计算语言学与中文信息处理研究近年来的发展综述

（2004——2008）

詹卫东
北京大学中文系
北京大学计算语言学教育部重点实验室

1 引言

一般来说，年鉴的内容中不少属于“记账式”的，即把近些年本领域出版的有影响的文献，发生过的重要事件等分类罗列。这种方式对于比较传统的、相对成熟的学科，不失为一种稳妥的做法。但对于新兴的交叉边缘学科，以这种方式完成的年鉴可能有一定的局限。因为多数读者往往是从自己所在的学科背景出发来了解交叉学科中的研究状况，如果仅仅罗列事实，而不对事实背后的学术理路加以分析和评论，可能难以帮助读者真正全面地认识一个新兴的交叉学科中已经完成的研究工作的价值，因而也难以把握该学科未来的发展方向。如果真是这样的话，也就达不到为一个学科整理出版年鉴的目的了。

计算语言学与中文信息处理，正是这样一个涉及到计算机科学、语言学、文字学、数学、逻辑、认知科学等多个学科的交叉研究领域[1]。本文打算在整理近年来该领域中的重要事实的同时，对研究工作中表现出的宏观上的突出特点加以分析和评论，希望由此可以对该领域未来的学术发展方向有更为清晰的认识。这样，有可能帮助不同学科背景（尤其是语言学背景）的研究人员参与这一交叉学科的探索时更好地进行研究工作的定位。

基于上述的指导思想。下文将分为四节来综述 2004——2008 年这一领域的研究状况。第二节是概貌性的描述。先从不同角度勾勒对计算语言学与中文信息处理这一领域的宏观认识，为之后的内容阐述提供一个合适的逻辑框架。然后对这五年本领域的大环境做概要的描述（以一些重要学术活动为主）。第三节是从信息处理的不同对象和不同层级的角度，说明这一领域在 2004——2008 年取得的技术层面的进展。第四节则是从学术内在的发展理路，特别是研究方法的角度，对这一学科近年来的发展特点加以分析和评论。第五节是结语。简要回顾这一领域的发展历史，并对未来的发展趋势提出我们的看法。以上第二、三节侧重对客观事实的描述，第四节侧重主观评论。希望本文这种“客观与主观兼顾，务实和务虚并重”的安排，对跨学科背景的读者，更主要的是语言学背景的读者，能有一定的参考价值。

2 概貌：从整体和外部环境角度看计算语言学与中文信息处理

2.1 对计算语言学与中文信息处理的整体格局的认识

为了更好地概括说明计算语言学与中文信息处理这一领域近年来的理论研究以及应用状况，本文首先为这一领域勾勒一个相对全面的框架（表 1 和图 1）。然后再针对这个框架中一些更值得重点关注的部分展开来加以分析和讨论。

---

[1] 根据中国国家标准《学科分类与代码表》（GB/T13745—1992），一级学科“语言学”下的二级学科“应用语言学”里包含有三级学科“计算语言学”（740.3550）。一级学科“计算机科学技术”下的二级学科“人工智能”里包含有三级学科“自然语言处理”（520.2020）和“机器翻译”（520.2030）。从学术界的实际生态情况来看，人们一般不大去区分“计算语言学”“自然语言处理”“机器翻译”“中文信息处理”等不同名称所指的研究范围。使用不同的名称，往往被看作是对同一个对象的不同侧面的强调。本文也采取这种宽泛的方式。

表 1：根据符号性质的差异对中文信息处理的对象进行分类

对象 / 任务 / 对象	**书面文本** ［视觉符号］	**口语语音** ［听觉符号］
处理**符号的意义**	文本理解 ［机器翻译 信息检索…］ 文本生成 ［文本摘要 问答系统…］	语音识别 ［口语翻译…］ 语音合成 ［口语问答…］
处理**符号的形式**	汉字输入、存储、输出 篇章版式分解与生成	语音信号采集、 波形特征抽取、波形生成

图 1：根据语言单位性质的差异对中文信息处理的对象和技术层级进行分类

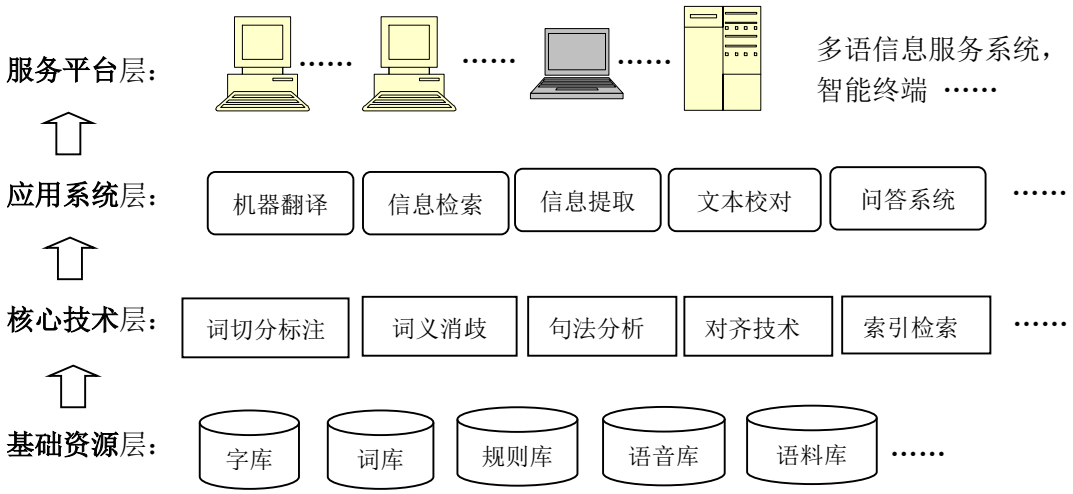
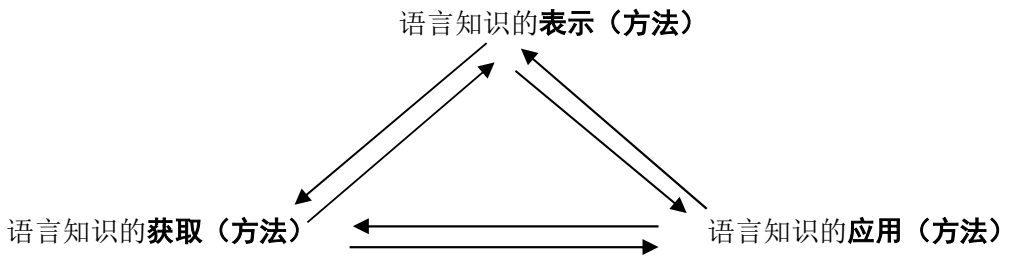


表 1 中的虚线表示不同的子领域有时候会发生相互作用，因而不同领域之间的界限并不总是截然分开的。图 1 可以看作是对表 1 中“符号的意义处理”这个层次的展开（“符号的形式处理”技术相对成熟且已经得到广泛应用，因此本文描述从简）。图 1 中提及的大多数概念都是针对书面文本信息处理的，但关于“基础资源”“核心技术”“应用系统”“服务平台”的层级划分，同样适用于口语语音信息处理的情况。本文第三节的综述基本是按照表 1 和图 1 这种认识框架出发来观察和描述计算语言学和中文信息处理近年来在技术层面所取得的进步。下面图 2 是从计算语言学的内在学术理路，即理论和方法的角度，来划分这一领域内的不同研究工作。本文第四节的综述就是取的这种认识角度，来考察和反思计算语言学近年来对语言知识的表示方式和知识的获取方式所发生的显著变化。

图 2：围绕“语言知识”看中文信息处理的研究内容



2.2 2004——2008 年中文信息处理大环境述要

伴随着我国整体国力的增强，中文在世界上的影响力与日俱增，计算技术水平的提高和互联网应用的迅猛发展，2004——2008 年间计算语言学与中文信息处理在国内迎来了一个非常好的发展环境。这可以从下面三方面的情况得到充分说明。

（一）发展中文信息处理技术的重要性被提升到国家科技发展战略的高度

2005 年年底的《瞭望新闻周刊》介绍了中国国家科技部对未来十年科技突破口的研究成果《中国技术前瞻报告》[2]，其中把“中文信息处理”列入信息领域下的一大突破方向。没过多久，2006 年年初国务院发布的长达 70 页的《国家中长期科学和技术发展规划纲要（2006—2020 年）》[3]中，列出了未来 15 年中国重点选择的前沿技术共八大方面，22 项技术，其中“中文信息处理”明确列入了第二大方面“信息技术”的“智能感知技术”[4]中。上述从国家战略角度做出的定位，直接的效果就是政府对自然语言（中文信息）处理技术相关课题的大力度支持。国家重大基础研究发展规划（通称 973 计划）、国家自然科学基金（NSF）、国家高技术发展规划（通称 863 计划）等都设立了重点支持这一领域的研究课题。其中 863 的“十一五”计划重点项目“中文为核心的多语言处理技术”（2006-2010）总经费达到 7000 万元。与此同时，在 2008 年北京举办奥运会这一重要历史事件作为中文信息处理大规模应用舞台的契机下，多项跨语言信息处理技术获得立项资助。使得中文信息处理研究人员比以往任何时候都有了更好的条件，来对这一领域的问题，从基础到应用，展开多层次的广泛而深入的研究。

这种大形势的变化也很快在信息科学领域的高层次学术出版物上得到了反映。中国计算机学会主编《2008 中国计算机科学技术发展报告》中，对计算语言学、机器翻译设置了专题介绍[文献 36，61]；《中国计算机学会通讯》2008 年第 2 期刊登了一期“中文信息处理”专辑[文献 40]。这些都说明信息科学界对中文信息处理的重视程度比以往有了显著提升。

（二）公开的技术评测及多层面的学术会议提供了更广泛的学术交流平台

由于自然语言信息处理任务本身的复杂性，对于各种理论方法在处理性能上的优劣，纯理论的论争很难给出有实质说服力的结论。此外，自然语言信息处理涉及的数据量非常大，不同的方法要进行比较，就需要在相同的大规模数据上进行可重复的测试。基于这两方面的原因，技术评测一直都是自然语言处理领域发展的重要推动力量。对此国际计算语言学界一直都对自然语言信息处理各项技术的公开评测给予非常高的重视[5][文献 38]，而国内以往的重视程度应该说是不够的。不过。这种情况在 2004——2008 年期间有了显著改观。可以从两个方面的表现来说明这种变化。

一是国际上有关中文信息处理的评测增多，同时国内研究人员参与的热情提高，并在一些项目上取得了较好的名次。有的研究单位在多次参加国际评测中，取得了明显的进步。

国际计算语言学会（ACL）下辖的汉语处理任务特殊兴趣小组（简称 SIGHAN）从 2003

---

[2] 参见新华网报道：http://news.xinhuanet.com/st/2005-12/15/content_3924650.htm。

[3] 参见中国政府门户网站：http://www.gov.cn/jrzg/2006-02/09/content_183787.htm。

[4] “智能感知技术”重点研究基于生物特征、以自然语言和动态图像的理解为基础的“以人为中心”的智能信息处理和控制技术，中文信息处理；研究生物特征识别、智能交通等相关领域的系统技术。

[5] 国际上目前已经有一大批有广泛影响力的定期举办的公开评测，涉及到自然语言信息处理的诸多子领域，这些评测往往能够极大地激发一个领域的技术创新。比如美国国防部高级计划研究署（DARPA）支持的“机器翻译语言信息识别、抽取及摘要”项目（TIDES）从 2002 年开始实施的“机器翻译评测”，在国际自然语言处理学界引发了一轮统计机器翻译的研究热潮（参见下文的分析）。再比如美国标准技术研究院（NIST）从 1992 年就开始组织的“文本检索评测”（TREC），到 2008 年已举办 17 届。会议的连续性和规模都能反映该项评测在本领域的影响力。读者可以从 TREC 网站上了解，2004 年的时候 TREC 的规模就达到参赛单位超过 100 个，评测子任务为 7 个，其中 Terabyte 子任务的数据量为 2500 万网页文档（460GB）。参见 http://trec.nist.gov/presentations/t2004.presentations.html。

年开始，举办针对中文分词的专项评测（SIGHAN Chinese Word Segmentation Bakeoff），从2003 年有 12 家单位参加第 1 届 Sighan 中文分词评测（bakeoff-2003）[6]到 2007 年第 4 届评测时有 42 家单位参加。该项评测的影响力逐年扩大，而且对中文分词技术水平的提高起到了极大的推动作用（参见下文 3.1 分析）。

表 2：第 1 —— 4 届 Sighan 中文信息处理评测参赛单位及比赛项目简表

年份	参赛单位数量	测试语料库数量	比赛项目	提交结果数量
2003	12	4	分词	38
2005	23	4	分词	130
2006	36	5	分词、命名实体识别	144（101/43）
2007	42	7	分词、命名实体识别、词性标注	263（166/33/64）

在 2006 年 Sighan 中文分词评测中，北京大学机器感知国家实验室采用最大熵模型，在微软研究院语料库上的封闭测试（MSRA-C，有 22 家单位提交结果）中取得了第一名的成绩[7]（精确率 0.961，召回率 0.964，综合得分，即 F 值为 0.963）[8]。

中科院计算所在美国 NIST 举办的第一届机器翻译评测中就参加了比赛。不过当时的基于规则的机器翻译系统的成绩很不理想。在训练语料不受限[9]的 4 个参赛系统中名列最后。但是，到了 2006 年，计算所的统计机器翻译系统取得了受限语料汉英项目的第 5 名（共 24 个参赛单位）的成绩，BLEU-4 的得分为 0.2913（第 1 名分数为 0.3393）。短短 4 年时间，使得中国机器翻译技术基本上跻身于国际一流水平的行列中。

二是国内的评测活动无论从质量还是从规模上都有了很大提升，并参照国际评测的办法，将评测活动和技术研讨会结合起来，使得通过技术评测推动学术研究更加有效。

国内中文信息处理相关的技术评测从 1991 年开始，主要由“863 计划”智能机主题专家组负责组织实施，直到 2005 年，共举办了 8 届评测。2005 年之后，“863 计划”暂停了对中文信息处理评测的直接支持。中文信息处理的相关评测开始以“民间”形式，由某个领域的骨干单位发起和倡导，并联合多家相关单位来组织实施。这方面成功的例子比如全国统计机器翻译评测及研讨会，从 2005 年开始，到 2008 年连续举办了 4 届。很好地推动了统计机器翻译研究在中国的开展。这类评测及相关研讨会很快得到了中文信息学会的支持，在学会的帮助下，之后陆续形成“中文信息学会系列评测”会议。尤其是对一些正在成为研究热点的自然语言处理子领域，这样的评测对形成更好的研究方向有非常积极的意义。比如美国的 TREC 会议在 2006 年新增了“博客检索评测任务”，具体任务是在博客（约 30GB，320 万篇数据量）中检索带有观点的文章。亚洲语言信息检索评测会议（NTGIR）也在 2006 年开始启动观点提取的评测任务，要求在新闻报道中提取带有主观观点的句子，并指出观点持有者。国际自然语言处理领域的这一前沿动向很快引起国内计算语言学界的重视。2008 年 8

[6] 2003 年第一届 Sighan Bakeoff 的网址：http://www.sighan.org/bakeoff2003/。

[7] 2006 年 Sighan 评测结果参见：http://sighan.cs.uchicago.edu/bakeoff2006/longstats.html。

[8] 精确率和召回率，以及 F 值的定义为：$\text{精确率(P)} = \frac{\text{自动分词结果中切分正确词的数目}}{\text{自动分词结果中词的数目}}$

$\text{召回率}(R) = \frac{\text{自动分词结果中切分正 确词的数目}}{\text{标准答案中词的数目}}$　　P 和 R 两个指标综合后的调和平均值 $F = \frac{2PR}{P+R}$

[9] NIST 机器翻译评测为将评测专注于方法而不是数据， 主要考察在语料受限情况下不同机器翻译系统的性能表现如何（类似于分词评测中主要是看封闭测试的结果而不是开放测试）。但基于规则的翻译系统往往不依赖具体语料。因而通常基于规则的翻译系统参加不受限语料的比赛项目。

月-10 月，中文信息学会就联合中科院自动化所、计算所和复旦大学共同组织了第一届中文倾向性分析评测（The first Chinese Opinion Analysis Evaluation，COAE2008，[文献 31]）。共设定了 6 个子任务。包括（1）中文情感词的识别；（2）中文情感词的褒贬度分析；（3）评价对象的抽取；（4）中文文本的主客观分析；（5）中文文本的褒贬度分析；（6）中文文本中的观点检索。涵盖了自然语言文本情感分析的各个层面的研究课题。其中任务 3 的语料规模为 473 篇文档，约 3000 个句子，语料主要来源于 4 个领域：汽车、笔记本电脑、手机、数码相机等。另外 5 项任务采用同一语料，规模约 4 万篇文档，其中超过十分之一的文档有观点倾向性。语料库来源于影视娱乐、财经、教育、房产、电脑、手机等 6 个领域的网页。国内外共 20 家单位参加了这次评测。评测之后，COAE2008 还在 11 月份就评测结果举办了研讨会。为从事这一前沿领域研究的学者们提供了一个很好的交流平台。除会议交流之外，评测积累的语料库和数据资源可以通过中文信息学会的中文语言资源联盟（Chinese LDC）[10]向科研人员免费或有偿提供使用。这种做法极大地改变了过去科研结果很难流通的状况，很好地推动了一个领域的技术创新和进步发展。

除开展评测活动及举办相关的研讨会之外，计算语言学与中文信息处理领域的定期的全国性学术会议，专题性学术会议，以及有关国际会议在 2004——2008 年也为数不少。限于篇幅这里不展开介绍。请读者看附录了解基本信息。

（三）学术研究的共享资源和交叉学科的人才培养有了更好的基础条件

上面提到的中文语言资源联盟是 2003 年在国家 973 计划的资助和相关课题研究的推动下成立的。该组织致力于语言资源规范和标准的建设以及建立合理有效的管理机制。目前中文语言资源联盟官方网站上已经列出了近 90 项语言资源，涉及分词和词性标注语料库、句法树库、词典（语法信息词典、内涵逻辑语义词典），语音语料库（语音合成、方言库），自动评测语料库、多语对齐语料库，等等。这些数据资源对于推动自然语言处理领域的快速发展无疑有着重大的意义。

从 2004 到 2008 年，由教育部语言文字信息管理司牵头，先后成立了国家语言资源监测与研究中心[11]的六个分中心。随着这六个分中心的启动与工作的展开，语言信息作为一种公共资源的意识将受到越来越多的关注。而这些中心所建设的大型动态流通语料库，无论是在信息处理领域，还是在语言研究与教学领域，也都将产生显著的辐射性影响。

此外，一些高等院校也开始加强自然语言信息处理领域人才的培养。比如 2006 年北大软件与微电子学院成立了语言信息工程系。下设“语言信息处理（NLP）”和“计算机辅助翻译（CAT）”两个专业方向。这些举措，都迎合了当前社会的发展对自然语言处理领域的工程技术型人才的需求日益加大这一明显趋势。

以上列举了 2004——2008 年间能够反映计算语言学与中文信息处理所处发展环境的一些重要事件。除此之外，还有不少大事。比如 2006 年开始设置的“钱伟长中文信息处理科学技术奖”[12] 。两年举办一次评奖活动，对于促进中文信息处理事业的发展无疑有着重大意义[13]。国家标准《信息处理用现代汉语词类标记规范》于 2006 年发布出版，对于中文信息处理中词一级单位的语法语义信息处理以及语料库标注，都有了更好的比较和参考依据。限于篇幅，这里无法一一列举更多的重要事件。但从以上举出的事项也应该不难看到，2004——2008 年是计算语言学与中文信息处理在中国蓬勃发展的五年。

---

[10] 参见中文语言资源联盟网站：http://www.chinip.csdb.cn/。

[11] 参见中国语言资源网：http://www.clr.org.cn/center.jsp。

[12] 详细内容可参见 http://www.cipsc.org.cn/ （中国中文信息学会网站）的报道介绍。

[13] 两届评奖中的近一半多的奖项都颁给了跟少数民族语言文字信息处理相关的研究工作。由此可以看出中文信息处理界对少数民族语言文字信息化工作的重视。

3 聚焦：2004——2008 年中文信息处理技术进步面面观

在一个良好的发展环境下，中文信息处理的技术水平和计算语言学的研究都有了显著进步。本节我们分别从信息处理的技术水平，应用研究的范围两个方面来看具体的进步表现，从而对中文信息处理在 2004——2008 年的发展有一个全貌性的认识。

3.1 中文信息处理技术的性能水平提高

（一）先看中文分词技术的性能情况

由于中文分词从 2003 年开始已经有国际性的技术大赛，使得我们可以方便地根据评测结果来了解该领域技术水平的状况以及技术进步的具体表现。从 2003 年到 2007 年四届 Sighan 中文分词评测都由来自不同单位的训练语料库和测试语料库。其中香港城市大学（City-U）是连续四届都提供语料的单位。下表列出了这四届评测中，在 City-U 语料库上封闭测试[14]F 值的最好成绩[文献 6，7，14，19][15]。

表 3：历届 Sighan 在 City-U 语料上评测结果 F 值最好成绩

	Recall	Precision	F-score	Roov	训练语料词数	测试语料词数
2007	0.9526	0.9493	0.9510	0.7495	1.04M/43K[16]	230K/23K
2006	0.9730	0.9720	0.9720	0.7870	1.6M/76K	220K/23K
2005	0.9410	0.9460	0.9430	0.6980	1.46M/69K	41K/9K
2003	0.9470	0.9340	0.9400	0.6250	240K	35K

上表反映了两个情况：（1）目前中文分词的整体准确率（按词数计算），在 95%左右。在 2004——2008 年间，准确率提升了大约 2 个百分点；（2）中文分词的进步，主要反映在未登录词（out of vocabulary，OOV）识别水平的提高上，未登录词的召回率（Roov）从 62%左右提升到 76%左右，提升了大约 14 个百分点。之所以在未登录识别方面取得重要的进展，是源于研究人员对中文分词的认识视角发生了很大变化，并在技术上把这种认识上的革新进行了充分实现，最终反映在中文分词的效果中[文献 29]）。2003 年的系统采用的是传统的基于词典匹配的最大概率法分词模型[文献 21]），2005 年的系统采用的是条件随机场（CRF）模型[文献 9]，2006 年的系统采用的是字聚类与 CRF 融合的模型[文献 15]，2007 年的系统采用的是无指导切分与 CRF 融合模型[文献 8]。后三个分词系统的共性都在于，把中文文本的分词问题看作是字的组合问题，而不是像[文献 21]所代表的传统做法那样，把分词问题看作是对句子的切分问题。也可以说，习惯上的“中文分词”问题被重新描述为“中文合词”问题[文献 13]。表 4 的示例说明了这种认识上的转变。例句中包含了中文分词中的两大难题：一是歧义问题，“和尚未”这个片段是所谓的交集型分词歧义，因为它既可以理解为“和/ 尚未/”，也可以理解为“和尚/ 未/”。二是未登录词问题，句中“计生办”这个缩略词可能在词典中或训练语料中没有出现过，但在分词处理时，应该分析为一个单位。从字组合为词的角度去看句中词语识别（tokenization）的问题，每个字根据它跟词的不同关系，可以区分为 4 种类型，表 4 中 B 代表一个字位于词首位置，E 代表字位于词尾位置，M 代表字位于词中

[14] 分词评测中的开放测试是指参评的分词系统不受限于主办方提供的训练语料库，可以利用任何知识源进行分词；封闭测试则要求参评系统只能利用赛会提供的训练语料库获取分词知识。

[15] 可以在 http://www.aclweb.org/anthology-new/sighan.html 下载历届 Sighan 研讨会的论文。

[16] 斜线分隔的两个数字中，左边的是语料库中词例（token）的个数，右边是词型（type）的个数。比如 1.04M/43K 指的是训练语料库的规模是 104 万词，其中不同的词为 4300 个（由语料库可抽取词表的规模）。

位置，S 代表该字独立成词。这样，分词问题就重新表述为如何对一个句中所有的字进行序列标注（分类）的问题。而解决这样的问题有不少统计模型可以发挥作用，如最大熵模型（ME），最大熵马尔可夫模型（MEMM），CRF 模型等等。参加历届 Sighan 分词大赛的系统也正是运用这些模型反复进行试验，才不断将中文分词处理的水平推向新的高度。

表 4：中文句子的“自然形式”“分词形式”和“字标注形式”

自然句形式	已结婚的和尚未结婚的都应该到计生办登记
词切分结果	已/ 结婚/ 的/ 和/ 尚未/ 结婚/ 的/ 都/ 应该/ 到/ 计生办/ 登记/
字标注结果	已 结 婚 的 和 尚 未 结 婚 的 都 应 该 到 计 生 办 登 记 S B E S S B E B E S S B E S B M E B E

这种认识上的转变带来的好处是：分词不再需要依赖词典。可以只基于对训练语料的统计进行。因此，在分词过程中，也不再需要专门去处理未登录词（区别于词表中的已登录词）。不管是未登录词还是已登录词，都是由字组合而来的。而字作为基本单元，是中文书面文本中毫无争议的天然基本单位（就分词任务而言，中文信息处理中没有“未登录字”的问题）。识别句中词单位的程序需要统一考虑的，就是各个汉字的组词能力。而每个字在不同环境下的组词特点，可以在很大程度上根据它以往使用时表现的特征（模式）来估计。人们通过改进统计模型，考虑更细微的特征，来更好地模拟一个字在不同环境中的组词特点，从而在面对新语料时，选出概率更大（因而在统计意义上也更合理）的分词（合词）结果。

（二）中文句法结构自动分析技术的性能

对于中文句法结构分析，2004——2008 年间还没有基于大规模语料的公开评测[17]。因而很难有大家一致接受的数据来说明句法分析技术所达到的水平。下面我们提供几个参考数据，可以由此大致了解中文句法结构分析技术的性能情况。

[文献 40]指出，目前在美国宾州中文树库（CTB）上的句法分析准确率为 80%左右。根据[文献 5]的报道，可以把近年来在 CTB 上做的一些句法分析试验结果列表如下[文献 3，4，5，16]，对于长度小于 40 个词的句子，标签精确率（LP）在 81%左右，标签召回率（LR）在 79%左右，F 值接近 80%。

表 5：基于 CTB 的中文短语结构分析性能试验数据

系统	LP	LR	F
Xiong（2005）	80.1%	78.7%	79.4%
Bikel（2004）	81.2%	78.0%	79.6%
Levy & Manning（2003）	78.4%	79.2%	78.8%
Chiang & Bikel（2002）	81.1%	78.8%	79.9%

不过，值得指出的是，一般研究报告在说明句法分析准确率的时候，通常都以短语结构的标签准确率作为度量指标。较少提及整句的分析准确率。为此，我们使用北大计算语言所的一个基于最大熵模型的句法分析器在 CTB1.0 版语料上做了实验。结果如表 6 所示。

[17] 要进行汉语句法结构分析技术的评测，首先要求有得到大家认可的汉语语法体系作为基础，并且以这样的语法系统为指导，对大规模真实语料进行相应的句法结构标注，由此形成的中文树库方可作为评测的客观依据，但目前这个条件显然还不够成熟。学术界目前有关中文句法结构分析比较通行的做法是以美国宾州大学中文树库（CTB）作为一个参照，来试验、比较各种句法分析方法的优劣。

表 6：基于最大熵模型的汉语完全句法分析实验数据[18]

开放测试			**封闭测试**		
句子数量	=	245	句子数量	=	119
短语结构召回率	=	0.7167	短语结构召回率	=	0.9084
短语结构精确率	=	0.7524	短语结构精确率	=	0.9518
整句匹配率	=	0.2653	整句匹配率	=	0.4538
平均结构边界交错率	=	0.2300	平均结构边界交错率	=	0.0036
无边界交错的句子比例	=	0.4776	无边界交错的句子比例	=	0.8487
边界交错数小于 2 的句子比例	=	0.6612	边界交错数小于 2 的句子比例	=	0.9580

可以看到，在分词和词性标注都正确的输入前提下，封闭测试的整句正确率在 45%左右。换言之，基本上有一半的句子无法完全分析正确。在开放测试条件[19]下，句子分析结果完全正确率目前不到 30%（即平均 100 个句子中完全分析正确的不到 30 句），还是比较低的[20]。

除采用短语结构语法进行句法分析外，研究人员还采用依存语法体系[21]对汉语句子进行分析。据[文献 11，40]，目前依存句法分析的封闭测试准确率在 84%左右。

因为完全句法分析的复杂度太高，研究人员不断在尝试进行浅层句法分析。表 7 是有关汉语语块分析（chunking）的实验结果数据[22]。可以在一定程度上反映汉语浅层句法分析的研究状况[23]。最好的分析结果接近 89%。需要说明的是，实验数据都是在分词和词性标注完全正确的基础上得到的。这在一定程度上降低了分析的难度[24]。

表 7：汉语语块分析实验数据[25]

模型	FMM	FMM+规则裁剪	PCFG	HMM1-gram	HMM3-gram
F-score	0.3588	0.6945	0.8144	0.8682	0.8839

（FMM：最大匹配法，PCFG：概率上下文无关文法；HMM3-gram:三元隐马尔可夫模型）

（三） 信息检索技术的性能情况

我们可以从 2005 年 11 月召开的国家“863 计划中文信息处理与智能人机接口技术评测”

---

18 试验所用的语料是宾州大学中文树库 1.0 版。该版本的树库语料含 325 个数据文件，4185 句，平均句长为每句 23.89 个词。

19 这里的开放测试指在测试语料集与训练语料集不同的情况下进行测试；相应的，封闭测试是指在测试语料是训练语料的一个子集的情况下进行测试。

20 下文第 4 节将谈到信息检索系统和机器翻译系统中加入句法分析技术很难带来系统整体性能的提高。除效率方面的原因，句法分析技术的整体性能还不理想是最主要的原因。

21 依存语法跟短语结构语法的关系，有点像汉语学界曾经争论过的“中心词分析法”和“层次分析法”之间的关系。在计算语言学的句法分析中，依存语法不在词和句子中间设置短语范畴，直接描述词与词之间的关系。因其更直观，符合人们对句中词语间关系的直觉，受到不少工程技术人员的青睐。

22 数据来自微软亚洲研究院黄昌宁教授的一份报告（http:// www.china-language.gov.cn/doc/NLP0/09.pps）。

23 语块分析是对句子做线性切割，类似词语切分，只不过切分单位更大了，不像完全句法分析，涉及到层次嵌套的复杂问题，因此一般认为语块分析的难度要低于完全句法分析。对于信息检索和提取等一些应用来说，浅层分析基本能满足应用需求。

24 一般计算分词正确率的时候，都是以词数计的。而对于句法结构分析（或语块分析）来说，分词正确率的计算单位应该是以整句来计更合理。举个简单的例子：一个句子（比如含 20 个词）中就算仅有一处分词错误，对句法结构分析的影响几乎都是致命的。在这种情况下，如果按词数计算分词正确率，则为 19/20，即 95%的正确率，而如果按句子数来计算分词正确率，则为 0！

25 试验所用的语料是 1998 年人民日报 1 月份的内容。该语料可从北大计算语言学研究所网站免费下载 http://icl.pku.edu.cn/icl_groups/corpus/dwldform1.asp。

研讨会公布的结果了解目前国内信息检索技术的性能情况。2005 年信息检索系统的技术评测只设置了 1 个子任务：相关网页检索，有 8 家单位报名参加（其中有 3 家单位未提交最终结果）。相关网页检索使用由北京大学提供的中文 Web 网页测试集，包含 5,712,710 个网页（90GB 数据），是 2004 年 6 月在中国范围内采样 17,683 个站点获得的。评测共有 50 个查询主题（topic）。系统提交查询时可以用人工输入查询，也可以由计算机程序自动产生查询（这种方式可以反映计算机扩展查询或者说理解查询主题的能力）。对每种查询方式，都给出平均准确率（MAP，Mean Average Precision）、相关文档篇数（R）确定后的平均精确率（R-Precision），以及前 10 个结果的平均精确率（P@10）三个指标来说明系统性能。这三个评测指标均是值越大越好。下表是参评系统中成绩突出的两个系统（“manual”代表人工构造查询，“auto”代表自动构造查询）的得分情况。

表 8：2005 年 863 评测信息检索系统部分得分情况

指标 / 系统	MAP		R-Precision		P@10	
	manual	auto	manual	auto	manual	auto
系统 α	0.3538	0.3107	0.4078	0.3672	0.6840	0.6240
系统 β	0.3671	0.2858	0.4140	0.3293	**0.7040**	0.6280

以 P@10 指标为例来说，大致相当于目前信息检索系统返回的前 10 个查询结果中有 70%左右的结果是相关度很高的。应该说，这个结果已经可以满足一般的信息检索需求。

（四）机器翻译技术的性能情况

2005 年 863 机器翻译评测采用人工评测和自动评测结合的方式。在此之后国内举行的历次机器翻译评测，都参照国际做法，完全采用自动评测，不设人工评分。不过因人工评分的结果更易于理解。这里我们仍先以 2005 年 863 机器翻译系统评测的结果来说明机器翻译技术的水平。然后再看 2008 年机器翻译的评测结果所反映出的变化情况。

2005 年 863 机器翻译评测项目设置了 6 个子评测项目，此外还设置了汉英词语对齐评测子任务。6 个子评测项目分别是英汉、汉英、汉日、日汉、日英、英日机器翻译。每个项目又根据语料性质不同分为对话翻译和篇章翻译两个小项目。机器翻译的结果按照人工打分和计算机自动打分两种方式进行。前者的评分标准如下表所示[文献 27]。

表 9：863 机器翻译评测人工打分的标准

评分	忠实度	流利度
0	完全没有译出来	完全不可理解
1	译文只有个别词符合原文	译文晦涩难懂
2	译文有少数内容符合原文	译文很不流畅
3	译文基本表达了原文的意思	译文基本流畅
4	译文表达了原文的绝大部分信息	译文流畅，但是在地道性方面有所不足
5	译文准确完整地表达了原文信息	译文是流畅而且地道的句子

计算机自动评测的指标包括 BLEU 评分、NIST 评分、一般文本匹配度（GTM）、词语位置

相关错误率（mWER）、词语位置无关错误率（mPER）等[26]。其中 NIST 分值、BLEU 分值、GTM 分值都是越高越好，mWER、mPER 的值则是越低越好。评测结果显示自动评测的排序跟人工评测的排序结果有很好的相关性。下表列出了在 2005 年 863 评测涉及汉语和英语的 4 个翻译评测项目中 BLEU 成绩排名第一的系统的具体得分情况[文献 27]。

表 10：2005 年 863 机器翻译评测的部分得分情况

语言	类别	NIST	BLEU	GTM	mWER	mPER	忠实度	流利度
汉英	对话	7.1392	**0.2506**	0.7158	0.6192	0.4843	**65.38**	**64.25**
	篇章	6.9015	**0.1843**	0.7053	0.7228	0.5337	**61.72**	**55.90**
英汉	对话	7.8703	**0.3776**	0.7470	0.5321	0.4156	**82.59**	**78.24**
	篇章	8.7453	**0.3709**	0.7930	0.6162	0.3934	**55.78**	**47.85**

2005 年参加评测机器翻译系统还是以基于规则方法的系统得分更高。到 2008 年的评测，则是统计方法的翻译系统全面胜出。2008 年第四届全国机器翻译评测（CWMT2008）[27]有 15 个单位参加。共提交了 73 个系统/翻译结果（按语言方向分）：英汉 40 个，汉英 33 个。评测设置了 4 项任务，其中一项任务是测试不同系统融合后对翻译质量的影响，除此之外包括 3 项纯粹的机器翻译评测：（1）汉英新闻语料翻译；（2）英汉新闻语料翻译；（3）英汉科技语料翻译。评测提供的训练语料包括新闻和科技领域公共的训练语料，规模为 868,947 汉英句对；以及科技领域独有的训练语料，规模为 620,985 句对。测试集的规模则在 1000 句左右。全部 3 项评测的冠军都是基于统计的机器翻译系统。基于规则的翻译系统在英汉新闻语料中最好成绩是第 2 名，在汉英新闻语料评测中是第 3 名。排名在最后的也主要是基于规则的系统。此外，规则系统在英汉科技语料评测中的成绩较差，在 9 个参赛系统中，3 个规则的系统分列第 6，8，9 名。有关机器翻译技术的性能情况很难靠上述评分有一个直观的印象。本文第 4 节在分析机器翻译技术近年来的研究范式转变时给出了机器翻译的实例，可供读者参考。

上文仅就中文分词、句法结构分析、信息检索、机器翻译等方面的技术发展水平做了介绍。限于篇幅，其他中文信息处理相关技术就不在这里展开说明了。值得一提的是，现在有不少中文信息处理系统都有在线展示的版本。有兴趣的读者可以在网上测试这些系统的性能表现。而且构造测试题本身，也包含了植入语言知识的过程，对语言学研究工作者，也可以算是一个不错的演练舞台。比如我们可以输入下面两个句子给一个语音合成系统：

例 1：这个村有三百多种树。

例 2：这个村有三百多人种树。

两例中只有一字（“人”）之差，但输出的语音流却差别很大。例 1 的“种”应发三声，而且音段上跟前字“多”更近。例 2 的“种”则发四声，跟后面的“树”构成一个音段。

3.2 中文信息处理技术的应用和研究范围拓宽

中文信息处理面对着不同层级的语言单位（对象），各个对象的处理难度并不一样，相

[26] 随着统计机器翻译技术的研究热潮兴起，各种机器翻译自动评测技术也是近年来国际自然语言处理领域研究的热点问题之一。这些评测指标是目前计算机自动评测机器翻译系统质量常见指标，其中 BLEU，NIST 指标都是基于 n-gram 语言模型的（在 2005 年 863 组织的评测中，BLEU 的 n 值取 4，NIST 的 n 值取 5）。NIST 举办的国际机器翻译评测也采用这些指标。关于 BLEU、NIST、GTM、mWER、mPER 的详细说明可参见[文献 1，10，17]。

[27] 参见 http://nlp.ict.ac.cn/demo/cwmt/year_pages/2008.html。

应地，中文信息处理各个子领域的应用和研究的发展状况也不同。通过考察期刊和会议中不同论文的分布情况，可以大致看出近年来研究领域的分布特点，以及相关应用和研究范围变化的情况。我们按照处理对象层级的不同，对国内中文信息领域的主要学术刊物《中文信息学报》（2004——2008 年）的论文分布情况进行了统计，表 11 是统计的结果。

表 11：《中文信息学报》2004——2008 年文章分布情况

年份	篇	句（短语）	词	字	音	图像	合计
2004	23	8	15	6	18	3	73
2005	32	25	10	12	5	4	88
2006	42	21	9	18	13	4	107
2007	42	37	18	8	10	3	118
2008	41	34	17	10	13	1	116
合计	180	125	69	54	59	15	502

从论文分布情况来看，主要研究工作是侧重语篇一级的信息处理。文字和语音方面的比重相对较少。从实际应用的情况来看，恰恰是字处理和语音处理中有一部分属于符号形式层的处理，难度相对较低，因而技术上取得的成果已经在实际应用逐渐普及。当然，这并不是说在这个层次上的问题都已经解决，不需要进一步发展了。实际上，汉语言文字符号的数字化仍有许多工作要做，也还有不少难关需要攻克。其中比较突出的问题来自两个方面：第一，在人们一般日常的文字信息处理已经完全数字化之外，目前还有相当多的“特殊”的文字内容有待数字化[文献 33]。比如中国浩如烟海的古籍内容在信息时代需要全面实现数字化，就涉及到大规模中文字库的研制[28]，涉及到汉字 OCR（光学字符识别）技术的改进；再如对大量手写内容和历史上的科技文献内容的数字化[文献 25]，以及视频图像中所包含文字信息的数字化，就会涉及到对复杂版面内容（包括图文、公式、表格等）以及图像信号的分析处理。这些都是在符号的形式层进行信息处理需要解决的问题。第二，随着信息产品的日益丰富和普及，越来越多的嵌入式设备和便携移动式信息设备（比如手机，固定电话的显示模块等）走进人们的生活，如何在这些微型设备中实现文字内容的数字化（即汉字的存储、传输等），也是科研人员面临的新挑战[文献 48]。显然，上述这两个方面的问题，要求人们从一“大”一“小”两个方向来拓宽，寻求更好地进行汉字符号形式层的处理。而且随着研究的深入，许多符号形式层的处理问题，需要在符号意义层取得进展后反作用于形式层的处理，比如汉字 OCR 汉字识别或者音字转换，要达到非常高的质量，就要求在后处理阶段，对识别出来的文字序列进行内容理解，从各种可能性中筛选出有意义的正确序列，排除无意义的错误序列，才可能得到更好的效果。

语用篇章级信息处理成为热点研究课题，主要表现在情感计算和隐喻识别两个方面。

当前中文文本的情感倾向性分析通常按照处理对象的不同，分为四个层次的研究，包括：（1）词语情感倾向性分析；（2）句子情感倾向分析；（3）篇章情感倾向性分析；（4）超大文本整体倾向性预测。其中每个层次的研究又可区分出不同的具体任务以及相应的不同的方法，比如词语情感倾向性分析就包括 3 个具体的任务：a. 情感词的识别；b. 情感词褒贬义的区分；c.情感词倾向义程度的度量。以情感词褒贬义的区分为例，中文中有的词，其词义本身从不同角度理解褒贬义不同，如“骄傲”，表示“自豪”义时为褒义，表示“自满”义

[28] 对此不难从汉字字符集的发展看出。比如作为国家标准的汉字字符集，从最早的 GB2312 只对常用（一、二级）的 6764 个汉字进行了编码，到后来的 GBK，GB18030，先后增加到 20902，27533 字。而一些 IT 企业研制的字库数量更是庞大，比如微软 Office XP，方正公司的宋体超大字符集字数都在 6 万以上。

时为贬义。再比如“提升”，跟“成本”结合在一起时，整体表示贬义（或不期望发生），跟“产量“结合在一起时，整体表示褒义（或期望发生）。不难看出，上述任务的难度有很大差异。前文已经提及的 COAE2008 的评测结果显示，在纯粹的情感词识别中，得分高的情感词准确率较高，P@100 能超过 91%。但整体准确率只有 50%左右。而计算机在判断情感词的褒贬时，正确率就大大降低了，P@100 的得分低于 50%。由此可见，对于文本中的主观信息的提取，目前计算机的能力还是相当有限的[文献 31]。从处理方法的角度，文本情感倾向性分析目前采用的方法可以分为两大类，一类是有指导（supervised）方法，一类是无指导（unsupervised）的方法。前者又可以根据知识源的不同分为基于情感倾向词典的方法和基于人工标注情感倾向性语料库的方法。词典方法一般依赖已有的大规模机读词典（比如 WordNet 和 HowNet 等），建立带有情感倾向标注的词语库，然后对现实文本中的词语的情感倾向性进行匹配和推断。语料库方法主要是利用词语的共现关系，利用机器学习去发现词语的搭配模式，从而识别主观性文本中的带有情感倾向性的词语及其搭配关系。无指导的方法先假设一些少量的种子词（比如“美、丑”分别作为褒义和贬义的种子词），然后到未经标注的语料库中，计算其他候选词跟种子词的相似度，通过“滚雪球”的方式获取更多的带有情感倾向性的词语。这种方法对初始的种子词集的依赖性很高。

计算机对文本中隐喻进行自动处理分为隐喻识别、隐喻理解和隐喻生成三个子任务[文献 32]。其中隐喻识别目前的主要方法有基于文本线索的方法、基于语义知识的方法和基于机器学习的方法这三类。基于文本线索的方法主要是通过文本中有可能提示存在隐喻表达的特征形式来激活隐喻识别过程。这些特征形式主要有领域信号或话题标志，元语言信号（比如“比方说”），比喻特征词（比如“像、仿佛”等），引号、破折号等标点符号。研究发现，90 %以上的隐喻是没有语言标记的。但是在有隐喻标记的表达中，隐喻则占了相当的比例，书面语文本中约占 50%。基于语义知识的方法主要是根据语义知识库（比如基于语言学中的论元结构理论建立的动名搭配限制关系库，名词的上下位概念关系库等）。基于机器学习的方法把隐喻的识别看作是一个纯粹的二值分类问题，即一个表达式在特定环境中是字面义用法还是隐喻义方法。研究人员采用常用的分类模型如最大熵模型、朴素贝叶斯模型、SVM 模型等，对诸如动词的隐喻、名词短语的隐喻，以及句子属于隐喻句还是非隐喻句等，进行了广泛的试验。在受限范围，取得了不错的结果。

此外，近年来国内学者在话题识别与跟踪[文献 28]，篇章结构分析[文献 53]也做了不少研究，拓展了中文信息处理的范围。限于篇幅。这里就不详细说明了。

4 透视：计算语言学近年来的发展特点及其原因分析

从中文信息处理的技术路线角度看，2004 到 2008 年多个领域都出现了一个共同的鲜明特征，就是原先的规则方法不再占据主要地位，统计方法开始在各方面崭露头角，用统计方法构建的系统在国际和国内的评测中表现出更为突出的性能。先看中文分词领域的情况。[文献 29]直率地指出，在中文分词领域，“如果说，像 Wu 这样的基于手工规则的自动分词系统还能在 2003 年 Bakeoff 的多项评测中名列前茅，那么，到了 2005 年和 2006 年的 Bakeoff 上，已经很难找到它们的身影了。取而代之的是基于词，尤其是基于字，的统计学习方法。”再看机器翻译技术，2005 年 863 组织的机器翻译评测，提交结果的 21 个系统中，19 个是基于规则或规则与实例相结合的系统，1 个纯基于实例的系统，只有 1 个是基于统计方法的系统，而当时国际上的机器翻译评测则以基于统计方法的系统居多，占到 60%以上。到 2008 年的 CWMT2008 评测时，这种情况发生了根本性的变化。在参加“汉英新闻翻译”项目的 12 个单位中，有 9 个是基于统计的系统，3 个基于规则的系统；在参加“英汉新闻翻译”项目的 11 个单位中，有 6 个是基于统计的系统，5 个基于规则的系统，在参加“英汉科技翻

译”项目的 9 个单位中，6 个是基于统计的系统，3 个基于规则的系统。而且在三项比赛中，都是基于统计的翻译系统得分最高。上述数字已经鲜明地反映出，经过 4 年左右的发展，国内基于统计的机器翻译系统取得了长足的进步。

除系统评测反映的情况外，相关学术著作的出版情况也能折射出上述技术路线的巨大变化。[文献 22]是 2004 年出版的，基本是关于规则方法的机器翻译技术的介绍，而到 2008 年出版的[文献 35]，则完全以统计方法为主来架构机器翻译系统。此外，2005 年国内翻译出版了三本国外自然语言处理教科书的中文版[文献 24，34，54]，从这些教科书的内容也不难看出。统计方法在当前国际计算语言学领域的主流地位。

在我们看来，统计方法和规则方法的核心区别是基于在哲学层面对语言的本质的不同认识，从而在技术层面形成了对语言知识的表示方式不同，获取语言知识的方法不同，当然相应的应用语言知识的算法模型也就不同。下面不妨概要地对比一下规则方法的机器翻译系统和统计方法机器翻译系统的宏观架构：

图 3：基于“要素合成原理”的机器翻译系统结构示意图[文献 50]

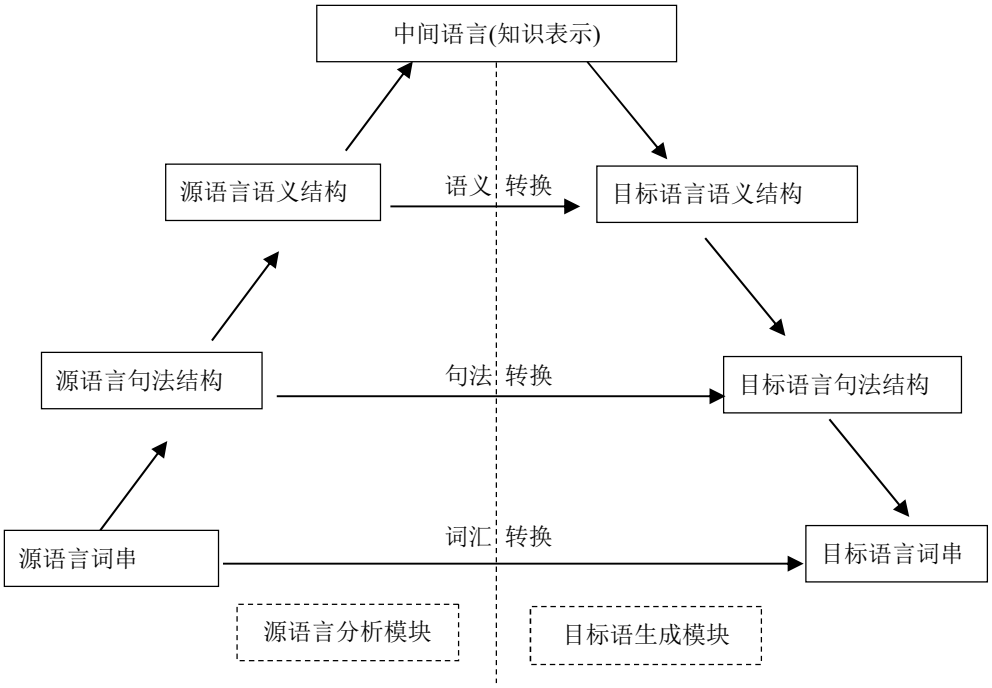
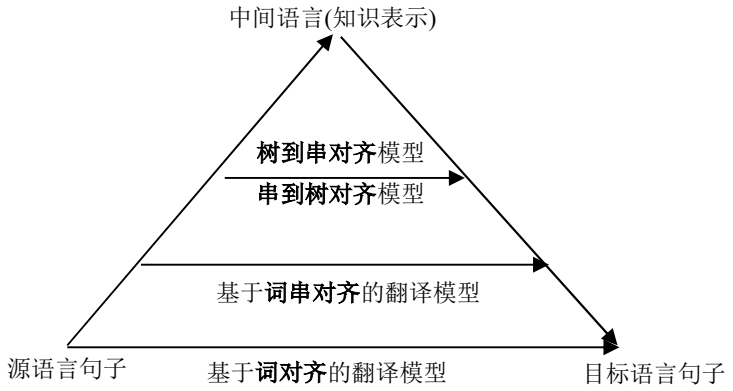


图 4：基于统计方法的机器翻译系统结构示意图[文献 35，37]



限于篇幅，本文无法从技术细节上去说明规则方法机器翻译跟统计方法机器翻译的差别。不过，从上面图示可以看到，规则方法是在用“先分解+后合成”的方式，模拟人的翻译过程，整个流程经历了源语言结构、语义分析和目标语言结构、词形生成，等等（因此，这些语言知识都必须由人总结出来再配置到机器翻译系统中）。而统计方法则是把具有对译关系的大小不等的各种语言片段（词和词串）统统“记忆”下来，在翻译的时候，通过搜索概率最大的对齐片段来实现翻译（这些知识表示为大量的统计参数，由机器扫描语料库后自动获取记录在系统中）。规则方法的机器翻译流程比较好理解。读者可以参看文献[22]了解其细节。不熟悉统计方法机器翻译技术的读者可以通过下面这个笑话来体会其核心思想（关于统计机器翻译技术的更为准确和全面的说明，请参考[文献 35]）。

在土楼长大的阿贵去英国留学，娶了个洋媳妇，带回土楼看望奶奶。
早上阿贵的媳妇碰到奶奶，就跟奶奶打招呼：“Good morning!”
奶奶听不懂英文，觉得很奇怪，就回答她：“我是阿贵阿妈!”
第二天早上阿贵的媳妇又碰到奶奶，仍然说：“Good morning!”
阿贵的奶奶虽然觉得很奇怪，不过仍然回答她：“我是阿贵阿妈!”。
晚上奶奶实在忍不住了，就问阿贵，他媳妇说的 Good morning 是啥意思。
阿贵回答：“那是她用英语向你问好，就是‘早安’的意思啦!”
到了第三天早上，阿贵的奶奶碰到洋媳妇，就赶紧说：“Good morning!”。
心里还美滋滋的。以为洋媳妇会赞美她的英语。
不料，洋媳也马上回答：“我是阿贵阿妈!”
奶奶当场愣住了……

从机器翻译的角度看上面这个笑话，就是“Good morning”跟“我是阿贵阿妈”这两个词串在语料库中反复出现，其对齐模式被机器发现，并认为二者具有统计意义上“非常般配”的翻译关系。

为了进一步从直观上体会统计机器翻译方法的效果，我们利用网上的三个在线机器翻译系统[29]，进行了英汉和汉英翻译的小测试，其中下面的翻译结果可以说明一些问题。

表 12：英——汉机器翻译示例

原文	U.S. President Barack Obama is angrily blasting oil industry officials who he says are failing to accept blame for the massive oil spill into the Gulf of Mexico. **The president** says all parties need to take responsibility, including **the government**.
MT1	U.S. 贝拉克・奥巴马总统愤怒抨击他说不承担巨型的漏油的责任入墨西哥湾的石油工业官员。这位总统宣称所有党需要承担责任，包括这个政府。
MT2	美国总统奥巴马愤怒爆破石油行业官员说，他是谁不接受为使墨西哥湾的大量石油泄漏责任。**布什总统**说，所有各方都必须承担责任，包括**政府**。
MT3	愤怒炸开石油工业官员美国总统 Barack Obama 存在,其他说正未能接受进入海湾墨西哥的很大漏油的责任的.**总统**说所有各方需要采取包含**政府**责任.

[29] MT1 是国外的规则机器翻译系统；MT2 是国外的统计机器翻译系统；MT3 是国内的规则机器翻译系统。机器翻译系统的效果涉及的方面很多，有时候通过一些细节可以看到一个系统的底蕴。比如 MT3 系统英译汉的译文，对标点符号就没有做处理，直接用了西文的标点符号。再比如从对冠词“The”的处理不同，也可以体会这些机器翻译系统之间的差别。

在表 12 的英——汉机器翻译示例中，“The president”在统计翻译系统 MT2 中竟然被翻译成了“布什总统”。从规则系统的角度讲，这是匪夷所思的错误。但是从统计机器翻译的角度，这是很可以理解的。可以想见，MT2 的训练语料库中有大量的“The president”和“布什总统”的对齐关系存在。再过一段时间，如果训练语料库随着时代的发展更新升级的话，“The president”就会更多地对应为“奥巴马总统”了。

表 13：汉——英机器翻译示例

原文	伊拉克政府声明称，伊拉克政府坚持要求美国军队按照美伊驻军地位协议，在今年 6 月 30 日前从伊拉克城镇全部撤出，这一期限“不可延长”。
MT1	The Iraqi government stated said that Iraqi government insisted requested the American Army according to the US-Iraqi garrison status agreement, withdrew before June 30 from the Iraqi cities completely, this deadline “cannot postpone”.
MT2	Iraqi government statement said the Iraqi government insisted the United States military forces under the US-Iraq Status of Forces Agreement, in this June 30 to withdraw from all Iraqi cities and towns, the term "can not be extended."
MT3	**He|she's** station troops position agreement the Iraq government is declared saying the Iraq government persists in demanding USA troops according to US,comply with Iraq city and town before this June 30 all withdraw, this one time limit "is not allowed to prolong ".

在表 13 的汉——英机器翻译示例中，MT3 把汉语“美伊驻军协议”中的“伊”理解为一个文言性质的人称代词来进行翻译，但又无法确定其性别，因而译文中出现了非常滑稽的“He|she”带上所有格标记“’s”的形式。这是规则机器翻译系统更容易出现的问题。

尽管上面给出的不同的机器译文各有缺点，但总体来说，基于统计方法开发机器翻译系统后来居上，超越基于规则方法的机器翻译系统，已是不争的事实。这不得不引起研究人员的深刻反思。尤其是惯于走规则路线的语言学背景的学者。更应面对这种情况做深入的思考。我们认为，计算语言学和中文信息处理领域近年来之所以出现统计方法占主流的形式，是有其深刻原因的。

（1）社会已经全面进入互联网时代。这个时代的特点是信息量大，信息传播速度快。自然语言的活跃程度远远高于以往任何一个时期[30]。这就意味着语言字符本身的不确定性在增强。而这种情况对基于理性主义的规则方法，构成了严重的挑战。但对于统计方法来说，发现不确定性对象背后的规律，正是它的强项。

（2）互联网的发展为统计方法准备了海量的数据。为统计方法大展拳脚提供了弹药。

（3）计算机的能力主要表现在“记忆”和“搜索”，而不是创新/演绎推理。统计方法在机器翻译以及中文分词等技术上的成绩，可以理解为计算机依靠其强大记忆能力，在海量数据和恰当的统计模型两驾马车的辅佐下取得的成功。完全人工的规则在语言知识的概括度和层级的系统性等方面可以表现出简洁的美感，但在工程应用层面，却缺乏对真实语料的有效覆盖，缺乏对具体而微的词语共现信息的准确刻画。人工规则更多的是在“类”的层面描述语言对象的性质，而统计方法（机器学习）则是在“例”的层面描述语言对象的分布、搭配、对齐等方面的性质。

正是上述因素的共同作用，将统计方法推到了自然语言处理历史舞台的中心位置。

[30] 近年来由网络而逐渐扩散到普通社会生活用语中的语言现象有明显增速加快的趋势。比如“山寨××”“被××”等等，涉及到语言中的字、词、句、篇各个层次。

除统计方法渐成主流研究范式外，计算语言学和中文信息处理领域近年来的另一个发展特点是研究人员更加关注各项技术的融合。无论是在方法研究中，还是在应用系统中，融合型的技术路线受到越来越多的重视。限于篇幅，下面仅通过一些有代表性的例子来概要说明这种情况。

在技术研究方面，[文献 46]全面地探讨了自然语言处理技术如何更好地融合到信息检索系统中，来提高检索效率和准确率。近年来，研究人员已经尝试将自然语言处理中的多项核心技术，包括分词、词性标注、短语识别、命名实体识别、概念抽取、指代消解、词义消歧等等，加入到信息检索系统中，大部分实验结果显示，将现有的自然语言处理技术融合到信息检索中，在一些特定应用中确实可以提高检索精度，但一些复杂的自然语言处理技术（比如词义消歧，指代消解），因为本身的精度不高，而且大大增加了处理和存储消耗，对检索结果基本帮助不大，甚至会产生负面影响。此外，由于中文的特殊性，一段时间以来，研究人员对中文文本的信息检索应该建立在字索引基础上还是建立在词索引基础上一直存在争论，这涉及到是否需要在中文文本的信息检索中引入词切分处理的问题。经过近年来对技术融合的研究，学术界基本已经取得一致的看法，对于中文文本的信息检索，将字索引和词索引结合起来能取得更好的检索效果，前者有利于获得比较高的召回率，后者则有利于提高检索的准确率。而且由于中文分词技术水平和性能的提高，在信息检索系统中加入中文词切分处理，并不会导致系统性能的下降。

在应用系统方面， [文献 60] 开发的人机结合模式的机器翻译平台，采取了将知识管理与机器翻译技术融合为一体的技术路线，在大规模科技资料翻译工程中取得了良好的应用效果。由 500 名用户协同使用该系统，对 2 亿字规模的专利文献进行翻译，在按照国家翻译质量标准认定的错误率不超过 1.5‰的前提下，平均翻译效率提高 2 到 4 倍。在人工抽样评测时，机构名称的汉英翻译正确率达到 96%，地址的翻译正确率达到 89%。这是规则和统计方法融合在特定领域的翻译中取得成功的一个实例。这项研究的成果，也获得了 2008 年 12 月“钱伟长中文信息处理科学技术奖一等奖”。

此外，语音处理和文本内容理解的融合，也是研究人员一直都非常关注的问题。特别是近年来自然语言处理越来越多地追求“面向内容理解的处理”（而非单纯的形式处理），口语机器翻译，口语信息检索，语音合成等应用，都需要将文本分析技术跟语音处理技术融合为一体，才能取得更好的效果。

我们认为，技术融合在近年来成为一个鲜明特点的原因有二。一方面是因为计算机技术和应用的发展推动的。从硬件角度看，当前电脑和手机等通信设备，计算机和传统家电的界限在日趋模糊。电子设备拥有的计算能力越来越强，由此导致人们的社会生活中涉及到的计算设备类型日趋多样化，从传统的基于个人计算机的自然语言处理技术，很自然地将向各种不同类型的设备和应用上过渡，在这一进程中，语音和文本，固定应用和移动应用，单机环境和网络环境的无缝连接，都将成为技术融合的推手。另一方面，从计算语言学和中文信息处理这一领域的研究发展状况看，近年来虽然统计方法和机器学习技术有了很大发展，使得 NLP 技术的性能离实用目标越来越近，但统计方法本身也已进入到一个平台期，发展速度出现趋缓的迹象。在还没有出现新的具有创新性的学术研究范型的情况下，研究人员通常会在应用层面，将已经比较成熟的方法（工具）用在不同的应用领域，也就是在基础研究相对停滞的情况下，转而追求技术层面的应用创新。

5 结语

本文采用的是由外观表象到内在理路的考察路线。纵观 2004——2008 年计算语言学和

中文信息处理的发展轨迹，并将这一轨迹的背景向后再拉伸 15 到 20 年，不难看到，从上世纪 90 年代以来，统计方法在国际计算语言学领域风生水起，但对国内计算语言学的影响甚微。直到 21 世纪初，统计方法在机器翻译这一计算语言学的制高点上开出绚烂奇葩，引得无数工程技术人员在自然语言处理的各个领域争相效仿。这一轮的强力波很快影响到国内计算语言学界。近年来国内计算语言学和中文信息处理中取得的主要进展，绝大多数都建立在大规模语料数据集以及各类统计机器学习方法的基础上。基于规则方法自然语言处理研究在这段时间相比而言，发展缓慢。不过，在这段时间，仍有一些语言学工作者，从各自的研究兴趣出发，面向计算机信息处理的需要，对汉语汉字信息处理各个层次的问题，从语言和认知计算的角度，进行了探索，比如[文献 49]对现代汉字的特征进行了全面的统计和计算分析。[文献 47]探讨了汉语的句块的韵律结构和句法语用的关联。[文献 55，56，57]对论元结构理论的扩展以及多义词义项判定的讨论。[文献 52]从认知语言学的角度审视汉语的计算问题。这些都是汉语学者尝试将汉语本体研究跟计算研究结合起来的努力。另外，值得注意的是，也有计算机科学背景的研究人员对当前汉语本体研究的结果和统计方法存在的问题持存疑态度，呼吁语言学家跟计算机工程技术人员一道努力，研究出新的能够真正服务于中文信息处理的“实用型”的汉语语言学体系。[文献 45，59]就是这方面研究工作的一个体现。

概括和分析计算语言学与中文信息处理近年来的发展特点，除了预示未来技术发展方向的意义外，对于语言学研究者，更大的意义可能是引发人们对语言性质更深层面的探问。简单回顾机器翻译从上个世纪中叶到现在 50 多年的发展历史，可以看到一个非常有趣的认识上的“螺旋式上升”。1950 年代，伴随着计算机的发明，人们提出了机器翻译的理想。当时机器翻译被看作是编码（加密）——解码（解密）的通信过程（信源——信道模型）。很快，这种简单的模型实现的词对词的翻译遭到了产业界和学术界的各方质疑。机器翻译的发展进入停滞，但却由此催生了一门学科——计算语言学。人们的想法很简单，先在理论层面把人类语言的计算性质研究清楚，再来从工程上实现机器翻译系统。在这之后的几十年里，机器翻译都被假定为“要素分解与合成模型”，即翻译要建立在对语言结构的理解基础上。这一信条在 90 年代初由 IBM 公司的研究人员打破[文献 35，36]，机器翻译重新被看作编码和解码的过程。但在当时的数据条件下，好的统计模型也缺少足够的样本来帮助计算机获取到“翻译经验”。直到 21 世纪，随着互联网上大规模双语语料数据的获取变得非常容易，同时统计模型本身也得到了大大改进。计算能力跟上世纪五十年代相比已经有了翻天覆地的变化。这些因素累积到一起，人们发现，机器翻译也可以不建立在对语言结构的理解基础上。通过从超大规模双语对齐语料获取的经验知识，利用快速的双语片段检索和重组技术，也可以构造出机器翻译系统，而且表现出的效果还好于传统基于规则的，建立在对语言结构进行分析基础上的翻译系统。这种变化的深刻性或许可以用一个简单的问题来类比：人学习外语到底是靠“死记硬背”，还是靠“规则生成”？这个问题现在当然还没有明确的答案，但或许在用计算机去模拟人类语言行为的研究和实践过程中，人们离揭开谜底会越来越近的。

作为一个年轻的交叉学科，人们对“机器翻译”“计算语言学”的认识，在近半个世纪里已经经历了不少变化[31]。而语言学作为一个相对传统一些的学科，人们对它的认识，也不是一成不变的。不仅如此，那些大的认识上的变化，更是意味着不同时期学科范式的革命。[文献 41]曾把现代语言学的发展概括为以下 4 个阶段：（1）传统语法——看作法律的语言学；（2）历史比较语言学——看作生物学的语言学；（3）结构主义语言学——看作化学的语言学；（4）转换生成语言学——看作数学的语言学。这种概括给我们很多启发。转换生成语法诞生的背景之一，就是计算机对自然语言处理能力的限制，促发人们去思考语言的深层本质为何。当乔姆斯基以公理化方法建立起语言的生成转换系统的时候，背后的哲学基础是理性主义。在这个背景下，语言被看作是确定性的数学对象。走规则路线的计算语言学研究，无

[31] 参看[文献 52]中冯志伟先生的序。

论其技术细节如何，都可以在这种公理化背景中找到其形式模型的根源。而在互联网时代，以海量数据为处理对象，以基于统计方法的机器学习技术为获取语言知识手段的新研究模式，则把语言看作非确定性的数学对象。概率论和信息论成了描述人类语言经验的有力工具。

通过归纳 2004——2008 年计算语言学和中文信息处理领域的发展特点并分析其成因，我们认识到，自然语言作为人类社会的本质属性之一，它跟社会本身发展的状况息息相关。互联网和计算技术的变革在新世纪已经深刻影响了人类社会，也必然影响到语言学相关学科的发展以及研究方式的更新。无论是对探究人类语言本质感兴趣的人文学者，还是对研发语言信息处理技术更感兴趣的工程人员，在思考未来的探索方向的时候，都不应忽视这一宏大背景。

参考文献：

[1] Chin-Yew Lin & Franz Josef Och , 2004, Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics, In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, Barcelona, Spain, July 21- 26, 2004.

[2] Chu-Ren Huang，Petr Simon，Shu-Kai Hsieh，Laurent Prévot，2007，Rethinking Chinese Word Segmentation: Tokenization, Character Classification, or Wordbreak Identification，In Proceedings of the 45th Annual Meeting of the ACL, 2007.

[3] Daniel M. Bikel. 2004. On the Parameter Space of Generative Lexicalized Statistical Parsing Models. PhD thesis, University of Pennsylvania.

[4] David Chiang & Daniel M. Bikel. 2002. Recovering Latent Information in Treebanks. In Proceedings of COLING 2002.

[5] Deyi Xiong, Shuanglong Li, Qun Liu, Shouxun Lin, and Yueliang Qian，2005，Parsing the Penn Chinese Treebank with Semantic Knowledge, In Robert Dale, Kam-Fai Wong, Jian Su, and Oi Yee Kwong (Eds.): *Natural Language Processing – Proceedings of IJCNLP 2005*, Spinger, pp.70-81.

[6] Gina-Anne Levow，2006，The Third International Chinese Language Processing Bakeoff: Word Segmentation and Named Entity Recognition，Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, pages 108–117, Sydney, Australia, July 2006.

[7] Guangjin Jin & Xiao Chen，2008，The Fourth International Chinese Language Processing Bakeoff: Chinese Word Segmentation Named Entity Recognition and Chinese POS Tagging，Proceedings of the Sixth SIGHAN Workshop for Chinese Language, pages: 69-81，Hyderabad, India, January, 2008.

[8] Hai Zhao & Chunyu Kit, 2008, Unsupervised Segmentation Helps Supervised Learning of Character Tagging forWord Segmentation and Named Entity Recognition, In Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing, 2008.

[9] Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky, and Christopher Manning, 2005, A Conditional Random Field Word Segmenter for Sighan Bakeoff 2005, In Proceedings of The Fourth SIGHAN Workshop on Chinese Language Processing, 2005.

[10] Joseph P. Turian, Luke Shen, & I. Dan Melamed, 2003, Evaluation of Machine Translation and its Evaluation. In Proceedings of MT Summit IX, New Orleans, LA, USA,2003.

[11] Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu，2001，Bleu: a Method for Automatic Evaluation of Machine Translation，In Proceedings of the 40th Annual Meeting on

Association for Computational Linguistics.
[12] Meixun Jin, Mi-Young Kim & Jong-Hyeok Lee, 2005, Two-Phase Shift-Reduce Deterministic Dependency Parser of Chinese, Proceedings of International Joint Conference on Natural Language Processing (IJCNLP).
[13] Nianwen Xue & Libin Shen，2003，Chinese Word Segmentation as LMR Tagging, In Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing, in conjunction with ACL'03. Sapporo, Japan.
[14] Richard Sproat & Thomas Emerson, 2003, The First International Chinese Word Segmentation Bakeoff，In Proceedings of the second SIGHAN workshop on Chinese language processing, Sapporo, Japan, July 2003.
[15] Richard Tzong-Han Tsai, Hsieh-Chuan Hung, Cheng-Lung Sung, Hong-Jie Dai, and Wen-Lian Hsu, On Closed Task of Chinese Word Segmentation: An Improved CRF Model Coupled with Character Clustering and Automatically Generated Template Matching, In Proceedings of SIGHAN Workshop, 2006.
[16] Roger Levy & Christopher Manning. 2003. Is it harder to parse Chinese, or the Chinese Treebank? In Proceedings of the 41st Annual Meeting on ACL, 2003.
[17] Roukos Papineni & Ward Zhu, 2001, Bleu: a Method for Automatic Evaluation of Machine Translation, (IBM Technical Report, Keyword. RC22176- W0109-022).
[18] Sheng Li, Tiejun Zhao，2006，Chinese Information Processing and Its Prospects，In Journal of Computer Science and Technology，2006 年第 5 期。
[19] Thomas Emerson，2005，The Second International Chinese Word Segmentation Bakeoff，Proceedings of the fouth SIGHAN workshop on Chinese language processing，pages: 123-132，Jeju Island, Korea, Oct. 2005.
[20] Tiejun Zhao, Guan Yi, Ting Liu, Qiang Wang，2007，Recent advances on NLP research in Harbin Institute of Technology，In Frontiers of Computer Science in China, 2007 年第 4 期。
[21] Wei-Yun Ma & Keh-Jiann Chen, 2003, Introduction to CKIP Chinese word segmentation system for the first international Chinese Word Segmentation Bakeoff, In Proceedings of the second SIGHAN workshop on Chinese language processing, 2003.
[22] 冯志伟，2004，《机器翻译研究》，中国对外翻译出版公司。
[23] 冯志伟，2007，自然语言处理中理性主义和经验主义的利弊得失，《长江学术》2007 年第 2 期。
[24] 冯志伟、孙乐译，2005，《自然语言处理综论》，电子工业出版社（原著 D.Jurafsky & J.H.Martin，2000）
[25] 顾绍通、马小虎、杨亦鸣，2008，基于字形拓扑结构的甲骨文输入编码研究，《中文信息学报》2008 年第 4 期。
[26] 洪宇、张宇、刘挺、李生，2007，话题检测与跟踪的评测及研究综述，《中文信息学报》2007 年第 6 期。
[27] 侯宏旭、刘群、张玉洁、井佐原均，2005 年度 863 机器翻译评测方法研究与实施，中文信息学报，2006 年 3 月第 20 卷增刊。
[28] 黄昌宁，2002，统计语言模型能做什么，《语言文字应用》2002 年第 1 期。
[29] 黄昌宁、赵海，2007，中文分词十年回顾，《中文信息学报》2007 年第 3 期。
[30] 黄瑾、刘洋、刘群，2007，机器翻译评测介绍，《信息技术快报》（计算所所刊，中国计算机学会会员赠送刊物），2007 年第 7 期。
[31] 黄萱菁、赵军，2008，中文文本情感倾向性分析，《中国计算机学会通讯》2008 年第 2

期。
[32] 贾玉祥、俞士汶、朱学锋，2009，隐喻自动处理研究进展，《中文信息学报》2009 年第 6 期。
[33] 李宇明，2003，搭建中华字符集大平台，《中文信息学报》2003 年第 2 期。
[34] 刘群 等译，2005，《自然语言理解》，电子工业出版社（原著 James Allen，1995）
[35] 刘群，2008，《汉英机器翻译若干关键技术研究》，清华大学出版社。
[36] 刘群，常宝宝，王厚峰，2009，机器翻译研究的发展与现状，中国计算机学会主编《2008 中国计算机科学技术发展报告》，机械工业出版社，2009 年。
[37] 刘群，2009，机器翻译研究新进展，《当代语言学》2009 年第 2 期。
[38] 刘群、钱跃良，2008，中文信息处理技术评测综述，《中国计算机学会通讯》2008 年第 2 期。
[39] 刘挺，2008，中文信息处理——“奇葩绽放”，《中国计算机学会通讯》2008 年第 2 期“中文信息处理”专辑前言。
[40] 刘挺、马金山，2009，汉语自动句法分析的理论与方法，《当代语言学》2009 年第 2 期。
[41] 钱峰，1990，《计算语言学引论》，学林出版社 1990 年版。
[42] 钱跃良、林守勋、刘群、刘宏，2005，2005 年度 863 计划中文信息处理与智能人机接口技术评测回顾，《中文信息学报》，2006 年 3 月第 20 卷增刊。
[43] 钱跃良、刘群、林守勋，2005，自然语言处理与人机交互技术评测综述，《信息技术快报》（计算所所刊，中国计算机学会会员赠送刊物），2005 年第 8 期。
[44] 宋柔，2003，统计和规范中的误区，载孙茂松等编《中文信息处理的若干重要问题》，科学出版社，2003 年。
[45] 宋柔，2008，现代汉语跨标点句句法关系的性质研究，《世界汉语教学》2008 年第 2 期。
[46] 王灿辉、张敏、马少平，2007，自然语言处理在信息检索中的应用综述，《中文信息学报》2007 年第 2 期。
[47] 王洪君，2005，普通话节律与句法语用关联之再探，《第八届全国人机语音通讯学术会议论文集》。
[48] 吴晓春、吴娴、李培峰、朱巧明，2008，一个手机整句输入算法的研究与实现，《中文信息学报》2008 年第 5 期。
[49] 邢红兵，2007，《现代汉字特征分析与计算研究》，商务印书馆 2007 年。
[50] 俞士汶，1999，自然语言处理与语法研究，载马庆株编《语法研究入门》，商务印书馆 1999 年版。
[51] 俞士汶、段慧明、朱学锋、孙斌，2002，北京大学现代汉语语料库基本加工规范，《中文信息学报》2002 年第 5 期。
[52] 袁毓林，2008，《基于认知的汉语计算语言学研究》，北京大学出版社 2008 年。
[53] 乐明，2008，汉语篇章修辞结构的标注研究，《中文信息学报》2008 年第 4 期。
[54] 苑春法 等译，2005，《统计自然语言处理基础》，电子工业出版社（原著 C.D.Manning & H.Schutze，1999）
[55] 詹卫东，2004，范围副词“都”的语义指向分析，《汉语学报》2004 年第 1 期。
[56] 詹卫东，2004，广义配价模式与汉语“把”字句的句法语义规则，《语言学论丛》第 29 辑。
[57] 詹卫东，2004，论元结构与句式变换，《中国语文》2004 年第 3 期。
[58] 张华平，2003，中文信息处理技术发展简史，http://www.nlp.org.cn（中文信息处理开放平台网站）
[59] 张瑞朋、宋柔，2007，否定词跨标点句管辖的判断，《中文信息学报》2007 年第 5 期。

[60] 张桂平、蔡东风，2008，基于知识管理和智能控制的协同翻译平台——知识管理和机器翻译的融合，《中文信息学报》，2008 年第 5 期。
[61] 赵铁军、郑德权、宗成庆，2009，中国计算语言学研究进展，中国计算机学会主编《2008 中国计算机科学技术发展报告》，机械工业出版社，2009 年。
[62] 周强，2007，汉语基本语块描述体系，《中文信息学报》2007 年第 3 期。
[63] 宗成庆，2008，《统计自然语言处理》，清华大学出版社。
[64] 宗成庆、曹右琦、俞士汶，2009，中文信息处理 60 年，《语言文字应用》2009 年第 4 期。

附录：2004 —— 2008 年若干重要学术活动（学术会议与技术评测）概览

2005 年，第八届全国计算语言学联合学术会议（JSCL-2005），2005 年 8 月 27—29 日 中国南京。会议论文集《自然语言理解与大规模内容计算》（孙茂松、陈群秀主编）由清华大学出版社出版。
2007 年，第九届全国计算语言学联合学术会议（CNCCL-2007），2007 年 8 月 6—8 日 中国大连。会议论文集《内容计算的研究与应用前沿》（孙茂松、陈群秀主编）由清华大学出版社出版。

2005 年，第八届全国人机语音通讯学术会议，2005 年 10 月 22-24 日，北京（中科院声学所）
2007 年，第九届全国人机语音通讯学术会议，2007 年 10 月 21-24 日，安徽（中国科技大学）

2005 年，全国第十届民族语言文字信息处理技术研讨会，2005 年 7 月 16 日—18 日，中国青海西宁（青海师范大学）。
2007 年，全国第十一届民族语言文字信息处理技术研讨会，2007 年 2 月 1—4 日，中国云南西双版纳。

2007 年，中文信息处理国际会议(ICCC2007) 2007 年 10 月 13 日—15 日 中国武汉。会议论文集《中文计算技术与语言问题研究》（萧国政、何炎祥、孙茂松主编）由电子工业出版社出版。

2006 年，中国中文信息学会二十五周年学术会议，2006 年 11 月 21 — 22 日 中国北京。会议论文集《中文信息处理前沿进展》（曹右琦、孙茂松主编）由清华大学出版社出版。
2008 年，中国中文信息学会二十七周年学术会议，2008 年 11 月 24 — 25 日 中国北京。

2004 年，第五届汉语词汇语义学研讨会（CLSW2004），2004 年 6 月 14 — 17 日，马来西亚、新加坡。
2005 年，第六届汉语词汇语义学研讨会（CLSW2005），2005 年 4 月 20 — 24 日，中国厦门。
2006 年，第七届汉语词汇语义学研讨会（CLSW2006），2006 年 5 月 22 — 23 日，中国台湾。
2007 年，第八届汉语词汇语义学研讨会（CLSW2007），2007 年 5 月 21 — 23 日，中国香港。
2008 年，第九届汉语词汇语义学研讨会（CLSW2008），2008 年 7 月 14 — 17 日，新加坡。

2007 年，2007 国家语言资源与应用语言学高峰论坛，2007 年 9 月 12 — 14 日，北京（北京语言大学）。

2005 年，第三届 HNC 与语言学研究学术研讨会，2005 年 12 月 21 — 22 日，中国北京。会议论文集《中文信息处理的探索与实践》（朱小健、张全、陈小盟主编）由北京师范大学出版社出版。

2008 年，第一届全国知网研讨会（The First National HowNet Workshop，NHW2008）。2008 年 5 月 18—21 日，北京（北京信息科技大学）。

2004 年，第四届中日自然语言处理专家研讨会，2004.11.10-15，中国香港。
2005 年，第五届中日自然语言处理专家研讨会，2005.11.8-11，日本东京。
2006 年，第六届中日自然语言处理专家研讨会，2006.11.11-13，中国上海。
2007 年，第七届中日自然语言处理专家研讨会，2007.11.10-12，日本名古屋。
2008 年，第八届中日自然语言处理专家研讨会，2008.11.7-9 日，中国北京。

2004 年， The First International Joint Conference On Natural Language Processing (IJCNLP-2004), Sanya City, China，Mar. 22-24，2004.
2005 年， The Second International Joint Conference On Natural Language Processing (IJCNLP-2005), Jeju Island, Republic of Korea，October 11-13, 2005.
2008 年， The Third International Joint Conference On Natural Language Processing (IJCNLP-2008), Hyderabad, India, January 7-12, 2008.

跟技术评测有关的会议

2005 年，第一届统计机器翻译研讨会（SSMT2005），2005 年 7 月 13 — 15 日，厦门（厦门大学）
2006 年，第二届统计机器翻译研讨会（SSMT2006），2006 年 10 月 17 — 18 日，北京（中科院计算所）
2007 年，第三届统计机器翻译研讨会（SSMT2007），2007 年 8 月 12 — 13 日，哈尔滨。（评测于 2007 年 7 月 15—20 日进行）
2008 年，第四届全国机器翻译研讨会（CWMT2008），2008 年 11 月 27 — 28 日，北京。（评测于 2008 年 8 月 31 日——10 月 30 日进行）

2005 年 863 评测：2005 年 3——9 月评测。11 月 28—29 日，评测研讨会，北京。
2005 年 Sighan 评测：2005 年 6——8 月评测。10 月 14—15 日，评测研讨会（The Fourth SIGHAN Workshop on Chinese Language Processing），韩国济州岛。
2006 年 Sighan 评测：2006 年 4——6 月评测。7 月 22-23 日，评测研讨会（The Fifth SIGHAN Workshop on Chinese Language Processing），澳大利亚悉尼。
2007 年 Sighan 评测：2007 年 7——10 月评测。2008 年 1 月 11—12 日，评测研讨会（The Sixth SIGHAN Workshop on Chinese Language Processing），印度，Hyderabad。

2008 年 COAE2008 评测：2008 年 8——10 月评测。11 月 16 日，评测研讨会，北京。