文章编号: 1003-0077(2007)03-0008-012

中文分词十年回顾

黄昌宁[1],赵 海[2]

(1. 微软亚洲研究院,北京 100080; 2. 香港城市大学,香港)

摘 要: 过去的十年间,尤其是 2003 年国际中文分词评测活动 Bakeoff 开展以来,中文自动分词技术有了可喜的进步。其主要表现为: (1)通过"分词规范 + 词表 + 分词语料库"的方法,使中文词语在真实文本中得到了可计算的定义,这是实现计算机自动分词和可比评测的基础; (2)实践证明,基于手工规则的分词系统在评测中不敌基于统计学习的分词系统; (3)在 Bakeoff 数据上的评估结果表明,未登录词造成的分词精度失落至少比分词歧义大 5 倍以上; (4)实验证明,能够大幅度提高未登录词识别性能的字标注统计学习方法优于以往的基于词(或词典)的方法,并使自动分词系统的精度达到了新高。

关键词: 计算机应用; 中文信息处理; 中文分词; 词语定义; 未登录词识别; 字标注分词方法

中图分类号: TP391 **文献标识码**: A

Chinese Word Segmentation: A Decade Review

HUANG Chang-ning[1], ZHAO Hai[2]

(1. Microsoft Research Asia, Beijing 100080, China; 2. City University of Hong Kong, Hong Kong, China)

Abstract: During the last decade, especially since the First International Chinese Word Segmentation Bakeoff was held in July 2003, the study in automatic Chinese word segmentation has been greatly improved. Those improvements could be summarized as following: (1) on the computation sense Chinese words in real text have been well-defined by "segmentation guidelines + lexicon + segmented corpus"; (2) practical results show that performance of statistic segmentation systems outperforms that of handcrafted rule-based systems; (3) the evaluation in terms of Bakeoff data shows that the accuracy drop caused by out-of-vocabulary (OOV) words is at least five times greater than that of segmentation ambiguities; (4) the better performance of OOV recognition the higher accuracy of the segmentation system in whole, and the accuracy of statistic segmentation systems with character-based tagging approach outperforms any other word-based system.

Key words: computer application; Chinese information processing; Chinese word segmentation (CWS); definition of words; out-of-vocabulary (OOV) word recognition; Character-based tagging approach of CWS

十年前,笔者受《语言文字应用》杂志主编费锦昌先生之托,主持了该杂志以自动分词为题的中文信息处理专题讨论,并为此以"中文信息处理中的分词问题"为题写了一篇短文,向语言学界的同行介绍计算机信息处理研究所面临的几个语言学问题[1]。根据当时的认识,笔者在这篇文章中提出了中文分词研究的四个难题: (1)"词"是否有清晰的界定; (2)分词和理解孰先孰后; (3)分词歧义消解; (4)未登录词(Out-of-vocabulary,简称 OOV)识别。如今十年过去了,经过国内外同行的不懈努力,在这四个问题上我们究竟取得了哪些进展呢?这正是本文要逐一回顾的问题。十年间,尤其是 2003 年 7 月首届国际中文分词评测活动 Bakeoff[2] 开展以来,中文自动分词技术有了可喜的进步。主要表现为: (1)通过"分词规范 + 词表 + 分词语料库"的方法,使中文词语在真实文本中得到可计算的定义,这是实现计算机自动分词和可比评测的基础; (2)实践证明,基于手工规则的分词系统在评测中不敌基于统计学习的分词系统; (3)在 Bakeoff 数据上的估算结果表明,未登录词(OOV)造成的分词精度失落

收稿日期: 2007-03-22 **定稿日期**: 2007-03-22

作者简介: 黄昌宁(1937—),男,微软亚洲研究院高级顾问,主要研究方向为计算语言学、中文信息处理。

至少比分词歧义大 5 倍以上; (4) 迄今为止的实验结果证明,能够大幅度提高未登录词识别性能的字标注统计学习方法优于以往的基于词(或词典)的方法,并使自动分词系统的精度达到了新高。

1 "词"是否有清晰的界定?

自动分词的一个重要前提是: 至少要在计算的意义上清楚界定真实文本中每个词语的边界。然而,这样一个起码的要求在十年前还是可望而不可及的奢想。

在每本汉语语法教科书中,都可以找到有关"词"的一条相当抽象的定义: 语言中有意义的能单说或用来造句的最小单位。在计算上,这种模棱两可的定义是不可操作的,或者说,是不可计算的。即使在母语为汉语的话者之间,中文词语的平均认同率也只有 0.76 左右[3]。

经过信息界和语言学界的共同努力,在 1993 年作为国家标准公布的《信息处理用现代汉语分词规范》[4]中,文本中的词语被称为"分词单位",以区别于语言学中更严格的"词"概念。国家标准按词类分别给出了各类分词单位的定义;然而,在许多地方无可奈何地把"结合紧密、使用稳定"视为分词单位的界定准则。众所周知,像"紧密"和"稳定"这样的判断是相当主观的,见仁见智。因此,无论在分词系统的实现上还是评测上都造成了极大的困惑。一句话,对文本中的词,人都没界定清楚,让计算机去做自动分词不是勉为其难了吗?

同许多同行一样,十年前笔者只认识到这是计算机自动分词面临的最大难题,而且寄希望于 1993 年公布的那个分词规范能够最终成为被公众普遍认同的标准。当时还认为"分词规范 + 词表"也许能更好地界定句子中的词语,即经过大规模语料的计算筛选,实现"结合紧密"和"使用稳定"的定量化[5]。大陆举办的历届 863、973 分词评测[6,7]也是遵照统一分词规范的思路来组织的。在这些评测中,组织者不公布词表和相关的分词语料,而参评系统输出的分词结果有时还允许有一定的"柔性"[8],即分词结果尽管同标准答案不一样,如果仍符合"结合紧密,使用稳定"的规范条款,就不算出错。这种评测方法的不足在于一定程度上引入了评测人员的主观判断,说到底还是缺少对什么是词的可计算定义。比如分词系统的"召回率"指标,其分母本该是标准答案中的总词次数,现在究竟是标准答案中的总词次数,还是待测系统输出中符合"柔性"答案的总词次数呢?如果是前者,就忽视了标准答案非同一性带来的偏误;如果是后者,又在一定程度上损失了可比性。

1.1　国际中文分词评测 Bakeoff

2003 年 7 月 SIGHAN①在日本札幌举办了首届国际中文分词评测 Bakeoff[2]。Bakeoff 采用了不同于国内 863、973 评测的另外一种分词评测方案。即事先在网上公布四种不同标准的训练语料(带标语料),一个月后公布与这四种标准相应的测试语料(原始语料)。表 1 是历届 Bakeoff 公布的 12 种分词语料库的统计数据。参评系统可以在这些语料中任意选择一种或多种标准来考评自己的分词系统。在每种语料库上又分封闭和开放两种测试: 封闭测试只允许使用从指定训练语料中获取的知识(如词表、N 元文法等)来从事自动分词学习;开放测试则不受这样的约束。换句话讲,Bakeoff 认识到短时期内各界不可能在一种分词标准上达成共识。那么不如换一个思路,让多个不同标准的分词语料同台测试。因为评测的主要目的是推动分词技术的进步,而不是制定统一的分词规范。由于为 Bakeoff 提供训练/测试语料库的单位都有各自的分词规范和词表,而且这些语料库都经过人工审定。因此,至少在每种语料库内部可以保证分词标准的一致性。

1.2　严格的质量控制

笔者认为,Bakeoff 通过不同标准的分词语料同台测试,完成了从"分词规范"到"规范 + 词表",再从"规范 + 词表"到"分词语料库"的"词语"定义过程。这是因为语料库的提供单位并不公布他们使用的词表(如果存在这样一个词表的话),所以在封闭测试中可供参评系统学习(或观察)的唯一材料就是分词语料库本身。从计算的意义上来说,一定规模的分词语料库(即训练集)不仅代表了一种特定的分词规范,而且体现了词语的一种可计算定义。然而,要制作高质量的分词语料库,分词规范和词表都是不可或缺的。

有些语料库提供单位对语料标注的质量重视不足,如 2003 年 PKU 和 AS 的测试语料出错率分别达到了 1.29 %和 2.26 %[11],造成了 Bakeoff 评测结

① SIGHAN 是国际计算语言学会(ACL)下属的"中文处理专业委员会"的简称,网址 http://www.sighan.org。

表1 历届 Bakeoff 公布的分词语料库一览表[2,9,10]

提供者	语料库	编码	训练集词次数	测试集词次数	OOV 率①
台湾"中研院"	AS2003	Big5	5.8M	12K	0.022
	AS2005		5.45M	122K	0.043
	AS2005		5.45M	91K	0.042
香港城市大学	CityU2003		240K	35K	0.071
	CityU2005		1.46M	41K	0.074
	CityU2006		1.64M	220K	0.040
美国宾州大学	CTB2003	GB	250K	40K	0.181
	CTB2006		508K	151K	0.088
微软亚洲研究院	MSRA2005		2.37M	107K	0.026
	MSRA2006		1.26M	100K	0.034
北京大学	PKU2003		1.1M	17K	0.069
	PKU2005		1.1M	104K	0.058

果的偏误[11,13]。其实,进一步提高分词语料的标注质量不仅是自动评测的需要,而且也是寻求中文词语可计算定义的必由之路。

笔者在标注和审定 MSRA 分词语料库的实践中体会到,语料标注的质量取决于以下三条:(1)严格执行"词表驱动"原则;(2)把人名、地名、机构名等命名实体和日期、时间等数字表达式的定义纳入分词规范;(3)把规范制定和语料标注两个过程紧密结合起来,务使规范达到词例化的详尽程度(详见下文对(2)、(3)的解释)。

所谓"词表驱动",就是在相关上下文中未见歧义的情况下,词表词应当作为一个完整的切分单位,决不许随意切碎或组合。必须杜绝所谓的"语法词"(比词表词短)和"心理词"(非词表词,又不属于新词)的干扰。孙茂松曾主张分词语料库以切成"心理词"为宜,并列举了像"酒瓶"、"烟厂"、"车门"、"雨水"、"大海"、"坏人"、"重建"、"分管"、"唱歌"、"吸烟"、"打断"、"缩短"、"发起"、"穿过"、"等于"、"取决于"、"国内外"、"工业化"等众多的"心理词"[12]。孙的结论并不乐观,他认为"心理词"的模糊性决定了严格意义上的切分一致性是不可能实现的。笔者不同意这种观点,而是主张事前认真判断这些词是否应该进入词表。这样的词一旦进入词表,就具有了词语的合法身份。在自动分词和人工审定的过程中如果能认真落实"词表驱动"原则,就一定能彻底消除"心理词"的干扰,保证切分结果的一致性。

前面提到 PKU 和 AS 语料中出现的分词错误,纠其原因就是在人工审定过程中违背了"词表驱动"原则。例如在 PKU2003 测试集中的以下词语:"如果说、交通部门、问明、大世界、全城、区内、庆春节、没想到、文化村、老同志、迎春花市、上图、西站、下图、冰上、回老家、县市区、极端分子"等,在训练集中在上下文未见歧义的情况下却不是词。例如:

测试集- # 331 等(3 次):西站/ 今年/ 春运/ 期间/ 新增/ 售票/ 窗口/ 24/ 个/ ,

训练集- # 10044 等(7 次):3/ 日/ ,/ 铁道部/ 副/ 部长/ 刘/ 志军/ 到/ 北京/ 西/ 站/ 现场/ 办公/ ,

训练集- # 4489(1 次):北京/ 铁路/ 公安处/ 与/ 西站/ 公安段/ 在/ 北京/ 西/ 站/ 开展/ 了/『雷/ 锋/ 在/ 我/ 心中/ ,/ 干警/ 在/ 您/ 身边/』/ 的/ 爱/ 民/ 便民/ 活动/ 。

反之,训练集中的下列词:"留学人员、贺岁、除夕夜、大酒店、同一天、新世纪、下雪、集团公司、西站、高等学校、羽毛球队、科教兴国、尽管如此、冰清玉洁"等,在测试集中在上下文未见歧义的情况下被切碎了。例如:

① 未登录词(OOV),泛指文本中出现的专有名词和新词等非词表词。这里 OOV 专指,在 Bakeoff 的一种语料库的测试集中出现而未曾在其训练集中观察到的那些词。OOV 率是指测试集的未登录词出现次数在该测试集总词次数中所占的比率。一般来说,OOV 率越高的语料切分难度也越高。

训练集- # 6941 等(17 次): 李/ 岚清/ 强调/ ,/ 科教兴国/ ,/ 教育为本/ 。

测试集- # 27 等(2 次): 大力/ 实施/ 科教/ 兴/ 国/ 战略/ 和/ 可/ 持续/ 发展/ 战略/ 。

其实,这种切分错误在一定程度上是可以通过自动检查程序(AutoCheck)来纠正的[11]。

保证分词标注质量的第二条措施是把人名、地名、机构名等命名实体和日期、时间等数字表达式的定义纳入分词规范。一方面,这是因为实体词的识别任务与自动分词任务,你中有我,我中有你,是不可分割的整体。另一方面,是因为这些实体词占了文本中未登录词的大约三分之二[14],把它们定义清楚了肯定有助于进一步提高标注的一致性。微软亚洲研究院制定的中文分词规范就涵盖了上述实体词的详尽定义。例如,"珠江"是一个无可争辩的地名,但"珠江流域"、"珠江三角洲"、"珠江中下游"算不算地名?"海淀区知春路"是一个地名,还是两个地名?天体、行星是不是地名?"奥运会"、"宋庆龄基金会"算不算机构名?这些细节如果不定义清楚,怎么能保证语料标注的一致性呢?

实施质量控制的第三条措施是,让分词规范的制定与分词语料的标注、审定过程交互进行。因为词表只是对"词语"的一种静态描写,没有说明每个词进入句子以后同周围的词发生的粘着、竞争、重组等复杂行为。换句话讲,当文本中动态出现未登录词(OOV)和交集型歧义(Overlapping Ambiguity, OAS)、组合型歧义(Combination Ambiguity, CAS)等现象时,需要在分词规范中引用带标语料库的大量实例来进一步完善相关词语的定义。下面就是这样的一些实例:

< 1a > 发行/ 公司/ 与/ 制作/ 公司/ 、/ 音乐/ 人/ 之间/ 正在/ 实现/ 集团/ 化/ 联合/ ,

< 1b > / 正/ 在/ 北京八十中学/ 就读/ 的/ 关序/ 指挥/ 八十/ 多/ 人/ 的/ 乐队/ 、/ 一百二十/ 多/ 人/ 的/ 合唱/ 队/ 在/ 政协礼堂/ 、/ 朝阳剧场/ 演出/ 交响乐/ 《/ 沙家浜/ 》/ 。

< 2a > 沈国放/ 在/ 会上/ 阐述/ 了/ 中国政府/ 的/ 有关/ 立场/ 。

< 2b > 在/ 全体/ 干部/ 会/ 上/ 宣布/ : / 李炳钦/ 为/ 竹林总公司/ 副/ 总经理/ 。

< 3a > 在/ 此/ 情形/ 下/ ,/ 军政府/ 和/ 反对派/ 将/ 怎样/ 应对/ 变/ 局/ 。

< 3b > 国际/ 社会/ 应/ 对/ 以/ 采取/ 严正/ 立场/ ,/ 迫使/ 其/ 履行/ 马德里/ 和会/ 确立/ 的/"/ 以/ 土地/ 换/ 和平/ "/ 的/ 原则/ 。

< 4a > 身/ 患/ 胃癌/ ,/ 还/ 经常/ 去/ 道班/ ,/ 和/ 道班/ 工人/ 吃/ 住/ 在/ 一起/ 。

< 4b > 最近/ ,/ 内蒙古/ 赤峰市/ 又/ 发生/ 一/ 起/ 小/ 煤窑/ 淹/ 井/ 事故/ ,/ 17/ 人/ 死亡/ 。

< 5a > 香港中旅/ 与/ 中国/ 旅行社/ 一道/ ,/ 努力/ 降低/ 内地/ 赴/ 港/ 旅游团/ 价格/ ,

< 5b > 小/ 红/ 帽/ 、/ 红/ 马甲/ 成为/ 沈阳/ 初夏/ 文化/ 市场/ 一/ 道/ 亮/ 丽/ 的/ 风景/ 线/ 。

< 6a > 儿童文学/ 原创/ 作品/ 对/ 整个/ 少年儿童/ 读物/ 出版/ 的/ 整体/ 走向/ 有着/ 重大/ 影响/ 。

< 6b > 她们/ 的/ 视野/ 已/ 从/ 庭院/ 式/ 的/ 经济/ 走/ 向/ 市场/ 。

2005-2006 年微软亚洲研究院(MSRA)提供给 Bakeoff 的语料库,由于严格实施质量控制,其百万词级训练语料库的出错率低于千分之一,十万词级测试语料库的出错率低于万分之五。这个结果也从一个侧面反映出我们在词语定义的规范化方面取得了实质性进步。

1.3　词语认同率对比

1996 年 Sproat 等通过六个母语为汉语的话者对同一篇文本的分词结果的对比,得出人与人之间的词语平均认同率只有 0.76 左右的结论[3]。而笔者利用 Bakeoff-2006 的四种分词标准不同的语料库进行了一次机器与机器之间的词语认同率对比,发现尽管不同语料库的分词标准不尽相同,机器之间的词语平均认同率却高达 0.90。这是不是也可以证明十年来业界在中文词语定义的规范化方面取得了实质性的进步呢?

在以下的两个实验中,我们遵循常规的测试方法,即每次测试轮流采用其中一个分词结果作为标准答案(Golden Standard)。评价指标是分词召回率 R、准确率 P 以及 R 和 P 的平均值 F①。召回率定义为给定分词结果中切分正确的词次数除以标准答案中的总词次数,准确率定义为给定分词结果中切分正确的词次数除以该分词结果中的总词次数。

表 2 示出六个母语为汉语的话者对同一篇中文文本的切分结果[3]。文本由 100 个句子组成,含 4 372 个字。M1-M3 和 T1-T3 分别代表三位大陆话者和三位台湾话者。从表 2 中可以看到,这六个人

① 表 2 采用 R 和 P 的算术平均值 $F = (R + P)/2$。表 3 采用 R 和 P 的调和平均值 $F = 2PR/(R + P)$。

之间的词语认同率最低为 0.69,最高为 0.89,平均认同率为 0.76。

表 2 不同人之间的词语认同率(F 值)

	M1	M2	M3	T1	T2	T3
M1		0.77	0.69	0.71	0.69	0.70
M2			0.72	0.73	0.71	0.70
M3				0.89	0.87	0.80
T1					0.88	0.82
T2						0.78

在不同机器的词语认同率对比实验中,我们以 Bakeoff-2006 的四种不同标准的分词语料库为对象。首先,用每个语料库的训练集分别训练出四个基于字标注的条件随机场(Conditional Random Field,简称 CRF)分词系统[15],每个分词系统都用相应语料库的名字命名。例如,用 AS 语料库训练的分词系统叫 AS,用 CityU 语料库训练的分词系统叫 CityU 等。我们的基本假设是:AS 分词系统最好地体现了 AS 的分词标准,而 CityU 的分词系统则最好地体现了 CityU 的分词标准,依此类推。第二步,用上述四个分词系统分别去切分不同语料库的测试集。这样在四个测试集上总共得到 16 个分词结果。第三步,假设每个分词系统在自己的测试集上得到的 F 值叫做 F_0。一般来说,F_0 会高于该分词系统在其他测试集上获得的分词精度,如 AS 分词系统在 AS 测试集上的 F_0 值最高。四个分词系统的 F_0 值分别是:0.953 8(AS),0.932 0(CTB),0.969 1(CityU),0.960 8(MSRA)。为了在不同分词系统之间进行词语认同率对比,我们对每个分词结果的 F 值实行归一化,即让每个分词系统在四个测试集上的 F 值都除以各自的 F_0 值,这样,一个分词系统在自己的测试集上的 F 值将等于 1.0(见表 3)。

表 3 不同分词系统之间的词语认同率(F 值)

测试语料库	分词系统			
	As	CTB	CityU	MSRA
AS2006	1.0	0.959 3	0.925 6	0.858 3
CTB2006	0.942 0	1.0	0.910 4	0.877 4
CityU2006	0.932 1	0.934 6	1.0	0.848 8
MSRA2006	0.857 0	0.886 6	0.848 3	1.0

从表 3 中可以看出,在四个不同分词系统之间词语认同率最低为 0.848 3,最高达 0.959 3,平均认同率达 0.90,大大高于 Sproat 统计的人间平均认同率 0.76。应当指出,由于 MSRA 的分词规范把命名实体(地名、机构名等专有名词和日期、时间等数字表达式)整体视为分词单位,所以 MSRA 语料库的平均词长略大于其他三个语料库(见表 7)。因而,如果只考虑其他三个分词系统,那么不同系统之间的词语平均认同率高达 0.93。

综上所述,原来很难精确定义的"词",通过"分词规范 + 词表 + 分词语料库"的方法得到了计算机所需要的可计算定义。这种体验能不能推广到像命名实体、词性、语块(Chunk)等其他语言对象上去,是一个值得探讨的问题。

2 理解和分词孰先孰后

十年前,笔者曾指出,由于自动分词是大部分中文信息处理系统的第一步(即前端),是对句子实施句法—语义分析的前提。也就是说,自动分词所依据的只能是文本的表层信息。所以,尽管人在识别句子中的词语时是以理解①为基础的,然而从实用的角度考虑,计算机自动分词系统不可能完全照搬人类的分词模式,而通常会选择"先分词后理解"的处理策略。

然而有些研究人员相信自然语言理解是一切文本分析,包括自动分词,的基础,所以提出了另一条技术路线——"先理解后分词"。把这种建议认真加以实现的是 Wu[16]。文献[16]重点分析了分词中的歧义消解问题。Wu 主张把中文分词建立在句法分析器 NLPwin 的基础上,即把分词的决定放在句法分析的过程中去解决,而不是过早地在句法分析前就做出决定。NLPwin 是一个强大的基于句法-语义规则的句子分析系统。它的词典是在北京大学的语法信息词典上开发的,当时词条已扩充到八万以上,并装备了大量歧义消解信息。Wu 认为,传统的最大匹配(Maximum Matching,简称 MM)算法缺少全局信息,而统计方法则缺少(句子)结构信息。句法分析器可以同时提供这两方面的信息,因此可以在整句"理解"的基础上达到最佳的歧义消解效果。由于当时还没有一个公开、可比的分词评测语料,Wu 的实验结果是在一个自选的小测试集上完成的。该测试集由 100 个句子组成,每个句子包含一个交集型或组合型歧义字串,它们大部分选自同

① 这里所谓的"理解"当然是指了解句子的意思,在计算上指句子句法—语义分析的结果,而不是指词或词性这样的表层信息。

行在文献中曾经用过的歧义例句。Wu 报告,句子分析的正确率达到 85 %(即 100 个句子中有 85 句分析结果正确)。分析结果正确的句子,其分词结果也是正确的。分析错误的句子则多半会包含一个切分错误。Wu 称,如果用切分正确的词次数除以被测句子中的总词次数来计算分词精度(Accuracy),那么,NL Pwin 达到的分词精度为 99 %。

然而,Wu 的上述实验结果的可信度是值得商榷的。原因有二: 一是测试集太小,只有 100 个句子,而且作者收集的这些歧义字串不具备统计意义上的采样随机性,所以不代表在真实文本中歧义字串实际出现的规律;二是每个歧义字串出现的上下文环境是固定的,因此,即使某个歧义字串在这个句子中切对了,不等于它在另一个句子里也一定能切对。换句话讲,Wu 报告的消歧成绩过于乐观了。

为了证实笔者的上述判断,我们不妨来看一下 NL Pwin 在组合歧义字串"才能"上的切分结果。选择字串"才能",是因为在文献[16]中作者一开始就用这个字串来解释基于句法分析的分词原理,而且在文后附录中给出的 20 个测试例句中就有两个句子含有这个歧义字串,它们是:

(a) 他有各种才能。

(b) 什么时候我才能克服这个困难?

我们用一个随机选自真实文本的 1 000 句测试集来考察 NL Pwin 的分词性能,下面是 NL Pwin 对该测试集中包括"才能"字串的五个句子的切分结果:

× (1) 股票/ 投资者/ 的/ 基本/ 权利/ 才能/ 得到/ 保障/ 。

× (2) 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/ 。

(3) 与/ 之/ 配套/ 的/ 软件/ 才/ 能/ 调试/ ,

× (4) 切实/ 纠正/ 有偿/ 新闻/ 等/ 不正之风/ ,/ 才能/ 更好/ 地/ 为/ 人民/ 服务/ 。

(5) 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/ 。

在上面五个句子中,NL Pwin 只切对了两句: (3)和(5),按句子计算正确率 40 %。意味深长的是,这五个句子中的"才能"都是要切开的,而且(1)和(3)、(4)和(5)的上下文或句法结构分别是相似的。为什么一个句法分析器会错对参半呢? 句法规则在这里出了什么毛病了吗? 根据笔者对大规模语料库的调查,在"才能"字串中"才/能"出现的概率是 0.94 左右(1 035 次/ 1 100 次)。所以可以用一个简单的基于先验概率的消歧方法: 见到"才能"字串统统切开,那么在这个测试集上的五个"才能"字串就都切对了。

为了进一步观察,下面是 NL Pwin 对 1 000 句测试集中更多组合型(6) —(10)和交集型(11) —(15)歧义字串的误切结果:

(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/ ,

(11) 决定/ 在/ 全/ 省/ 戒/ 玩/ 风/ ,/ 兴学/ 风/ ,

(12) 最大/ 限度/ 地/ 防止/ / 有害/ 信息流/ 入/ 和/ 传播/

(13) 保/ 修/ 条款/ 亦/ 不详/ 尽/ ,

(14) 挽救/ 一/ 个/ 人生/ 命/ 的/ 义务/ 将/ 凌驾/ 于/ 不/ 侵犯/ 别人/ 隐私/ 的/ 义务/ 。

(15) 改变/ "/ 一 手/ 硬/ ,/ 一/ 手 软/ "的/ 状态/ ,/ 有/ 新闻界/ 的/ 一/ 份/ 功劳/ 。

以上 NL Pwin 误切实例说明: 一方面,句法分析器认为分析"正确"的句子(即输出的句子),其分词结果未必是正确的。因为迄今还没有一部语法能保证不符合该语法(Ungrammatical)的句子都能被分析器所拒绝(即不生成句法树)。何况当前及今后一段时间里,句法分析器本身的正确率远低于自动分词系统。因此,靠句法分析器这样的高端技术来解决自动分词这样相对初级的文本处理问题,似乎在逻辑上就有违于常理。另一方面,许多分词错误单靠通常意义上的句法-语义知识也是鞭长莫及的。例如,句(7)中词表词"对路"是一个形容词,分析系统根据上下文的词类信息所得出的结构 ——"/ 对路/ 的/ 渴望",大概不能算错吧? 又例如,句(8)中词表词"群山"是个名词,句法分析器根据上下文得到的局部结构 ——"这/ 群山/ 里/ 的/ 女/ 娃娃",也属合情合理。再如,句(11)中"兴学"和"学风"同是词表词,除非采用一条词例化规则: "兴学风 →兴/ 学风",否则要靠语言知识来正确切分这个交集型字串真是太难了。因为孤立地看另一种切法"兴学/ 风"也

不无道理,不是常见"抢购风"、"吃喝风"什么的吗?

应当指出,Bakeoff 对中文分词技术的一大贡献就在于它提供了一个公开、可比的分词性能测试平台,包括多种训练、测试语料和多视角的分词评测指标。多年来在自动分词上难分仲伯的某些分词技术,在相同的测试环境下往往顿显高下。

2003 年 Wu 的 NLPwin 系统参加了第一届 Bakeoff 评测,取得很好的成绩:一个第一(PKU 开放)、两个第二(PKU 封闭,CTB 开放)和一个第三(CTB 封闭)。但面对 Bakeoff 的实验数据,Wu 承认句法分析器对分词性能的影响十分有限,在 CTB 语料的封闭测试中,采用句法分析器的分词精度甚至低于没有句法分析器的情况。他同意为实用目的(Practical Purpose)将在分词中放弃句法分析①[17]。

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。

3 未登录词对分词精度的影响

长期以来,研究人员一直把未登录词和分词歧义并列为影响分词精度的两大因素。十年前,笔者自己的认识也是这样的。十年来,研究人员在这两个问题上倾注了大量的精力,探索过各种各样的解决方案。其中,对交集型歧义字串进行的大规模语料库调查,以及明确提出把分词歧义消解过程分解为侦察和消歧两个子过程的认识[7],都是近十年来分词研究的重大收获。然而未登录词和分词歧义两者究竟孰重孰轻,亟需有一个定量的分析。因为这个问题其实影响着自动分词系统的总体设计思路。2003 年笔者对 Bakeoff 的关注也在很大程度上跟这个问题有关。

每届 Bakeoff 都用正向最大匹配(Forward Maximum Matching,简称 FMM)算法对每个语料库进行带有未登录词的基线(Baseline)和不含未登录词的顶线(Topline)两种切分,并分别形成两套性能指标。其中,F_{base} 和 F_{top} 分别表示基线和顶线的 F 值。Xue 曾用(F_{top} - F_{base})表示未登录词单独给分词系统带来的精度失落[18]。笔者进一步用(1 - F_{top})表示分词歧义单独造成的分词精度失落[19]。

表 4 示出 Bakeoff-2003 四个语料库不含和含有未登录词的 FMM 分词性能对比。数据显示,CTB 语料库的未登录词率为 0.181(见表 1),是四个语料库中最高的。在这个语料库上,未登录词单独造成的分词精度失落(F_{Top} - F_{Base})最高,0.260;歧义切分单独造成的精度失落(1 - F_{Top})也最高,0.015。而在 PKU 语料库上,未登录词造成的分词精度失落 0.128,在四个语料库中仅次于 CTB;但其歧义切分造成的精度失落 0.005 是最低的;而未登录词造成的分词精度失落比歧义切分造成的精度失落(F_{Top} - F_{Base})/(1 - F_{Top})大 25.6 倍,为四个语料库之冠。这个统计结果表明,在 Bakeoff-2003 的四个语料库中,未登录词造成的分词精度失落比歧义切分造成的精度失落至少大 10 倍左右。

表 4 Bakeoff-2003 四个语料库不带 OOV 和带 OOV 的 FMM 分词性能对比

语料库	AS2003	CityU2003	CTB2003	PKU2003
顶线 F_{top}	0.992	0.989	0.985	0.995
基线 F_{base}	0.915	0.867	0.725	0.867
F_{top} - F_{base}	0.077	0.122	0.260	0.128
1 - F_{top}	0.008	0.011	0.015	0.005
比率	9.6	11.1	17.3	25.6

为了观察更多的材料,表 5 给出了 Bakeoff-2005 的另外四个语料库的统计数据,说明未登录词造成的分词精度失落比歧义切分造成的精度失落大 5.6-14.2 倍之间。综合表 4 和表 5 的统计我们说,在大规模真实文本中未登录词造成的分词精度失落比歧义切分造成的精度失落至少大 5 倍以上②是可

表 5 Bakeoff2005 语料库不带 OOV 和带 OOV 的 FMM 分词性能对比

语料库	AS2005	CityU2005	MSRA2005	PKU2005
顶线 F_{top}	0.982	0.989	0.991	0.987
基线 F_{base}	0.882	0.833	0.933	0.869
F_{top} - F_{base}	0.100	0.156	0.058	0.118
1 - F_{top}	0.018	0.011	0.009	0.013
比率	5.6	14.2	6.4	9.1

① 在同笔者的通信中 Wu 指出,他的分词策略是"理解与分词互动",而不是"先理解后分词"。他认为,"理解离不开分词,分词也离不开理解;很多歧义需要理解才能消歧。这种理念是好的,但目前技术尚未成熟,所以还不具有实用价值"。详见第 20 页本文的附录,该附录并不代表笔者的观点,仅供读者参考。

② 当然,具体的数值可能随着所用语料的改变而有所波动。

信的。这个结论给我们的重要启示是: 在考虑自动分词系统的总体方案时,那些能够大幅度提升未登录词识别性能的分词方法,一般来讲,也将提高分词系统的总体性能。Bakeoff-2003 及其后的分词技术发展趋势完全证实了这样一个推断。

4　基于字的分词新方法

4.1　新方法崭露头角

在 2002 年之前,自动分词方法基本上是基于词(或词典①)的,在此基础上可进一步分成基于规则和基于统计的两大类。第一篇基于字标注(Character-based Tagging)的分词论文发表在 2002 年第一届 SIGHAN 研讨会上[18],当时并未引起学界的重视。一年后,Xue 在最大熵(Maximum Entropy, ME)模型上实现的基于字的分词系统参加了 Bakeoff-2003 的评测[20],在 AS 语料库的封闭测试项目上获得第二名(见表 6),然而其 OOV 召回率 R_{OOV} (0.729)却位居榜首。Xue 还在 CityU 语料库的封闭测试中获得第三名,其 R_{OOV} (0.670)仍然是该项比赛中最高的[2]。尽管在 Bakeoff2003 中各种分词技术的优劣尚难分仲伯,但既然未登录词对分词精度的影响比分词歧义至少大 5 倍以上(见上节),我们自然看好这种能获致最高 OOV 召回率②的分词方法。这一预测果然在 Bakeoff2005 上得到了证实。

表 6　AS2003 封闭测试前三名的正式成绩

参赛队	召回率 R	精确率 P	调和均值 F	R_{OOV}	R_{IV}
UC Berkley	0.966	0.956	0.961	0.364	0.980
Nianwen Xue	0.961	0.958	0.959	0.729	0.966
Nara IST Japan	0.944	0.945	0.945	0.574	0.952

基于字标注的分词系统在 Bakeoff-2005 上崭露头角。其中 Low[21]的系统采用最大熵模型,在四项开放测试中夺得三项冠军(AS, CityU, PKU)和一项亚军(MSRA)。Tseng 的系统采用条件随机场模型,在四项封闭测试中取得两项冠军(CityU, MSRA)、一项亚军(PKU)和一项季军(AS)[22]。到了 Bakeoff-2006,基于字的分词系统已遍地开花。其中,笔者用条件随机场模型实现的基于字标注的分词系统,在参加的六项分词评测中,夺得四个第一(CityU 开放,AS 开放,AS 封闭,CTB 封闭)和两个第三(CTB 开放,CityU 封闭)[15]。

以往的分词方法,无论是基于规则的还是基于统计的,一般都依赖于一个事先编制的词表(词典)。自动分词过程就是通过词表和相关信息来做出词语切分的决策。与此相反,基于字标注的分词方法实际上是构词方法。即把分词过程视为字在字串中的标注问题。由于每个字在构造一个特定的词语时都占据着一个确定的构词位置(即词位),假如规定每个字最多只有四个构词位置: 即 B(词首),M(词中),E(词尾)和 S(单独成词),那么下面句子(甲)的分词结果就可以直接表示成如(乙)所示的逐字标注形式:

(甲) 分词结果: / 上海/ 计划/ 到/ 本/ 世纪/ 末/ 实现/ 人均/ 国内/ 生产/ 总值/ 五千美元/ 。

(乙) 字标注形式: 上/B 海/E 计/B 划/E 到/S 本/S 世/B 纪/E 末/S 实/B 现/E 人/B 均/E 国/B 内/E 生/B 产/E 总/B 值/E 五/B 千/M 美/M 元/E。/S

首先需要说明,这里说到的"字"不只限于汉字。考虑到中文真实文本中不可避免地会包含一定数量的非汉字字符,本文所说的"字",也包括外文字母、阿拉伯数字和标点符号等字符。所有这些字符都是构词的基本单元。当然,汉字依然是这个单元集合中数量最多的一类字符。

把分词过程视为字的标注问题的一个重要优势在于,它能够平衡地看待词表词和未登录词的识别问题。在这种分词技术中,文本中的词表词和未登录词都是用统一的字标注过程来实现的。在学习架构上,既可以不必专门强调词表词信息,也不用专门设计特定的未登录词(如人名、地名、机构名)识别模块。这使得分词系统的设计大大简化。在字标注过程中,所有的字根据预定义的特征进行词位特性的学习,获得一个概率模型。然后,在待分字串上,根据字与字之间的结合紧密程度,得到一个词位的标注结果。最后,根据词位定义直接获得最终的分词结果。总而言之,在这样一个分词过程中,分词成为字重组的简单过程。然而这一简单处理带来的分词结果却是令人满意的。

4.2　字的词位分类及其所用的基本特征

现代机器学习的主要方法,包括支持向量机

① 尽管有所谓的无词典自动切词方法,但由于分词精度太低,在要求较高的应用中并未形成实用价值。

② 更准确地说,应该采用未登录词识别的 F 值而不是召回率 R。

(Support Vector Machine, SVM)、最大熵和条件随机场,都已经被研究人员用于由字构词的词位学习中。事实上,由于分词在中文信息处理中的初级地位,可供选用的特征也非常少。迄今为止,最常用的两类特征是字本身以及词位(状态)转移概率(这里我们沿用隐马尔科夫模型(HMM)中的术语)。

对于支持向量机和最大熵方法来说,需要设计独立的状态转移特征来表达词位的转化。但是对于一阶线性链条件随机场学习来说,这一转移过程将被自动集成到系统中来,而无需专门指定。这样,对于采用条件随机场建模的分词系统来说,需要考虑的仅仅是字特征。

词位学习中确定字特征的主要参数是上下文窗口的宽度,也就是使用距当前字多远的字来作为当前字标注的依据。相关工作表明,使用前后各两个字(即 5 个字的窗口宽度)是比较理想的。实际上,根据历届 Bakeoff 提交的报告,很少有系统使用超过 5 个字的窗口宽度。这是具有统计学依据的。笔者统计了 Bakeoff-2003 和 Bakeoff-2005 的全部 8 个训练语料库词长的频率分布,结果见表 7。从中可以看到,在所有语料库中 90 %的词次是 1～2 字词,95 %的词次是 3 字或 3 字以下词,99 %以上的词次都是 5 字或 5 字以下词。因此,使用宽度为 5 个字的上下文窗口足以覆盖真实文本中绝大多数的构词情形。

笔者在文献[23]中给出了一个确定有效词位标注集的定量标准 —平均加权词长。其定义为:

$$L_k = \frac{1}{N_i} \sum_{i}^{K} i N_k \qquad (1)$$

上式中,L_k 是 $i \geq k$ 时的平均加权词长,N_k 是语料中词长为 k 的词次数,K 是语料中出现过的最大词长,N 是语料库的总词次数。如果 $k = 1$,那么 L_1 代表整个语料的平均词长。Bakeoff-2003 和 Bakeoff-2005 各训练语料库的平均加权词长分布数据见表 8。从统计中可以看到,所有语料库的平均加权词长在 1.51～1.71 之间。因此,5 字长的上下文窗口恰好大致表达了前后各一个词的上下文(确切范围是 4.53～5.13)。从这个意义上来说,5 字宽的上下文窗口具备了字和词的双重含义。

表 7 Bakeoff-2003 和 Bakeoff-2005 各训练语料库词长的频率分布

词长	AS2003	AS2005	CityU2003	CityU2005	CTB2003	MSRA2005	PKU2003	PKU2005
1	0.544 7	0.571 2	0.494 0	0.468 9	0.436 7	0.471 5	0.472 1	0.472 7
2	0.393 8	0.378 7	0.427 1	0.455 4	0.471 9	0.438 7	0.450 8	0.449 9
3	0.046 3	0.035 8	0.058 7	0.059 7	0.067 2	0.047 5	0.049 5	0.0495 3
4	0.010 7	0.009 9	0.015 9	0.013 4	0.011 6	0.024 2	0.020 4	0.020 5
5	0.001 8	0.001 9	0.002 4	0.001 6	0.007 6	0.008 9	0.005 7	0.005 6
6	0.000 8	0.000 7	0.001 0	0.000 5	0.002 4	0.003 7	0.000 7	0.000 7
≤7	0.998 7	0.998 6	0.999 6	0.999 8	0.999 2	0.996 2	0.999 7	0.999 5

表 8 Bakeoff-2003 和 Bakeoff-2005 各语料的平均加权词长分布

k	AS2003	AS2005	CityU2003	CityU2005	CTB2003	MSRA2005	PKU2003	PKU2005
1	1.545 8	1.509 0	1.613 0	1.627 5	1.701 6	1.710 2	1.642 9	1.645 5
2	1.001 1	0.937 8	1.119 0	1.158 7	1.264 9	1.240 1	1.170 8	1.172 8
3	0.213 5	0.180 5	0.264 8	0.247 9	0.321 1	0.361 9	0.269 2	0.273 0
4	0.074 7	0.073 1	0.088 7	0.068 8	0.119 5	0.219 3	0.120 8	0.124 4
5	0.032 0	0.033 4	0.025 2	0.015 0	0.073 2	0.122 3	0.039 0	0.042 3
6	0.022 8	0.024 1	0.013 3	0.007 2	0.035 1	0.077 6	0.010 5	0.014 2
7	0.017 8	0.019 9	0.007 2	0.004 4	0.020 7	0.055 2	0.006 3	0.009 9
8	0.013 6	0.016 6	0.003 8	0.002 0	0.013 3	0.037 4	0.002 9	0.006 5

在最大熵或条件随机场学习中,用于语言特征表达的特征函数起到了核心作用。一般来说,特征函数定义在一个加氏集 $H \times T$ 上,其中 H 是可能的上下文或者任意的预定义条件的集合, T 是一组可选的标注集。特征函数通常可以表示如下:

$$f(h,t) = \begin{cases} 1, & \text{if } h = h_i \text{ and } t = t_j \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

上式中, $h_i \in H, t_j \in T$。习惯上,我们把一组上下文特征按照共同的属性分为若干组,称之为特征模板。比如, C_2 代表所有当前字后面的第二个字,就是一个一元(Unigram)特征模板。

MSRA 使用的 N 元(N-gram)特征模板集如表 9 所示。表中的 C_n 代表当前字或者和当前字相距若干位的字。例如, C_0 代表当前字, C_1 代表当前字的后一个字, C_{-1} 代表当前字的前一个字,依此类推。

表 9　MSRA 的 n 元特征模板集

模板集	类　型	特　征
MSRA 特征模板	Unigram(一元)	C_n, $n = -1, 0, 1$
	Bigram(二元)	$C_n C_{n+1}$, $n = -1, 0$
		$C_{-1} C_1$

由于分词本质上是对一个字串中的每一个字作切分与否的二值决策过程,因此大多数基于字的分词方法使用 2 词位的词位标注集。在最大熵模型中,广泛使用的是 4 词位的词位标注集。微软亚洲研究院在 Bakeoff-2006 的参赛系统中,首次使用了 6 词位的词位标注集[15]。这三类标注集的定义如表 10 所示。

表 10　三类词位标注集的定义

标注集	标　记	单字与多字词的词位标注举例
2 词位	B, E	B, BE, BEE, …
4 词位	B, M, E, S	S, BE, BME, BMME, …
6 词位	B, B_2, B_3, M, E, S	S, BE, BB_2E, BB_2B_3E, BB_2B_3ME, BB_2B_3MME, …

和大多数基于 2 词位标注集的系统不同,笔者的工作表明,使用 6 词位标注集,在适当的特征模板配合下,能够更有效地标注每个字的词位信息,从而在总体上获得更好的分词结果。据了解,几乎所有的基于条件随机场的分词系统都使用了 2 词位标注集,因此他们大多需要使用复杂的特征来弥补标注集在表达能力上的不足。与此相反,笔者使用 6 词位标注集,因而即使使用相对简单的 N 元特征模板集,依然取得了领先的分词结果。这一事实表明,尽管分词过程本质上是一个二值决策过程,然而,统筹选择词位标注集和特征模板集通常能获得更好的分词性能。

4.3　MSRA 分词系统的实验结果

如前所述,Bakeoff 的评测分为开放和封闭两种测试。由于开放测试涉及的方法和语言资源变化多样,不便于对分词技术本身做出有效评价,因此我们仅在封闭测试条件下进行实验比较①。表 11 到 13 列出了 MSRA 分词系统的实验结果及其同所有三届 Bakeoff 上最佳结果的对比[2,9,10]。MSRA 分词系统采用条件随机场(CRF)模型[15],本文的所有实验结果都是在 N 元特征模板集(见表 9)和 6 词位标注集上实现的。

表 11　Bakeoff-2003 语料封闭测试的实验结果比较

参与者	F 值			
	AS2003	CityU2003	CTB2003	PKU2003
Peng, 2004	0.956	0.928	0.849	0.941
Tseng, 2005	0.970	**0.947**	0.863	0.953
Bakeoff2003 第一名	0.961	0.940	**0.881** ②	0.951
MSRA 分词系统	**0.973**	**0.947**	0.872	**0.956**

表 11 的实验结果对比,说明今天的分词系统在分词精度上比 2003 年有了较大的提升,这不能不归功于分词方法上的进步。表 12 和表 13 反映的基本上都是基于字的分词方法的实验结果,各系统之间的精度差异虽然不大,但说明各种统计模型和参数

表 12　Bakeoff-2005 语料封闭测试的实验结果比较

参与者	F 值			
	AS2005	CityU2005	PKU2005	MSRA2005
Ng, 2005	**0.953**	**0.950**	0.948	0.960
Tseng, 2005	0.947	0.943	0.950	0.964
Bakeoff2005 第一名	0.952	0.943	0.950	0.964
MSRA 分词系统	**0.953**	0.948	**0.952**	**0.974**

① 由于本文的讨论只限于 Bakeoff 封闭测试,因此没有对语言资源(知识)在未登录词识别和歧义消解方面的贡献展开讨论。对开放测试有兴趣的读者请参阅文献[15,21]。

② 这项成绩与该参赛队的开放测试成绩相同,其合理性有待进一步工作的验证。

选择在起作用。总的来说,实验结果表明,微软亚洲研究院的基于字标注的分词系统能够在历届 Bakeoff 语料库上达到最佳或接近最佳(State-of-the-art)的分词精度。

表 13 Bakeoff-2006 语料封闭测试的实验结果比较①

参与者(数字为参赛者代号)	F 值			
	AS2006	CityU2006	CTB2006	MSRA2006
Bakeoff2006 第一名	0.958	0.972	0.933	0.963
32(PKU, Beijing)	0.953	0.970	0.930	**0.963**
15(AS, Taipei)	**0.957**	**0.972**	/	0.954
26(IIR, Singapore)	0.949	0.965	0.926	0.957
MSRA 分词系统	0.954	0.969	**0.932**	0.961

5 结论

十年来,尤其是 2003 年 Bakeoff 分词评测开展以来,中文分词技术获得了长足的进步。其主要表现为:(1)通过"分词规范 + 词表 + 分词语料库"的方法,使中文词语在真实文本中得到了可计算的定义,这是实现计算机自动分词和可比评测的基础;(2)基于手工规则的分词方法在评测中不敌统计学习方法;(3)在 Bakeoff 数据上的估算表明,未登录词造成的分词精度失落至少比分词歧义大 5 倍以上;(4)因此能够大幅度提高未登录词识别性能的分词方法必将带动分词系统整体性能的提升。基于字标注的统计学习方法正是在这种背景下崭露头角的。Bakeoff 评测数据证明,这种基于字标注的分词系统优于以往的基于词(或词典)的分词系统。

回顾这十年来分词技术的进步,有什么是可供其他自然语言处理技术借鉴的经验呢?笔者认为,由于自然语言的模糊性和复杂性,一方面,对于任何进入计算的语言对象都应当为其寻求一种可计算的定义;另一方面,对于推动任何一种应用技术的进步来说,公开、可比的评测都是至关重要的。语言对象的定义和有关这种对象的自动评测是紧密关联的,没有可计算的定义,也就不会有可信的评测。

参考文献:

[1] 黄昌宁. 中文信息处理的分词问题[J]. 语言文字应用, 1997,(1): 72-78.

[2] Sproat, R. and Emerson, T. The First International Chinese Word Segmentation Bakeoff[A]. In: Proceedings of the Second SIGHAN Workshop on Chinese Language Processing[C]. Sapporo, Japan: July 11-12, 2003,133-143.

[3] Sproat R., Shi, C. et al. A stochastic finite-state word segmentation algorithm for Chinese[J]. Computational Linguistics, 1996, 22(3): 377-404.

[4] 国家技术监督局. 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[S]. 北京: 中国标准出版社, 1993.

[5] 孙茂松,张磊. 人机并存,"质"量"合一[J]. 语言文字应用, 1997,(1): 79-86.

[6] 刘开瑛. 现代汉语自动分词评测研究[J]. 语言文字应用,1997,(1): 101-106.

[7] 孙茂松,邹嘉彦. 汉语自动分词综述[J]. 当代语言学, 2001,3(1),22-32.

[8] 杨尔弘,方莹,等. 汉语自动分词和词形评测[J]. 中文信息学报, 2006,20(1): 44-49.

[9] Emerson, T. The Second International Chinese Word Segmentation Bakeoff[A]. In: Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing[C]. Jeju Island, Korea: 2005,123-133.

[10] Levow, G. The Third International Chinese Language Processing Bakeoff: Word segmentation and named entity recognition[A]. In: Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing[C]. Sydney: July 2006, 108-117.

[11] Chengjie Sun, Chang-Ning Huang et al. Detecting segmentation errors in Chinese annotated corpus[A]. In: Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing[C]. Jeju Island, Korea: 2005, 1-8.

[12] 孙茂松. 谈谈汉语分词语料库的一致性问题[J]. 语言文字应用, 1999,(2).

[13] Aitao Chen. Chinese word segmentation using minimal linguistic Knowledge[A]. In: Proceedings of the Second SIGHAN Workshop on Chinese Language Processing[C]. Sapporo, Japan: July 11-12, 2003, 172-175.

[14] Hongqiao Li, Chang-Ning Huang et al. The use of SVM for Chinese new word identification[A]. In: Proceedings of the 1st International Joint Conference on Natural Language Processing (IJCNLP2004)[C]. Hainan Island, China: March 22-24, 2004, 723-732.

[15] Hai Zhao, Chang-Ning Huang and Mu Li. An improved Chinese word segmentation system with conditional random field[A]. In: Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing

① MSRA 分词系统的封闭测试结果列低于 Bakeoff-2006 的正式成绩[11],是因为在本文中使用了更简单的特征模版。

[C]. Sydney: July 2006, 108-117.

[16] Andi Wu and Zhixin Jiang. Word segmentation in sentence analysis[A]. In: Proceedings of 1998 International Conference on Chinese Information Processing[C]. Beijing, China: 1998, 169-180.

[17] Andi Wu. Chinese word segmentation in MSR-NLP [A]. In: Proceedings of the Second SIGHAN Workshop on Chinese Language Processing[C]. Sapporo, Japan: July 11-12, 2003, 172-175.

[18] Neinwen Xue and Susan P. Converse. Combining classifiers for Chinese word segmentation [A]. In: Proceedings of the First SIGHAN Workshop on Chinese Language Processing [C], Taipei, Taiwan: 2002, 63-70.

[19] 黄昌宁. 聚焦 Bakeoff[A]. 张普,蔺荪,等编. 数字化汉语教学的研究与应用[C]. 香港城市大学: 2006 年 7 月 19-22 日, 20-27.

[20] Neinwen Xue and Libin Shen. Chinese word segmentation as LMR tagging[A]. In: Proceedings of the Second SIGHAN Workshop on Chinese Language Processing[C]. Sapporo, Japan: July 11-12, 2003, 176-179.

[21] Jin Kiat Low, Hwee Tou Ng and Wenyuan Guo. A maximum entropy approach to Chinese words Segmentation [A]. In: Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing [C]. Jeju Island, Korea: 2005, 161-164.

[22] Huihsin Tseng, Pichuan Chang et al. A conditional random field word segmenter for SIGHAN Bakeoff 2005[A]. In: Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing[C]. Jeju Island, Korea: 2005, 168-171.

[23] Hai Zhao, Changning Huang et al. Effective tag set selection in Chinese word segmentation via conditional random field modeling[A]. In: Proceedings of PACLIC-20[C]. Wuhan, China: November 1-3, 2006, 87-94.

第二届"'语言与国家'高层论坛"在绍兴举行

2007 年 4 月 18—19 日,第二届"'语言与国家'高层论坛"在浙江绍兴举行。会议由教育部语言文字信息管理司主办,绍兴文理学院承办。

河北师范大学校长苏宝荣、云南师范大学校长骆小所、全国科技名词审定委员会副主任刘青、绍兴文理学院院长王建华、中国社会科学院民族学与人类学研究所副所长黄行等 20 余位语言学专家参加了论坛。国家语委副主任、教育部语信司司长李宇明教授作了主题发言。会议就汉语作为母语教育、少数民族的汉语教育、华文教学中的汉语教育、对外汉语教育等四个议题进行了讨论。会后将由商务印书馆出版论文集。

乔永