汉语言文字信息处理状况分析

詹卫东
北京大学

摘　要　本文第一节概括说明了汉语言文字信息处理的整体态势，以及本文选择哪些内容作为重点分析对象的理由；第二节集中分析了核心技术的现状；第三节分析了应用系统的现状；第四节评述语言资源建设的情况；第五节是结语，指出了本领域值得注意的新动向。

关键词　汉语　信息处理　技术评测　信息检索　机器翻译　语料库　语言资源

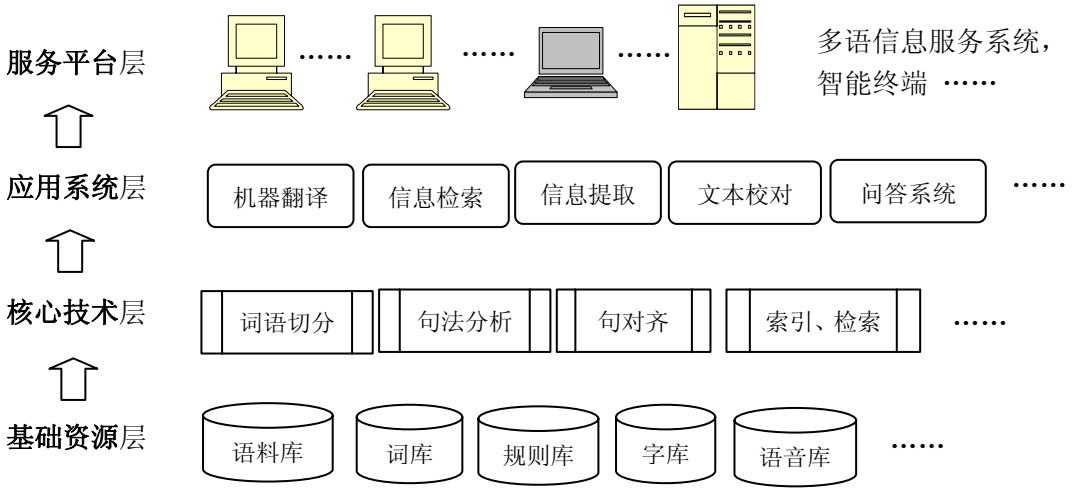
一　汉语言文字信息处理现状概述

自然语言（在本文中具体关注的是汉语语言文字）的信息处理，是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。对于这样一个交叉特点鲜明的领域，可以从不同视角，在不同层次上来认识。为了更好地概括说明这一领域目前的理论研究以及实际应用状况，本文首先为这一领域勾勒一个相对全面的框架（表一和图一）。然后再针对这个框架中“相对更值得一说”的部分展开来加以分析和讨论。

表一：汉语言文字信息处理的对象、层次和任务（虚线表示并不总是有严格界限）

对象 / 任务 / 层次	**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

图一：汉语言文字信息处理的宏观架构[1]



[1] 图一基本上可以看作是对表一中“符号的意义处理”这个层次的展开（“符号的形式处理”已经得到普遍应用，因此本文描述从简）。图一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。

从上面一表一图出发，可以将当前汉语言文字信息处理的总体发展状况概括为：

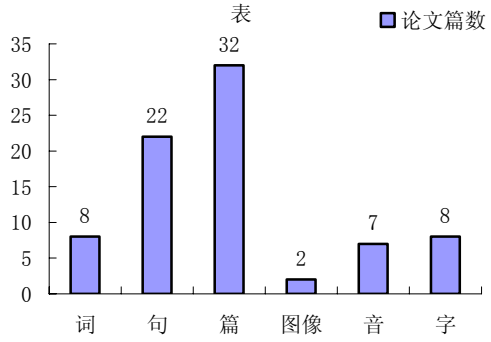
（1）对于符号形式层的处理，已经取得很大成功，并且在社会生活中得到广泛应用。

（2）对于符号意义层的处理，一些相对浅层的分析技术已经有很大发展并进入实用，比如中文词语切分技术已经应用于互联网信息检索系统，语音识别技术已经应用于语音电信增值服务（参见第三节），等等；而一些需要对自然语言进行深层分析的技术，比如句法分析、机器翻译等，仍然没有取得突破性进展，离真正走向大规模实际应用还有较大距离。

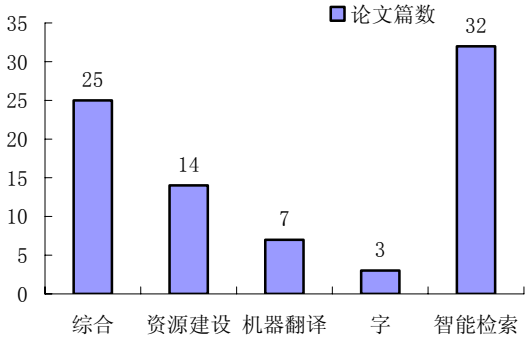
对于上述概括，需要说明的是，尽管符号形式层的处理已经得到普遍应用，但并不是说在这个层次上就没有可研究的问题，不需要进一步发展了。实际上，汉语言文字符号的数字化仍有许多工作要做，也还有不少难关需要攻克。其中比较突出的问题来自两个方面：第一，在人们一般日常的文字信息处理已经完全数字化之外，目前还有相当多的“特殊”的文字内容有待数字化（李宇明，2003）。比如中国浩如烟海的古籍内容在信息时代需要全面实现数字化，就涉及到大规模中文字库的研制[2]，涉及到汉字OCR（光学字符识别）技术的改进；再如对大量手写内容和历史上的科技文献内容的数字化，以及视频图像中所包含文字信息的数字化，就会涉及到对复杂版面内容（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。显然，上述这两个方面的问题，要求人们从一“大”一“小”两个方向来寻求如何更好地进行汉字符号形式层的处理。

尽管如此，鉴于汉字符号形式层的信息处理在相当大的范围内已经达到实用程度，下文将重点分析符号意义层的信息处理状况，这一方面是受篇幅的限制；另一方面也是因为，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。此外，从这一领域学术刊物和学术会议上发表的论文的分布情况看，也显示当前的汉语言文字信息处理研究，是以符号意义层的信息处理研究为重点和热点，而对符号形式层的信息处理研究，关注度相对较少一些。下面图二、图三基本显示了这一现状。

图二：中文信息学报2005年论文分布情况表



图三：JSCL-2005论文分布情况表



《中文信息学报》（双月刊）是中国中文信息学会会刊，该刊刊登的论文应该说能够基本反映中国语言信息处理目前的整体发展水平和研究态势。2005 年《中文信息学报》6 期共

[2] 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。

刊登论文 88 篇，其中 9 篇是涉及少数民族语言文字的（占 10%），其余 79 篇是有关汉语言文字的信息处理的，如果以各篇论文所研究的语言单位层级来区分，可以得到如图二所示的论文的分布情况。关于语音和文字的论文合在一起不到 20%。大量研究集中在词、句、篇[3]等语言单位上。JSCL-2005 是第八届全国计算语言学联合学术会议（两年一届）。该会议是国内自然语言信息处理的综合性会议，也代表了当前语言信息处理研究的发展水平。在会议正式论文和特邀报告中，除 5 篇有关少数民族语言文字的论文外，有关汉语言文字信息处理的文章共 81 篇，论文分布情况如图三所示（其中“综合”类是在汉语理论层次或信息处理核心技术层次上对汉语词、句、篇三级单位有关问题加以研讨）。在此次会议中，有关汉字符号形式层处理的论文只有 3 篇[4]，而各种篇章级的应用研究，包括信息检索、提取、文本分类、摘要、过滤等等，统称为“智能检索”，共 32 篇。很明显，当前本领域的研究重点和热点集中在篇章级的应用系统上；资源建设和机器翻译技术的研究，以及在汉语各级语言单位层级上的信息处理研究也受到广泛关注。而有关汉字层次上的研究，则明显较少。就符号意义层的处理来说，目前的工作主要是在图一所示的下面三个层次上，最高层的“服务平台”层需要下面的基础打牢之后才可能真正搭建起来。

基于上述情况，下文先从核心技术说起，再延及应用系统和基础资源：第二节将着重分析核心技术的发展水平；第三节谈应用系统的发展水平；第四节分析语言基础资源建设的现状；第五节是结语，指出当前本领域若干值得注意的新趋势。

二　核心技术的发展现状

除了图一列举的词语切分（也简称分词技术）、句法分析、句对齐、索引及检索等之外，汉语信息处理的核心技术还应该包括词性标注技术、词义消歧技术、词和短语对齐技术、句子相似度计算技术，等等。限于篇幅这里仅对分词技术和句法分析技术的现状做概要分析[5]。

要了解语言信息处理技术的现状，显然应该是通过公正的评测来说明问题。近年来国际上对 NLP 技术的大规模评测越来越重视，国内这方面的工作也在积极推进（钱跃良等 2005）。目前国际上 NLP 技术评测的共同特点是（1）完全公开（2）用大规模真实语料数据进行测试（3）由计算机程序自动打分来评价系统的性能。从 NLP 技术近年来发展的情况来看，这样的评测在推进技术进步方面起到了显著作用（黄昌宁 2002）。

2005 年国内没有举办有关中文分词的评测。而国际计算语言学联合会（ACL）下设的中文信息处理兴趣组（SIGHAN）从 2003 年开始举办第一届国际性的中文分词评测（Bakeoff1），2005 年举办了第二届（Bakeoff2）。因此我们对目前中文分词系统的技术水平的考察，主要基于从 SIGHAN 网站上获取的评测结果数据。需要说明的是，一般采用计算精确率（precision），召回率（recall）的办法来评估一个中文自动分词系统的性能。精确率和召回率分别定义如下：

---

[3] 这里“词”指是汉语分词、词性标注，词义排歧等方面的研究，“句”指的是句法分析相关研究，“篇”指的是篇章指代研究，以及各种以篇章单位为处理对象的应用系统的研究，包括信息检索、提取、文本分类、话题发现，……等等

[4] JSCL上没有语音处理技术方面的论文，2005 年语音技术方面的论文都集中到“第八届全国人机语音通信学术会议”上发表了（134 篇）。

[5] 分词几乎是所有中文信息处理的基础，句法分析则是通向真正的语言理解的关键一步，一直一来都是自然语言处理中的核心问题，因而本文将这二者作为主要考察对象。对齐技术主要用于双语（多语）平行语料库的建设，进而应用于基于记忆或基于统计的机器翻译等系统中。目前汉英句子对齐的正确率一般在 95% 以上（张艳、柏冈秀纪, 2005），词对齐和短语对齐的相关研究近期比较少（中文信息学报 2005 年没有一篇有关词对齐的研究）。因此本文对这方面的工作不做具体分析。下面是今年“863 计划中文信息处理与智能人机接口技术评测”（参见第三节）中机器翻译项目的技术评测任务中一家单位参加汉英词语对齐评测的结果：精确率为 0.8087，召回率为 0.7220，F-Score为 0.7629，对齐错误率为 0.2348。读者可以从中大致了解目前词语对齐的技术水平。

$$精确率(P)=\frac{自动分词结果中切分正确词的数目}{自动分词结果中词的数目}*100\%$$

$$召回率(R)=\frac{自动分词结果中切分正确词的数目}{标准答案中词的数目}*100\%$$

通常将 P 和 R 两个指标综合为二者的调和平均值 F-Score 来反映一个系统的整体性能。F-Score 可以有不同的定义公式，通常采用的一个是（SIGHAN 的 Bakeoff 采用的也是这个公式）：$F=\frac{2PR}{P+R}$。下面表二就是这两届评测各子项目中调和平均成绩（F-Score）排名第一的结果（本文关注的主要是技术，而不特别关注具体的参评单位）。

表二：SIGHAN Bakeoff1 和 Bakeoff2 的部分结果（数据来源：http://www.sighan.org/）

项目	F-score	R-oov	R-iv	时间	来自
AS-o	0.904	0.426	0.926	2003	美国
AS-c	0.961	0.364	0.980	2003	美国
CTB-o	0.912	0.766	0.949	2003	中国大陆
CTB-c	0.881	0.705	0.927	2003	中国大陆
HK-o	0.956	0.788	0.971	2003	中国台湾
HK-c	0.940	0.625	0.972	2003	中国台湾
PK-o	0.959	0.799	0.975	2003	美国
PK-c	0.951	0.724	0.979	2003	中国大陆
AS-o	0.956	0.684	0.975	2005	新加坡
AS-c	0.952	0.696	0.963	2005	日本
MSR-o	0.972	0.590	0.990	2005	中国大陆
MSR-c	0.964	0.717	0.968	2005	美国
HK-o	0.962	0.806	0.980	2005	新加坡
HK-c	0.943	0.698	0.961	2005	美国
PK-o	0.969	0.838	0.976	2005	新加坡
PK-c	0.950	0.636	0.972	2005	美国

说明：SIGHAN举办的两届Bakeoff的评测方式基本一样，都是选取了四种语料库，每种语料库上参评系统可以选择开放测试（Open test）和封闭测试（Close test）两种方式[6]。第一届提供评测语料的四家单位是北京大学（PK），香港城市大学（HK），台湾中研院（AS）和美国宾州大学（CTB）。其中前三家单位继续为第二届Bakeoff提供语料，而美国宾州大学没有为第二届Bakeoff提供语料，改为由微软研究院（MSR）提供语料。表二中第一列的评测项目即由提供语料的单位名称缩写加开放（O）或封闭（C）两种方式组成。除基本的P，R，F-score成绩外，Bakeoff还给出了各参评系统的未登录词召回率（R-OOV）和词表词召回率（R-IV）指标。表二把两届评测中各项目F-Score的最好成绩列出。在一定程度上可以反映目前中文分词系统的质量水平（当然也存在这样的可能性：质量更好的分词系统没有参评）。其中对于未登录词的识别，召回率在 60%-80%之间。这意味着如果计算机处理的语料中包含较多生词时，分词系统的性能将受到明显影响。

[6] 所谓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，可以利用任何知识进行分词；封闭测试则要求参评系统只能利用训练语料库获取分词知识。

中文句法结构分析目前还没有基于大规模语料的公开评测[7]。因而很难有大家一致接受的数据来说明问题。下面我们提供两个方面的数据，表三是北大计算语言所常宝宝博士所做的完全句法分析（full parsing）的实验结果数据。表四是来自微软亚洲研究院黄昌宁教授的一份报告[8]中有关汉语语块分析（chunking）的实验结果数据。前者在一定程度上反映汉语深层句法结构分析的研究状况；后者则在一定程度上反映汉语浅层句法分析的研究状况[9]。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[10]。

表三：基于最大熵模型的汉语完全句法分析实验数据（语料：宾州大学中文树库 1.0 版[11]）

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

表四：汉语语块分析实验数据（语料：1998 年人民日报 1 月份语料[12]）

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）

对于完全句法分析来说，如果按照“整句匹配率”（complete match）指标作为评判标准，可以看到，在开放测试条件[13]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。这样也就很容易理解机器翻译系统为什么性能很难上去了（连句法结构分析都不正确，如何得到正确的翻译结果呢？）。下面第三节中机器翻译系统评测数据也反映了这一现状。

三　应用系统的发展现状

需要说明的是，这里所说的“应用系统”，主要还是侧重于中文信息处理研究单位在实

---

[7] 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作为一个参照，来试验、比较各种句法分析方法的优劣。最近的相关研究可参看Xiong, Deyi, et al.(2005)。

[8] 来自中国语言文字网http:// www.china-language.gov.cn/doc/NLP0/09.pps。

[9] 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；而对于机器翻译等应用系统来说，一般需要深层句法结构分析才能满足需要。

[10] 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！

[11] 该版本的树库语料含 325 个数据文件，4185 句，平均句长 23.89 词。

[12] 该语料可从北大计算语言学研究所网站下载http://icl.pku.edu.cn/icl_groups/corpus/dwldform1.asp。

[13] 这里的开放测试不同于上文Bakeoff中的含义，而是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。

验室环境下进行相关应用技术的研发和探索时所开发的系统，对于真正运营中的商业应用系统，本文基本不展开讨论。一方面，目前从互联网媒体上比较容易找到有关应用系统的非技术特征的调查数据，以近年来最引人注目的搜索引擎为例，新浪（Sina）、赛迪（CCID）和中国互联网信息中心（CNNIC）在今年都公布过 2004 年和 2005 年中国搜索引擎市场调查报告。对诸如百度、Google、雅虎、新浪、搜狐等多家搜索引擎网站和门户网站提供的搜索服务情况做了调查。读者可以通过这些报告了解到目前各家搜索引擎的市场占有率、用户群特征等等情况（相关网页：http://tech.sina.com.cn/focus/serch_05/index.shtml 或 http://www.cnnic.cn/html/Dir/2005/08/30/3085.htm）。另一方面，也有第三方技术评测单位对市场上的搜索引擎产品进行过综合质量评估，比如清华大学 IT 可用性实验室就先后在 2004 年 6 月和 2005 年 9 月两次对中国市场的搜索引擎质量进行过评估，评测内容包括标准搜索方式下检索结果相关性、网页覆盖率、死链率、作弊率、中文分词质量等等。2005 年 9 月的评测结果发表在《计算机世界报》2005 年 11 月 14 日第 44 期 E7 版。读者也可以从计算机世界网站下载全文（http://www2.ccw.com.cn/05/0544/f/0544f20_2.asp）。报告在结论部分指出，本土搜索引擎（比如百度）在中文分词技术质量方面有明显优势。下表是引自该文的对各家搜索引擎的分词技术所做的评测结果：

表五：

系统	Google	一搜	百度	中搜	爱问	搜狗
分词成绩	80%	77.80%	90%	81.10%	74.40	76.70%

从表中数据可以看出，中文分词系统在面对互联网海量数据时，分词正确率比 Bakeoff 评测中的成绩要低不少（Bakeoff 评测的数据量在 10 万词次以内）。但必须指出，尽管如此，中文分词系统对搜索引擎的质量改进还是比较明显的，比如在上一代搜索引擎中查找“和服”“市政”这些存在分词歧义的词语，返回的结果网页经常是“和服务”“市政府”“市政协”等不相关网页。但经过中文分词处理后，还是在相当程度上提高了查询结果的主题相关性，现在的搜索引擎已经可以在很大程度上避免返回这些无关网页。

限于篇幅，下面仅以汉语信息处理应用中最具代表性的信息检索系统和机器翻译系统为例来说明目前的技术所达到的水平。前者因为近年来基于互联网的搜索引擎市场持续升温从而也备受学术研究者的青睐，后者则因为可以说是综合反映计算机对自然语言真正意义上的“理解”水平，因而值得一谈。

关于信息检索系统和机器翻译系统的技术水平，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”研讨会公布的结果了解最新的情况[14]。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）[15]。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（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

[14] 863 计划从 1990 开始尝试进行对自然语言信息处理技术进行公开评测，1991 年正式实施，此后虽在有些年份中断（比如 1993，1996，1997，1999-2002），但这项工作基本还是延续下来了。特别是近年来国际上利用评测来推动技术进步越来越成为大家公认的一种有效做法，有关中文信息处理技术的评测工作也得到了 863 专家组和学术界的重视和支持。有关评测详情可访问http://www.863data.org.cn/。

[15] 相比之下，国际上最具影响力的信息检索评测TREC（由美国国家标准局NIST和美国军方的国防部高级研究计划署DARPA组织）的规模要大很多。TREC评测从 1992 年开始，每年一次。从TREC网站上可以了解到 2004 年TREC的规模(2005 年的总结尚未公布)，参加单位超过 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），详见TREC网站 http://trec.nist.gov/presentations/t2004.presentations.html。

（“manual”代表人工构造查询，“auto”代表自动构造查询）的得分情况。

表六：（数据来源：2005 年 11 月召开的 863 评测研讨会）。

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

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 个查询结果中有接近 70%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。这也正好说明目前Web搜索引擎在互联网时代确确实实能为人们更快捷地获取信息提供便利，因而成就了一个巨大的市场[16]。

2005 年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。这里我们只关注前四个跟中文有关的项目。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。

表七：（引自 863 评测网站）

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[17]。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。下表列出了在今年 863 评测的各个项目中BLEU成绩排名第一的系统的具体得分情况[18]。

表八：（数据来源：2005 年 11 月召开的 863 评测研讨会）

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**

[16] 国内数据调查研究机构赛迪顾问在 2005 年 12 月 27 日举行的 2005 中国搜索年会上，发布了《2005-2006 年中国搜索引擎市场及投资机会研究年度报告》。该报告显示，2005 年中国搜索引擎市场规模实现了 42.2%的增长，达到 11.8 亿元人民币（详见http://search4.ccidnet.com/nianhui.htm）。

[17] 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（在今年 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Papineni et al.(2001)，Joseph P. Turian et al.(2003)，Chin-Yew Lin and Franz Josef Och (2004)。

[18] NIST在其网站上公布了 2005 年机器翻译评测的结果，其中汉英翻译大数据集测试和无限制文本数据集测试两项排名第一的都是Google公司的系统，BLEU（N=4）得分分别为 0.3531 和 0.3516。显示出该系统的稳定性特色。具体细节可访问NIST网站查询：
http://www.nist.gov/speech/tests/mt/mt05eval_official_results_release_20050801_v3.html。

	篇章	8.7453	**0.3709**	0.7930	0.6162	0.3934	**55.78**	**47.85**
汉日	对话	7.1158	**0.3512**	0.7792	0.6483	0.4421	**53.44**	**44.87**
	篇章	8.5858	**0.3750**	0.8265	0.6450	0.3886	**44.74**	**35.29**
日汉	对话	7.7098	**0.3302**	0.7302	0.6030	0.4430	**67.94**	**67.03**
	篇章	7.9797	**0.3007**	0.7170	0.6748	0.4636	**50.41**	**44.58**

总的来看，完全自动的机器翻译的质量目前离实用还有明显的距离。这也再次提示人们应该从实际出发来定位机器翻译系统的设计目标。2000 年中国科学技术基金会、中国科学院科技翻译协会等单位曾联合在互联网上做过一项有关机器翻译发展现状与未来的调查[19]。调查显示：计算机可以帮助翻译人员解决的前三位问题是提高工作效率、查字典、自动记忆翻译结果。对于翻译软件的功能，非常需要的前三位功能是：“大容量的专业词库”、“交互翻译”和“记忆功能”（所占比例分别为 18.3%、14.8%和 13.7%）。交叉分析的结果显示，不同行业的回答者对翻译软件的功能需求方面不存在差异。调查得出的结论是，未来市场需要的机器翻译软件是具有大容量的专业词库、交互翻译、记忆功能的翻译软件。从目前全自动机器翻译系统的实际表现来说，五年前的这个调查还是很有参考价值的。

此外，近年来语音技术在实际应用中的发展非常迅速，比如捷通华声、科大迅飞[20]等公司的语音技术在金融系统、保险、电力以及政府部门等涉及到公众生活领域的语音服务已经切实发挥了作用，产生了巨大的经济效益。同时，智能语音识别技术在普通话教学与水平测试中的应用，也越来越受到关注。目前的研究表明，计算机自动进行普通话水平测试已经达到较高精度，预期会产生显著的社会效益。限于篇幅，本文不展开讨论，详情可访问相关网站查询。

四　汉语语言资源的建设

汉语信息处理技术和应用系统要达到实用目标，没有大规模高质量的语言基础资源，是难以想像的。随着计算机软硬件环境的不断改善，研究的深入，目前语言资源建设的条件比以往有了很大的提高。比如自动分词和词性标注软件的性能改进，可以帮助人们在更短时间内，花费更少的人力建设更大规模的分词和词性标注语料库，而句法分析器性能的改进以及辅助编辑工具的使用则可以帮助人们更富于效率地建设中文句法树库。这些进步已经明显地带动了语料库的建设规模向广度和深度两个方向上的拓展（可以预见，经过深加工的大规模中文树库，经过词、短语对齐的双语平行语料库将成为今后语言资源建设的重点）。此外，随着开放源代码软件（Open Source Software）的不断增多，互联网上也出现了越来越多的可以利用的程序资源[21]（创立于 2002 年 8 月，由中国科学院计算技术研究所软件研究室自然语言处理课题组发起并主办的“中文自然语言处理开放平台”http://www.nlp.org.cn是国内自然语言处理领域开源网站的代表），从而大大缩短了语料库检索系统的开发时间。应该说，语言资源建设目前正处于一个比较好的大环境中。

如果结合语言资源建设的历史发展来看现状，有两点是很明显的：（1）无论是语料库，还是词库，语言基础资源的建设都是费时费力，投资成本相当高的工程。因为要达到实用目标，语言资源库就必需要有相当大的规模，而且要求有相当高的质量。这一“大”、一“高”，就决定了语言基础资源的建设绝非朝夕之功，而常常是经年累月精雕细刻的结果。以目前有

[19] 相关报道见赛迪网http://www0.ccidnet.com/news/buss/2000/11/17/81_10770.html。

[20] 参见http://www.sinovoice.com.cn/ 和 http://www.iflytek.com/

[21] 比如基于Java的公开源代码搜索引擎Lucene（http://lucene.apache.org/java/docs/index.html），以及基于XML格式语料库和C++程序语言开发的语料库Concordance系统XAIRA（http://www.oucs.ox.ac.uk/rts/xaira/ 或 http://sourceforge.net/projects/xaira/）。北京大学汉语语言学研究中心（CCL）在网站上提供的面向汉语研究的语料库在线检索系统就基于Lucene开发（http://ccl.pku.edu.cn）。

代表性的国家语委的现代汉语通用语料库为例[22]，该语料库 1991 年立项，到 2005 年，历时近 15 年，生语料规模超过 1 亿字。其中五千万字进行了分词和词性标注，100 万字（5 万句）进行了短语结构标注。目前部分语料已经提供网上查询服务。作为大型的国家级语料库，该语料库的加工仍在继续。再看信息处理用电子词典中最有代表性的成果，北京大学计算语言学研究所和北京大学中文系联合研制的“现代汉语语法信息词典”（俞士汶等 2003a，2005），该词典自 1986 年国家“七五”科技攻关项目立项，历经近 20 年，发展至今，已收录超过 8 万词，共 32 个数据库，总信息量超过 250 万，词库数据超过 16MB。以该词典为基础，北大计算语言所进行了一系列的中文信息处理基础语言资源的建设，包括现代汉语语义词典（王惠等 2003）、中文概念词典（于江生等 2003，刘扬 2005）、现代汉语基本标注语料库（俞士汶等 2003b，2004），等等。从以上两个单位富有代表性的语言资源建设历程中，不难体会到资源建设的工程之艰巨。(2) 语言资源确实在信息处理的发展中发挥了巨大的推动作用。比如北大计算语言所 2001 年在其网站上公布了 1998 年 1 月份人民日报标注语料（200 多万词次）供免费下载。至今已有超过 7800 次下载。此后学术杂志和会议论文中常可以看到基于该语料库所做的研究(比如中科院计算所开发的汉语分词和词性标注系统就基于北大人民日报语料库获取参数，该系统作为开放源代码软件目前已有超过 3 万次下载[23])。再如清华大学智能技术与系统国家重点实验室在研制汉语分词系统的过程中，积累了超过 8 亿字的现代汉语生语料，为定量确定一个语言单位是否成词提供统计数据，在制订“信息处理用现代汉语分词词表”的工作中起到了非常基础性的作用。

以上两点是就语料库、词库等语言基础资源本身而论的。事实上，随着语料库建设和应用的深入，人们越来越感觉到，语言基础资源之外的一些因素，在很大程度上影响了语言资源的建设与利用。这其中突出的问题有两个，一是在语言资源建设之初，如何解决语言原始资料的版权或者说是授权加工的问题；二是在语言资源库建成之后，如何最大地发挥其效用，让更多的人可以使用，同时又能保障语言资源开发者的合理利益的问题。应该说，这两个非常实际的问题因为涉及到国家法规政策以及科研单位的经济利益等诸多因素，目前还没有得到很好的解决。但越来越多的有识之士已经意识到这些问题并开始着手寻求解决的途径。近年来人们为此做出的努力可以从两个代表性的事情上看出。一个是以国内信息处理的学术界力量为主导，引进国外先进的语言资源管理机制，从 2003 年开始，在国家 973 计划的资助和相关课题研究的推动下，成立了中文语言资源联盟Chinese LDC（赵军等 2003），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，到 2005 年，中文语言资源联盟官方网站上已经列出了 41 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。其定价模式区分两个因素：商用/研究用；中国大陆地区使用/在境外国外使用。尽管跟美国UPenn的LDC[24]相比，中国在这方面起步已经晚了十年。但毕竟是起步了。我们希望，随着机制的不断完善，中文语言资源联盟在中文信息处理和语言基础资源建设方面将发挥越来越大的推动作用。第二个有代表性的事情是以政府力量为主导，由高等院校的研究单位来参与实施的。从 2004 到 2005 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的五个分中心，包括教育部语言文字信息管理司与北京语言大学共建的平面媒体监测与研究分中心，与华中师范大学共建的网络媒体监测与研究分中心，与中国传媒大学共建的有声媒体监测与研究分中心，与暨南大学共建的海外

[22] 其他国内知名的语言知识库、语料库也都是长年累积建设而成的，比如董振东先生开发的“知网”（HowNet），清华大学的语义词典、语料库等等。

[23] 参见中文自然语言处理开放平台网站http://www.nlp.org.cn/project/project.php?proj_id=6。

[24] 美国语言资源联盟（LDC）成立于 1992 年，发展至今，在自然语言信息处理领域已经产生深远影响。是目前国际上最大的语言资料库集中地。该组织发布的语言资源数目为 313 项。详情可访问其官方网站查询http://www.ldc.upenn.edu/。网站上按年列出了语言资源目录。

华文媒体监测与研究分中心，与厦门大学共建的教育、教材媒体监测与研究分中心。随着这五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，都将产生显著的辐射性影响。

值得一题的是，尽管人们在语料库的传播和共享机制方面已经做出了努力，但在语料库的知识版权问题上，目前还没有出台有效的法规和举措[25]。这在很大程度上也是因为语言资源库的建设对于许多人来说还是“陌生的新事物”。因此需要政府有关部门出面加强协调，同时国家法律法规制订部门应该注意到语言资源建设中的特殊的知识版权问题，以促进科学研究，推动信息技术进步为出发点，制订更合理的相关法律。

五　结语：兼谈汉语言文字信息处理值得注意的新动向

要在一篇文章中巨细无遗地展现 2005 年汉语言文字信息处理的全貌是不可能的。上文试图通过对一些“重点内容”的透视，来努力勾勒出目前汉语言文字信息处理所达到的水平，基本上是“抓住一点、不及其余”，其局限性也是相当的明显。事实上，如果我们将镜头拉远拉宽，2005 年汉语言文字信息处理领域还有很多重要的事情没有进入上文的分析视野，比如国家 863 计划、973 计划、国家自然科学基金资助的重大课题中，对汉语信息处理相关研究都给予了很大的支持，显示了国家和政府相关部门对汉语信息处理领域的高度重视。再比如 2005 年一系列国际知名的自然语言处理教材译成中文出版（冯志伟等 2005，刘群等 2005，苑春法等 2005），则体现出这一领域的教学工作得到了更多的关注。注意到这些或者宏观或者微观但都非常重要的事实，显然有助于我们更好地认识这一领域的现状。但限于篇幅，在权衡之下，本文还是选择了把有限的笔墨统统集中在关于汉语言文字信息处理技术和资源本体内容的描述上，相应地也就把许多“外围的大事”背景化或者干脆淡化了[26]。

通观 2005 年汉语信息处理的进展情况，可以看到，伴随近年来互联网的热潮不断，人们工作生活中的信息处理量以加速度方式在急剧膨胀。这使得智能化的搜索引擎的需求现实性日益明显，从而大力驱动着信息处理的相关研究，包括信息检索、信息提取、文本分类、垃圾邮件过滤在内的诸多应用技术成为当前研究的热点。除这些在“浅层”进行信息处理的应用系统外，人们也在开始加大对于“深层”信息理解的关注度，比如文本褒贬色彩的评价研究，文本隐喻的发现，等等。以上是从信息接收方“理解”信息的角度来看信息处理所能察觉到的研究新动向。如果从信息发出方的角度来看如何“制造信息”，则近年来的热点莫过于对Ontology（知识本体）的研究（黄居仁 2005）。广义地说，已有的语言基础资源库在某种程度上都可以看作是一个具体的Ontology。而随着像WordNet[27]等免费语言资源和像Protégé [28]这样的开放源代码Ontology构建工具软件在国际信息处理界的影响力加大，以及XML等新一代文本内容标示语言的兴起和普及，越来越多的信息处理学者开始把自己的工作跟Ontology的设计联系起来，比如在术语提取、术语知识库管理等领域，相关研究已经成为新的趋势。人们希望，随着越来越多具体的人类知识（当然也包括语言知识）被搭建成一个个的Ontology系统，对自然语言文本中多义词的消歧，对文本内容的理解，进而对基于内容理解的信息搜索和信息提取，都会带来质量上可观的改进。

如果说以上都是循着原有的信息处理发展轨迹继续向前的话，那么，将信息处理研究中

---

[25] 冯志伟（2002）教授撰文介绍中国语料库的状况时就指出过这一问题，并建议国家对语料库的版权问题制订专门法规。

[26] 比如上文仅提到了JSCL-2005 会议，但如果能综合考察本领域在 2005 年召开的诸多国际国内重要学术会议，也可更全面更准确地反映这一领域学术研究的活跃程度。

[27] http://wordnet.princeton.edu/

[28] http://protege.stanford.edu/

积累起来的技术用于词典编纂，用于语言教学，可以说是大大拓展了信息处理研究的应用范围。比如在汉英（英汉）双语对齐的语料库基础上，开发双语词典编纂平台，就是突出的例子。2005 年北大计算语言学研究所与外语教学与研究出版社合作，将计算语言所开发的双语句子对齐软件、大规模双语语料库检索和搭配统计软件等整合为一个基于 Web 界面的词典编纂平台。目前已经实现原型系统，待进一步调试后即可投入词典编纂的具体应用。像国外著名的词典出版机构 Oxford、Longman、Collins 等，都是语料库词典编纂方法的积极倡导者和实践者，它们推出的词典产品风行全球，语料库及现代计算机信息处理技术在其中的作用功不可没。现在，国内这方面的条件也已逐渐成熟，利用已有的信息处理技术的成果，在大规模语料库的支撑下，加快中文词典编纂现代化的进程，无疑将是未来词典出版业的一个方向。而汉语信息处理的研究者，在这个进程中，可以扮演积极而重要的角色。

总起来看，2005 年在汉语信息处理的发展历程中不能说有多少特别之处，自然语言处理的困难无时无刻不在限制着这条路上的行进者的步伐。但执著的追求者们还是在坚定地向前：就算是一小步，那也是向前的一小步。

致谢：北京大学计算语言学研究所俞士汶教授、常宝宝博士、胡俊峰博士，李素建博士、中科院计算所刘群博士对本文初稿提出了宝贵意见和修改建议。常宝宝博士、刘群博士和北大计算语言所刘扬博士为作者提供了许多参考资料。在此一并表示诚挚的谢意。文中错谬盖由作者本人负责。

参考文献

冯志伟（2002），中国语料库研究的历史与现状，载《汉语语言与计算学报》（新加坡）, 2002, 第 12 卷，1 期。

冯志伟、孙乐 译（2005）《自然语言处理综论》，电子工业出版社。译自 Daniel Jurafsky & James H. Martin, 2000, *Speech and Language Processing: An Introduction to Natural Language Processing*, Computational Linguistics, and Speech Recognition. Prentice Hall. 1999。

黄昌宁 等（2005） 何谓金本位，载孙茂松、陈群秀 主编《自然语言理解与大规模内容计算》，清华大学出版社 2005 年版。

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）。

黄居仁（2005）语意网与中文信息化的前瞻：知识本体与自然语言处理，载孙茂松、陈群秀主编《自然语言理解与大规模内容计算》，清华大学出版社。

教育部语言文字应用研究所，2005，国家语委现代汉语语料库介绍，“语言学研究手段现代化学术研讨会”，北京大学 2005.11.12-13。（语料库在线检索系统网址：http://219.238.40.213:8080/）

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。

刘扬（2005）中文概念词典的研究与开发，“语言研究现代化手段问题学术研讨会”，2005 年 11.12-13 日，北京大学。

刘群、张华平、骆卫华、孙健 译（2005），刘群审校，《自然语言理解（第二版）》，电子工业出版社，北京，2005.1，译自 James Allen, 1995, *Natural Language Understanding (Second Edition)*, The Benjamin / Cummings Publishing Company, Inc., 1995。

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（中国科学院计算技术研究所内部刊物，中国计算机学会赠阅会员刊物），第 3 卷第 8 期，2005 年 8 月，网址：http://www.ict.ac.cn/5-3.asp。

孙茂松（2001）汉语自动分词研究的若干最新进展，《中文信息学会二十周年学术会议论文

集》，清华大学出版社。

孙茂松（2003）对统计语言模型的若干认识。载 徐波、孙茂松、靳光谨 主编，中文信息处理若干重要问题，科学出版社 2003 年版。

王惠、詹卫东、俞士汶（2003）现代汉语语义词典规格说明书，载《汉语语言与计算学报》（新加坡），2003 年 6 月，第 13 卷 2 期。

易绵竹 南振兴 编（2005）计算语言学，上海外语教育出版社，“迈向 21 世纪的语言学”丛书。

于江生、刘扬、俞士汶（2003）中文概念词典规格说明，载《汉语语言与计算学报》（新加坡），2003 年 6 月，第 13 卷 2 期。

俞士汶、朱学峰、王惠、张化瑞、张芸芸、朱德熙、陆俭明、郭锐（2003a）《现代汉语语法信息词典详解》（第二版），清华大学出版社 2003 年版。

俞士汶、段慧明、朱学峰、孙斌、常宝宝（2003b）北大语料库加工规范：切分·词性标注·注音.《汉语语言与计算学报》（新加坡），2003 年 6 月，第 13 卷 2 期。

俞士汶、段慧明、朱学峰（2004）综合型语言知识库的建设与利用，《中文信息学报》2004 年第 5 期。

俞士汶、段慧明、朱学峰（2005）词语兼类暨动词向名词漂移现象的计量分析，载孙茂松、陈群秀主编《自然语言理解与大规模内容计算》，清华大学出版社 2005 年版。

苑春法、李庆中、王昀 等译（2005）《统计自然语言处理基础》译自 Christopher D. Manning & Hinrich Schutze, 1999, *Foundations of Statistical Nautral Language Processing*, MIT. 1999。

詹卫东（2003），面向自然语言处理的大规模语义知识库研究述要，载徐波、孙茂松、靳光谨主编《中文信息处理若干重要问题》，科学出版社 2003 年版。

张艳 柏冈秀纪（2005）基于长度的扩展方法的汉英句子对齐，《中文信息学报》2005 年第 5 期。

赵军 徐波 孙茂松 靳光谨（2003）中文语言资源联盟的建设和发展，载徐波、孙茂松、靳光谨主编《中文信息处理若干重要问题》，科学出版社 2003 年版。

Huang, Chur-Ren & Winfried Lenders, 2004,ed., *Computational Linguistics and Beyond,* Institute of Linguistics, Academia Sinica. Language & Linguisitcs Monograph Series B: Frontiers in Linguistics I. 中译本：俞士汶、黄居仁 主编 （2005） 计算语言学前瞻，商务印书馆 2005 年版。

Lin,Chin-Yew & 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.

Papineni, Roukos, Ward, Zhu (2001). Bleu: a Method for Automatic Evaluation of Machine Translation, (IBM Technical Report, Keyword. RC22176- W0109-022).

Turian,Joseph P., Luke Shen, and I. Dan Melamed(2003), Evaluation of Machine Translation and its Evaluation. In *Proceedings of MT Summit IX,* New Orleans, LA, USA,2003.

Xiong, Deyi, 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.