由字构词——
中文分词新方法

黄昌宁　　赵　海
微软亚洲研究院
cnhuang@msrchina.research.microsoft.com

大纲

- “由字构词”方法的来龙去脉
- MSRA的“由字构词”分词系统
- 技术进步背后的理念更新
- 结束语

“由字构词”方法的来龙去脉

- 把分词视为字的词位分类问题
  - ◆字的构词位置（词位）：

词首B（占领）　　词尾E（抢占）

词中M（独占鳌头）　单字词S（已 占 全国）

◆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/。/

◆词位标注：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

“由字构词”方法的来龙去脉

- 第一篇由字构词的分词论文发表在第一届SIGHAN研讨会上 [Xue,2002]。
- Xue在最大熵模型上实现的由字构词分词系统，在Bakeoff2003 的2项封闭测试上获得第二、三名，然而其未登录词召回率R_{OOV}（0.729 和 0.670 ）却位居榜首。
- 统计表明，使F_{OOV}更高的分词方法将整体提升分词系统的F值（见论文2.1节）。

“由字构词”方法的来龙去脉

- Bakeoff2005：
  - Ng 基于最大熵模型的分词系统在4项开放测试中获3个第一、1个第二。
  - Tseng 基于条件随机场的系统在4项封闭测试中获2个第一、1个第二和1个第三。
- Bakeoff2006：
  - 微软亚洲研究院采用6词位标记的分词系统在6项测试中获4个第一、2个第三。

Bakeoff十二种语料库的概况

提供者	语料库	编码	训练集词次数	测试集词次数	OOV率
台湾中央研究院	AS2003	Big5	5.8M	12K	**0.022**
	AS2005		5.45M	122K	0.043
	AS2006		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

大纲

- “由字构词”方法的来龙去脉
- MSRA的“由字构词”分词系统
- 技术进步背后的理念更新
- 结束语

MSRA分词系统：特征选择-1

- 设H 是可能的上下文或预定义条件的集合，T是一组可选标注集，条件随机场的特征函数定义为

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

$$h_i \in H, t_j \in T$$

MSRA分词系统：特征选择-2

- 两种常用于封闭测试的特征模板集

模板集	特征类型	特征
TMPT-6（MSRA）	一元（unigram）	C_n, n=-1,0,1
	二元（bigram）	C_nC_{n+1}, n=-1,0
		$C_{-1}C_1$
TMPT-10	一元	C_n, n=-2,-1,0,1,2
	二元	C_nC_{n+1}, n=-2,-1,0,1
		$C_{-1}C_1$

例如：“我们在北京”， 若当前字C_0=在，则被激活的特征包括：“们”、“在“、”北“，”们在“、”在北“，”们北“。

MSRA分词系统：标注集选择

标注集	标记	单字与多字词词位标注举例
2词位	B, I	B, BI, BII, …
3词位	B, I, O	O, BI, BII, …
4词位	B, M, E, S	S, BE, BME, BMME, …
6词位 (MSRA)	B, B_2, B_3, M, E, S	S, BE, BB_2E, BB_2B_3E, BB_2B_3ME, BB_2B_3MME, …

注：选择6词位的实验根据是语料库的平均加权词长（见2.3节）。

6词位选择依据：平均加权词长

- 语料库的平均加权词长定义为 $L_k = \frac{1}{N}\sum_{i=k}^{K} iN_k$

k	AS		CTB	CityU		PKU		MSRA
	2003	2005	2003	2003	2005	2003	2005	2005
1	1.5458	1.5089	1.7016	1.6130	1.6275	1.6429	1.6455	**1.7101**
2	1.0010	0.9378	1.2649	1.1190	1.1586	1.1708	1.1728	**1.2401**
3	0.2135	0.1804	0.3211	0.2648	0.2479	0.2692	0.2730	**0.3619**
4	0.0747	0.0730	0.1195	0.0887	*0.0688*	0.1208	0.1244	**0.2193**
5	0.0320	0.0334	0.0732	0.0252	*0.0150*	0.0390	0.0423	**0.1223**
6	0.0228	0.0241	0.0351	0.0133	*0.0072*	0.0105	0.0142	**0.0776**

在Bakeoff2003语料上的封闭测试

参与者	*F* 值			
	AS	CityU	CTB	PKU
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***	0.951
微软/TMPT-6	***0.973***	***0.947***	0.872	***0.956***

注：由于参赛者当年报告的封闭测试成绩等同于其开放测试成绩，CTB2003的最佳封闭成绩（0.881）迄今未曾打破。

在Bakeoff2005语料上的封闭测试

参与者	*F* 值			
	AS	CityU	PKU	MSRA
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	0.95	0.964
微软/TMPT-6	***0.953***	0.948	***0.952***	***0.974***

注：Ng除采用特征集TMPT10（5字窗宽）外，还用了标点符号、数字、日期（年、月、日）和英文字母等特征。

在Bakeoff2006语料上的封闭测试

参与者	*F* 值			
	AS	CityU	CTB	MSRA
Bakeoff2006第一名	***0.958***	***0.972***	***0.933***	***0.963***
#32	0.953	0.970	0.930	***0.963***
#15-b	0.957	***0.972***	——	0.954
#26	0.949	0.965	0.926	0.957
微软/TMPT-6	0.954	0.969	0.932	0.961

注：我们的结果略低于Bakeoff2006公布的数字，这是由于我们在这里只采用*n*元特征模板TMPT-6。

不同标注集和特征模板集的比较

词位标注集	特征模板集	*F* 值			
		AS-06	CityU-06	CTB-06	MSRA-06
6词位	TMPT-6	*0.953845*	*0.969117*	*0.932015*	*0.960820*
	TMPT-10	0.952865	0.967052	0.931438	0.958384
4词位	TMPT-6	0.953218	0.967833	0.930709	0.953231
	TMPT-10	0.952476	0.966564	0.930161	0.955222

注：所有最高性能都出现在6词位标注集和三字窗宽的TMPT-6特征模板上。

MSRA分词系统的错误分析

1.答案：/更/好/地/促进/中华民族/的/大团结/、
/大联合/，
输出：/更/好/地/促进/中华民族/的/大团结/、
/大/联合/，

2.答案：/民革/同/中国共产党/一道/经受/考验/，
输出：/民革/同/中国共产党/一/道/经受/考验/，

3. 答案：/女孩子/家/当/演员/丢人/，
输出：/女孩子/家当/演员/丢人/，

4. 答案：/总后嫩江基地/的/先进事迹/，
输出：/总/后/嫩江基地/的/先进事迹/，

MSRA分词系统的错误分析

5. 答案：/德国/对/西班牙/北部/小/镇/格尔尼卡/进行/狂/轰/滥/炸/

输出：/德国/对/西班牙/北部/小/镇/格尔尼卡/进行/狂轰滥炸/

6. 答案：/对/支/委/以上/党员/干部/除/正常/教育/外/，

输出：/对/支委/以上/党员/干部/除/正常/教育/外/，

7. 答案：/党校/、/成/教/、/农民/学校/三块/牌子/一套/班子/，

输出：/党校/、/成教/、/农民/学校/三块/牌子/一套/班子/，

大纲

- “由字构词”方法的来龙去脉
- MSRA的“由字构词”分词系统
- 技术进步背后的理念更新
- 结束语

技术进步背后的理念更新

- 中文的词语只应有一个标准，还是可以有多个标准并存？
- 中文词语是怎样被精良定义的？

  —— 规范 + 词表 + 大规模标注语料库
- 未登录词识别和歧义消解，孰重孰轻？
- 字本位，还是词本位？
- 知识中心，还是数据中心（驱动）？

不同分词者之间的一致性比较

- [Sproat,1996] M1-M3，T1-T3 分别代表三位大陆和三位台湾被试者。测试语料含100个句子，4,372个字。评价指标为精确率P 和召回率R 的算术平均值。被试者之间的平均一致性只有 0.76 。

	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

不同分词标准之间的一致性比较

测试语料库	训练语料库（ *F*值）			
	As2006	CTB2006	CityU2006	MSRA2006
AS2006	1.0	**0.959300**	0.925638	0.858262
CTB2006	0.942003	1.0	0.910410	0.877422
CityU2006	0.932136	0.934643	1.0	0.848826
MSRA2006	0.856964	0.886632	**0.848003**	1.0

注1：MSRA系统采用特征模版TMPT-6 和 6词位标注集。
注2：四种语料库之间的平均一致性为 **0.90** 。

“由字构词”方法的构词法基础

- 令$T=\{B,B_2,B_3,M,E,S\}$，任意字C_i在词位t_j的能产度为

$$P_{C_i}(t_j)=\frac{\mathrm{count}(C_i.t_j)}{\sum_{t_j\in T}\mathrm{count}(c_i,t_j)}$$

◆ 若任意字C_i在某词位上的能产度高于0.5，就称这个词位是该字的主词位。

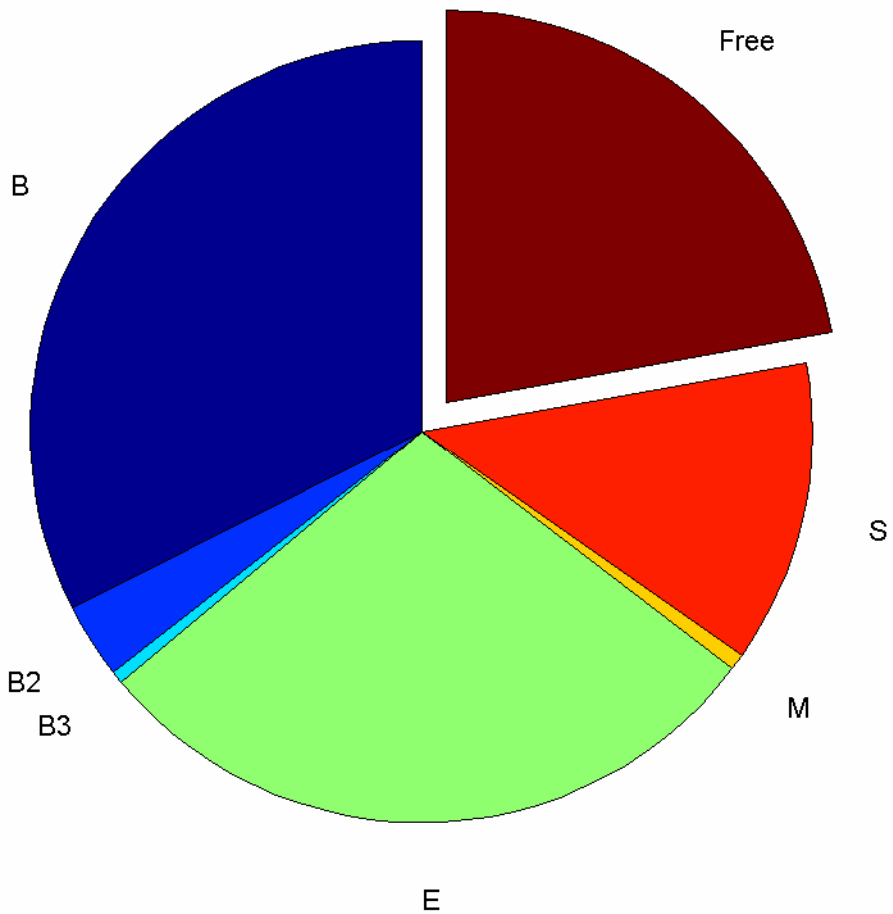
◆ 那些没有主词位的字被叫做自由字。

各词位的字量分布统计

- 数据来源：MSRA2005 训练语料库，237万词次。
- 语料库总字量：　5,147
- 有主词位的字量：3,920（76.16%）
- 自由字的字量：　1,227（23.84%）

标记	*B*	*B2*	*B3*	*M*	*E*	*S*	总字量
字量	1634	156	27	33	1438	632	3920
百分比(%)	31.74	3.03	0.52	0.64	27.94	12.28	76.16

各词位的字量分布统计

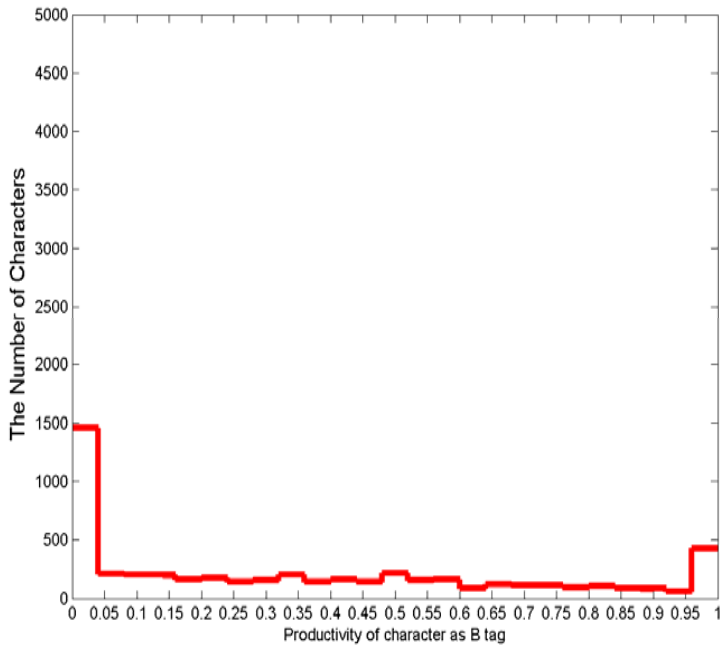


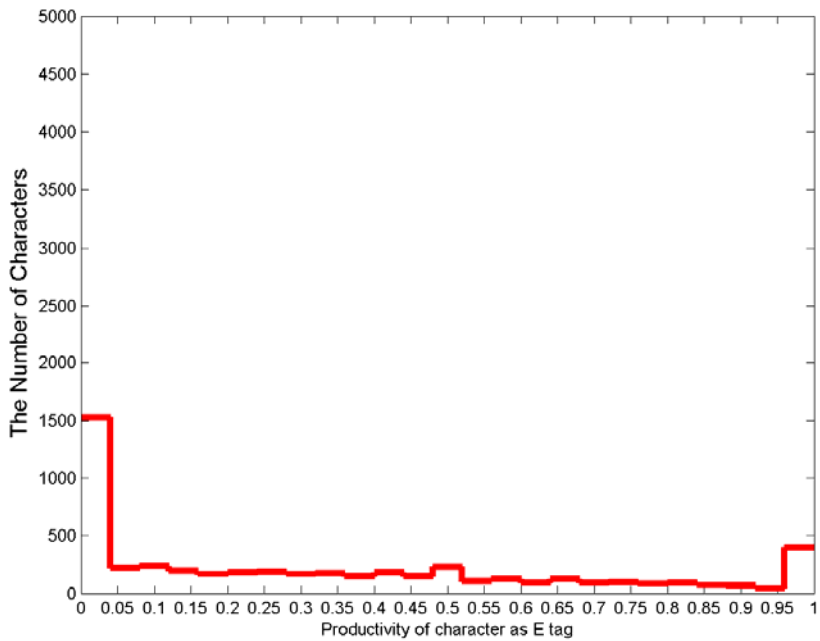
10个高频字及其能产度分布

Characters	Frequency	B	E	S	B_2	B_3	M
的	129132	0.001169	0.010338	**0.987679**	0.000519	0.000163	0.000132
一	40189	**0.540023**	0.058648	0.285650	0.086889	0.019408	0.009381
国	40091	0.310070	**0.468609**	0.020828	0.151206	0.024968	0.024320
在	32594	0.024821	0.099742	**0.869485**	0.003712	0.002178	0.000061
中	29762	**0.490558**	0.093609	0.315570	0.032424	0.032323	0.035515
了	29305	0.026480	0.052346	**0.919980**	0.000478	0.000682	0.000034
是	28020	0.015703	0.338829	**0.641113**	0.001642	0.002712	0.000000
人	27260	**0.355026**	0.304952	0.228833	0.023844	0.063243	0.024101
和	26328	0.047820	0.008356	**0.922440**	0.007710	0.001785	0.011888
有	26196	0.268133	0.313597	**0.376661**	0.018934	0.008207	0.014468

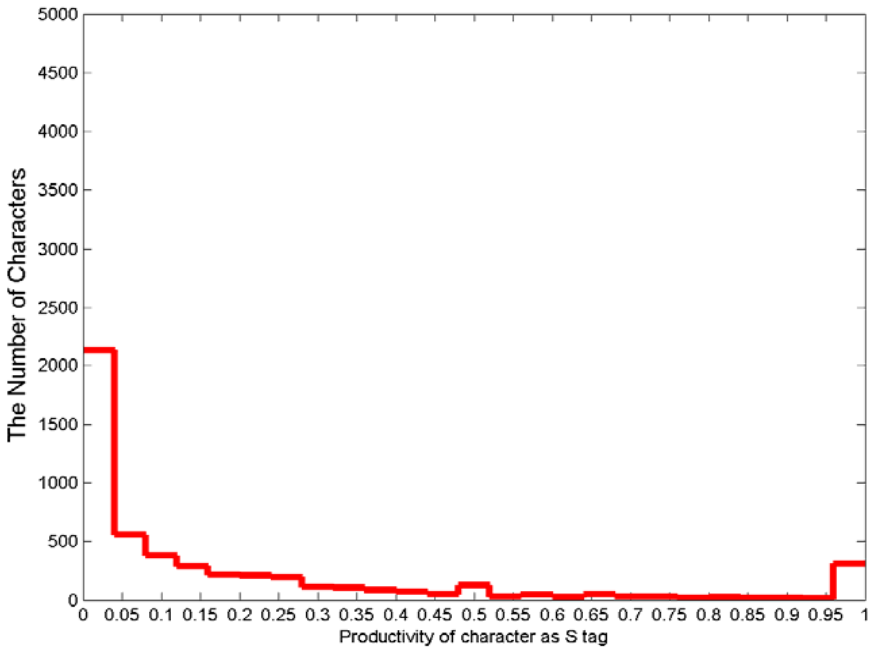
注：主词位为 S 的字：的，在，了，是，和。
主词位为 B 的字：一。
自由字：国，中，人，有。

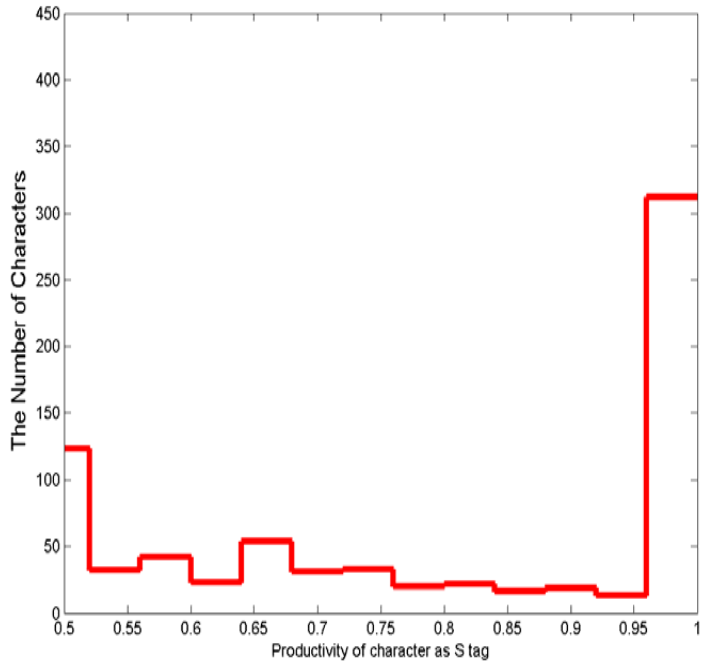
词位*B*和*E*的字量分布统计



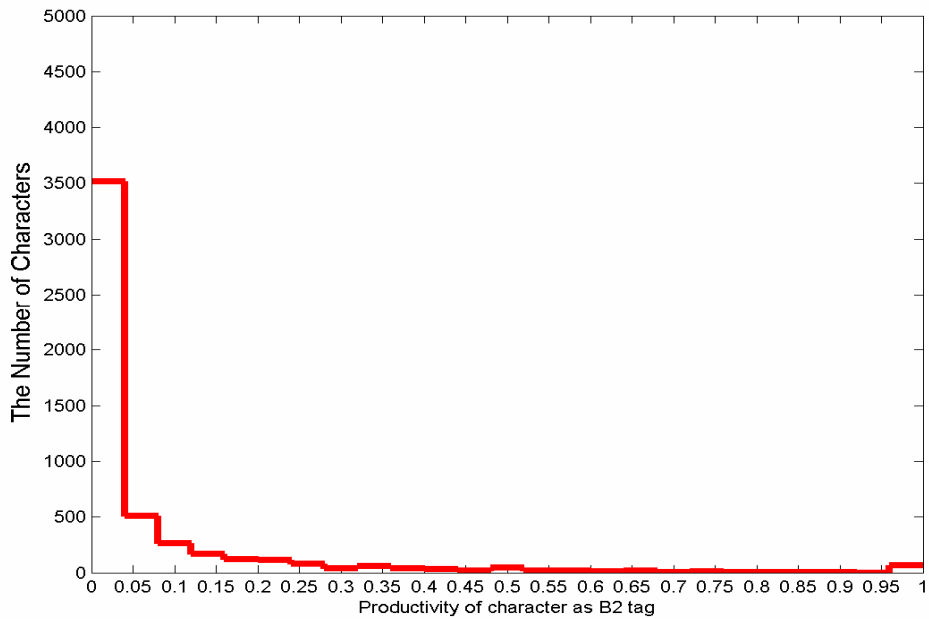


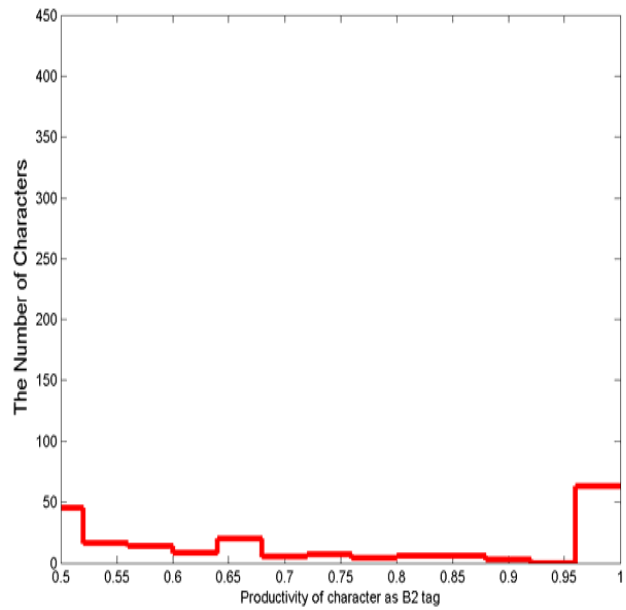
词位S的字量分布统计


The Number of Characters
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
0 0.05 0.1 0.15 0.2 0.25 0.3 0.35 0.4 0.45 0.5 0.55 0.6 0.65 0.7 0.75 0.8 0.85 0.9 0.95 1
Productivity of character as S tag


The Number of Characters
450
400
350
300
250
200
150
100
50
0
0.5 0.55 0.6 0.65 0.7 0.75 0.8 0.85 0.9 0.95 1
Productivity of character as S tag

词位B_2的字量分布统计

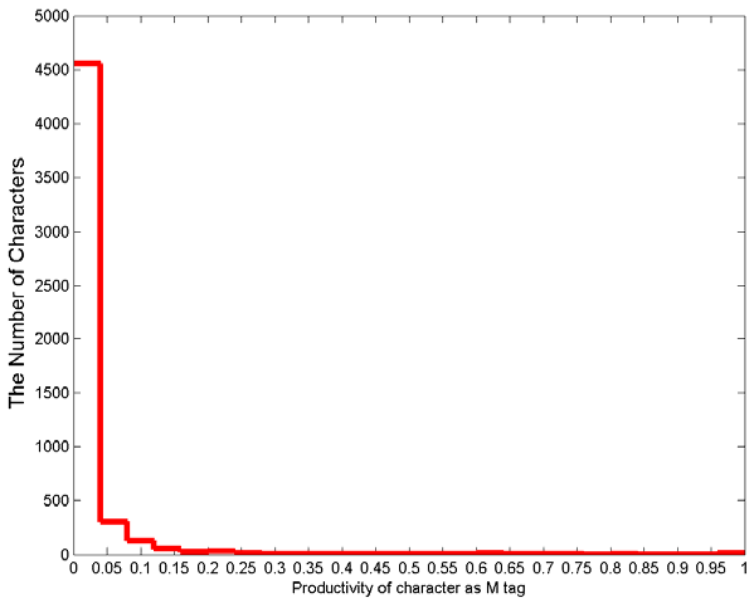


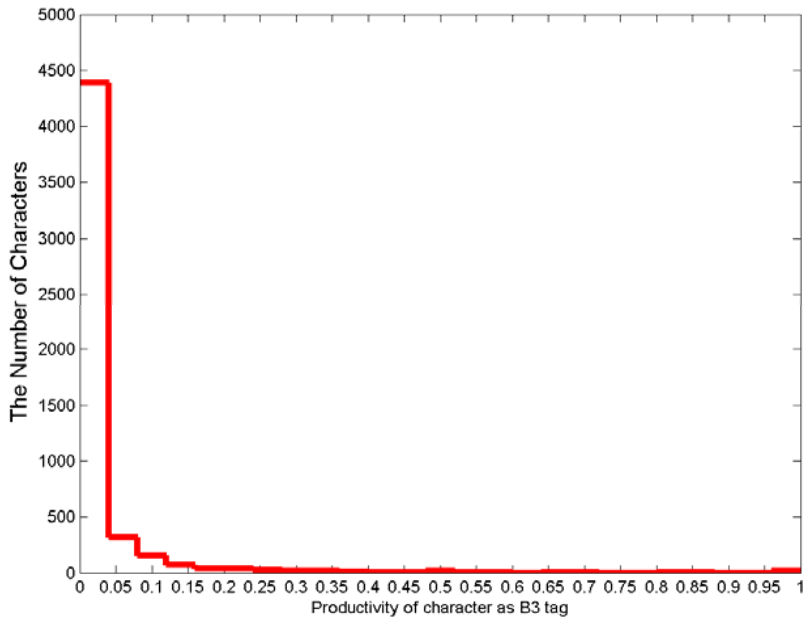


主词位为B_2的某些字的用法

	字	字频	*B*	*E*	*S*	B_2	例子（词/词频）
中国人名用字	泽	1137	0.014952	0.058047	0.006157	0.903254	毛泽东/136　宫泽喜一/4
	晓	238	0.054622	0.289916	0.042017	0.600840	马晓春/19　陈晓钟/9
	彦	89	0.067416	0.359551	0.000000	0.573034	刘彦彬/33　王彦田/3
	秉	56	0.357143	0.017857	0.000000	0.625000	戴秉国/16　王秉岐/8
	钊	54	0.000000	0.333333	0.000000	0.666667	叶钊颖/35　汤钊猷/1
	淑	50	0.020000	0.200000	0.000000	0.720000	于淑芳/3　李淑转/3
	肇	47	0.404255	0.000000	0.000000	0.553191	李肇星/10　韩肇康/5
	岱	20	0.000000	0.250000	0.000000	0.750000	张岱年/13　林岱峰/1
译名用字	勒	567	0.038801	0.275132	0.000000	0.539683	巴勒斯坦/178　格勒尔/4
	洛	477	0.220126	0.134172	0.002096	0.519916	摩洛哥/41　卡洛斯/8
	瀚	102	0.029412	0.058824	0.009804	0.774510	约翰内斯堡/35　约翰逊/4

词位M和B_3的字量分布统计





大纲

- “由字构词”方法的来龙去脉
- MSRA的“由字构词”分词系统
- 技术进步背后的理念更新
- 结束语

结束语

- 四年来，国际中文分词评测活动令自动分词技术面目一新。
- “由字构词”的分词新方法在公开评测中异军突起。它通过改进未登录词识别能力大幅度提升了分词系统的总体性能。
- 微软分词系统得益于6词位标注集和*n*元特征模板集的巧妙配合。

谢谢！