分割的意义的最小单位”。知网假设“所有的概念都可以分解成各种各样的义原。同时我们也设想应该有一个有限的义原集合，其中的义原组合成一个无限的概念集合”。知网从大约 6000 个汉字中提取出了有限的义原集合。知网 2.0 版包括 50220 个汉语词，62174 个汉语概念，55422 个英语词，以及 72994 个英语概念。

（四）HNC

黄曾阳的《HNC（概念层次网络）理论》认为，语句及自然语言的理解，就是从语言空间向语言概念空间的映射过程。结合汉语“字义基元化，词义组合化”的特点，HNC 理论以概念化、层次化、网络化的语义表达为基础，建立自然语言的概念空间，形成了 HNC 概念符号体系。目前，该研究已将 32570 个词语捆绑于 HNC 概念节点上，并标注完成 100 万字以上的熟语料（http：//www. hncnlp. com）。

董振东在 HowNet 中提出的概念关联及概念—属性表示理论，以及黄曾阳提出的概念层次网络理论，都是汉语语言知识表示理论研究的代表性成果。

（五）同义词词林

梅家驹等的《同义词词林》是第一部汉语类义词典[①]，共收录七万个汉语词语、词素、词组及成语等。该词典按照语义为主兼顾词类的分类方法，构建了汉语语义体系，为包括机器翻译在内的多方面的语言学研究提供基础[②]。

---

① 梅家驹、竺一鸣、高蕴琦、殷鸿翔：《编纂汉语类义词典的尝试——〈同义词词林〉简介》，《辞书研究》1983 年第 1 期，第 133—138、47 页。

② 鲍克怡：《汉语类义词典探索——〈同义词词林〉编后》，《辞书研究》1983 年第 2 期，第 64—70、152 页。

上述语言知识库的建立，普遍受到复杂特征集、词汇主义以及格语法等思想的影响。从理论框架的确立到每个词项的把握，研究者都需要结合汉语实际，进行深入思考。特别是有很多问题在语言学界长期存在争论，在实践中更加难以把握。这些研究出于语言工程的考虑，推出具有一定规模的探索性成品，具有广泛的理论和应用价值。

四　学术活动

随着汉语计算语言学研究的深入和人才队伍的壮大，中国中文信息研究会于 1981 年成立，1986 年更名为中国中文信息学会。1986 年，其会刊《中文信息学报》创刊。1991 年起，开始举办全国计算语言学联合学术会议（JSCL）[2008 年更名为全国计算语言学学术会议（CNCCL）]，是国内计算语言学领域最具影响力的全国性学术会议。

《国外语言学》（现为《当代语言学》）和《语言文字应用》曾大量介绍国外语法理论，为规则时期的语法研究提供了参考借鉴。

20 世纪 90 年代初，国内出版了多部计算语言学相关的著作，包括钱锋《计算语言学引论》（1990）、陆致极《计算语言学导论》（1990）、刘开瑛和郭炳炎《自然语言处理》（1991）以及冯志伟《中文信息处理与汉语研究》（1992）等。这些著作一方面反映了当时国外计算语言学研究的基本面貌，另一方面跟踪了国内汉语计算语言学研究的相关进展。

五　小结

在规则时期的后期，研究者普遍意识到复杂特征集和词汇主义思想可以有效地约束语法规则，因而编制出的约束规则已相当复杂。然而，由于自然语言固有的歧义性，随着规则的增多，规则之间的冲突愈演愈烈，规则系统存在的一致性低、可维护性低等缺陷也逐渐暴露出来。在这种情形之下，经验主义方法应运而生，为解决规则系统的瓶颈提供了思路。

第四节　统计时期

20 世纪 90 年代，随着计算机在存储方面性能的提升，语料库不断涌现。计算语言学研究由理性主义方法进入经验主义方法占主导的时期。

另一方面，随着计算机在计算方面性能的提升，机器学习方法崭露头角。在机器学习中，语言问题被抽象为数学模型，以数据驱动的方式，通过优化模型求解参数。传统机器学习系统的性能很大程度上依赖于专家的特征工程水平。特征是对规则的进一步抽象，而特征工程指对语料库中纷繁复杂的语言现象进行抽象，将其转换为机器学习算法的特征值输入的过程。在这个时期，研究者关注的是特征，而不再是具体的规则，转而由机器学习算法负责从语料库中自动地学习出规则及其相关的概率。

20 世纪 90 年代至 21 世纪 10 年代初期，我国计算语言学研究以大规模真实文本为研究对象，普遍采用的是基于传统机器学习的统计方法。

一　机器翻译

20 世纪 90 年代之后，统计机器翻译逐渐成为机器翻译研究的主流。统计机器翻译的思想[①]是将翻译理解为一个随机过程，使用一个语言模型和一个翻译模型对翻译进行建模。直观地说，语言模型刻画了译文的流利度，而翻译模型刻画了翻译的忠实度。统计机器翻译不需要人工构造词典和规则，从句子对齐的双语语料库中自动获得语言知识。因此，统计机器翻译具有开发成本低、周期短，可以

① Brown, P. F., Cocke J., Della Pietra S. A., et al., 1990, "A statistical approach to machine translation", *Proceedings of the Workshop on Speech and Natural Language-ACL*, pp. 146 – 151.

迅速迁移到新语种和领域的优点。

如图 13—1 所示[①]，统计翻译模型的发展，经历了基于词的模型、基于短语的模型和基于句法的模型三个阶段[②]。基于句法的模型可以根据句法形式的不同，分为基于形式化句法和基于语言学句法两类。

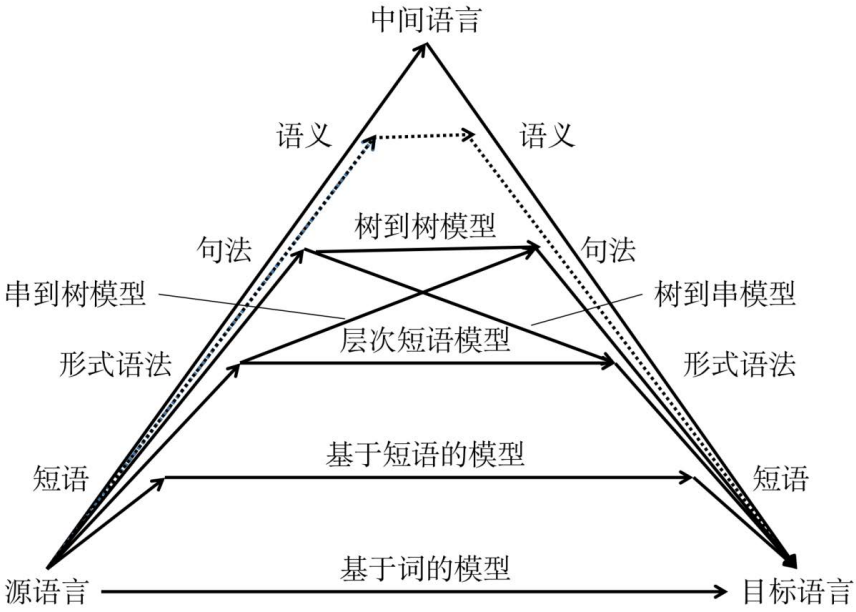

图 13—1　机器翻译金字塔

（一）基于形式化句法的统计翻译模型

与基于单词或短语的模型相比，这类模型在规则中引入了变量，从而使得翻译过程是有结构、层次化的。但是与通常语言学意义上的句法不同，这些变量并不涉及短语类型（NP、VP）和句法功能（主语、谓语）等概念。这方面的代表性工作包括 D. Xiong 等在

---

① 中国中文信息学会：《中文信息处理发展报告（2016）》，2016 年 12 月，http：//www. cipsc. org. cn/download. php? file = cips2016. pdf。

② 熊德意、刘群、林守勋：《基于句法的统计机器翻译综述》，《中文信息学报》2008 年第 2 期，第 28—39 页。

Maximum Entropy Based Phrase Reordering Model for Statistical Machine Translation 中提出的最大熵括号转录语法模型等。

这种方法要求为两种不同语言的规则之间建立对应关系，并服从相同的概率分布，这与自然语言的真实分布规律存在较大差异。

（二）基于语言学句法的统计翻译模型

根据所采用的句法结构树形式的不同，可以将模型分为以下两类。

1. 基于短语结构树的模型

短语结构树描述了句子的组成成分及各成分之间的句法关系。

（1）基于树的统计机器翻译模型

从2006年开始，中国科学院计算所在基于短语结构树的模型方面进行了大量探索。Y. Liu等在“Tree-to-String Alignment Template for Statistical Machine Translation”一文中提出一种基于树到串对齐模板(TAT)的统计翻译模型。这种方法首先用句法分析器获得源语言的短语结构树，然后利用自动获取的树到串对齐模板将源语言的树映射到目标语言上。树到串对齐模板有效地捕捉了词、短语以及子句的重排序信息。在源语言端句法分析正确的情况下，该模型可以较好地解决机器翻译中的长距离调序问题。

这是典型的树到串的统计机器翻译模型，只在源语言端利用句法结构，而目标语言端仅考虑词语序列。

树到串翻译模型为统计机器翻译提供了新的研究框架，在 COLING-ACL 2006 上获得了 Meritorious Asian NLP Paper Award，在国际上具有较大影响。

（2）基于森林的统计机器翻译模型

基于树的机器翻译模型强烈依赖于句法分析的正确性。在句法分析正确率低的情况下，直观的做法是在句法分析时输出多个最优结果。然而，这些结果中存在大量冗余，机器翻译性能提升因此相当有限。

针对这一问题，H. Mi 等在 Forest-Based Translation 中首次将句法压缩森林[①]的结构表示应用于统计机器翻译，后来的研究普遍采用了这种做法。

（3）基于串的统计机器翻译模型

通常情况下，训练句法分析器和翻译模型的语料在规模上和领域上均存在差异。训练源语言句法分析器的句法树库一般由人工标注，规模只有几万句；而训练翻译模型的双语语料库规模则可达几百万对句子。

为了适应这种差异，Y. Liu 和 Q. Liu 在“Joint Parsing and Translation”中提出了基于串的翻译方法。基本思想是源语言句法分析模型与翻译模型共享相同的概率空间，在句法分析的同时完成翻译解码。实验表明，在树到串模型中分别采用基于树、森林和串的三种方法进行解码时，系统性能逐步提高。

2. 基于依存树的模型

依存树描述了词与词之间的语义依存关系。熊德意等在《基于句法的统计机器翻译综述》中认为，与短语结构语法相比，依存语法的许多特性使其更适合机器翻译。基于依存树的统计机器翻译模

① Huang, L., Chiang D., 2005, “Better k-best Parsing”, *Proceedings of the 9th International Workshop on Parsing Technologies*, Vancouver, B. C, October.

型基本上是树到树模型。

D. Xiong 等在“A Dependency Treelet String Correspondence Model for Statistical Machine Translation”中提出了基于依存树权的机器翻译模型。这种方法不要求规则中每个非叶子节点都必须带上其子节点，只要是依存树上的联通子图即可用于建模。

除中国科学院计算所之外，哈尔滨工业大学、东北大学、厦门大学及中国科学院自动化所等也在统计机器翻译领域进行了富有成效的探索。

2012 年，中国中文信息学会召开了第八届全国机器翻译研讨会[①]。这届会议不同以往，没有继续举办机器翻译的公开评测活动。根据历届评测结果的统计发现，翻译质量已由显著提高发展到提高不明显。而且，很多单位依赖开源代码搭建系统，缺乏技术创新点。此时，长距离依赖和调序问题困扰着统计机器翻译模型，研究逐渐陷入瓶颈，需要寻找新的技术突破口。

二　分词

在规则时期分词研究遇到的三个困难，在统计时期依然是研究的重点。

（一）词的概念

分词规范、词表以及语料库可以从不同的角度对汉语词的概念进行界定。研究者对三者的关系存在不同看法。孙茂松《谈谈汉语分词语料库的一致性问题》认为，《信息处理用现代汉语分词规范》可操作性低，难以构造出高度一致的词表和分词语料库。梁南元等在《制定〈信息处理用现代汉语常用词词表〉的原则与问题讨论》以及孙茂松、张磊在《人机共存，质量合一——谈谈制定信息处理用汉语词表的策略》中，根据“人机结合、定性与定量并举”的思

① 杜金华、张萌、宗成庆等：《中国机器翻译研究的机遇与挑战——第八届全国机器翻译研讨会总结与展望》，《中文信息学报》2013 年第 4 期，第 1—8 页。

路，主张词表的构造应该并且必须依赖语料库。具体地，从生语料库出发，利用相邻汉字的互信息等统计量，生成候选词表；然后根据《规范》以及汉语构词法，对候选词表进行人工筛选，迭代地生成领域词表。

黄昌宁、赵海　《中文分词十年回顾》则强调词表对语料库标注的影响。他们认为，从计算的意义上而言，根据分词规范定义词表，再根据词表标注分词语料库的方法，使得原本很难精确定义的词，通过分词语料库体现出词语的一种可计算定义。他们主张在标注过程中严格执行　"词表驱动"原则，即在上下文未见歧义的情况下，将词表词作为一个完整的分词单位。杜绝所谓　"语法词"和　"心理词"的干扰。

2003年，国际计算语言学学会汉语兴趣小组(SIGHAN)举办了首届汉语分词国际评测Bakeoff①。因为Bakeoff分词评测中语料库的提供者并不公布他们使用的分词规范和词表，所以，在封闭测试中，分词语料库是可供机器学习的唯一资源。

（二）切分歧义

---

① Sproat, R. and Emerson, T., 2003, "The First InternationalChinese Word Segmentation Bakeoff", In: *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*. Sapporo, Japan: July 11 – 12, 133 – 143.

② 梁南元：《书面汉语自动分词系统—CDWS》，《中文信息学报》1987年第2期，第44—52页。

③ 徐辉、何克抗、孙波：《书面汉语自动分词专家系统的实现》，《中文信息学报》1991年第3期，第38—47页。

（1）无监督方法

基于“词是稳定的汉字的组合”这一观察，无监督分词方法认为汉字与汉字在上下文中相邻共现的概率能够较好地反映成词的可能性[5]，以此作为分词的定量依据。这些方法从生语料库出发，利用相邻汉字的串频[6]、互信息[7]、t 测试差[8]、卡方和广义似然比[9]等统计量，通过阈值进行分词。

---

① Lai, T. B., Sun, M, T'sou B. K., et al., 1997, "Chinese word segmentation and part-of-speech tagging in one step". *ROCLING* 1997 *Poster Papers*, 229 – 236.

② 沈达阳、孙茂松、黄昌宁：《汉语分词系统中的信息集成和最佳路径搜索方法》，《中文信息学报》1997 年第 2 期，第 34—47 页。

③ 白拴虎：《汉语词切分及词性标注一体化方法》，《计算语言学进展与应用》，清华大学出版社 1995 年版，第 56—61 页。

④ Wu, A., Jiang, Z., 1998, "Word segmentation in sentence analysis", *Proceedings of the* 1998 *International Conference on Chinese Information Processing*, pp. 169 – 180.

⑤ 刘迁、贾惠波：《中文信息处理中自动分词技术的研究与展望》，《计算机工程与应用》2006 年第 3 期，第 175—177、182 页。

⑥ 刘挺、吴岩、王开铸：《串频统计和词形匹配相结合的汉语自动分词系统》，《中文信息学报》1998 年第 1 期，第 18—26 页。

⑦ Sproat, R, Shih, C. L., 1993, "A statistical method for finding word boundaries in Chinese text", *Computer processing of Chinese and Oriental Language.*

⑧ 李蓉、刘少辉、叶世伟等：《基于 SVM 和 k-NN 结合的汉语交集型歧义切分方法》，《中文信息学报》2001 年第 6 期，第 13—18 页。

⑨ 黄萱菁、吴立德、王文欣等：《基于机器学习的无需人工编制词典的切词系统》，《模式识别与人工智能》1996 年第 4 期，第 297—303 页。

无监督分词方法的优点是无须人工标注，缺点是会抽出一些常用词组，例如“有的”，“这是”。

（2）有监督方法

孙茂松、邹嘉彦在《汉语自动分词研究中的若干理论问题》中指出，基于词频的切分消歧可抽象建模为与最少分词法类似的求解有向图两点间最优路径问题，不同之处在于边的权重为从人工标注语料库中统计的词频。这种方法的缺点是分词结果无法随上下文发生变化。

序列标注模型则具备了一定的上下文相关的消岐能力，它是有监督中文分词的经典模型。由于分词在文本理解过程中的初始地位，可供使用的语言特征通常仅限于滑动窗口内的 n-gram 及其状态转移概率，其中 n-gram 单元可以为字或者词。赵海在《中文分词十年又回顾：2007—2017》中指出，字特征多采取 5 字的滑动窗口；词特征多采用 3 词的滑动窗口。根据 n-gram 特征的抽取是否依赖词典，分词的序列标注模型可以划分为基于词和基于字等类型。

1990 年代，有研究者[①②③]使用隐马尔可夫模型，将自动分词和词性标注结合起来。这是一种基于词的序列标注模型。

Xue N. 等在 Combining classifiers for Chinese word segmentation 中提出的基于字标注的分词方法在 Bakeoff 评测中表现出色[④⑤]。基于字

---

① Lai, T. B., Sun, M, T'sou B. K., et al., 1997, “Chinese word segmentation and part-of-speech tagging in one step”. *ROCLING* 1997 *Poster Papers*, 229 – 236.

② 沈达阳、孙茂松、黄昌宁：《汉语分词系统中的信息集成和最佳路径搜索方法》，《中文信息学报》，1997 年第 2 期，第 34—47 页。

③ 白拴虎：《汉语词切分及词性标注一体化方法》，《计算语言学进展与应用》，北京：清华大学出版社，1995 年，56—61 页。

④ Xue, N. and Shen, L., 2003, “Chinese word segmentation as LMR tagging”. In: *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing.* Sapporo, Japan: July 11 – 12, 176 – 179.

⑤ Zhao, H., Huang, C. and Li, M., 2006, “An improved Chinese word segmentation system with condi-tional randomfield”. In: *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing.* Sydney: July, 108 – 117.

标注的分词方法将分词视为字序列标注问题，对句子逐字标注词位[①]。这里，词位是指一个字在构造词语时所占据的构词位置。以 2 词位的词位标注集为例，句子（A）的分词结果可以表示成如（B）所示的形式。其中，B 表示词首，E 表示词尾。

（A）分词结果：我/喜欢/唱/歌/。

（B）字标注形式：我/B 喜/B 欢/E 唱/B 歌/B。/B

F. Peng 等在 Chinese segmentation and new word detection using conditional random fields 的研究中进一步地将条件随机场（Conditional Random Field，CRF）引入分词学习，成为统计时期的主流分词模型。随后，又有研究者[②]引入了 semi-CRF。基于字标注的 CRF 模型通常基于马尔可夫过程建模，处理时每步只对输入序列的一个单元进行预测。而基于词的 semi-CRF 则基于半马尔可夫过程建模，处理时每步将序列中的连续多个单元标注成相同标签。

字序列标注模型输出的标签除了词位之外，还有研究参考句法分析中的移进规约等操作，尝试使用基于转移的分词模型。郭振等在《基于字符的中文分词、词性标注和依存句法分析联合模型》中定义了四种转移操作，用以实现基于字符的中文分词、词性标注和依存句法分析联合模型。

（三）未登录词

研究发现，在大规模真实文本中，未登录词对分词精度的影响

---

① Ng，H. T. and Low，J. K. 2004. " Chinese part-of-speech tagging：One-at-a-time or all-at-once? word-based or character-based? . " Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing.

② Andrew G. "A hybrid Markov/semi-Markov conditional random field for sequence segmentation". In Proceedings of the Conference on Empirical Methods in Natural Language Processing，2006，pp. 465 – 472.

比切分歧义至少大 5 倍以上[①②]。未登录词包括两类[③]：（1）新词、专业术语等；（2）专有名词，如人名、地名、机构名等。

在基于字标注的分词方法成为主流之前，针对第一种未登录词，主要利用无监督分词方法，在大规模生语料库上生成候选词表，人工筛选出新词补充到词表。针对第二种未登录词，主要利用从各种专有名词库中总结出的统计量（如姓氏频率）和专有名词的构词规则进行识别。

2002 年之后，主流的做法是把分词过程视为字序列标注问题，对词表词和未登录词一视同仁，简化了分词系统的设计，并且取得了良好效果。

（四）评测

刘开瑛在《现代汉语自动分词评测技术研究》报告中指出，1995 年，国家 863 智能机接口自然语言处理评测工作组组织了现代汉语书面语自动分词的评测，其中分词、交集型歧义、多义型歧义的处理正确率分别最高为 89.4%、78% 和 59%。在未登录词方面，人名和地名的正确率分别最高为 58% 和 65%。

2002 年，973 专家组评测了国内主要的汉语词法分析系统。中国科学院计算所采用统计与规则方法相结合的汉语词法分析系统 ICTCLAS[④] 获得最好成绩。2010 年，该系统获得了钱伟长中文信息处理科学技术奖一等奖。

国内举办的历届 863、973 分词评测，评测标准以《信息处

---

① 孙茂松、邹嘉彦：《汉语自动分词研究中的若干理论问题》，《语言文字应用》1995 年第 4 期，第 40—46 页。

② 黄昌宁、赵海：《中文分词十年回顾》，《中文信息学报》2007 年第 3 期，第 8—19 页。

③ 黄昌宁、李玉梅、朱晓丹：《中文文本标注规范（5.0 版）》，http://sighan.cs.uchicago.edu/bakeoff2006/MSRAsamp/msra-spec.pdf。

④ 刘群、张华平、俞鸿魁等：《基于层叠隐马模型的汉语词法分析》，《计算机研究与发展》2004 年第 8 期，第 1421—1429 页。

理用现代汉语分词规范》为主，以词表或参考语料为辅。在执行中，或者如刘开瑛在《现代汉语自动分词评测技术研究》中提倡的依据“从宽”的原则，在标准答案中标出多种形式的正确答案；或者如杨尔弘等在《汉语自动分词和词性标注评测》中主张的允许一定的“柔性”，在分词结果不同于标准答案的情况下，如果仍符合“结合紧密，使用稳定”的准则，就认为正确。黄昌宁、赵海在《中文分词十年回顾》指出，这种评测方案的弊端是人为的主观判断被引入了评测，同时影响了评测指标的直观理解。

鉴于在短期内各界难以在分词规范上达成共识，研究者们意识到不如换个思路，将问题解耦。在Bakeoff评测中，各个分词规范对应的语料库各自独立地进行评测，互不干扰。这样至少可以保证训练和测试语料是遵循相同的分词规范标注的。表 13—2 是历届 Bakeoff 公布的分词语料库的统计数据[①]。

表 13—2　　历届 Bakeoff 分词语料库一览表

提供者	语料库	编码	训练集词次数	测试集词次数	未登录词占比
台湾“中央研究院”	AS2003	Big 5	5. 8M	12K	0. 022
	AS2005		5. 45M	122K	0. 043
	AS2006		5. 45M	91K	0. 042
香港城市大学	CityU 2003		240K	35K	0. 071
	CityU 2005		1. 46M	41K	0. 074
	CityU 2006		1. 64M	220K	0. 040

① 黄昌宁、赵海：《中文分词十年回顾》，《中文信息学报》2007 年第 3 期，第 8—19 页。

续表

提供者	语料库	编码	训练集词次数	测试集词次数	未登录词占比
美国宾州大学	CTB2003	GB	250K	40K	0.181
	CTB2006		508K	151K	0.088
微软亚洲研究院	MSRA2005		237M	107 K	0.026
	MSRA2006		1.26M	100K	0.034
北京大学	PKU 2003		1.1M	17K	0.069
	PKU 2005		1.1M	104K	0.058

在统计时期，中文分词技术取得了长足的进步。首先，大规模分词语料库为汉语词的概念赋予了更高的可计算性。其次，基于字标注的机器学习方法能够显著提高未登录词识别的性能，已成主流，而基于规则的分词系统则逐渐退出历史舞台。

三　基础资源建设

（一）语料库

1979 年，武汉大学搭建的现代汉语文学作品语料库[①]，527 万字，是我国最早的机读语料库。

1992 年，北京大学计算语言学研究所与富士通公司合作，对 2700 万字的 1998 年《人民日报》进行了标注。根据俞士汶等《北京大学现代汉语语料库基本加工规范》介绍，标注内容包括词语切分、词性标注、专有名词标注，以及多音词注音。

2003 年，北京大学中国语言学研究中心推出的 CCL 语料库（http：//ccl. pku. edu. cn：8080/ccl_ corpus）包括现代汉语语料 5 亿字，古代汉语语料 2 亿字，以及含有 23 万个句对的英汉双语对齐语料库。

荀恩东等在《大数据背景下 BCC 语料库的研制》中介绍的北京

① 刘涌泉：《中国计算机和自然语言处理的新进展》，《情报科学》1987 年第 1 期，第 64—70、95 页。

语言大学的 BCC 汉语语料库，包括报刊、文学、微博、科技、综合和古汉语等多领域语料 150 亿字。

进入 21 世纪，随着互联网的普及，语料的来源从书刊报纸逐渐转移到互联网。常用的互联网语料有网页、微博、百科等几类。2012 年，搜狗公司发布的 SogouT 互联网语料库[①]，由来自互联网各种类型的 1.3 亿个原始网页组成，压缩前的大小超过了 5TB。

此外，《中文信息处理发展报告（2016）》总结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、内蒙古大学切分和标注后的蒙古语语料库等。

（二）知识图谱

与 HowNet 等手工研制的知识库和本体项目不同，统计时期的知识获取采用自动的方式，特别是以互联网语料为主要来源。随着大规模维基百科类富结构知识资源的出现，2012 年，谷歌提出知识图谱的概念。知识图谱以结构化的形式描述概念、实体及其之间的关系，本质上是语义网络。

在工业界，国内各大互联网企业纷纷开始着手构建自己的知识图谱，如百度“知心”，搜狗“知立方”。在学术界，上海交通大学建立的 Zhishi. me[②]（http：//zhishi. me）从百度百科、互动百科、中文维基等三大百科中抽取结构化数据进行融合，拥有一千万个实体数据和一亿两千万个 RDF 三元组。复旦大学的 CN-DBpedia[③]（ht-

① Liu, Y., Chen, F., Kong, W., et al. 2012. “Identifying Web Spam with the Wisdom of the Crowds”, *ACM Transaction on the Web*. Volume 6, Issue 1.

② Niu, X., Sun, X., Wang, H., et al., 2011, “Zhishi. me-Weaving Chinese Linking Open Data”, *Semantic Web In-Use track*, *The* 10*th International Semantic Web Conference* (*ISWC* 2011), Bonn, Germany.

③ Xu, B., Xu, Y., Liang, J., et al., 2017, “CN-DBpedia: a never-ending Chinese knowledge extraction system”, *International Conference on Industrial*, *Engineering and Other Applications of Applied Intelligent Systems*. Springer, Cham, pp. 428 – 438.

tp：//kw. fudan. edu. cn）从百科类网站的纯文本页面中抽取信息，经过滤、融合、推断等操作后形成结构化数据。清华大学建立的 XLore[①]（http：//xlore. org）则是基于中英文维基百科、百度百科以及互动百科等构建的大规模中英文平衡知识图谱。这些通用领域的知识图谱包含了大量现实世界中的常识性知识，覆盖面广，可以看作“结构化的百科知识库”。

四　学术活动

政府通过国家自然科学基金、国家社会科学基金、国家高科技研究规划发展项目（863）以及国家重点基础研究发展规划项目（973）等多种方式投入大量经费，积极推动语言技术发展[②]。

2006 年，中国中文信息学会首次举办钱伟长中文信息处理科学技术奖颁奖，这是我国中文信息处理领域最高科学技术奖。

中文信息学会在机器翻译、信息检索、知识图谱、社会计算、健康信息处理、少数民族语言文字信息处理等各个领域均举办了年会。值得一提的是，自 2002 年开始举办的全国学生计算语言学研讨会（SWCL），是面向汉语计算语言学领域学生的学术会议，整个会议由学生组织，打造学科生力军。自 2005 年开始每年举办的中文信息学会暑期学校暨 前沿技术讲习班 ，会对过去一年全球最新发表的相关论文进行归纳整理，帮助学员追踪学术动向，有助于提升整体科研水平。

随着来自中国的论文投稿在各大国际学术会议上的比例逐年增加，具有国际影响力的学术会议开始选择在中国举行。2004 年，首届国际自然语言处理联合会议（IJCNLP）在海南举行。2010 年，国

---

① Wang, Z., Li, J., Wang, Z., et al., 2013, “XLore: A large-scale english-chinese bilingual knowledge graph”, *Proc of the* 12*th Int Semantic Web Conf*, New York: ACM, pp. 121 – 124.

② 宗成庆、曹右琦、俞士汶：《中文信息处理 60 年》，《语言文字应用》2009 年第 4 期，第 53—61 页。

际计算语言学领域的权威学术会议之一 COLING 在北京举办。

在统计时期，计算语言学的方法论日趋成熟，陆续有多部教科书性质的学术专著及译著出版，包括冯志伟《计算语言学基础》（2001），黄昌宁、李涓子《语料库语言学》（2002），翁富良、王野翊《计算语言学导论》（2005），宗成庆《统计自然语言处理》（2008），冯志伟《自然语言处理简明教程》（2012），以及多部译著，包括苑春法等译《统计自然语言处理基础》（2005）[①] 以及冯志伟、孙乐译《自然语言处理综论》（2005）[②] 等。相关的机器学习方面的著作已成经典，包括李航《统计学习方法》（2012）和周志华《机器学习》（2016）。另外值得一提的是吴军的《数学之美》（2012），这是一本集趣味性与知识性的计算语言学科普书籍，获得读者高度评价。

统计和机器学习方法普遍依赖数据，而研究者往往面临数据匮乏的窘境。在这个时期，研究者们逐渐意识到资源共享对学术研究起到的巨大推动作用。资源共享的合作机制能从根本上解决研究中数据匮乏、技术封闭的弊病，避免低水平重复，缩短开发周期，形成良性循环。这是研究者们在意识层面的一次革新。结巴分词、中国科学院计算所 ICTLAS、哈尔滨工业大学刘挺等研发的语言技术平台LTP 等开源软件已使大量研究者受益。

2003 年，中文信息学会下属中文语言资源联盟（CLDC）（http：//www. chineseldc. org）[③] 成立。它是为推动我国语言资源共享成立的第一个联盟性学术组织。该联盟已拥有包括词典、分词词性标注语料库、句法树库、双语平行语料库、少数民族语言语料库、863 评测语料库等在内的多种语言资源。

---

① Manning, C., Schütze H., 2005.

② Jurafsky, D., Martin, J. H., 2005.

③ 陶建华：《中文语言资源联盟简介》，《术语标准化与信息技术》2010 年第 4 期，第 46—47 页。

第五节　深度学习时期

进入 21 世纪以来，高性能计算和海量数据为人工智能的崛起提供了引擎和燃料，深度学习已经在图像和语音领域取得了空前的进展。随后，深度学习的热潮也席卷了计算语言学。随着 word2vec[①②] 的诞生，语言的表示得以分布于神经网络相互关联的神经元中，联结主义在计算语言学中得以深刻体现。神经网络具备逐层抽象的能力，避免了传统机器学习方法中依赖专家的特征工程，使得传统机器学习中的流水线模型得以在深度学习中以端到端的形式呈现。

一　机器翻译

J. Zhang 和 C. Zong 的“Deep neural networks in machine translation: An overview”综述指出，在深度学习热潮中，神经网络最初登场是用于改进统计机器翻译。2013 年，基于神经网络的机器翻译方法被重新提出[③]。2014 年之后，端到端神经机器翻译[④]获得重视，其

---

① Mikolov, T., Chen, K., Corrado, G., et al., 2013, “Efficient estimation of word representations in vector space”, arXiv preprint arXiv: 1301.3781.

② Mikolov, T., Sutskever, I., Chen, K., et al., 2013, “Distributed representations of words and phrases and their compositionality”, *Advances in neural information processing systems*, pp. 3111 – 3119.

③ Kalchbrenner, N., Blunsom, P., 2013. “Recurrent continuous translation models”, *Proceedings of the ACL Conference on Empirical Methods in Natural Language Processing* (*EMNLP* 2013). Seattle, USA, pp. 1700 – 1709.

④ Sutskever, I., Vinyals, O., Le Q., 2014, “Sequence to sequence learning with neural networks”, *Proc of the 28th NIPS*. Red Hook, NY: Curran Associates Inc, pp. 3104 – 3112.

基本思想是在翻译建模上，撇开统计机器翻译的经典步骤，不需要词对齐、句法分析、翻译规则抽取等多层次的语言学抽象，由神经网络模型直接实现从源语言到目标语言的映射。

注意力机制[①]是对经典神经机器翻译模型的重大改进。它将源语言句子的编码由固定向量扩展为向量序列，使得在生成目标语言词语时，能够动态地参考与生成该词相关的源语言词语信息。目前，基于注意力机制的编码器—解码器模型已成为神经机器翻译的主流架构。

神经机器翻译可以明显改善统计机器翻译难以有效处理的长距离依赖和调序等问题，在译文流利度上要优于统计机器翻译。但是存在过度翻译和翻译不充分的情况，在忠实度上尚略逊一筹。

我国机器翻译研究者的贡献大致可以归纳为如下几个方面。

1. 对编码器、解码器的改进

以词语作为翻译基本单元的机器翻译系统面临着未登录词、词语切分、词语形态变化、数据稀疏等问题。针对这些问题，研究者们尝试采用更细的翻译粒度。J. Su 等[②]和 Z. Yang 等[③]采用汉字序列作为源语言端输入，由编码器端的神经网络自动抽象出词汇信息用于翻译。

以符号形式存储的双语词典和翻译规则等是重要的翻译知识。大量研究尝试通过扩展编码器和解码器的结构，将语言学知识融入神经机器翻译模型。W. He 等[④]提出在解码时加入词语翻译表和语言

---

① Bahdanau, D., Cho, K., Bengio, Y., 2014, "Neural machine translation by jointly learning to align and translate", arXiv: 1409.0473.

② Su, J., Tan, Z., Xiong, D., et al., 2017, "Lattice based recurrent neural network encoders for neural machine translation", *Proceedings of the* 31*st AAAI Conference on Artificial Intelligence* (*AAAI* 2017). San Francisco, USA, pp. 3302 – 3308.

③ Yang, Z., Chen, W., Wang, F., et al., 2016, "A character aware encoder for neural machine translation", *Proceedings of the COLING* 2016, *the* 26*th International Conference on Computational Linguistics*. Osaka, Japan, pp. 3063 – 3070.

④ He, W., He, Z., Wu, H., et al., 2016, "Improved neural machine translation with SMT features", *Proceedings of the* 30*th AAAI Conference on Artificial Intelligence* (*AAAI*2016). Phoenix, USA, pp. 151 – 157

模型。进一步地，X. Wang 等[①]提出在解码时由统计机器翻译提供目标语言候选词列表，以此提高目标语言的生成质量。

在句法特征方面，J. Li 等[②]、K. Chen 等[③]以及 S. Wu[④] 分别将源语言的短语结构特征和依存句法特征融合到编码器中；H. Chen 等[⑤]则将源语言句法信息融合到编码器和解码器双端。受统计机器翻译发展历程的启发，有研究探索将神经机器翻译从序列到序列模型扩展至基于句法树的形式。S. Wu 等在 Sequence-to-Dependency neural machine translation 中提出序列到依存神经机器翻译模型。

2. 对注意力机制的改进

Z. Tu 等在“Context gates for neural machine translation”中观察到，在翻译过程中，源语言上下文对翻译忠实度的影响较大，而目

---

① Wang, X., Lu, Z., Tu, Z., et al., 2017, “Neural machine translation advised by statistical machine translation”, *Proceedings of the* 31*st AAAI Conference on Artificial Intelligence* (*AAAI* 2017). San Francisco, USA, pp. 3330 – 3336.

② Li, J., Xiong, D., Tu, Z., et al., 2017, “Modeling source syntax for neural machine translation”, *Proceedings of the* 55*th Annual Meeting of the Association for Computational Linguistics* (*ACL* 2017). Vancouver, Canada, pp. 688 – 697

③ Chen, K., Wang, R., Utiyama, M., et al., 2017, “Neural machine translation with source dependency representation”, *Proceedings of the* 2017 *Conference on Empirical Methods in Natural Language Processing* (*EMNLP* 2017). Copenhagen, Denmark, pp. 2836 – 3842.

④ Wu, S., Zhou, M., Zhang, D., 2017, “Improved neural machine translation with source syntax”, *Proceedings of the* 26*th International Joint Conference on Artificial Intelligence* (*IJCAI* 2017). Melbourne, Australia, pp. 4179 – 4185

⑤ Chen, H., Huang, S., CHIANG D, et al., 2017, “Improved neural machine translation with a syntax-aware encoder and decoder”, *Proceedings of the* 55*th Annual Meeting of the Association for Computational Linguistics* (*ACL* 2017). Vancouver, Canada, pp. 1936 – 1945

标语言上下文对流利度的影响较大。在他们提出的上下文门方法中，在生成实词时更多关注源语言上下文，生成虚词时则给予目标语言上下文更多关注。

Y. Cheng 等[①]发现，神经网络机器翻译在源语言到目标语言，或目标语言到源语言的单向训练注意力机制时存在不足，因此提出对注意力机制进行双向联合训练，相互弥补，提高对齐和翻译性能。

有些研究尝试将统计机器翻译中广泛采用的各种特征引入注意力机制。Z. Tu 等在 Modeling coverage for neural machine translation 中使用覆盖向量记录翻译过程中的注意力历史，引导注意力机制更多地关注未翻译词语。通过这种方式，将统计机器翻译中的覆盖率引入注意力机制，有效缓解神经机器翻译普遍存在的过度翻译和翻译不充分问题。S. Feng 等[②]以及 J. Zhang 等[③]将统计机器翻译中的位变模型、繁衍模型等思想引入注意力机制，提高词对齐质量，缓解过度翻译问题。

3. 对外部记忆的改进

外部记忆应用在神经机器翻译的重要工作是华为诺亚方舟实验室的研究者 M. Wang 等在“Memory enhanced decoder for neural machine translation”中提出的 MEMDEC 解码方法。它通过外部记忆对循环神经网络解码器进行扩展，在一定程度上弥补了注意力机制的

---

① Cheng, Y., Shen, S., He, Z., et al., 2016, “Agreement-based joint training for bidirectional attention-based neural machine translation,” *Proc of the 25th IJCAI*. Palo Alto, CA: IJCAI, pp. 2761 – 2767

② Feng, S., Liu, S., Li, M., et al. 2016, “Implicit distortion and fertility models for attention-based encoder-decoder NMT model”. arXiv preprint/1601.03317v3.

③ Zhang, J., Wang, M., Liu, Q., et al. 2017, “Incorporating word reordering knowledge into attention-based neural machine translation”, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (*ACL* 2017). Vancouver, Canada, pp. 1524 – 1534.

不足。

有些研究将语言学知识存储在外部记忆里。Y. Feng 等[①]在记忆里存储低频词的翻译规则。Y. Tang 等的“Neural machine translation with external phrase memory”将符号形式的双语短语对存储在短语记忆里。解码时可以生成短语，突破了神经机器翻译解码时一次只能生成一个词语的限制。缺点是双语短语对仅支持一对一翻译。类似地，X. Wang 等[②]通过一个基于短语的统计机器翻译模型动态生成短语记忆。

4. 对模型架构的改进

D. He 等[③]利用对偶学习显著降低平行语料使用量。Z. Yang 等[④]以及 L. Wu 等[⑤]分别独立地将生成对抗网络应用到神经机器翻译中。B. Zhang 等[⑥]则采用变分神经机器翻译。

多语言机器翻译指采用一个模型完成多种语言之间的翻译。D. Dong 等的研究“Multi-task learning for multiple language translation”将一对多机器翻译的任务建模为多任务学习。通过共享源语言编码

---

① Feng, Y., Zhang, S., Zhang, A., et al., 2017, “Memoryaugmented neural machine translation”, *Proceedings of the* 2017 *Conference on Empirical Methods in Natural Language Processing* (*EMNLP* 2017). Copenhagen, Denmark, pp. 1401 – 1410.

② Wang, X., Tu, Z., Xiong, D., et al. 2017, “Translating phrases in neural machine translation”, *Proceedings of the*2017*Conference on Empirical Methods in Natural Language Processing* (*EMNLP* 2017). Copenhagen, Denmark, pp. 1432 – 1442.

③ He, D., Xia, Y., Qin, T., et al.. 2016, “Dual learning for machine translation”, *Proceedings of the* 30*th Conference on Neural Information Processing Systems* (*NIPS* 2016). Barcelona, Spain, pp. 1 – 9.

④ Yang, Z., Chen, W., Wang, F., et al., 2017, “Improving neural machine translation with conditional sequence generative adversarial nets”. arXiv preprint/1703. 04887v2.

⑤ Wu, L., Xia, Y., Zhao, L., et al., 2017, “Adversarial neural machine translation.” arXiv preprint/1704. 06933v3.

⑥ Zhang, B., Xiong, D., Su, J., 2017, “Variational neural machine translation”, *Proceedings of the* 2016*Conference on Empirical Methods in Natural Language Processing* (*EMNLP*2016). Austin, USA, pp. 521 – 530.

器，提高资源稀缺语言对的翻译质量。

神经机器翻译由于参数规模巨大，只有当训练语料库具备足够规模，才会显著超越统计机器翻译[①]。Y. Cheng 等在“Joint training for pivot-based neural machine translation”中提出的联合训练方法、Y. Chen 等在“A teacher student framework for zero-resource neural machine translation”中提出的“老师—学生”框架，以及 H. Zheng 等在“Maximum expected likelihood estimation for zero-resource neural machine translation”中提出的最大期望似然估计方法，都是采用枢轴语言实现资源稀缺语言之间的翻译。

L. Zhou 等[②]提出一种系统融合框架，通过多个注意力机制，对神经机器翻译和统计机器翻译的翻译结果进行融合。J. Zhang 等[③]通过将双语词典、短语表和覆盖惩罚等先验知识表示为对数线性模型的特征，集成到神经机器翻译中。

5. 对损失估计的改进

文献[④]指出神经机器翻译传统的训练准则极大似然估计存在问题。S. Shen 等的“Minimum risk training for neural machine translation”将最小风险训练方法引入神经机器翻译，获得稳定且显著的性能提升。

神经机器翻译在训练时，通常使用训练语料中的真实目标上下文预测目标词；而在推理时只能使用存在误差的已预测上下文预测

---

① Zoph, B., Yuret, D., May, J., et al., 2016, “Transfer learning for low-resource neural machine translation”, arXiv preprint arXiv: 1604.02201.

② Zhou, L., Hu, W., Zhang, J., et al., 2017, “Neural system combination for machine translation”, *Proceedings of the*55*th Annual Meeting of the Association for Computational Linguistics* (*ACL* 2017). Vancouver, Canada, pp. 378 – 384.

③ Zhang, J., Liu, Y., Luan, H., et al., 2017, “Prior knowledge integration for neural machine translation using posterior regularization”, *Proceedings of the* 55*th Annual Meeting of the Association for Computational Linguistics* (*ACL* 2017). Vancouver, Canada, pp. 1514 – 1523.

④ Ranzato, M., Chopra, S., Auli, M., et al., 2015, “Sequence level training with recurrent neural networks”, arXiv: 1511.06732.

新的目标词。W. Zhang 等在 Bridging the gap between training and inference for neural machine translation 中提出在训练时，不仅考虑真实序列，也从预测序列中采样获取上下文，取得显著的性能提升。该论文缓解了序列到序列模型长期存在的暴露偏差问题，摘得 ACL 2019 最佳长论文奖。

6. 对语料处理的改进

与双语语料相比，单语语料具有规模大、易获取等特点。J. Zhang 和 C. Zong 发表的“Exploiting source-side monolingual data in neural machine translation”中，通过自学习等方式使用源语言单语语料构造大规模双语平行语料。

Y. Cheng 等在“Semi-supervised learning for neural machine translation”中提出使用半监督学习方法同时利用源语言和目标语言单语语料，基本思想是引入自编码器训练双向神经机器翻译。

目前，神经机器翻译已经取代统计机器翻译，成为学术界和工业界商用在线机器翻译系统的主流方法，在在线翻译、跨语言检索等方面有着广泛应用[①]。在 WMT 2014 英语到法语翻译任务上，在百度大规模计算能力的支撑下，J. Zhou 等的“Deep recurrent models with fast-forward connections for neural machine translation”采用深层长短时记忆网络架构取得该任务最好成绩，首次超越统计机器翻译方法。小牛翻译 NiuTrans 由东北大学研发，支持 118 种语言互译，包含维蒙藏哈朝彝壮等七大少数民族语言，覆盖“一带一路”沿线国家所有官方语言。2016 年，NiuTrans 系统获得钱伟长中文信息处理科学技术一等奖。

李亚超等对多个在线神经机器翻译模型展开了评测。结果如表

① 王海峰、吴华、刘占一：《互联网机器翻译》，《中文信息学报》2011 年第 6 期，第 72—80 页。

13—3 所示[①]，所有翻译系统译文质量均达到较高水平。不过，由于测试集本身有可能已经被包含在这些神经机器翻译模型的训练语料中，这个结果只能作为参考。

表 13—3　　神经机器翻译系统性能对比

翻译系统	BLEU4	BLEU5	BLEU6	BLEU7
百度翻译	48. 96	41. 63	35. 36	29. 97
Google 翻译	50. 18	42. 94	36. 65	31. 25
小牛翻译	51. 67	44. 30	37. 81	32. 18
搜狗翻译	60. 72	53. 74	47. 47	41. 88

二　分词

在深度学习阶段，大量研究尝试使用神经网络模型对中文分词使用的传统机器学习模型进行替换。X. Zheng et al. 的 Deep learning for Chinese word segmentation and POS tagging 首次将深度学习方法应用于中文分词任务。该研究以预训练的字向量作为输入，用神经网络模型替换了 Low et al. [②] 的最大熵模型，进行序列标注。类似地，Y. Liu et al. [③] 将神经网络用于基于 semi-CRF 的分词。

X. Chen et al. 在 Long short-term memory neural networks for Chinese word segmentation 中提出利用长短期记忆神经网络捕捉长距离依赖，缓解了固定大小的滑动窗口在特征抽取方面的不足。Z. Huang et al. 在 Bidirectional LSTM-CRF models for sequence tag-

① 李亚超、熊德意、张民：《神经机器翻译综述》，《计算机学报》2018 年第 12 期，第 2734—2755 页。

② Low J. K., Ng H. T., and Guo W. 2005. A maximum entropy approach to Chinese word segmentation. In Proceedings of the SIGHAN Workshop on Chinese Language Processing, 2005, pp. 448 – 455.

③ Liu Y., Che W., Guo J., et al. 2016. Exploring segment representations for neural segmentation models. In Proceedings of the International Joint Conference on Artificial Intelligence, pp. 2880 – 2886.

ging 中首次提出使用“字向量 + 双向长短时记忆网络 + 条件随机场”模型进行中文分词，这种架构成为深度学习时期分词任务的主流架构。

长期以来，中文分词存在着多种标注规范及语料。X. Chen 等的“Adversarial multi-criteria learning for Chinese word segmentation”尝试利用生成对抗网络发掘这些语料中的共性。他们把使用多种语料的训练过程建模为多任务学习。在所有语料共享一个长短时记忆网络的基础上，每种语料均有单独的长短时记忆网络，这些模块共同构成特征抽取层。然后，在上述多任务框架基础上加入生成对抗网络，由判别器负责检查共享网络中是否混入了属于某种特定语料的特征，从而把私有特征从共享网络中剥离出去，保证了共享网络特征的单纯性。该论文获得了 ACL 2017 杰出论文奖。

赵海在《中文分词十年又回顾：2007—2017》中指出，深度学习时期的中文分词依然需要平衡地考虑未登录词和词典词的识别。实验表明，在未登录词的识别上，基于字的模型比基于词的模型更具有优势。D. Cai et al. 的Fast and accurate neural word segmentation for Chinese 研究，为训练集中的高频词直接计算词向量，而低频词或者未登录词的词向量则由字向量生成，为分词任务在深度学习时期的研究意义提供了新的思路。

总体而言，在分词任务上，深度学习与传统机器学习相比，无论是精度还是速度，并未显示出显著的优势。在深度学习时期，一方面分词技术日臻成熟，另一方面由于词向量技术，特别是上下文敏感的词向量技术[①]的发展，许多研究纷纷绕开分词的步骤，直接使用汉字作为输入。香侬科技 Y. Meng 等在“Is Word Segmentation Necessary for Deep Learning of Chinese Representations?”中指出，对很多使用深度学习的计算语言学任务而言，分词的必要性正在下降。

---

① Peters, M. E., Neumann, M., Iyyer, M., et al. 2018, “Deep contextualized word representations”, arXiv preprint arXiv: 1802.05365.

三　基础资源建设

C. Manning[①] 指出，深度学习带给计算语言学的最大改变源自词向量。北京师范大学中文信息处理研究所等机构的研究者[②]开源了「中文词向量语料库」，该库包含经过数十种用各领域语料（百度百科、维基百科、人民日报 1947—2017 年、知乎、微博、文学、金融、古汉语等）训练的词向量，涵盖各领域，且包含多种训练设置。此外，腾讯公司也发布了包含 800 多万个中文词语的词向量数据[③]。

谷歌于 2018 年底发布 BERT 预训练模型[④]，刷新了多项自然语言处理任务的最好成绩，被认为是继词向量之后，深度学习在自然语言处理方向的最大进展。2019 年，百度提出知识增强的语义表示模型 ERNIE（Enhanced Representation through Knowledge Integration）[⑤]，发布了基于 PaddlePaddle 的开源代码与模型。相较于 BERT 以汉字作为语言建模的单元，ERNIE 以字作为预训练的输入，对词、实体等语义单元进行语言建模，并使用大量知识类的中文语料进行预训练。ERNIE 模型在包括语言推断、语义相似度、命名实体识别、情感分析、问答匹配等多项任务中，均超越了 BERT 的性能。

---

① Manning, C., 2015, "Computational linguistics and deep learning", *Computational Linguistics*, Vol. 4, pp. 701 – 707.

② Li, S., Zhao, Z., Hu, R. et al., 2018, "Analogical Reasoning on Chinese Morphological and Semantic Relations", *ACL*.

③ Song, Y., Shi, S., Li, J., et al. 2018, "Directional Skip-Gram: Explicitly Distinguishing Left and Right Context for Word Embeddings", *NAACL* 2018.

④ Devlin, J., Chang, M. W., Lee, K., et al., 2018, "Bert: Pre-training of deep bidirectional transformers for language understanding", arXiv preprint arXiv: 1810.04805.

⑤ Sun, Y., Wang, S., Li, Y., et al., 2019, "ERNIE: Enhanced Representation through Knowledge Integration", arXiv preprint arXiv: 1904.09223.

四　学术活动

进入 21 世纪以来，国内研究者在国际上发表的论文数量呈爆发式增长。据《2018 自然语言处理研究报告》统计，针对 2014—2018 年 ACL、NAACL、COLING、EMNLP 等 4 个顶级国际会议的统计数据表明，累计发表 10 篇以上论文的国内学者包括中国科学院计算所刘群，哈尔滨工业大学刘挺，微软周明，北京大学常宝宝、李素建、万小军和穗志方，复旦大学黄萱菁和邱锡鹏、清华大学刘洋和孙茂松，软件所孙乐等。有鉴于此，中文信息学会青年工作委员会每年都组织国际顶会接收的论文作者在国内举办预讲会。预讲会不仅为演讲者提供了练兵的机会，而且为无法参加正式会议的研究者提供了与最新研究成果面对面沟通交流的机会。

2015 年，计算语言学领域顶级的国际会议 ACL-IJCNLP 在北京召开。2019 年，EMNLP-IJCNLP 在香港召开。值得一提的是，由中科院计算所等单位组成的国内研究团队在 ACL 2019 上首次斩获最佳长论文奖。这些表明国内研究正在逐步得到国际认可。

随着开源思想深入人心，国内研究者开源了大量神经网络相关的算法。例如，复旦大学开发的基于深度学习的中文自然语言处理系统[①]，该系统目前可用于中文分词、命名识别、词性标注、句子分类、句法分析、语义分析、知识库访问、对话问答、文本聚类分类、文本摘要、信息抽取、情感分析、三元组抽取等多项任务。

计算语言学同人工智能其他领域一样，技术更新迭代的速度非常之快。以往学术论文的发表周期较长，而且出于版权保护的考虑，论文的获取存在难度。在开源思想的引领下，许多研究者首选将论文公开发表在 arxiv 网站（https：//arxiv. org），然后再考虑向会议及传

① Zheng, X. , Chen, H. , Xu, T. , 2013, "Deep learning for Chinese word Segmentation and POS tagging", In *Proc. Conference on Empirical Methods on Natural Language Processing* (*EMNLP*' 13), October 18 – 21, pp. 647 – 657.

统论文出版渠道投稿。arxiv 是一个公开社区，全球的研究者可以免费上传、下载并引用其中的论文，其中不乏高引用率论文。

百度、阿里巴巴、腾讯、科大讯飞、搜狗、微软亚洲研究院、TRS 等公司均投入了大量资源从事计算语言学相关研发。这些公司与研究机构相比，在数据和计算资源方面具有优势。

自媒体也逐渐成为计算语言学推广的中坚力量。这些自媒体实时追踪国内外研究动态，通过订阅推送，定期组织线上直播，线下分享会和培训等方式，为学术圈注入活力。

五　小结

深度学习时期，神经网络对计算语言学的巨大推动作用主要体现在两个层面。一是语言的表示，词向量以及预训练模型对语言中蕴含的词法、句法、语义等信息进行建模，提高了语言的可计算性。目前，汉语已经具备了一定的语言表示资源。二是模型的计算。与传统机器学习相比，神经网络的建模更多地参考了人类的认知过程。在神经网络强大的学习能力之下，汉语与其他语言在抽象建模方面正在日趋同质化。

此外，如何将联结主义与日渐式微的符号主义有机地结合起来，是目前计算语言学面临的难题之一。一方面，不同于符号逻辑，神经网络的推理更类似于人类的直觉，经常受人诟病的一点是缺乏可解释性。缺乏合理的语言学解释，使得研究人员对神经网络模型进行分析和调试变得尤为困难。目前在可解释性上已出现一些进展，例如 Y. Ding 等发表的“Visualizing and understanding neural machine translation”可以在一定程度上对神经机器翻译的翻译过程进行可视化和解释。从神经网络模型中抽取语言学知识来解释和改进模型，将是未来重要的研究方向。另一方面，目前先验知识，特别是以离散符号表示的语言学知识，与神经网络融合时经常采用 ad hoc 的方式，缺乏通用性。

与此同时，人工智能的另一流派，行为主义（actionism）方法

开始登上计算语言学研究的舞台。强化学习、生成对抗网络等方法通过试错自主学习，在某些计算语言学任务上已经取得了不错的效果，期待未来有更大用武之地。

第六节　结语

纵观计算语言学的发展，我们有几点思考。

（1）几十年来，计算语言学研究对象的变化不大，依然是试图解决图灵测试的问题；然而，研究范式已然发生了巨变。从最初的符号主义和理性主义方法，演化到了联结主义和基于语料库的经验主义方法。

（2）面对自然语言的模糊性和复杂性，计算语言学一直在语言计算的复杂度和性能中寻找平衡。历年来，计算语言学最重要的研究成果无不源于提高了语言的可计算性。从最初的离散计算方式，已发展到使用概率、机器学习模型参数等实数值进行运算。

（3）开放、共享的开源思想已被研究者普遍接受，出现了一大批颇具影响的开源代码、公开的语料库及论文。这使得研究者们得以站在巨人的肩膀上，同时这也是计算语言学发展如此迅速的原因之一。

（4）公开的技术评测已然成为促进计算语言学研究的有效手段。一方面，它可以推动研究单位间的实质性交流。另一方面，对研究者而言，评测规范、数据及工具都是有价值的研究资源。快速、低成本的评测可以加速推进技术进步。

回顾新中国成立 70 年来计算语言学的发展，汉字信息处理技术、分词技术已经成熟，搜索引擎、机器翻译、智能问答等已成为日常生活的一部分。技术的进步使得我们正在逐渐接近图灵测试的理想场景。

2018 年，清华大学发布的《2018 自然语言处理研究报告》统计

指出，华人研究者发表论文的整体水平低于国际上自然语言处理领域头部的学者，论文引用率偏低。这说明虽然我国计算语言学研究已经紧跟国际潮流，但是仍然缺乏具有影响力的开创性研究。很多研究往往局限于对一些边缘问题的修补，或者只是针对特定条件下特定的解决方案，有待拓宽研究视野。尤其如何针对汉语自身的特点和规律，建立真正适合汉语的理论体系和实践方法，将是汉语计算语言学研究者长期面临的严峻挑战。

参考文献

常宝宝、张伟：《机器翻译研究的现状和发展趋势》，《术语标准化与信息技术》1998 年第 2 期。

陈敏、王翠叶：《中文信息处理的现状与展望》，《语言文字应用》1995 年第 4 期。

陈群秀：《汉语自然语言理解研究概况及前景》，《语文建设》1992 年第 9 期。

戴新宇、尹存燕、陈家骏等：《机器翻译研究现状与展望》，《计算机科学》2004 年第 11 期。

冯志伟：《我国机器翻译研究工作的发展》，《情报学报》1985 年第 3 期。

冯志伟：《我国机器翻译研究工作的回顾》，《语文建设》1990 年第 5 期。

冯志伟：《机器翻译发展的曲折道路（一）》，《术语标准化与信息技术》1996 年第 3 期。

冯志伟：《机器翻译发展的曲折道路（二）》，《术语标准化与信息技术》1996 年第 4 期。

冯志伟：《机器翻译——从实验室走向市场》，《语言文字应用》1997 年第 3 期。

冯志伟：《汉字和汉语的计算机处理》，《当代语言学》2001 年第 1 期。

冯志伟：《自然语言处理的历史与现状》，《中国外语》2008 年第 1 期。

冯志伟：《基于语料库的机器翻译系统》，《术语标准化与信息技术》2010 年第 1 期。

冯志伟：《计算语言学的历史回顾与现状分析》，《外国语（上海外国语大学学报）》2011 年第 1 期。

龚滨良：《建国以来中文信息处理技术大事记》，《中国科技史料》1985 年第 2 期。

黄昌宁：《中文信息处理中的分词问题》，《语言文字应用》1997 年第 1 期。

亢世勇：《计算机时代汉语语法研究的特点》，《术语标准化与信息技术》1999 年第 2 期。

刘峤、李杨、段宏等：《知识图谱构建技术综述》，《计算机研究与发展》2016 年第 3 期。

刘群：《统计机器翻译综述》，《中文信息学报》2003 年第 4 期。

刘群：《机器翻译研究新进展》，《当代语言学》2009 年第 2 期。

刘群：《基于句法的统计机器翻译模型与方法》，《中文信息学报》2011 年第 6 期。

刘群：《机器翻译技术现状与展望》，《集成技术》2012 年第 1 期。

刘洋：《神经机器翻译前沿进展》，《计算机研究与发展》2017 年第 6 期。

刘倬：《我国机器翻译研究的历史和现状》，《中国翻译》1983 年第 11 期。

清华大学计算机系—中国工程科技知识中心、知识智能联合研究中心（K&I）《2018 机器翻译与人工智能研究报告》2018 年 5 月，https：//static. aminer. cn/misc/article/translation. pdf。

清华大学人工智能研究院、北京智源人工智能研究院、清华—工程院知识智能联合研究中心：《人工智能之数据挖掘》2019 年 1

月，https：//static. aminer. cn/misc/pdf/datamining. pdf。

清华大学人工智能研究院、北京智源人工智能研究院、清华—工程院知识智能联合研究中心：《人工智能之知识图谱》2019 年 1 月，https：//static. aminer. cn/misc/pdf/knowledgegraph. pdf。

王献昌、史晓东、陈火旺：《机器翻译与自然语言处理的现状与趋势》，《计算机科学》1992 年第 3 期。

袁毓林：《计算语言学的理论方法和研究取向》，《中国社会科学》2001 年第 4 期。

张普：《共和国的中文信息处理 60 年》，《语言文字应用》2009 年第 3 期。

中国中文信息学会：《我国中文信息处理技术的发展与展望》，《中国科学技术协会．科技进步与学科发展——“科学技术面向新世纪”学术年会论文集》，1998 年。

中国中文信息学会语言与知识计算专委会：《知识图谱发展报告（2018）》http：//www. cipsc. org. cn/download. php? file = KGDevReport2018. pdf。

宗成庆、高庆狮：《中国语言技术进展》，《中国计算机学会通讯》2008 年第 8 期。

（作者　胡钦谙、顾曰国）