汉语自动分词研究评述

孙茂松 邹嘉彦

孙茂松，北京 清华大学计算机系
邹嘉彦，香港 香港城市大学 语言资讯科学研究中心

提要 本文首先阐述了汉语自动分词研究的现实性和可能性，接着围绕该研究中的三个基本问题（切分歧义消解、未登录词处理和语言资源建设）展开了重点讨论，并扼要评介了十几年来产生的各种方法。最后就这个领域未来的研究要点发表了一些个人意见。

关键词 中文信息处理　汉语自动分词 切分歧义消解　未登录词处理　语言资源建设

1. 汉语自动分词的现实性与可能性

众所周知，中文文本没有类似英文空格之类的显式表标示词的边界标志。汉语自动分词的任务，通俗地说，就是要由机器在中文文本中词与词之间自动加上空格。一提到自动分词，通常会遇到两种比较典型的质疑。一种质疑是来自外行人的：这件事看上去平凡之极，好像一点儿也不“热闹”，会有什么用呢？另一种质疑则是来自内行人的：自动分词研究已经紧锣密鼓地进行了十几年了，可到现在也未见一个经得起考验的系统推出来（与此形成鲜明对照的是，日语同样也存在分词问题，但已经有了圈内人士广泛认同的日语分词系统），这几乎成了中文信息处理中一个“永恒”的话题，那么，到底还有没有希望搞出真正意义上的“门道”来？

第一种质疑关心的是自动分词的现实性问题，其答案是十分明确的。当前的大环境令人鼓舞：中国正在向信息化社会迅速前进，其突出表征是 Internet 上中文网页的急剧增加和中文电子出版物、中文数字图书馆的迅速普及。以非受限文本为主要对象的中文自然语言处理研究于是也水涨船高，重要性日益显著。而汉语自动分词是任何中文自然语言处理系统都难以回避的第一道基本“工序”，其作用是怎么估计都不会过分。只有逾越这个障碍，中文处理系统才称得上初步打上了“智能”的印记，构建于词平面之上的各种后续语言分析手段才有展示身手的舞台。否则，系统便只能被束缚在字平面上，成不了太大气候。具体来说，自动分词在很多现实应用领域（中文文本的自动检索、过滤、分类及摘要，中文文本的自动校对，汉外机器翻译，汉字识别与汉语语音识别的后处理，汉语语音合成，以句子为单位的汉字键盘输入，汉字简繁体转换等）中都扮演着极为重要的角色（Wu Z.M.and Tseng G.1993;Wu Z. M. and Tseng G.1995; Nie J.Y.and Brisebois M.et al.1996;Sun M. S. andLin F.Z.,et al.1996）。我们举两个例子直观说明一下。

［文本检索］

设文本 A 含句子（1a）而文本 B 含句子（1b）：

（1）a.和服｜务｜于三日后裁制完毕，并呈送将军府中。

b.王府饭店的设施｜和｜服务｜是一流的。

显然，文本 A 讲的是日本“和服”，文本 B 则与酒店的“服务”有关，两者风马牛不相干。如果不分词或者“和服务”分词有误，都会导致荒谬的检索结果。

［文语转换］

注意句子（2a）、（2b）中的“查金泰”：

（2）a.他们是来｜查｜金泰｜撞人那件事的。

b.行侠仗义的｜查金泰｜远近闻名。

句子（2a）中“查”为动词，应读 cha，句子（2b）中则为姓氏，应读 zha。

第二种质疑直指自动分词的可能性问题。虽然迄今为止我们尚不能下一个完全肯定的结论，但经过圈内学者十几年不懈的探索，这个答案的轮廓还是大体凸显出来了。毕竟词平面上的研究与句法平面和语义平面相比照，本身难度要小得多，并且无论是在计算语言学方面还是在普通语言学方面，所取得的成果也要成熟、扎实得多。现有的工作积累已经达到了可以厚积薄发的程度。如果说面向非受限文本的汉语句法、语义自动分析还是可望而不可即的话，那么，面对相同对象的汉语自动分词，则距凯歌初奏只有几步之遥了（当然即使达到了那个目标，也还不是功德圆满）。Sproat R. and Shih C.L.,et al.（1996）及 Sun M. S.and Shen D.Y.,et al.（1997 ）的汉语自动分词原型系统已初具处理非受限文本所需的种种功能，他们沿着正确方向跨了一大步。

本文的重点是第 2 节，将集中讨论汉语自动分词中的基本问题， 并扼要评介十几年来产生的各种方法（文后的参考文献基本囊括了这一领域比较有代表性的论文）。第 3 节则就今后的研究要点发表一些个人意见。

2. 汉语自动分词中的基本问题和主要解决方法

2.1 切分歧义及其处理方法

2.1.1 切分歧义的基本类型

切分歧义是汉语自动分词研究中的一个“拦路虎”。梁南元（1987）最早对这个现象进行了比较系统的考察。他定义了两种基本的切分歧义类型：

定义 1 汉字串 AJB 被称作交集型切分歧义，如果满足 AJ、JB 同时为词（A、J、B 分别为汉字串）。此时汉字串 J 被称作交集串。

［例］交集型切分歧义：“结合成”

（3）a.结合｜成

b.结｜合成

其中 A＝“结”，J＝“合”，B＝“成”。

定义 2 汉字串 AB 被称作多义组合型切分歧义，如果满足 A、B、 AB 同时为词。

［例］多义组合型切分歧义：“起身”

（4）a.他站｜起｜身｜来。

b.他明天｜起身｜去北京。

对交集型切分歧义，他还定义了链长：

定义 3 一个交集型切分歧义所拥有的交集串的集合称为交集串链，它的个数称为链长。

如，交集型切分歧义“结合成分子”、“结合”、“合成”、“成分”、“分子”均成词，交集串的集合为｛“合”，“成”，“分”｝，链长为 3。

这些定义所涉及的几个概念，基本刻画了汉语切分歧义的结构特点，因而一直沿用下来。

梁南元（1987）对一个 48,092 字的自然科学、社会科学样本进行了统计：交集型切分歧义 518 个，多义组合型切分歧义 42 个。据此推断，中文文本中切分歧义的出现频度约为 1.2 次／100 字，交集型切分歧义与多义组合型切分歧义的出现比例约为 12∶1。

有意思的是，刘挺、王开铸（1998）的调查却显示了与梁南元截然相反的结果：中文文本中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。造成这种情形的原因在于，定义 2 有疏漏。Sun M. S.and Benjamin K.T.（1995）猜测， 加上一条限制才真正反映

了梁的本意：

定义 2' 汉字串 AB 被称作多义组合型切分歧义，如果满足（1）A、 B、AB 同时为词；（2）中文文本中至少存在一个前后语境 C，在 C 的约束下，A、B 在语法和语义上都成立。

例如，汉字串“平淡”符合定义 2，但不符合定义 2' （因为“平｜淡”在文本中不可能成立）。刘、王将“平淡”计入了多义组合型切分歧义，梁并未计入。由于符合定义 2 的汉字串数量远远大于符合定义 2'的汉字串数量，出现“乾坤颠倒”也就不足为怪了。

仔细分析一下，定义 1 和定义 2 都是完全从机器角度加以形式定义的，定义 2'则增加了人的判断。孙茂松、黄昌宁等（1997）认为， 定义 2 中给出的名称“多义组合型切分歧义”是不太科学的（实际上，某些交集型切分歧义也是多义组合的），易引起混淆，与“交集型”这个纯形式的名称相呼应，称作“包孕型”或者“覆盖型”可能更恰当。

董振东（1997）采用了另外一套名称：称交集型切分歧义为“偶发歧义”，称多义组合型切分歧义为“固有歧义”。“两者的区别在于：造成前者歧义的前后语境是非常个性化的、偶然的、难以预测的”，“而后者是可以预测的”。这个表述相当深刻地点出了两类歧义的性质，耐人寻味。但名称的准确性仍有可斟酌之处。

	视角	真歧义类	伪歧义类
交集型切分歧义	定义	定义 1	
	性质	偶发歧义	
	数量	少量	大量
	例子	地面积,和平等,的确定	和软件,在建设,部门对
覆盖型切分歧义	定义	定义 2'	定义 2 扣除定义 2'的外延
	性质	固有歧义	偶发歧义
	数量	少量	大量
	例子	起身,把手,一行,三角	平淡,高度,词条,结论

表 1 切分歧义类型表

孙茂松、左正平（1998）指出，切分歧义应进一步区别“真切分歧义”和“伪切分歧义”。譬如：同属交集型，“地面积”为真歧义（“这几块｜地｜面积｜还真不小”“地面｜积｜了厚厚的雪”），“和软件”则为伪歧义（虽然存在两种不同的切分形式“和软｜件”和“和｜软件”，但在真实文本中，无一例外地应被切分为“和｜软件”）；同属覆盖型，“起身”为真歧义，“平淡”则为伪歧义。

归纳以上论述，本文整理出一张切分歧义类型表（见表 1）， 希望对澄清概念上流传已久的混乱有所帮助。

关于切分歧义，还有两点基本观察：

1）根据孙茂松、左正平（1998）对一个 1 亿字语料库的穷尽式统计，交集型切分歧义长度变化范围为 3～14 个字（“提高人民群众生活水平息息相关”），交集串长度变化范围为 1～3 个字（“如箭在弦上”），链长变化范围为 1～9 个字（“中国人民生活水平和美化”）；

2） 交集型和覆盖型常常会相互纠缠在一起，这就更增加了变数。如图 1 中的“提高人民生活水平”共可衍生出 19 种可能的形式切分（弧线表示可成词）。

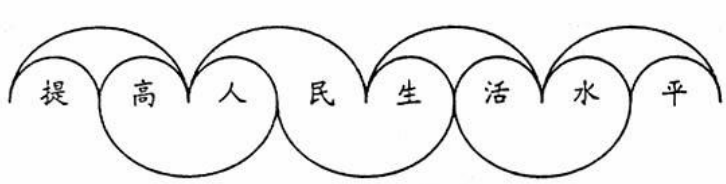

图1 若干基本类型的混合

2.1.2 切分歧义的检测与消解

切分歧义处理包括两部分内容：（1）切分歧义的检测；（2）切分歧义的消解。这两部分在逻辑关系上可分成两个相对独立的步骤。

首先谈谈切分歧义的检测问题。“最大匹配法”（精确的说法应该叫“最长词优先匹配法”）是最早出现、同时也是最基本的汉语自动分词方法，1963 年就在《文字改革》杂志上被介绍过（刘涌泉 1988）。刘源、梁南元（1986）首次将这个方法大规模应用到汉语自动分词系统中。依扫描句子的方向，又分正向最大匹配 MM（从左向右）和逆向最大匹配 RMM（从右向左）两种。 最大匹配法实际上将切分歧义检测与消解这两个过程合二为一，对输入句子给出唯一的切分可能性，并以之为解。据梁南元（1987）的实验结果，在词典完备、没有任何其它知识的条件下，最大匹配法的错误切分率为 1 次 / 169 字～1 次 / 245 字，并且具有简单、快速的优点。Guo J.（1997）更对最大匹配法的工作原理作了严格的形式解释。此外，揭春雨、刘源等（1989）比较完整地分析了最大匹配法的结构及其时间效率。

从最大匹配法出发导出了“双向最大匹配法”，即 MM＋ RMM。SunM.S. and Benjamin K.T.（1995）注意到：汉语文本中 90.0％左右的句子，MM 和 RMM 的切分完全重合且正确，9.0％左右的句子 MM 和 RMM 切分不同，但其中必有一个是正确的（歧义检测成功），只有不到 1.0 ％的句子，或者 MM 和 RMM 的切分虽重合却是错的，或者 MM 和 RMM 切分不同但两个都不对（歧义检测失败）。这正是双向最大匹配法在实用中文信息处理系统中得以广泛使用的原因所在。

显然，双向最大匹配法存在着切分歧义检测盲区。针对切分歧义检测，另外两个有价值的工作是，王晓龙、王开铸等（1989）的“最少分词法”（歧义检测能力较双向最大匹配法要强些，产生的可能切分个数仅略有增加）和马晏（1996）的“全切分法”（穷举所有可能切分，实现了无盲区的切分歧义检测，但代价是导致大量的切分“垃圾”）。这个问题直到今天也没有完全解决--如果把双向最大匹配法视作一个极端（最简单）而全切分法视作另一个极端（最繁杂）的话，我们的目标应该是：在这两极之间寻找一个“删繁就简”的折衷方案，既（几乎）排除了检测盲区，又抑制了可能切分个数的无理膨胀。

接下来讨论切分歧义的消解问题。十几年来，研究人员几乎调动了人工智能领域所有“时髦”的计算手段来对付切分歧义，堪称“八仙过海，各显神通”。典型的手段包括：“松弛法”（Fan C.K. and Tsai W. H. 1988），“扩充转移网络”（黄祥喜 1989）， “短语结构文法”（梁南元 1990；姚天顺、张桂平等 1990；Yeh C.L. and Lee H. J. 1991；韩世欣、王开铸 1992），“专家系统”（徐辉、 何克抗等 1991），“神经网络”（徐秉铮、詹剑等 1993 ）， “有限状态自动机”（Sproat R. and Shih C.L., et al. 1996），“隐 Markov 模型”（LaiB.Y. and Sun M.S., et al. 1997；沈达阳、孙茂松等 1997a； 孙茂松、左正平等 1999a），“Brill 式转换法”（Palmer D.D.1997）等。 这些新的探索体现了切分歧义消解计算的不同侧面，在一定范围内取得了各自的效果，但从总体上看，还都嫌粗糙；或者虽然研究比较充分，但模型本身的计算能力偏弱；或者仅仅搭起了一个框架，浅尝辄止；或者实验规模太小，说服力不足。

通过不断的实践，人们越来越深刻地认识到，如果没有足够的语言知识作为支撑，再先进的计算手段也只能是“银样蜡枪头--中看不中用”。切分歧义消解经历了一个由浅及深、

由简单到复杂的语言知识利用的演变过程：

1 ）一些系统（尤其是早期系统）主要利用词频以及语素（自由抑或约束）、切分歧义表层结构等简单信息（Fan C.K. and Tsai W. H.1988；李国臣、刘开瑛等 1988；王永成、苏海菊等 1990；Chen K. J.and Liu S.H. 1992；马晏 1996）。

2）Sun M.S. and Lai B.Y., et al. (1992) 揭示了音节信息在自动分词中的作用。

3）何克抗，徐辉等（1991）断言，95.0 ％左右的切分歧义可以借重句法以下的知识解决，只有 5.0％必须诉诸语义和语用知识。 基于规则的几个分词系统（黄祥喜 1989；梁南元 1990；姚天顺、张桂平等 1990；Yeh C.L. and Lee H.J.1991；韩世欣、王开铸 1992；徐辉、何克抗等 1991）都自觉或不自觉地受到这个结论的支配，切分歧义消解主要诉诸词法与句法规则。存在的缺陷是，规则集由人凭主观编制而成，会受到系统性、有效性、一致性、可维护性等“天然”问题困扰。

4）为克服人工句法规则集的弊端， 一些研究人员开始尝试另一种途径一句法统计。Lai B.Y. and Sun M.S., et al.(1992; 1997) 、Chang C.H. and Chen C.D.(1993)、白拴虎（1995）等将自动分词和基于 Markov 链的词性自动标注技术结合起来，利用从人工标注语料库中提取出的词性二元统计规律来消解切分歧义（词性标注对分词有反馈作用，两者并行）。初步实验（Lai B.Y.and Sun M.S., et al.1997）表明，同“先做最大匹配分词，再作词性自动标注”（词性标注对分词无反馈作用，两者串行）相比，这种做法的分词精度和词性标注精度分别提高了 1.3％和 1.4％。

（5）他俩儿谈恋爱是从头年元月开始的。

切分 a.… 是 ｜ 从头 ｜ 年 ｜ 元月 ｜ …

动词 副词 时间量词 时间词

切分 b.… 是 ｜ 从 ｜ 头年 ｜ 元月 ｜ …

动词 介词 时间词 时间词

虽然“从头”、“年”的词频之积大于“从”、“头年”的词频之积，但词性序列“动词＋副词＋时间量词＋时间词”的概率远小于“动词＋介词＋时间词＋时间词”的概率，所以选择切分 b 作为结果。

5）Wu A.D. and Jiang Z.X.(1998)走得更远。他们相信， 多数情况下，切分歧义可以在输入句子的局部范围内得到妥善处理，但有些比较复杂的切分歧义，必须在句中更大的范围内才能解决。当遇到这种情况时，他们的系统将对句子做完整的句法分析，如果分析失败，则拒绝相应的切分：

（6）在这些企业中国有企业有十个。

切分 a. 在｜这些｜企业｜中｜国有｜企业｜有｜十｜个｜。

切分 b. 在｜这些｜企业｜中国｜有｜企业｜有｜十｜个｜。

切分 b 得不到可信的句法树，因而被拒绝。

当然，分析的层次越深，机器对知识库质量、规模等的依赖性就越强，所需要的时间、空间代价也就越大(况且面向真实文本的汉语句法分析器在可预期的将来几乎没有实现的可能，这也是应予考虑的因素)。有时不免使人产生一种陷入因果循环般的困惑：消解切分歧义这一相对“简单”的任务似乎不得不倚仗比分词本身困难得多的句法分析才得以完成。这个“悖论”里面其实蕴涵着深刻的“潜台词”，对中文自然语言处理系统的设计很有启发，囿于篇幅，这里就不展开了。

另一个值得一提的工作是，孙茂松、左正平等（1999b）发现， 从一个 1 亿字真实汉语语料库中抽取出的前 4，619 个高频交集型歧义切分覆盖了该语料库中全部交集型歧义切分的 59.20 ％(它们对另一个完全独立的语料库的覆盖率为 50.85％， 说明高频交集型切分的分布相对不同的领域是比较稳定的)，其中 4，279 个属伪歧义（如“和软件”、“充分发挥”、

“情不自禁地”），覆盖率高达 53.35％。 鉴于伪歧义的消解与上下文无关，于是他们提出了一个简单却很有效的策略：对伪歧义型高频交集型歧义切分，可以把它们的正确（唯一）切分形式预先记录在一张表中，其歧义消解通过直接查表即可实现。本质上，这是一个基于记忆的模型。

2.2 未登录词及其处理

未登录词大致包含两大类：1）新涌现的通用词或专业术语等；2）专有名词，如中国人名、外国译名、地名、机构名（泛指机关、团体和其它企事业单位）等。前一种未登录词理论上是可预期的，能够人工预先添加到词表中（但这也只是理想状态，在真实环境下并不易做到）；后一种未登录词则完全不可预期，无论词表多么庞大，也无法囊括。

孙茂松、邹嘉彦（1995）指出，真实文本中（即便是大众通用领域），未登录词对分词精度的影响超过了歧义切分。未登录词处理在实用型分词系统中占的份量举足轻重。

对第一种未登录词的处理，一般是在大规模语料库的支持下，先由机器根据某种算法自动生成一张候选词表（无监督的机器学习策略），再人工筛选出其中的新词并补充到词表中。鉴于经过精加工的千万字、甚至亿字级的汉语分词语料库目前还是水月镜花，所以这个方向上现有的研究无一不以从极大规模生语料库中提炼出的 n 元汉字串之分布（n≥2）为基础。Sproat R. and Shih C.L. (1993) 借用信息论中的“互信息”定量描述任意两个汉字之间的结合力。Sun M.S. and Shen D.Y.,et al. (1998)沿这个思路前进了一步，提出了汉字间 t- 测试差的概念作为互信息的有益补充。黄萱菁、吴立德等（1996）则引入经典统计论中的“四分联立表”及检验联立表独立性的皮尔逊 x[2]－统计量，对长度分别为 2 字、3 字和 4 字的任意汉字串做内部关联性分析， 继而获得候选词表。Nie J.Y. and Jin W.Y., et al.(1994) ， 刘挺、 吴岩等（1998）的工作仅利用了相对简单的字串频信息。这里提到的几个统计量（互信息、t-测试差、x[2]－统计量、字串频）都是依赖于极大规模语料库的，孙茂松、邹嘉彦（1995）故而称之为全局统计量。

处理第二种未登录词的做法通常是：首先依据从各类专有名词库中总结出的统计知识（如姓氏用字及其频度）和人工归纳出的专有名词的某些结构规则，在输入句子中猜测可能成为专有名词的汉字串并给出其置信度，之后利用对该类专有名词有标识意义的紧邻上下文信息（如称谓），以及全局统计量和局部统计量（参见下文），进行进一步的鉴定。已有的工作涉及了四种常见的专有名词：中国人名的识别（张俊盛、陈舜德等 1992；宋柔、朱宏等 1993；孙茂松、黄昌宁等 1995）、外国译名的识别（孙茂松、张维杰 1993）、 中国地名的识别（沈达阳、 孙茂松 1995）及机构名的识别（Chen H.H. and Lee J.C. 1994；张小衡、 王玲玲 1997）。从各家报告的实验结果来看，外国译名的识别效果最好，中国人名次之，中国地名再次之，机构名最差。而任务本身的难度实质上也是循这个顺序由小增大。

沈达阳、孙茂松等（1997b ）特别强调了局部统计量在未登录词处理中的价值。局部统计量是相对全局统计量而言的，是指从当前文章得到且其有效范围一般仅限于该文章的统计量（通常为字串频）。孙茂松、邹嘉彦（1995）通过下例演示了局部统计量的功效：

（7）河南会员冯俊发愿无偿赠送百日红 1000 株。

切分 a.河南｜会员｜冯俊发｜愿｜无偿｜赠送｜百日红｜1000｜株｜。

切分 b.河南｜会员｜冯俊｜发愿｜无偿｜赠送｜百日红｜1000｜株｜。

孤立地看句子（7）， 即使进行句法甚至语义分析也不能判断到底是切分 a 还是切分 b（两者都具合理性）。只有跳出句子界限的束缚，在比句子更大的单位--篇章内才能定夺。譬如，若下文出现“冯俊发”如何如何，则取切分 a；出现“冯俊”如何如何，则取切分 b。显然，局部统计量与心理学中的“短时记忆”机制或计算机技术中的“缓冲区”机制是“心

有灵犀一点通”的。

一般地，未登录词的介入会引起新的切分歧义，从而使分词系统所面临的形势更加复杂化。Sun M.S. and Shen D.Y., et al. (1997) 将切分歧义明确地细分为：1）普通词与普通词之间的切分歧义（第 2.1 节）；2）普通词与未登录词之间的切分歧义；3）未登录词与未登录词之间的切分歧义。

观察句子（8）：

（8）王林江爱踢足球。

中国人名识别模块猜出的候选者为“王林”、“王林江”、“林江”、“林江爱”、“江爱”，中国地名识别模块猜出的候选者为“林江”。其中中国人名“王林”与“王林江”、“王林”与“林江”、“王林”与“林江爱”、“王林江”与“林江”、“王林江”与“林江爱”、“王林江”与“江爱”、“林江”与“林江爱”、“林江”与“江爱”、“林江爱”与“江爱”之间以及中国人名“林江”与中国地名“林江”之间产生了未登录词与未登录词之间的切分歧义，普通词“爱”与“江爱”、“林江爱”之间则产生了普通词与未登录词之间的切分歧义。

必须说明，目前关于未登录词处理的研究，总的来说还是比较初步，在方法上特别是在局部统计量的计算模型上还要下大气力。这里不加说明地列出两组例子，读者不妨仔细体会个中滋味：

（9） a. 刘清楚楚动人。 => 刘清 | 楚楚动人 | 。

b. 刘华清楚这件事。 => 刘华 | 清楚 | 这 | 件 | 事 | 。

c. 刘华清楚地重游。 => 刘华清 | 楚地 | 重游 | 。

d. 刘华清楚地记得 ... => 刘华 | 清楚 | 地 | 记得 | ...

（10） a. 你老张着什么急呀。 => 你 | 老张 | 着 | 什么 | 急 | 呀 | 。

b. 你老张着什么嘴呀？ => 你 | 老 | 张 | 着 | 什么 | 嘴 | 呀 | ？

2.3 语言资源建设

一个好的自动分词系统离不开必要的语言资源的支持。涉及到的最主要的资源有三个：通用词表、经过分词和词性标注的语料库以及极大规模生语料库。一方面，它们为开采分词系统所需要的各类知识提供了“矿藏”丰富的宝山（如：切分歧义的静态分布与采用什么样的词表有关，切分歧义的动态分布及其句法消解模式，乃至隐 Markov 模型的统计参数，都可从分词和词性标注的语料库中习得，全局统计量则可由极大规模生语料库自动转化而来）；另一方面，分词和词性标注的语料库又可作为测试材料对自动分词系统的性能进行定量评估。因此，语言资源的构造同样是自动分词研究不可或缺的一环。

这个环节上面临的主要困难其实源自汉语语言学研究中悬而未决的一些“经典”问题，如词与语素及短语的界限、词类划分体系及词的具体归类等等。受文章长度的制约，不打算多谈了。这里仅想对第一个问题（其实就是所谓的分词规范）简单讲几句。分词规范直接影响到词表和分词语料库的质量，虽然已经有了国家标准（国家技术监督局 1993；刘源等 1994），有的单位也制定了自己的规范（黄居仁、陈克健等 1997），但这些规范的可操作性都不太强（如国家标准中多次出现的关于“什么是词”的表述：“结合紧密、使用稳定”，就无法操作），很难据之构造出一致性好的词表和分词语料库来（孙茂松 1999）。针对这一点，梁南元、刘源等（1991）和孙茂松、张磊（1997）提出了“人机结合、定性与定量并举”的解决思路，并进行了一定规模的实验，但这个思路是否真的可操作，尚言之过早。

顺带提一下，在这个环节上，语言学是大有用武之地的，计算语言学正在以一种迫切、坦诚的心情张开双臂期待着与语言学的拥抱。反过来，语言计算的性质（系统必须覆盖拟处理的一切语言现象）也会逼迫语言学更多地以全面、系统的观点解释、分析语言，从中升华

出来的理论可能更贴近语言的真实面貌，更经得起推敲。

3. 今后的研究要点

1995 年 12 月，国家科委组织了 863 智能机专题自动分词评测， 国内有几个系统参加。开放测试条件下的评测结果是：分词精度最高为 89.4%；交集型切分歧义处理的正确率最高为 78.0%，覆盖型切分歧义处理的正确率最高为 59.0%；而未登录词识别的正确率，人名最高为 58.0%，地名最高为 65.0%（刘开瑛 1997）。1998 年 3 月， 国家科委又搞了第二次评测，结果与第一次差不多。这意味着，即使是对汉语分析最低级、最简单的任务--自动分词，距真正意义的实用还有距离，我们还须付出艰苦、细致的努力。

这个不容乐观的现状并不影响我们在第 1 节中对汉语自动分词的可行性做出比较乐观的估计，因为虽然有待完成的工程量还很大，但在任务难度的性质上，自动分词毕竟不属于“挟泰山以超北海”--“非不为也，乃不能也”一类。那么，今后的研究应着重在哪几点上“有所为”，才能有助于达至我们的理想境界呢？结合自己的研究经验，笔者认为大概要抓以下一些工作：1）尽快建立一个广为接受的、 高质量的通用词表。这是保证其它一切自动分词研究是否扎实、可靠的先决条件； 2）建立一套为学界同仁认同并遵守的汉语自动分词规范和词性标注规范，研制百万字级的经分词、词性标注的平衡语料库以及千万字级的甚至亿字级经分词的通用语料库。各家的工作成果应尽量共享，避免简单重复；3）在通用词表及极大规模语料库的支持下，系统地发现那些频度高、稳定性好（指与领域基本无关）的切分歧义（或可称为通用切分歧义）并有针对性地给出解决办法；4 ）对覆盖型切分歧义的研究目前十分薄弱，统计手段似乎鞭长莫及，宜探讨新的对策；5 ）使已有的各种专有名词识别机制更加精细化，并增设日本人名、少数民族人名识别机制；6）研究各种专有名词之间的冲突处理机制；7）继续发掘全局统计量和局部统计量的潜力，同时注意克服其副作用；8）研究融合词法、 句法甚至部分语义信息，集经验主义（统计形式）与理性主义（规则形式）于一体的分词算法；9 ）以已有工作为基础（曹焕光、 郑家恒 1992 ），构造更加合理的自动分词评测模型，争取评测工作的权威化、公开化、持续化；10）在机器学习理论的指导下，研究从线性或半结构化语言单位序列中获取结构化语言知识的途径，以及有监督学习和无监督学习的互补互动策略，最大限度地提高自动分词系统对复杂开放环境的自适应能力。

【参考文献】

Chang, C.H. and Chen C.D. 1993. A study on integrating Chinese word segmentation and part-of-speech tagging. Communications of COLIPS 3.2.69-77.

Chen, H.H. and Lee J. C. 1994. The identification oforganization names in Chinese texts. Communications of COLIPS4.2.131-142.

Chen, K. J. and Liu S.H. 1992. Word identification for Mandarin Chinese sentences. Proceedings of the 14th International Conference on Computational Linguistics, 101-107.Nantes.

Fan, C. K. and Tsai W. H. 1988. Automatic wordidentification in Chinese sentences by the relaxationtechnique. Computer Processing of Chinese and OrientalLanguages 4.1.33-56.

Guo, J. 1997. Critical tokenization and its properties. Computational Linguistics 23.4.569-59.

Lai,B.Y.,Sun M.S.,et al.1992.Tagging- based first orderMarkov model approach to Chinese word identification, Proceedings of 1992 International Conference onComputer Processing of Chinese and Oriental Languages, Florida.

----.1997.Chinese word segmentation and part- of- speechtagging in one step.Proceedings of

International Conference:1997 Research on Computational Linguistics,229-236.Taipei.

Nie,J.Y. , Brisebois M. , et al. 1996. On Chinese wordsegmentation and word- based text retrieval. Proceedings ofInternational Conference on Chinese Computing 1996, 405-412.Singapore.

Nie,J.Y.,Jin W.Y.,et al.1994.A hybrid approach to unknownword detection and segmentation of Chinese. Proceedings of International Conference on Chinese Computing 1994,405-412.Singapore.

Palmer,D.D.1997.A trainable rule- based Algorithm for word segmentation.Proceedings of the 35th Annual Meeting of ACL and 8th Conference of the European Chapter of ACL.Madrid.

Sproat,R. and Shih C. L. 1993. A statistical method forfinding word boundaries in Chinese text. Computer Processing of Chinese and Oriental Languages 4.4.336-249.

Sproat,R. and Shih C.L.,et al.1996.A stochastic finite-stateword segmentation algorithm for Chinese. ComputationalLinguistics 22.3.377-404.

Sun,M.S.and Benjamin K. T. 1995. Ambiguity resolution inChinese word segmentation. Proceedings of the 10th AsiaConference on Language,Information and Computation, 121 -126.Hong Kong.

Sun, M.S., Lai B.Y., et al. 1992. Some issues onstatistical approach to Chinese word identification.Proceedings of the 3rd International Conference on ChineseInformation Processing, 246-253. Beijing.

Sun, M.S., Lin F.Z., et al. 1996. Linguistic processingfor Chinese OCR & TTS. Proceedings of the 2nd InternationalConference of Virtual Systems and Multimedia,27-42.Gifu.

Sun,M.S.,Shen D.Y.,et al.1997.Cseg & Tag 1.0: A practicalword segmenter and POS tagger for Chinese texts. Proceedingsof the 5th Conference on Applied Natural Language Processing, 119-126.Washington D.C.

----.1998.Chinese word segmentation without using lexiconand hand-crafted training data.Proceedings of the 36th AnnualMeeting of Association of Computational Linguistics and the17th International Conference on Computational Linguistics,1265-1271.Montreal.

Wu,A.D.and Jiang Z.X.1998. Word segmentation in sentenceanalysis.Proceedings of the 1998 International Conference onChinese Information Processing,169-180.Beijing.

Wu,Z.M.and Tseng G. 1993. Chinese text segmentation fortext retrieval: achievements and problems. Journal of theAmerican Society for Information Science 44.9.532-542.

----.1995.ACTS: An automatic Chinese text segmentationsystem for full text retrieval. Journal of the AmericanSociety for Information Science 46.1.83-96.

Yeh,C.L.and Lee H.J.1991.Rule- based word identificationfor Mandarin Chinese sentences - a unification approach.Computer Processing of Chinese and Oriental Languages 5.2. 97-118.

白拴虎，1995，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，56-61 页。

曹焕光、郑家恒，1992，自动分词软件质量的评价模型。《中文信息学报》第 4 期，57-61 页。

董振东， 1997，汉语分词研究漫谈。 《语言文字应用》第 1 期，107-112 页。

国家技术监督局，1993，中华人民共和国国家标准 GB/T 13715-92。《信息处理用现代汉语分词规范》北京：中国标准出版社。

黄居仁、陈克健等，1997，“资讯处理用中文分词规范”设计理念及规范内容。《语言文字应用》第 1 期，92-100 页。

黄萱菁、吴立德等，1996，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》第 4 期，297-303 页。

黄祥喜，1989，书面汉语自动分词的“生成-测试”方法。《中文信息学报》第 4 期，42-49 页。

韩世欣、王开铸，1992，基于短语结构文法的分词研究。《中文信息学报》第 3 期，48-53 页。

何克抗、徐辉等，1991，书面汉语自动分词专家系统设计原理。《中文信息学报》第 2 期，1-14 页。

揭春雨、刘源等，1989，论汉语自动分词方法。《中文信息学报》第 1 期，1-9 页。

李国臣、刘开瑛等，1988，汉语自动分词及歧义组合结构的处理。《中文信息学报》第 3 期，27-33 页。

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

--------，1990，汉语计算机自动分词知识。《中文信息学报》第 2 期，29-33 页。

梁南元、刘源等，1991，制订《信息处理用现代汉语常用词词表》的原则与问题讨论。《中文信息学报》第 3 期，26-37 页。

刘开瑛，1997，现代汉语自动分词评测技术研究。《语言文字应用》第 1 期，101-106 页。

刘挺、吴岩等，1998，串频统计和词匹配相结合的汉语自动分词系统。《中文信息学报》第 1 期，17-25 页。

刘挺、王开铸，1998，关于歧义字段切分的思考与实验。《中文信息学报》第 2 期，63-64 页。

刘涌泉，1988，再谈词的问题。《中文信息学报》第 2 期，47 -50 页。

刘源、梁南元，1986，汉语处理的基础工程--现代汉语词频统计。《中文信息学报》第 1 期，17-25 页。

刘源等，1994，《信息处理用现代汉语分词规范及自动分词方法》北京：清华大学出版社及广西科学技术出版社。

马晏，1996，基于评价的汉语自动分词系统的研究与实现。《语言信息处理专论》北京：清华大学出版社及广西科学技术出版社，2 -36 页。

沈达阳、孙茂松，1995，中国地名的自动辨识。《计算语言学进展与应用》北京：清华大学出版社，68-74 页。

沈达阳、孙茂松等，1997a， 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》第 2 期，34-47 页。

--， 1997b，局部统计在汉语未登录词辨识中应用和实现方法。《语言工程》北京：清华大学出版社，127-132 页。

宋柔、朱宏等，1993，基于语料库和规则库的人名识别法。《计算语言学研究与应用》北京：北京语言学院出版社，150-154 页。

孙茂松，1999，谈谈汉语分词语料库的一致性问题。《语言文字应用》第 2 期，87-90 页。

孙茂松、黄昌宁等，1995，中文姓名的自动辨识。《中文信息学报》第 2 期，16-27 页。

--，1997，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。《计算机研究与发展》第 5 期，332-339 页。

孙茂松、张维杰，1993，英语姓名译名的自动识别。《计算语言学研究与应用》，北京：北京语言学院出版社，144-149 页。

孙茂松、张磊，1997，人机共存，质量合一--谈谈制定信息处理用汉语词表的策略。《语

言文字应用》第 1 期，79-86 页。

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

孙茂松、左正平，1998，汉语真实文本中的交集型切分歧义。《汉语计量与计算研究》香港：香港城市大学出版社，323-338 页。

--，1999a，消解中文三字长交集型分词歧义的算法。《清华大学学报》第 5 期，101-103 页。

孙茂松、左正平等，1999b， 高频最大交集型歧义切分字段在汉语自动分词中的作用。《中文信息学报》第 1 期，27-34 页。

王晓龙、王开铸等，1989，最少分词问题及其解法。《科学通报》第 13 期，1030-1032 页。

王永成、苏海菊等，1990，中文词的自动处理。《中文信息学报》第 4 期，1-10 页。

姚天顺、张桂平等，1990，基于规则的汉语自动分词系统。《中文信息学报》第 1 期，37-43 页。

徐秉铮、詹剑等，1993，基于神经网络的分词方法。《中文信息学报》第 2 期，36-44 页。

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

张俊盛、陈舜德等，1992，多语料库作法之中文姓名辨识。《中文信息学报》第 3 期，7-15 页。

张小衡、王玲玲，1997，中文机构名称的识别与分析。《中文信息学报》第 4 期，21-32 页。

本文发表在《当代语言学》2001 年第 1 期。pp. 22-32。