中文信息处理研究现状分析*

宗成庆

（中国科学院自动化研究所　北京　100190）

提　要　60多年来中文信息处理研究取得了令人瞩目的成就。但是，这一领域也面临问题和挑战。本文在对中文信息处理研究成就简要归纳的基础上，分析这一领域的技术现状，直面存在的问题，并对未来发展的方向提出一些看法。希望本文指出的问题能够引起中国国内同行的关注，为未来的中文信息处理研究提供有益的参考。

关键词　中文信息处理；自然语言处理；自然语言理解；计算语言学

Chinese Language Processing: Achievements and Problems

Zong Chengqing

Abstract　In the past over 60 years, research on Chinese language processing has made great achievements. With the rapid development and popularization of the Internet and communication technology, Chinese language processing technology has attracted worldwide attention in recent years. This article summarizes the achievements of Chinese language processing and analyzes the present status of the technology in this field, particularly the problems that the field may face in term of development. The author argues that it is still difficult for artificial intelligence to "understand" rather than "process" naturally produced Chinese because of the following three reasons: (1) the current information processing technology is inadequate in processing grammatically complex Chinese sentences; (2) there are unsolved problems in machine learning technologies; and (3) our understanding of how human brain processes language is still very limited. This paper concludes that we need a better understanding of how the Chinese language is decoded in human brain and build a computational model that specifically targets at the Chinese language in order for artificial intelligence to understand naturally produced Chinese.

Key words　Chinese language processing; natural language processing; natural language understanding; computational linguistics

一、引　言

自1956年人工智能（artificial intelligence，简称AI）概念被提出以来，自然语言理解（natural language understanding，简称NLU）就一直是这一领域研究的核心问题之一。尽管20世纪60年代提出的计算语言学（computational linguistics，简称CL）和80年代衍生的自然语言处理（natural language processing，简称NLP）概念分别从数学建模和语言工程角度各自诠释了不同的外延，但NLU、CL和NLP这三个术语的实质内容和共同面对的科学问题并无本质的差异，尤其从实际应用的角度看，几乎一样。因此，在不引起混淆的情况下人们常以"人类语言技术"（human language technology，

作者简介：宗成庆，男，中国科学院自动化研究所研究员、博士生导师，主要研究方向包括自然语言处理、机器翻译和文本分类等。电子邮箱：cqzong@nlpr.ia.ac.cn

*　作者衷心地感谢杨尔弘教授对初稿提出的修改建议。感谢匿名审稿专家提出的宝贵意见。

简称 HLT）泛指这一语言学、计算机科学和人工智能等多学科交叉的研究领域（宗成庆 2013）。

中文信息处理（Chinese language processing，简称 CLP）是指针对中国的语言文字开展相关研究的一个专属领域，是自然语言处理的一个具体分支。广义上讲，“中文”是中国各民族使用的语言文字的总称，在不引起误解的情况下，“中文”与“汉语”指的是同一概念。随着中国综合国力的增强，以互联网为纽带的经济和信息全球化趋势，尤其是中国“一带一路”战略的实施，向包括中文信息处理在内的人类语言技术提出了前所未有的挑战，巨大的技术市场吸引着全球科学家和企业家的目光（宗成庆等 2009）。

与其他语言的处理技术相比，中文信息处理处于怎样的技术水平？近年来，中文信息处理从资源库建设、理论建树，到技术研发和人才队伍培养，有哪些根本性的变化？在相关学科快速发展的新形势下，中文信息处理研究又将何去何从？本文将在简要归纳中文[①]信息处理研究所取得成就的基础上，分析当前的技术状况，直面存在的问题，并对未来发展的方向提出看法。希望本文指出的问题能够引起中国国内同行的关注，为未来的中文信息处理研究提供有益的参考。

二、中文信息处理研究的进展与现状

从 1949 年新中国成立前后的语言文字改革算起，到 20 世纪 70 年代中期开始的汉字编码和输入法研究，再到今天网络时代的全方位、大规模中文信息处理技术研究、开发和应用，中文信息处理走过了 60 多年的曲折历程。在半个多世纪的发展过程中几代人付出了艰苦的努力，一系列国家标准、规范和理论模型及应用系统应运而生。概括起来，这些成果可以归纳为如下几个方面（宗成庆、高庆狮 2008；宗成庆等 2009）：

（1）汉字简化与规范化工作基本完成，汉语拼音方案被国际标准化组织（ISO）接纳，汉语拼音正词法规则已成为国家标准。

（2）汉字编码、输入 / 输出、编辑、排版等相关技术已经解决，亚伟中文速录机和汉字激光照排、印刷系统已被大规模产业化应用。

（3）面向信息处理的汉语分词规范已经制定，以“综合型语言知识库”和知网（HowNet）[②]为典型代表的一批汉语资源库（包括语料库、词汇知识库、语法信息词典等）相继建成。

（4）汉语词语自动切分、命名实体识别、句法分析、词义消歧、语义角色标注和篇章分析等自然语言处理的基础问题得到全面研究和推进，一系列不断改进的模型和方法被相继提出，一大批高质量的研究论文发表在国际一流的学术会议和权威期刊上。

（5）机器翻译、信息检索、舆情监测、语音识别和语音合成等应用技术在众多互联网企业、国家特定领域和机构中得到实际应用，对推动国民经济发展、提高信息化服务水平和维护国家安全发挥了重要作用。

另外值得提及的是，由国家语言文字工作委员会发布的“中国语言生活绿皮书”[③]正在为国家语言文字工作方针政策提供参考，为语言文字研究者、语言文字产品研发者和社会其他人士提供语言服务，引领社会语言生活走向和谐（李宇明 2007）。

随着计算机和互联网技术的快速发展和普及，中文信息处理遇到了前所未有的大好时机。根据联合国对世界主要语种、分布与应用力调查的结果，世界十大语言依次是：英语、汉语、德语、法语、俄语、西班牙语、日语、阿拉伯语、韩语（朝鲜语）、葡萄牙语。而中国互联网络信息中心（CNNIC）发布的《第 21 次中国互联网络发展状况统计报告》表明，中国互联网上有 87.8% 的内容是文本。2014 年 7 月 21 日 CNNIC 发布的《第 34 次中国互联网络发展状况统计报告》显示，截止到 2014 年 6 月，中国网民规模达 6.32 亿。这些数据清楚地告诉我们这样一个不争的事实：无论从政治、经济、文化、军事和安全等政府关注的角度看，还是从商贸、旅游和信息服务等商业市场因素考虑，中文信息处理已经成为国际互联网和移动通信平台上获取和传递信息难以绕开的技术结点。不仅 IBM、

微软、谷歌等世界巨头公司投入了大量的人力和财力瞄准中国市场开展相关技术研究，斯坦福大学、宾夕法尼亚大学、加州大学伯克利分校等国际一流大学也为中文信息处理研究做出了卓著贡献，他们开发的汉语分词系统、句法分析器和命名实体识别工具等，以及 LDC 汉语语料库[④]（包括分词、句法树和篇章语料库等）得到广泛应用。这意味着，中文信息处理不仅是中国学者关注的问题，而且已经成为国际学术界和企业界共同研究的课题。

近年来中国的自然语言处理研究水平迅速提升，大陆学者在 HLT 相关领域的国际一流学术会议和期刊上发表的论文数量不断增长。图 1 是 2015 年第 53 届国际计算语言学学会年会与第 7 届自然语言处理国际联合会议（ACL-IJCNLP）[⑤]投稿和被接受的论文数量按国家或地区分布的直方图：

ACL-IJCNLP'2015 分为主会和专题研讨会两种。其中，主会是 ACL 大会的主体，它以论文质量高、录用率低、影响力大而著称。每年该会录用论文的数量通常被看作是一个国家或地区在本领域整体水平和实力的象征。ACL-IJCNLP' 2015 主会共收到长文投稿 692 篇，录用 173 篇；收到短文投稿 648 篇，录用 145 篇。也就是说，长文和短文合计投稿量为 1340 篇，录用 318 篇，录用率约为 23.7%。从图 1 可以看出，在 1340 篇投稿中第一作

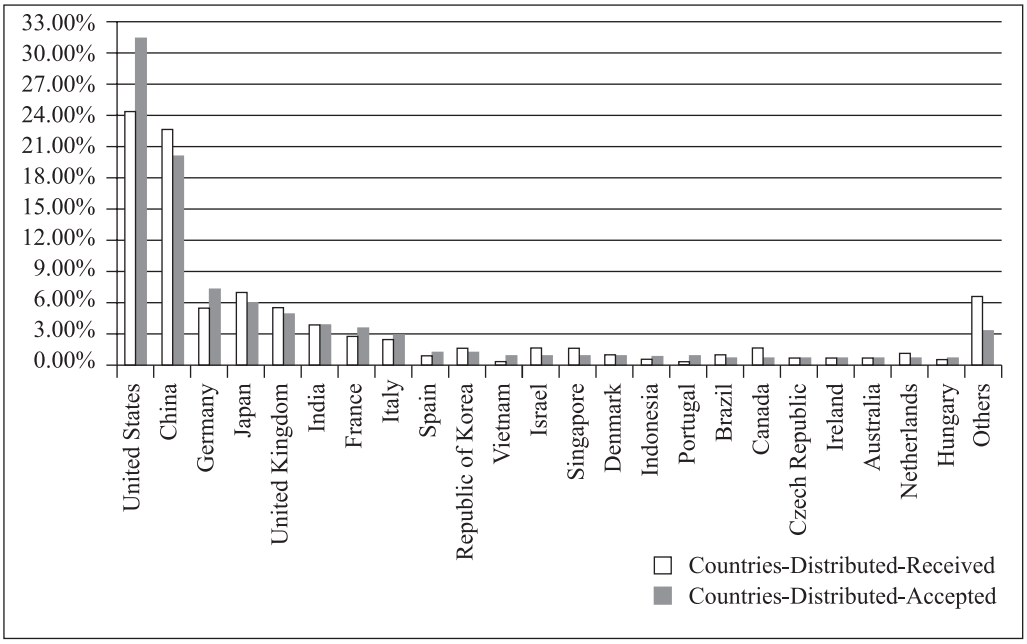

图 1　ACL–IJCNLP'2015 论文投稿的情况

者来自中国大陆的论文数量占到了 22.7%，仅次于美国（24.5%）。值得注意的是，即使是来自美国的投稿，第一作者也有可能是中国学者，包括众多留美的中国学生。据统计，在被录用的 318 篇论文中第一作者为中国人的论文数量约占 37.1%。换句话说，超过三分之一被录用的论文出自中国人之手。

除了 ACL 会议以外，国际计算语言学大会（International Conference on Computational Linguistics, 简称 COLING）[⑥]、国际人工智能联合会议（International Joint Conference on Artificial Intelligence, 简称 IJCAI）、ACM 信息检索大会（Special Interest Group on Information Retrieval, 简称 SIGIR）和 ACM 信息与知识管理国际会议（International Conference on Information and Knowledge Management，简称 CIKM）等其他相关的一流学术会议都已登陆中国。

与此同时，中国的自然语言处理人才队伍迅速成长，一批优秀的学者在国际一流学术会议和权威学术机构中担任重要职务。2013 年王海峰博士出任

ACL 主席，同年宗成庆当选国际计算语言学委员会[⑦]委员，2014 年和 2015 年吴华博士和宗成庆分别担任第 52 届和 53 届 ACL 大会程序委员会共同主席，2016 年赵世奇博士出任 ACL 秘书长。还有一大批优秀的中国学者在各类一流国际学术会议上担任组委会主席、领域主席、讲座主席和出版主席等。

毋庸置疑，中国学者已经成为国际 HLT 领域一支举足轻重的生力军。除了自身的努力以外，很重要的一个原因是国家综合实力的增强。国家不断增加的科研经费投入使更多的学者有机会走出国门，并把更多优秀的国外学者（包括学有所成的海外华人）请到中国来。当然，互联网技术起了非常重要的作用。借助于互联网，任何人都可以随时随地地查阅学术资料，实时了解和跟踪最新的国际研究动态，从而把握正确的研究方向。另外，以 IBM、微软公司、谷歌等为代表的国际大公司在中国大陆开设的研究机构，也对相关领域的技术发展和人才培养起到了推波助澜的作用。他们与中国科研机构和高校的密切交流与合作，使更多的青年学生有机会在高水平的技术平台上利用公司特有的计算资源和数据资源快速地学习和实践先进的技术。当然，这些公司是人才培养和市场开拓的受益者。

三、现状分析与问题思考

从中文信息处理发展现状来看，近 20 年是该领域迅速崛起和中国学者在国际舞台发挥作用的黄金时期。那么，这些丰硕的成果是否意味着中文信息处理的理论方法已经具有根本性的建树呢？

众所周知，自然语言处理方法有理性主义方法和经验主义方法两大流派。理性主义方法通常以乔姆斯基（Noam Chomsky）的语法理论为基础，建立基于规则和知识库的逻辑推理系统。而经验主义方法则以数理统计和信息论为基础，实现基于大规模语料库的统计机器学习方法。两种方法的融合正在成为人们探索的第三条路径。这些方法在目前的自然语言处理系统中都发挥了重要作用，但是，计算机要从中文信息“处理”走向真正的“理解”还有很长的路要走，在这条遥远的征途上至少需要跨越三条鸿沟：（1）建立符合中文（这里尤指汉语）语言特点的自然语言处理理论体系；（2）设计更加有效的机器学习算法和模型；（3）揭示和发现人类大脑理解语言的基本机理。

（一）现有中文信息处理方法的局限性

目前采用的中文信息处理方法和评价标准大都是从英语等西方语言的处理方法中借鉴过来的，无论是基于规则的方法，还是基于统计的方法，从来都没有针对汉语本身的特点“量身定做”。例如，传统的自然语言处理方法通常从词法分析（汉语词语自动切分）开始，到句法分析、语义分析，分阶段逐步进行，不同层次的任务往往是独立完成的。句法分析（syntactic parsing）是其中的关键环节，其任务是将给定的句子自动解析成完整的句法分析树。它的基本假设是每一个句子的句法结构都能够用一棵完整的句法分析树表示，如图 2 所示。

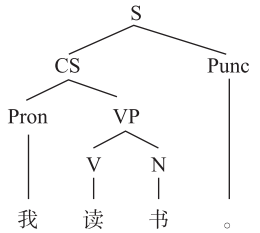

图 2　句子“我读书。”的句法分析树

但是，这一假设对于汉语而言往往不能成立，至少是非常苛刻的。汉语句子中通常不使用标识结构信息的专用词汇（如英语复句中的 which, that, where 等引导词），是一种语义驱动的松散结构，句法和语义之间存在着千丝万缕的关系，而且汉语中标点的使用也不像英语那样有严格的限制。例如：

（1）我喜欢在春天去观赏桃花，在夏天去欣赏荷花，在秋天去观赏红叶，但更喜欢在冬天去欣赏雪景。

这是一个典型的流水句。根据我们对随机抽取出的 4431 个长度超过 20 个词的句子的统计，有 1830 个流水复句，占全部长句的 41.3%（李幸、宗成庆 2006）。流水句结构看起来比较松散，但语义上却有紧密的联系。如果非要用一棵完整的句法树

表示这种句子的结构，不仅在实现上非常困难，而且对达到语言理解的目标几乎没有太多帮助。过去几十年里，人们提出了大量自动句法分析的算法，目前比较著名的句法分析工具有：Collins Parser、Bikel Parser、Charniak Parser、Berkeley Parser、Stanford Parser、MST Parser、MaltParser 和 MINIPAR Parser 等。但这些系统在规范的汉语文本上最好的句法分析性能（短语准确率）也只有 86% 左右，而日语和英语的句法分析性能已经超过 90%。即使 C. Dyer 和 M. Ballesteros 等人近期实现的基于神经网络的句法分析方法的性能得到了进一步提升（Ballesteros *et al*. 2015；Dyer *et al*. 2015），汉语句法分析器的性能仍然比英语的低 5 个百分点左右。

对于篇章结构分析来说，目前广泛采用的篇章理论包括修辞结构理论、中心理论、脉络理论、篇章表示理论和言语行为理论等（宗成庆 2013），而这些理论无一例外地来自西方语言学。汉语的篇章结构与英语有明显的区别，这是大家所共知的事实。根据我们对 2016 年国际计算自然语言学习会议（Conference on Computational Natural Language Learning，简称 CoNLL）发布的汉英篇章论元关系分析评测任务的语料统计，汉语中非显式的篇章单元之间的关系占到了 78.3%，远远超过了英语篇章中 54.5% 的比例。汉语中篇章单元之间可使用的连接词有 385 个之多，而英文中只有 100 个左右（Kang *et al*. 2016）。而且汉语中的标点逗号可以隐含地表示某种篇章单元关系，例如表示前后两个单元之间隐含的转折、让步、因果等关系，而英语的标点不具备这样的功能。所有这些差异都清楚地提醒我们，汉语需要建立自己的篇章分析理论。

值得庆幸的是，国内已有专家在汉语篇章分析理论研究方面进行卓有成效的探索，如宋柔（2012）提出的“广义话题结构理论”、王德亮（2004）研究的“篇章向心理论”等，但离建立相对成熟和完善的汉语篇章理论体系还有较远的距离。

另外，汉语中的指代消歧也是中文信息处理面临的棘手问题。请看如下两个例句：

（2）夫人穿着很得体，举止优雅，左臂上挂着一个暗黄色的皮包，右手领着一只白色的小狗，据说是京巴。

（3）夫人穿着很得体，举止优雅，左臂上挂着一个暗黄色的皮包，右手领着一只白色的小狗，据说是局长的太太。

在这两个句子中除下划线标识的部分以外，其余部分完全一样，但“据说”的所指完全不同，一个是指“小狗是京巴”，而另一个则是指“夫人是局长的太太”。这种表达方式在英文中是不可能出现的。

综上所述，不同语言具有不同的特点，无论在词法、句法、语义等不同的层面上，还是在词汇、短语、句子和篇章等不同的语言单位上，有共性，也有差异，尤其语义与语言的文化背景密切相关。我们认为，不存在与语言无关的自然语言处理方法和全世界语种通用的自然语言处理理论体系。最终要解决中文信息处理的问题，使其真正实用化，必需建立适合中文语言特点的理论体系。

（二）现有机器学习方法的缺陷

20 世纪 80 年代末期、90 年代初期以来，统计机器学习方法逐渐兴起，并成为当前自然语言处理领域的主流方法。其基本思路是，基于大规模人工标注的语料样本建立数学模型，通过调试模型的参数使其达到最优（这一过程称作模型的训练过程）。所建的数学模型就像一个小学生，标注的语料则是老师为学生提供的样例，而训练过程则类似于老师教小学生如何按照样例学习句子分析方法或完成其他任务的过程。最终小学生的成绩如何取决于学生本身的能力、样例规模的大小和学生学习的技巧，对应地，统计模型的性能好坏取决于数学模型本身、训练样本规模的大小和模型参数的调试情况。

序列标注方法是自然语言处理中常用的一种典型的机器学习方法。以汉语自动分词为例，序列标注方法的基本思路是：每个“字”（包括字符、数字、标点等文本中出现的任何符号）只有 4 种可能的身份出现在文本中，即词首字（B）、词尾字（E）、词中间字（M）和单字词（S）。对于给定的文本，如果能够对每个“字”打上一个标签（B、

E、M 或 S 中的任意一个），那么分词任务就完成了。被标记为 B 和 E 的“字”及其之间标以 M 的“字”（如果有的话）构成一个分词单位，被标记为 S 的“字”独立成词。例如，句子“我喜欢读书。”的序列标注结果为：我 /S 喜 /B 欢 /E 读 /S 书 /S 。最终的分词结果就是：我 / 喜欢 / 读 / 书 / 。

在为每个“字”打标签的过程中，依据当前“字”的上下文计算对当前“字”贴上某种标签的条件概率，选择概率最大的候选标签。实际上这是一种通过上下文分类进行标签选择的方法，称为区分式方法。确定上下文多大范围内、哪些因素可作为计算概率的条件的过程，则称作特征选择。

类似地，命名实体识别、语块识别和篇章单元识别等，都可采用这种方法实现。

统计方法的优点不言而喻，它避免了基于规则的方法中由于人工编写规则的主观性因素可能导致的语言现象覆盖面小甚至错误的情况。有些自然语言处理任务（如机器翻译）并不需要人工标注语料，这就大大地减少了系统对人的依赖性，极大地提高了系统开发的效率。这也是统计方法备受青睐的重要原因之一。但是，目前的统计方法仍然存在若干问题和不足。归纳起来，这些缺陷包括：

1. 模型性能过于依赖训练样本

根据上面的介绍，训练样本的质量和规模对模型最终的性能起着至关重要的作用。一般而言，如果样本的规模太小，或者样本的质量太差，模型的性能肯定不好。人工标注大规模训练样本同样是一件艰苦的工作，而且标注样本往往难以随着语言使用情况的变化而自动调整。即使机器翻译等任务不需要人工标注的训练样本，但仍然需要样本的数量达到足够的规模，这对于有些领域或语言对来说是无法做到的。例如，波斯语与汉语之间的自动翻译系统就很难收集到大规模波斯语与汉语句子级双语平行语料，即使在新闻等公共领域，收集几十万句对都是困难的，更不必说在某些特定领域。

2. 固化的模型参数导致模型无法处理“陌生”的语言现象

在统计方法中模型一旦被训练完成，参数是被固化的，对于超出特征预设范围的语言现象完全无能为力。例如，在词义消歧任务中我们通常根据歧义词出现的上下文建立分类模型，由上下文决定词语的语义。以“打”字的词义消歧为例，“打”字做实词用时有多个含义，“打毛衣”“打电话”和“打篮球”等不同表达中“打”字的含义各不相同，因此可以设定“打”字前后一定范围内的上下文词作为分类特征构建分类模型。假如设定上下文窗口范围为 ±1（即在当前词前后一个词的窗口范围内），大多数情况下“打”字的含义都可以区分出来。但是，对于超出窗口范围的情况模型便无能为力了。例如，在句子“张三打了一壶绍兴老酒。”中，“打”字与“老酒”之间间隔 4 个词，这就很可能导致模型误判“打”的词义。

3. 缺乏领域自适应能力

模型对训练语料所在领域的语言现象处理可能表现出较好的性能，但一旦超出领域范围或测试集与训练样本有较大差异，模型性能将大幅度下降。例如，在标注的大规模《人民日报》分词语料上训练出来的汉语词语自动切分模型的准确率可达 96% 左右，甚至更高，但在微博等非规范文本基础上训练出的分词性能至少要低 5 个百分点左右。在 LDC 汉语树库上训练出来的句法分析系统准确率可达 86% 左右，但在非规范网络文本上的分析准确率只有 60% 左右（宗成庆 2013）。统计模型对领域自适应能力的缺乏严重制约了该方法的应用。

4. 难以通过人机交互自动完成参数更新

人类在语言学习中可以通过人际之间和人与自然界之间的不断交互主动学习新的知识（包括语言知识和生活常识等），从而不断提高语言学习和理解的能力，但对于目前的统计自然语言处理系统而言却无法做到这一点。如何使系统通过人机交互过程，自动根据语用信息判别和提取有用的知识，完成模型参数的自动更新，以达到模型性能不断提高的效果，到目前为止还需探索。

5. 常识学习与归纳推理能力亟待提高

现有的统计学习方法在局部问题求解上可以达到较好的技术水平，但是在整体归纳和全局抽象方面却显得力不从心。例如，有如下一则新闻报道：

张小五从警 20 多年来，历尽千辛万苦，立下

无数战功，曾被誉为孤胆英雄。然而，谁也未曾想到，就是这样一位曾让毒贩闻风丧胆的铁骨英雄竟然为了区区小利而精神崩溃，悔恨之下昨晚在家开枪自毙。

对这则新闻目前的词语自动切分准确率可达96%以上，命名实体（人名“张小五”）识别和句间关系分析（关键词“然而”引起的转折），甚至语义角色标注等，都没有太大问题，准确率至少可达85%以上。但是，对于一个自动问答系统来说，要正确地回答“张小五是什么警察？死了没有？”等，恐怕非常困难，因为它无法建立起“毒贩”与“缉毒警察”之间的对应关系，也不会知道“自毙”与“死亡”的必然联系。当前中文信息处理系统的常识学习和归纳推理能力亟待提高。

宏观上讲，统计是一种“赌博”方法，决策的依据是概率值大小，一定程度上有点“撞大运”的味道。其基本假设是：样本中蕴含着全部与特定自然语言处理任务相关的知识，而且处理任务（测试集）与训练样本符合同样的规律，只要有足够多的训练样本，模型就能够学习到相应的知识，并对待处理集进行正确的分析。且不说如何拥有“足够多”、多到多大规模的训练样本，只就模型本身的学习能力、区分能力和自适应能力等方面而言，还远无法与人脑的自然语言理解能力相比较。

（三）自然语言研究需要与脑神经科学和认知科学相结合

近年来，类人智能和类脑计算备受瞩目，尤其AlphaGo围棋系统战胜人类选手以来，人工智能被再度推向媒体舆论和学术研究的风口浪尖。但是，对于人脑是如何完成自然语言理解过程的，比如为什么一个三岁的儿童在学习一个新的词项时，父母只需做简单的解释，给出一两个例子，孩子就可以理解并使用所学的词项，而且基本不会用错，根本不需要大量的训练样本，目前尚无法给出非常清楚、合理的解释。

近年来基于神经网络的深度学习方法备受推崇，它在某种意义上的确模拟了人脑的认知功能，但是，这种方法只是对神经元结构和信号传递方式给出的形式化数学描述，并非是基于人脑的工作机理建立起来的数学模型，同样难以摆脱对大规模训练样本的依赖。

目前人们只是在宏观上大致了解脑区的划分和在语言理解过程中所起的不同作用，但在介观和微观层面，语言理解的生物过程与神经元信号传递的关系，以及信号与语义、概念和物理世界之间的对应与联系等，都是未知的。如何打通宏观、介观和微观层面的联系并给出清晰的解释，将是未来需解决的问题。从微观层面进一步研究人脑的结构，发现和揭示人脑理解语言的机理，借鉴或模拟人脑的工作机理并建立形式化的数学模型才是最终解决自然语言理解问题的根本出路。这需要与语言学家、脑神经科学家和认知科学家的共同努力和协作。

30多年来自然语言处理研究成绩斐然，但中文信息处理的理论研究和技术创新却有弱化之势。近年来中文信息处理技术性能的提高在很大程度上源自数据规模的扩大和计算机硬件性能的提高，在理论方法和数学模型上并没有太多的建树，真正面向汉语的计算理论和实现技术似乎并不多见。

在ACL-IJCNLP’2015录用的318篇论文中，115篇是关于深度学习方法的，约占36.2%。而深度学习方法的热度仍在持续升高，2016年会议录用的论文中与深度学习方法相关的论文比例再创新高。但是，如此大量的论文中，有多少还在关注汉语呢？据对ACL-IJCNLP’2015投稿论文的统计，在形态分析专题领域的28篇投稿（包括长文和短文）中，关于中文词语切分（中文信息处理的经典问题）的论文仅有6篇，其中包括一篇关于藏语分词的论文，而句法分析专题领域的全部108篇投稿中，只有22篇是研究汉语句法分析方法的。所有这些稿件都无一例外地采用了统计方法，它们的贡献基本是在别人提出的模型的基础上，做些特征选择和参数调整等方面的改进工作，在中文信息处理的理论创新方面鲜有建树。

近几年来随着国内指标（SCI/SSCI论文数量、引用次数、高被引论文数等）导向的各种学术评估愈演愈烈，很多研究开始一味地跟踪热点、追逐新潮，只是为了早出成果、快发论文，而最终忘记了解决中文语言理解这一问题的根本目标。这正是我

们担忧的关键所在。

四、结束语

过去 60 多年中，中文信息处理取得了令人振奋的成果，尤其在统计方法成为主流方法之前，老一代学者创建了一系列面向汉语特点的理论方法和实用技术，并为中文语言资源库建设做出了卓越贡献，人才培养和队伍建设成就显著。而当统计方法一统天下之后，对语言学特性和认知规律的研究在自然语言处理领域并没有得到应有的重视。其实，早在 10 多年前有关专家就已经通过脑功能成像技术研究证明，汉英两种语言的名词和动词在人脑中的表征并不完全一样（Li *et al*. 2004）。如何针对汉语自身的特点和规律建立专用的模型和算法，恐怕才是最终解决汉语理解问题的正确出路。

总体而言，目前计算机处理自然语言的能力仅仅停留在“处理”层面，还远不能达到“理解”的水平，未来的任务艰巨而充满挑战。跟踪国际前沿是每一位科研工作者应有的素质和理念，但是，在学习和跟踪国际先进技术的同时，无论如何都不应该丧失以解决我们母语问题为目标的创新意识。

注　释

① 本文接下来讨论的中文信息处理研究现状和趋势，主要指汉语信息处理的技术状况。

② 参见 http://www.keenage.com/html/c_index.html。

③ 第一部“中国语言生活绿皮书”——《中国语言生活状况报告（2005）》于 2006 年 9 月 18 日正式出版。此后每年发布一次，持续至今。

④ https://www.ldc.upenn.edu/。

⑤ ACL 是国际计算语言学学会（Association for Computational Linguistics）的缩写。该学会成立于 1962 年，第一届 ACL 年会于 1963 年 8 月在美国召开，目前是本领域最具影响力和权威性最高的顶级学术会议，被中国计算机学会（CCF）认定为 A 类会议。第 53 届 ACL 年会与亚洲自然语言处理联合会（The Asian Federation of Natural Language Processing，简称 AFNLP）第 7 届自然语言处理国际联合会议（The 7th International Joint Conference on Natural Language Processing，简称 IJCNLP）于 2015 年 7 月 26 日至 31 日在北京举办，会议名称通常简写为：ACL-IJCNLP’ 2015。

⑥ COLING 创办于 1965 年，每两年召开一次，是本领域最具权威性和影响力的一流学术会议之一。

⑦ International Committee on Computational Linguistics，简称 ICCL。网址：http://nlp.shef.ac.uk/iccl/。

参考文献

李　幸、宗成庆　2006 《引入标点处理的层次化汉语长句句法分析方法》，《中文信息学报》第 4 期。

李宇明　2007 《关于〈中国语言生活绿皮书〉》，《语言文字应用》第 1 期。

宋　柔　2012 《汉语篇章广义话题结构研究》，北京语言大学语言信息处理研究所研究报告。

王德亮　2004 《汉语零形回指解析——基于向心理论的研究》，《现代外语》第 4 期。

宗成庆　2013 《统计自然语言处理》，北京：清华大学出版社。

宗成庆、曹右琦、俞士汶　2009 《中文信息处理 60 年》，《语言文字应用》第 4 期。

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

Ballesteros, Miguel, Chris Dyer, and Noah A. Smith. 2015. Improved Transition-Based Parsing by Modeling Characters instead of Words with LSTMs. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*).

Dyer, Chris, Miguel Ballesteros, Wang Ling, Austin Matthews, and Noah A. Smith. 2015. Transition-Based Dependency Parsing with Stack Long Short-Term Memory. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (*ACL-IJCNLP*).

Kang, Xiaomian, Haoran Li, Long Zhou, Jiajun Zhang, and Chengqing Zong. 2016. An End-to-End Chinese Discourse Parser with Adaptation to Explicit and Non-Explicit Relation Recognition. *Proceedings of the SIGNLL Conference on Computational Natural Language Learning* (*CoNLL*).

Li, Ping, Zhen Jin, and Li Hai Tan. 2014. Neural Representations of Nouns and Verbs in Chinese: An fMRI Study. *Neuroimage* 21, 1533-1541.

责任编辑：戴　燃