对自动分词的反思

黄昌宁 高剑峰 李沐
微软亚洲研究院
Email:{cnhuang,jfgao,t-muli}@microsoft.com

摘要：自动分词是中文信息处理诸多应用系统的一个不可或缺的模块。二十年来国内外许多研究人员曾经在这块土地上辛勤耕耘，并取得了一定的成果，但从实用化的角度上来考察仍不尽人意。本文通过对自动分词任务的定义，分词歧义消解知识的调查，以及在统计语言模型的统一平台上实现自动分词，说明面向计算机的语言知识颗粒度极细、颗粒数量极大，在本质上不同于面向人的语言知识，也不是传统的句法-语义知识所能覆盖的。重新审视我们在自动分词乃至中文信息处理整个领域的思维模式和技术路线，实属必要。

关键词：自动分词，切分歧义，统计语言模型，面向计算机的语言知识

Reconsidering on Chinese Word Segmentation

Changning Huang, Jianfeng Gao and Mu Li
Microsoft Research Asia
E-mail:{cnhuang, jfgao, t-muli}@microsoft.com

ABSTRACT: Automatic word segmentation is an indispensable component of many Chinese information processing (CIP) systems. In last two decades many researchers have made large amounts of efforts in this area and achieved, in some cases, promising results. Unfortunately, from the practical engineering point of view, the performance of Chinese word segmentation is still under expectation. In this paper we will give a definition of the task of automatic word segmentation, investigate the knowledge for resolving ambiguities in Chinese word segmentation. and describe in detail an implementation of Chinese word segmentation system which is based on a unified platform of statistical language modeling. The purpose of this paper is to certify that the linguistic knowledge which is applied in computer systems is essentially different from that used by human being, because in the former the granularity of knowledge is much more fine, and the amount is much larger. Such kind of linguistic knowledge is far beyond the scope of traditional syntactic-semantic knowledge. Therefore, it is necessary to change our mind set and reconsider our technical strategies in Chinese word segmentation and even in the entire area of CIP.

Keywords: Chinese word segmentation, segmentation ambiguities, statistical language modeling, linguistic knowledge for computer

1. 一个没有明确定义的任务

本文不打算对自动分词的研究历史作一番详尽的综述，有兴趣的读者请参阅 [1]、[2]。一般人对自动分词的认识是：系统输入一个中文文本 T，输出文本 T'，T' 是 T 的分词形式。然而什么是中文的词，迄今没有一个被广泛接受的定义。调查表明，在母语为汉语的被试者之间对中文文本中出现的词语的认同率只有大约 70%。从计算的严格意义上说，自动分词是一个没有明确定义的问题。1992 年中国国家标准局颁布了作为国家标准的《信息处理用现代汉语分词规范》[3]。《规范》的大部分规定是通过举例和定性描述来体现的。如“结合紧密，使用稳定的二字或三字词组，一律为分词单位”。何谓“紧密”，何谓“稳定”，人们在实际工作中很难界定。这些原因使得《规范》并没有从根本上统一国人对词的认识，哪怕只是在信息处理界。在这种情况下，建立公平公开的自动分词评测的努力也一样步履维艰。

为摆脱这种困境，目前实用的自动分词系统通常是在一个给定的分词词表的基础上实现的，叫做基于特定词表的分词。分词词表的规模一般在 5 万到 25 万词条之间。词表之外的词一律叫做未登录词（OOV）。然而在给定词表的情况下，在一个待切分的文本中究竟有多少交集型歧义字段（OAS）和覆盖型歧义字段（CAS）[1]、[2]？什么样的语言知识能有效地从文本中识别和排解 OAS 和 CAS 歧义？

此外，在真实文本的切分中，未登录词总数的大约九成是专有名词（人名、地名、机构名），其余的是新词(包括专业术语)。每一个专有名词，如地名，都是一个词，还比较容易取得共识。但对新词来说，就会出现词的界定问题。即文本中被认定为新词的一个字段，究竟是一个词还是多个词？于是绕了一圈，又回到分词规范的老问题上来了。何况未登录词的辨识一般依赖于上下文。那么，这里又需要什么样的语言知识？需要多少知识？

如果我们不能明确地回答上面的这些问题，又怎么能期望计算机将圆满地实现自动分词任务呢？历经近二十年的努力，中文自动分词系统在实用化上似乎仍不尽人意，一个跟本原因就是：与自动分词有关的这些问题从来就没有一个明确的答案。

2. 分词需要什么样的语言知识？

何克抗、徐辉等（1991）[4] 曾断言，95% 左右的切分歧义可以通过句法或句法以下的语言知识来解决，只有 5% 必须依仗语义和语用知识。王永成（1991）[5] 也认为，“从理论上说，中文自动分词的基本方法可分成三类：形式分词方法、语法分词方法和语义分词方法。” Wu, A.D. and Jiang, Z.X. (1998) [6] 相信，通过句法分析才能最终解决切分歧义问题。孙茂松 [1] 对切分歧义的检测与消解作了公允的综述。他认为，问题的解决有赖于足够的语言知识的支撑；并且对 [6] 和其他雷同的主张提出了质疑。孙认为，消解切分歧义这样一个相对 ‘简单’ 的任务却要依仗比分词本身困难得多的句法分析才能完成，似乎是一个悖论。笔者认为这个问题反映了一个深刻的理念，就是大家经常挂在嘴边的所谓句法-语义方法，究竟指什么？

其实，客观上存在两种本质不同的语言知识：面向人的语言知识和面向计算机的语言知识。面向人的语言知识，体现为以人为对象的各类辞书，以及词法学、句法学、语义学和语用学等各种研究成果。在中文信息界，人们通常提到的句法分词、语义分词方法，就是打算利用面向人的语言知识来解决分词问题。笔者认为，面向计算机的语言知识具有完全不同于前一种语言知识的特质。其特征反映在知识表示、颗粒度、数量和知识获取方法

等诸方面。下面结合自动分词任务来具体讨论一下这个问题。首先来看一看歧义切分（OAS和CAS）的几个实例，它们是从一个基于句法分析的自动分词系统（简称PBWS）的分词结果中选出来的。

(1) OAS误切实例：

①决定在全省戒玩风，/兴学/风/，

②最大限度地防止有害/信息流/入/和/传播

③我们希望明天在世界的/领奖/台上/有更多的中国年轻人出现。

④挽救一/个/人生/命/的义务将凌驾于不侵犯别人隐私的义务。

⑤这些“志愿军”最初出现在英国/足球赛/场上/

(2) CAS误切实例：

⑥东/中西部/地区要按照优势互补、互惠互利、真诚合作的原则，加强联合。

⑦过去思想封闭的赞皇人，/对路/的渴望竟如此强烈，

⑧你们这/群山/里的女娃娃有了学本领、闯世界的志气。

⑨希望你们再/创新/的业绩。

⑩进书店/跟进/超市买柴米油盐/一/样/，

“才能”是一个高频的CAS字段。在总共933句的测试集中出现了5次，正确切分都是“才/能”。PBWS切对了2次，切错了3次。下面是PBWS对该字段的切分结果：

□股票投资者的基本权利/才能/得到保障。（错）

□怎样在安装等待过程中设计出活动的画面/才能/让用户不致焦躁。（错）

□切实纠正有偿新闻等不正之风，/才能/更好地为人民服务。（错）

□由此入手，/才/能/更深刻地洞察信息时代教育改革发展的趋势与前景。（对）

□与之配套的软件/才/能/调试通，...（对）

就“才能”这个测试点来说，标准答案是10个词，PBSW输出了7个词，其中4个词与答案匹配。所以精确率是0.57 (4/7)，召回率是0.40 (4/10)。我们不知道PBSW的语法规则是怎么写的。只从句型上看，⑬和⑭两句很相似，可是为什么会一错、一对呢？这些切分实例让我们看到了，词作为造句的备用单位一旦进入文本时会形成一幅怎样扑朔迷离的图画。众多的词在动态生成语句时，由于邻接、粘连和碰撞产生了千姿百态的歧义。这些现象，不论是OAS还是CAS，都是偶发的，难以预测的。上面这些切分实例还说明，即使拥有像句法分析这样的深层语言知识的支持，歧义消解也并没有达到人们预期的效果。因此完全有必要重新估量句法-语义手段在自然语言歧义消解问题上的有效性。

回过头来再看看以人为对象的词典。这类词典只收集一个个静态的、抽象的词项[1]，对每个词项的读音、词性和词义加以注释，如此而已。从来也不会去关心这些词项在真实文本中被动态调用时，怎样随机地生成切分歧义，更不会针对可能出现的歧义给出相应的消解对策。这正是两种语言知识之间的巨大差异。自动分词系统必须装备用来消解分词歧义的专门知识，它们来自大规模语料库的OAS和CAS实例调查，搞清楚每一条OAS或CAS的频率分布，必要时逐条编制歧义消解的对策。面向计算机的语言研究必须解决这个任务。

[1] 只有当一个词项进入一个特定的句子时，它才成为一个具体的词例，并受到上下文的约束。

下面分别介绍 OAS 和 CAS 的语料库调查。

(1) OAS 调查

[2] 曾指出 OAS 的两个重要特性：(a) 如果正向最大匹配 (FMM) 的切分结果 O_F 等于反向最大匹配 (BMM) 的切分结果 O_B，即 $O_F = O_B$，例如“动人/情景”，则最大匹配 (MM) 切分结果正确的概率是 99%；否则 (b) 如果 FMM 的切分结果不同于 BMM，$O_F \neq O_B$，如 O_F = “兴学/风”，O_B = “兴/学风”，则 MM 切分中至少有一个是正确切分的概率也是 99%。利用这个方法可以实现文本中 OAS 的有效侦察。Mu Li, et al.(2003) [7] 在更大规模上验证了上述结论。他采用的分词词表有 93,700 词条。测试语料来自 1997 年人民日报，内含汉字 460,000 字次，或 247,000 词次 (见附录 1)。在该测试集中，共找到 5,759 条最大 OAS (types)，称为 OAS 集合 C。然后根据 OAS 特性把集合 C 分成 A 和 B 两个子集。其结果如表 1 所示。从子集 A ($O_F = O_B$) 可见，MM 切分的精确率可达 0.988 (2,731/2,763)。在子集 B ($O_F \neq O_B$) 中，召回率下限为 0.957 (2,866/2,996)，而在整个 OAS 集合 C 中为 0.972 (5,597/5,759)。

表 1　OAS 在测试集中的分布

A　$O_F = O_B$ 47.98% (2763)		B　$O_F \neq O_B$ 52.02% (2996)	
$O_F = O_B$ = 正确切分	$O_F = O_B \neq$ 正确切分	O_F =正确 $\vee$ O_B = 正确	$O_F \neq$正确 $\wedge$ $O_B \neq$正确
47.42%(2731)	0.56%(32)	49.77% (2866)	2.26% (130)

孙茂松、左正平等 (1999) [8] 对一个 1 亿字的新闻语料库进行了最大交集型歧义切分字段 (为了便于说明，我们仍称之为 OAS) 的调查。他们使用的分词词表含 112,967 个词条。从语料库中共提取出 233,888 个不同的 OAS (types)。这些 OAS 在语料库中累计出现 1,793,317 次 (tokens)。报告称，前 2,500 个高频 OAS (types) 覆盖了语料库中全部 OAS 的 50%；前 4,619 个高频 OAS(types)覆盖了全部 OAS 的 59.2%。报告建议，通过人工干与方式把其中属于伪歧义[2]的 4,279 条 OAS 的唯一切分形式事先记录在一张分词知识表中，就可以解决 OAS 消歧的约 60% (types)。我们在一个约 6.5 亿字次的人民日报语料库上完成了类似调查 [7]。首先检出 $O_F \neq O_B$ 的 OAS 730,000 条，从中选出高频的 OAS 47,000 条。然后为每一条 OAS 随机地从语料中抽出 20 个例句，它们经过人工切分以后，形成 41,000

[2]这些 OAS 实际上只有唯一的一种切分方式，与上下文无关。

条词例化的消歧规则。它们覆盖语料中全部 OAS (types) 的约 80%，这个结果同 [8] 的调查很接近。下面是一条这样的消歧规则："信心地 => 信心 |地"。它表示，在包含字段"信心地"的 20 个句子中，至少有 19 次应切成"信心/地"。

(2) CAS 调查

与 OAS 不同，目前还没有有效的手段来侦察文本中的 CAS。CAS 调查所用的语料是 88 MB（兆字节）的新闻语料库。全部语料经过一个基于统计的自动分词系统切分，未经过人工校对。为了使问题尽可能简化，调查中只考察词表中由二字词产生的 CAS。实验用的词表中共有二字词 49,778 条，参考北京大学句法信息词典提供的 1,716 个单字词，凡完全由这些单字词组成的二字词就被视为 CAS 候选。由此检得二字长 CAS 候选 26,073 条（types，下同）。其中有 24,728 条 (94.8% = 24,728/26,073) 在上述语料库中出现过。如果把 "真 CAS" 定义为那些不仅在语料库中出现过、而且确曾被分词系统切分成两个单字词的 CAS 候选。那么在这个语料库中一共找到 6,119 条真 CAS[3]，是词表中 CAS 候选总数的 23.5% (6,119/26,073)。如果从中选出 100 条左右高频的真 CAS，用某种机器学习方法为每一条 CAS 定做一个上下文相关的二值分类器，就可以在一定程度上解决 CAS 的消歧问题 [9]。但这少量的 CAS 消歧知识在真实文本中的覆盖率不会太高。

以上调查说明，由于真实文本中偶发的 OAS 和 CAS 歧义涉及成千上万个词，因此消歧所需的语言知识必然是词例化的，知识的数量是巨大的。

从以上讨论可以看到，解决切分歧义所需的语言知识有两个特点：一是动态性，它们在真实文本中的出现纯属偶然，难于预测；二是颗粒度极细，颗粒数量极大。传统的句法-语义知识不能反映词例的动态行为，而且无论在颗粒度上还是在数量上都不可能满足自动分词的需要。下面介绍微软亚洲研究院在统计语言模型的框架上实现的一个自动分词系统，目的是把各类分词知识尽可能地整合在一个统一的平台上。

3. 一个基于统计语言模型的分词系统

3.1 定义文本中的词

如第 1 节所述，为了给自动分词任务一个明确的定义，首先必须给文本中的词下一个可操作的定义。下面是 Jianfeng Gao (2003) [10] 给出的定义：

(1) 待切文本中能和分词词表中任意一个词相匹配的字段是一个词；

(2) 文本中任意一个经词法派生出来的词/短语是一个词，如词的重叠形式(干干净净、研究研究、说说话、天天)，前缀派生（非党员、副部长），后缀派生（全面性、朋友们、受教育者），中缀派生（看得出、看不出），动词+时态助词（克服了、蚕食着），动词+趋向动词（走出、走出来），动词的分离形式（长度不超过 3 个字，如：洗了澡、洗个澡、洗过澡）等等；

(3) 文本中被明确定义的任意一个实体名词（如日期、时间、时段、货币、百分数、

[3]由于语料的分词结果未经人工校对，这个数字可能偏低。

温度、长度、面积、体积、重量、地址、电话、传真号、电子邮箱等）是一个词；

(4) 文本中任意一个专有名词（人名、地名、机构名）是一个词。

上述定义没有把文本中的新词计入自动分词任务，原因是：未登录词由专有名词(95%)和新词（5%）等两大部分组成，只有专有名词才有明确定义。

消解 OAS 和 CAS 所涉及的知识已在第 2 节中阐述。专有名词和实体名词的识别在下面的统计语言模型平台上加以解决。

3.2 统计语言模型 (SLM)

令随机变量 S 为一个汉字序列，W是在 S上所有可能被切分出来的词序列，将下式中条件概率值最高的词序列 W* 选出，它应当就是系统将输出的分词结果：

$$W^* = \arg\max_W P(W/S) \tag{1}$$

根据贝叶斯公式，(1) 可改写成：

$$W^* = \arg\max_w \frac{P(W)P(S|W)}{P(S)} \tag{2}$$

由于分母 $P(S)$为常数，不影响极大值的计算，可略去。于是有

$$W^* = \arg\max_w P(W)P(S/W) \tag{3}$$

为了把第 3.1 节中定义的 4 类词都纳入同一个 SLM 框架，我们把专有名词的人名 *PN*、地名 *LN*、机构名 *ON* 分别设为三类，实体名词中的日期 *dat*、时间 *tim*、百分数 *per*、货币 *mon* 等各设一类，对词法派生词 *MW* 和词表词 *LW* 则每词单独设一类。这样按照表 2 可以把一个可能的词序列 W转换成一个可能的词类序列 $C = c_1 c_2 \ldots c_N$，则

$$C^* = \arg\max_C P(C)P(S/C) \tag{4}$$

式中 $P(C)$是大家熟悉的三元模型，即

$$P(C) = P(c_1)\ P(c_2/c_1)\prod_{i=3}^{N} P(c_i \mid c_{i-2}c_{i-1}) \tag{5}$$

三元模型的参数通过最大似然估计（见式（6））在一个带有词类（word class）标记的训练语料上计算，并采用回退平滑算法解决数据稀疏问题 [11]。

$$P(c_i/c_{i-2}c_{i-1}) = \frac{count(c_{i-2}c_{i-1}c_i)}{count(c_{i-2}c_{i-1})} \tag{6}$$

$P(C)$ 被用来估计任意一个词类序列 C出现的概率。以人名识别为例，人名类 *PN* 很可能在词类串“*LN*=清华大学/*lw*=教授”之后出现，其条件概率可表为：*P(PN/ LN=清华大学 lw=教授)*，所以又被称为上下文模型。

$P(S/C)$ 叫做生成模型。在满足独立性假设的条件下，它可以近似为：

$$P(S/C) \approx \prod_{i=1}^{N} P(s_i/c_i) \tag{7}$$

即认为任意一个词类 c_i 生成汉字串 s_i 的概率只同 c_i 自身有关，而与其上下文无关。

设“教授”是词表词，则 $P(s_i=教授/c_i=lw)=1$ （见表 2）。

表 2　生成模型 $P(S/C)$[10]

词类	生成模型 $P(S/C)$	语言知识
词表词 (LW)	若 S 是词表词，$P(S\|LW)=1$；否则 0	分词词表
词法派生词 (MW)	若 S 是派生词，$P(S\|MW)=1$；否则 0	派生词词表
人名 (PN)	基于字的二元模型	姓氏表，中文人名模板
地名 (LN)	基于字的二元模型	地名表，地名关键词表，地名简称表
机构名 (ON)	基于词类的二元模型	机构名关键词表，机构名简称表
实体名词 (FT)	若 S 可用实体名词规则集 G 识别，$P(S/G)=1$；否则为 0	实体名词规则集

3.3 模型的训练

系统的词表含 98,668 词条，词法派生词词表收入派生词 59,285 条。训练预料由 88MB 新闻文本构成。模型的训练由以下三步组成：（1）在上述两个词表的基础上，用正向最大匹配法 (FMM) 切分训练语料，专有名词通过一个专门模块标注 [12]，实体名词通过相应的规则和有限状态自动机（FSA）标注，由此产生一个带词类（word class）标记的初始语料；（2）用带词类标记的初始语料估计统计语言模型的概率参数，如公式 (6)；（3）用所得语言模型对训练语料重新进行切分和标注（公式 (4)、(5)、(7)），得到一个刷新的训练语料。重复步骤（2）和（3），直至系统的性能不再有明显提高为止。

(1) OAS 消歧：通过 MM 法检测训练语料中的 OAS，用一个特定的类 〈GAP〉 取代全体 OAS，以此来训练上下文模型 $P(C)$。类 〈GAP〉 的生成模型的参数可以通过 OAS 消歧规则（见第 2 节）或某种机器学习方法来估算 [7]。

(2) CAS 消歧：通过训练语料库的调查（见第 2 节），选出最高频、且其切分分布比较均衡的 70 条 CAS，用机器学习方法为每一条 CAS 训练一个二值分类器。再用这些分类器在训练语料中消解这些 CAS 的歧义。

4. 性能评测

评测对推动自动分词技术的进步具有重要意义，下面给予扼要的阐述。

4.1 实施分词评测的主要步骤

（1） 根据自动分词任务中给出的的定义（见 3.1 节），制备规模适当的分词词表和词法派生词词表，同时制定专有名词和实体名词的详尽规范；

（2） 收集一个题材和体裁分布平衡的测试文本集（简称测试集）。测试集的规模一般在 50 万至 100 万字次左右。

（3） 用人工对测试集实施分词和词类（word class）标注，所得标注结果叫做标准文本。标准文本的文本内容和排列顺序同测试集完全一样，不同的是前者带有分词词类标记，后者是生语料。

表 3 机构名的标注示例

	情况	标记方法	例子
1	普通名字 + 机构名	整体标出	[0 海尔集团]
2	地名 + 机构名	整体标出机构名 不含地名嵌套的标记	[0 北京市电信局] [0 保定一棉厂] [0 东直门中学] [0 北京航空航天大学] [0 上海博物馆] [L 首都国际机场] 机构名关键词包括： 超市|饭店|百货大楼|商场|植物园| 动物园|幼儿园|大使馆|领事馆|图书馆| 酒店|旅馆，等
3	人名 + 机构名	整体标出机构名 不含嵌套标记	[0 李嘉诚基金会] [0 中山医学院]
4	简称	一律整体标注	[0 北约] [0 上轮集团] ----指上海轮胎集团 [0 白宫]

（4） 编制一个分词评测软件。软件的输入是两个文本：(a) 被测系统对测试集实施自动分词的输出结果；(b) 标准文本。测试软件对这两个文本进行逐词对比，然后分别输出被测系统分词整体和指定单项的评测结果。自动分词的性能指标包括：精确率、召回率和 F 指标等。

应当指出，为了进行评测制定专有名词、实体名词的详尽规范是十分必要的。下面是微软亚洲研究院制定的《命名实体标注规范》[4]中有关机构名的示例（见前页表 3）。

微软亚洲研究院采用的分词测试集选自 1997 年人民日报，包含经济、文化、政治、科

[4] [4]该规范参照了美国 NIST 制定的相应标准 MET-2 和 IEER-99，
http://www.itl.nist.gov/iaui/894.02/related_projects/muc/
http://www.nist.gov/speech/tests/ie-er/er_99/er_99.htm

技、法律、体育等 10 种题材和描写文、叙述文、说明文、应用文、口语等 5 种体裁。测试集的构成及其门类分布详见附录 1。该标准文本总共有 247,039 个词次（下同），其中属于词表或派生词词表的词 205,162 个，*PN* 4,347 个，*LN* 5,311 个，*ON* 3,850 个和实体名词 6,630 个（详见附录 2）。

4.2 自动分词的性能指标

自动分词系统的主要性能指标是分词的精确率 *P* 和召回率 *R*。它们的定义如下：

$$\text{精确率 } P = \frac{\text{分词结果中切分正确的总词数}}{\text{分词结果中的总词数}} * 100\ \% \tag{8}$$

$$\text{召回率 } R = \frac{\text{分词结果中切分正确的总词数}}{\text{标准文本中的总词数}} * 100\ \% \tag{9}$$

一般来说，精确率 *P* 和召回率 *R* 是相互制约的，提高召回率的措施往往导致精确率的下降，反之亦然。因此在比较两个分词系统的性能时，只考察这两个指标中的一个往往是片面的。为了对比的方便，学术界经常采用 *P* 和 R 的调和平均值 *F* 作为第三个性能评价指标。下面是 *F* 的计算公式，式中 β 是 *P* 的加权因子（一般令 $\beta=1$）：

$$F\text{-指标} = \frac{(\beta^2+1)*P*R}{(\beta^2*P)+R} = \frac{2*P*R}{P+R} \quad (\text{当}\beta=1) \tag{10}$$

对于人名、地名、机构名及其他实体名词的识别结果，也可以类似地分别定义其单项的精确率、召回率和 *F*-指标。以地名为例，地名识别的 *P* 和 *R* 的公式分别如（11）、（12）所示：

$$\text{精确率 P} = \frac{\text{分词结果中正确识别的地名总数}}{\text{分词结果中的地名总数}} * 100\ \% \tag{11}$$

$$\text{召回率 R} = \frac{\text{分词结果中正确识别的地名总数}}{\text{标准文本中的地名总数}} * 100\ \% \tag{12}$$

应当指出，过去国内曾经使用过各式各样的分词指标，其中有些是明显不合适的。例如，采用如下的“平均出错率”作为自动分词的性能指标：

$$\text{平均出错率 } E = \frac{\text{分词结果中错误切分的总次数}}{\text{测试文本中的总字数}} * k \quad [\text{次/百字}] \tag{13}$$

式中常数 *k* 是一个换算因子，目的是使 E 的单位变成［次/百字］。更有甚者，直接利用“平均出错率”E 定义分词“正确率”为：

$$\text{正确率 } A = (1-E)\ \% \tag{14}$$

其实仅从以上两个公式的量纲，就可看出公式（14）的换算是完全错误的。

冯志伟（2001）[12]介绍的“863 评测”，不论对分词整体还是对专有名词等单项评测，一律只用“精确率”（或“正确率”）一个指标，而且没有给出计算公式。以地名单项测试为例，如果被测系统 A 和 B 输出的正确地名总数 m 相同，那么是否应当认定两个分词系统的地名识别 “精确率” 相同呢？其实不然。如果进一步考察发现：系统 A 输出的地名总数为 n1，系统 B 输出的地名总数为 n2，且 n2 >> n1。那么根据公式（11），不难判断系统 A 实际的地名识别精确率（P = m/n1）要比系统 B 的精确率（P = m/n2）高得多；两个系统相同的仅是召回率（见公式（12））。系统 B 靠过多地报出文本中的地名来提高其地名召回率，后果必然会损害其自身的地名识别精确率。

5. 实验结果

5.1 SLM 分词系统评测

测试集和标准文本如 4.1 节所述（见附录 1 和 2）。

实验内容：

(1) 在给定词表的条件下，仅用 FMM 在测试集上得到的切分结果；
(2) 在没有专有名词和实体名词识别的情况下，单独用统计语言模型所得的分词结果；
(3) (2)+实体名词识别；
(4) (3)+人名识别；
(5) (4)+地名识别；
(6) (5)+机构名识别。

实验结果如表 4 所示。实验表明，实体名词和专有名词等未登录词的识别不仅本身有很高的应用价值，而且在很大程度上提高了分词系统的精确率和召回率。尽管如此，错误分析显示，实体名词和专有名词只占标准文本总词次的 8.7%（附录 2），而它们引起的分词错误却占分词错误总数的 59.2%。尤其是机构名 ON 占标准文本总词次的 1.6%，引起的分词错误却占分词错误总数的 20.6%。这说明，自动分词系统的性能仍有较大的改进空间。

表 4　SLM 分词系统的测试结果

	系统	自动分词		实体名词		PN		LN		ON	
		P%	R%	P%	R%	P%	R%	P%	R%	P%	R%
1	正像最大匹配 FMM	83.7	92.7								
2	统计语言模型 SLM	84.4	93.8								
3	2 +实体名词	89.9	95.5	84.4	80.0						
4	3 + PN	94.1	96.7	84.5	80.0	81.0	90.0				
5	4 + LN	94.7	97.0	84.5	80.0	86.4	90.0	79.4	86.0		
6	5 + ON	96.3	97.4	85.2	80.0	87.5	90.0	89.2	85.4	81.4	65.6

5.2 跨系统性能对比

为了比较不同分词方法对分词性能的影响，我们把 SLM 系统同其他三个分词系统作了性能对比。这三个系统分别是：

（1） MSWS 系统：这是微软公司（Microsoft）发布的一个产品，可作为 Windows 的 API（应用程序接口）调用。除了 FMM 和 BMM 方法之外，该系统在 OAS/CAS 消歧和专名、实体名词识别方面采用的是启发式规则；

（2） LCWS 系统：是大陆最好的分词系统之一。系统的工作原理类似于 MSWS，但拥有较大的人名和地名词典；

（3） BPWS 系统（参见第 4 节）：这是一个基于句法分析器的分词系统，它在分词过程中尽可能地利用深层的句法-语义知识。

由于这些分词系统所用的词表不同，对专有名词的标注规范不统一，使得在它们之间实行分词性能的对比不能够自动进行。为此我们从微软分词测试集（附录 1）中随机抽出 933 个句子，作为跨系统对比用的小测试集。它总共包含 22,833 词次，其中 PN 329 个，LN 617 个，ON 435 个。对比时，只考察 OAS 和专有名词等指定的测试点，这是因为在这些测试点上的切分正误容易作出公断。各个系统输出的结果都须经过人工判定，判定时采用标准从宽原则。对比结果如表 5 所示。结果显示 SLM 系统在专有名词识别和 OAS 消歧的性能上全面超过了其他分词系统，证明把不同的分词知识整合在一个统计语言模型上的尝试是成功的。

表 5　跨系统性能对比

系统	OAS	LN			PN			ON		
	错切数	P %	R %	$F_{\beta=1}$	P %	R %	$F_{\beta=1}$	P %	R %	$F_{\beta=1}$
MSWS	63	93.5	44.2	60.0	90.7	74.4	81.8	64.2	46.9	60.0
LCWS	49	85.4	72.0	78.2	94.5	78.1	85.6	71.3	13.1	22.2
PBWS	20	76.7	73.6	75.2	78.0	78.7	78.4	81.7	21.6	34.2
SLM	10	87.6	86.4	87.0	83.0	89.7	86.2	79.9	61.7	69.6

6. 结论

本文通过自动分词这样一个传统节目，试图说明自然语言处理所需的语言知识有别于面向人的语言知识。传统的句法-语义方法在切分歧义消解这个老问题上没有达到人们原先预期的效果，这是一个意味深长的事实。表 6 是面向人和面向计算机两种语言研究的一个对比。希望它对关心中文信息处理技术方法论的研究者有所启发。

表 6　两种语言研究的对比

面向计算机的语言研究	面向人的语言研究
1）大规模真实文本	受限的规范语言
2）概率化参数模型	句法-语义分析
3）上下文相关	多半与上下文无关
4）知识词例化（lexicalized）	基于词类(POS)/短语类
5）颗粒度极细	颗粒度粗
6）覆盖面宽	覆盖面窄
7）从语料中自动/半自动获取	语言学家的语感/直觉
8）以定量化评测驱动研究	例不十，法不立；例外不十，法不破

参 考 文 献

[1] 孙茂松、邹嘉彦，2001，汉语自动分词研究综述。《当代语言学》2001 年第 1 期，22-32 页。

[2] 黄昌宁，1997，中文信息处理中的分词问题。《语言文字应用》1997 年第 1 期，72-78 页

[3] 《信息处理用现代汉语分词规范（GB/T1375-92）》，中国标准出版社，1992

[4] 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。

[5] 王永成等，1991，《中文信息处理技术及其基础》，上海交通大学出版社，89-114 页

[6] Wu, A.D. and Jiang, Z.X. 1998. Chinese word segmentation in sentence analysis. *Proceedings of the 1998 International Conference on Chinese Information Processing*, 169-180. Beijing.

[7] Mu Li, Jianfeng Gao, Changning Huang and Jianfeng Li. 2003, Unsupervised training for overlapping ambiguity resolution in Chinese word segmentation. In *Proceedings of the Second SIGHAN Workshop attached to ACL-2003*, Japan.

[8] 孙茂松、左正平、邹嘉彦，1999，高频最大交集型歧义切分字段在汉语自动分词中的应用。《中文信息学报》1999 年第 1 期，27-34 页

[9] Xiao Luo, Maosong Sun and Benjamin K. Tsou. 2002. Covering Ambiguity Resolution in Chinese Word Segmentation Based on Contextual Information. In *Proceedings of 19th International Conference on Computational Linguistics (COLING 2002)*. 598-604. Taiwan.

[10] Jianfeng Gao, Mu Li and Chang-Ning Huang, 2003. Improved source-channel models for Chinese word segmentation. In *Proceedings of 41th Annual Meeting of Association for Computational Linguistics*, Japan.

[11] Jianfeng Gao, Joshua Goodman, Mingjing Li and Kai-Fu Lee. 2002. Toward a unified approach to statistical language modeling for Chinese. *ACM Transactions on Asian Language Information Processing*, Vol. 1, No. 1, pp 3-33.

[12] Jian Sun, Jianfeng Gao, Lei Zhang, Ming Zhou, and Changning Huang. 2002. Chinese named entity identification using class-based language model. In *Proceedings of 19th International Conference on Computational Linguistics (COLING 2002)*. 967-973. Taiwan.

[13] 冯志伟，2001，汉字和汉语的计算机处理，《当代语言学》2001 年第 1 期，1-21 页

附录 1　微软亚洲研究院分词测试集的题材和体裁分布

题材\体裁(MB)	描写文	说明文	叙述文	应用文	口语	总计	%
文化		2.2	49.6	12.2		64	6%
经济		10.6	102.6	55.1	12.7	181	17%
文学	33.1	13.2	6.3			52.6	5%
军事			42.1			42.1	4%
政治		36.2	102.9	88.8	100.8	328.7	31%
科技	7.3	14.7	85.8	2.1	9.4	119.3	11%
社会	4.5	6.2	56.9	23.9		91.5	8%
体育		10.5	33.7		8.3	52.5	5%
计算机		24.8	65.6			90.4	9%
法律		2.1	26.3			28.4	3%
总计	44.9	120.5	571.8	182.1	131.2	1,051	
%	4%	12%	54%	17%	13%		100%

附录 2　各类词在标准文本中的分布

词的类型		词次数	百分比
词表词		205,162	83.0%
专有名词（共 13,508 ）	人名	4,347	1.8%
	地名	5,311	2.1%
	机构名	3,850	1.6%
实体名词（共 6,630 个）	日期	2,310	3.2%
	时间	227	
	百分数	311	
	货币	350	
	数字	3387	
	度量	451	
	电子邮箱	0	
	电话	13	
	Web 网址	0	
标点符号		20,250	8.2%
其他		265	1.0%
合计		247,039	100%