文章编号：1003-0077（2007）05-0000-00

基于有效子串标注的中文分词*

赵海[1]，揭春雨[1]

（1.香港城市大学中文翻译及语言学系，香港九龙达之路 83 号）

摘要：由于基于已切分语料的学习方法和体系的兴起，中文分词在本世纪的头几年取得了显著的突破。尤其是 2003 年国际中文分词评测活动 Bakeoff 开展以来，基于字标注的统计学习方法引起了广泛关注。本文探讨这一学习框架的推广问题，以一种更为可靠的算法寻找更长的标注单元来实现中文分词的大规模语料学习，同时改进已有工作的不足。我们提出子串标注的一般化框架，包括两个步骤，一是确定有效子串词典的迭代最大匹配过滤算法，二是在给定文本上实现子串单元识别的双词典最大匹配算法。该方法的有效性在 Bakeoff-2005 评测语料上获得了验证。

关键词：中文分词；基于子串标注的分词

中图分类号：TP391　　**文献标识码：**A

Effective Subsequence-based Tagging for Chinese Word Segmentation

Hai Zhao[1], Chunyu Kit[1]

(1. Department of Chinese, Translation and Linguistics, City University of Hong Kong, 83 Tat Avenue, Kowloon, Hong Kong SAR, China)

Abstract: The research of automatic Chinese word segmentation has been advancing rapidly in recent years, especially since the First International Chinese Word Segmentation Bakeoff held in 2003. In particular, character-based tagging has claimed a great success in this field. In this paper, we attempt to generalize this method to subsequence-based tagging. Our goal is to find longer tagging units through a reliable algorithm. We propose a two-step framework to serve this purpose. In the first step, an iterative maximum matching filtering algorithm is applied to obtain an effective subsequence lexicon, while in the second step, a bi-lexicon based maximum matching algorithm is used for identifying subsequence units. The effectiveness of this approach is verified by our experiments using two closed test data sets from Bakeoff-2005.

Key words: Chinese word segmentation (CWS), subsequence-based tagging approach of CWS

1 引言

中文分词技术[13][14][21]随着基于切分语料的机器学习方法的兴起而在最近几年获得了显著突破。特别是 SIGHAN[1]举办的国际中文分词评测(International Chinese Word Segmentation Bakeoff, 简称 Bakeoff)活动[1]，提供多标准的训练和测试语料，让研究者们得以搁置困扰学界多年的切分标准问题[15][16]，把研究集中到机器学习方法的改进上来[1][2][3]。Bakeoff 活动中，基于字标注的机器学习方法获得了广泛注意[7][19]。此类方法在 Bakeoff-2005 以及 2006 上获得了巨大成功[4][18]，性能领先的系统几乎无一例外都应用了类似的标注学习的思想[6][8][17]，形成中文分词研究中新的主流技术。

本文继续致力于这一技术的深化，考虑使用更长的子串作为基本的标注单元来实现更充分的分词知识学习。尽管已有一些工作考虑了这一思想，但是它们不能在单一的学习过程中

* **收稿日期：**2007 年 6 月 25 日　**定稿日期：**2007 年 6 月 25 日

基金项目：香港城市大学 SRG 项目 7002037 和香港特别行政区资 局 （ UGC）的 CERG 研究项目 9040861(CityU 1318/03H)

作者简介：赵海（1976——），男，博士，博士后研究员，主要研究方向为自然语言处理和机器学习；揭春雨（1964——），男，博士，助理教授，博、硕士导师，主要研究方向为计算语言学、机器学习、计算术语学和计算诗学。

[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。

获得理想的分词性能，而是依赖于附加的集成技术支撑。我们的改进是将子串单元的获取分解为两个步骤，提出使用改进的最大匹配算法来获得有效的子串标注单元。在 Bakeoff 语料上，所提出方法的有效性得到了验证。

本文后续内容组织如下：第 2 节介绍分词标注学习的基本模型，包括学习算法和特征标注集；第 3 节介绍子串标注以及词典生成的算法细节；第 4 节提供评估结果；最后一节是本文小结。

2 学习模型

基于字标注[2]的分词方法实际上是将分词知识的学习转换成字串的标注过程。由于每个字在构造一个特定的词语时都占据一个构词位置，即字位，因此，可以将分词过程看成学习这个字位信息的机器学习过程。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。

2．1 条件随机场模型

条件随机场（Conditional Random Fields, CRFs）是一个无向图上概率分布的学习框架，由 Lafferty 等首先引入到自然语言处理的串标引学习任务中来[5]。最常用的一类 CRF 是线性链 CRF，适用于我们的分词学习。记观测串为 $W=w_1w_2\cdots w_n$，标记串（状态）序列 $Y=y_1y_2\cdots y_n$，线性链 CRF 对一个给定串的标注，其概率定义为：

$$p_\lambda(Y\mid W)=\frac{1}{Z(W)}\exp(\sum_{t\in T}\sum_{k}\lambda_k f_k(y_{t-1},y_t,W,t))$$

其中，Y 是串的标注序列，W 是待标记的字符，f_k 是特征函数，λ_k是对应的特征函数的权值，而 t 是标记，$Z(W)$ 是归一化因子，使得上式成为概率分布。

CRF 模型的参数估计通常使用 L-BFGS 算法来完成[11]。CRF 的解码过程，也就是求解未知串标注的过程，需要搜索计算该串上的一个最大联合概率，即

$$Y^*=\arg\max_Y P(Y\mid W)$$

在线性链 CRF 上，这个计算任务可以用一般的 Viterbi 算法来有效地完成。

2．2 标注集和特征模板

分词本质上是对字串中的每一个字相应作出一个在该处切分与否的二值决策过程，已有的基于字标注的 CRF 分词系统大多使用二字位标注集[8]。在基于最大熵模型的分词系统中，广泛使用的是四字位标注集[7]。我们在 Bakeoff-2006 的参赛系统中，首次使用了六字位标注集[6]。已有的结果表明，较之于其他标注集，六字位标注集搭配适当的特征模板，能够获得更佳性能[9]。

本文继续使用六字位标注集进行标注。我们记该集合为 $T=\{B, B_2, B_3, M, E, S\}$，其中，$B$、$B_2$ 和 B_3分别标示一个词的前三字位置，M标示更后但非词尾的位置，E标示词尾，S表示单字词。表 1 中给出了六字位标注集对不同长度的词的标注例示。

表 1　三字位和六字位标注集的定义

标注集	标记	单字与多字词的字位标注举例
三字位	B, E, S	S, BE, BEE, BEEE, …
六字位	B, B_2, B_3, M, E, S	S, BE, BB_2E, BB_2B_3E, BB_2B_3ME, BB_2B_3MME,…

条件随机场或最大熵学习中，用于表达语言特性的特征函数起核心作用。通常，所用的

[2]这里说到的“字”不只限于汉字，也包括外文字母、阿拉伯数字和标点符号等各种字符。当然，汉字是这个集合中数量最多的一类字符。

特征会按照某种定义被适当分组，称之为特征模板。在中文分词学习中，最重要也是最基本的特征模板，就是当前字符本身及其上下文各字符。我们使用的基本特征模板将使用 6 个字符组合：C_{-1}，C_0，C_1，$C_{-1}C_0$，C_0C_1，以及 $C_{-1}C_1$。这里的下标 0、-1 和 1 分别指示当前及其前后一个字符的位置。我们记这组特征模板为 TMPT-H。为了便于比较，我们也将使用[10]中的一组模板，它包括 10 组字符组合，C_{-2}，C_{-1}，C_0，C_1，C_2，$C_{-2}C_{-1}$，$C_{-1}C_0$，C_0C_1，C_0C_2以及 $C_{-1}C_1$。记这组模板为 TMPT-R。[10]使用三标注集进行串标注，该标注集定义详见表 1。

3 从基于字到基于子串的标注：算法描述

关于分词的串标注学习中，迄今为止的工作绝大多数都是基于字的。诚然，在一般情形下，字是最基本的构词单元，但是，这种做法也会忽略掉很多有意义的组合信息。例如下面的句子：

(a) 以/北京市/为/例/，/国有/粮食/企业/有/２８家

(b) 笔者/近日/采访了/位于/北京/朝阳区/的/几家/粮食/批发/市场/和/集贸/市场

(c) 烟波/浩渺/的/密云水库/，/是/首都/北京/的/重要/水源

这三个句子都包含有“北京”一词，高频而且固定，但是基于字标注的学习算法不能有效利用这一信息，依然会对“北”和“京”这两个字进行独立的标注学习。在解码的时候，算法也同样无法利用这一信息，而依然需要在系统规定的特征之外对这两个字的组合特性作出它们相互独立的概率假设。

因此，如果能够捕捉这种有效的高频子串，那么这种有意义的信息应该能用来改进分词的串学习的效果。在已有的工作中，高频词直接被抽取，作最终所用的子串词典。子串标注方法，在训练语料上以词内最大匹配来标注子串单元，在无标记的测试文本上直接使用最大匹配算法[10]。然而，由于所用的子串词典以及确定标注单元的切分算法不佳，导致了大量的标记跨越现象，影响最终性能。下面解释一下切分标记跨越。假设正确切分句子为：

(d) 风湿性/心脏病/的/中医/疗法/介绍/如下

使用某个初始的子串词典进行最大匹配算法操作后，切分为

(e) 风/湿/性/心脏/病/的/中/医疗/法/介绍/如/下

该切分中，“医疗”一词跨越了正确的切分标记“中医/疗法”。在这种情形，除非采用复杂的后处理技术进行弥补，否则，难以纠正这个标记跨越在学习训练中最终导致的性能损失。

为此，我们这里提出一个子串标注框架，分两个步骤，一个用来构造较为理想的子串词典，一个用来有效地从原始子串中标出所需的子串单元。针对第一个步骤，我们提出一种称之为迭代最大匹配过滤的算法，用来构造子串词典，其算法描述如下：

1. 从训练用的已切分语料中按照某个截断频率抽取高频词构成初始的子串词典。
2. 用这个词典对训练语料所有的字串进行最大匹配切分。
3. 如果某个子串词典词跨越了训练语料中的切分标记，则从词典中去掉该词。
4. 重复 2-3 直到最大匹配切分在整个训练语料上不再导致任何切分标记跨越，此时获得的子串词典，即是我们所构造的词典，用于下一步的操作。

针对第二个步骤，我们提出一种称之为双词典最大匹配算法，将基于字的串转为基于子串的标注单元，该算法描述如下：

1. 使用第一个步骤获得子串词典（称之为主词典）以及另外一个辅助词典。
2. 设置匹配位置 $p=1$，表示从第一个字之前的位置开始匹配操作。
3. 如果主词典中存在一个词长为 L 的词，能匹配从 p 到 $p+L$ 的子串，并且，没有任何一个来自辅助词典的词跨越切分位置 $p+L$，则在 $p+L$ 处设置切分标记，匹配位置更新为 $p=p+L$。否则，在 $p+1$ 处设置切分标记，匹配位置更新为 $p=p+1$。
4. 重复 3 直到整个串被匹配完毕。
5. 由 4 切分出来的字串的各个部分，用作为子串标注单元。

我们以上述的句子(d)(e)为例，如果获得的切分结果为

(f) 风/湿/性/心脏/病/的/中/医/疗/法/介绍/如/下

则在(d)的标准切分下，我们所获得训练用的切分标注串为

风	湿	性	心脏	病	的	中	医	疗	法	介绍	如	下
B	B_2	E	B	E	S	B	B_2	B_3	E	S	B	E

这样，上节为字标注设计的标注集和特征模板，可以不加修改直接移植到基于子串的标注上来。所不同的是，标注单元全部由单个字符换成了以上算法标出的子串。

4 评估

我们使用第二届国际分词竞赛(Bakeoff-2005)中的两组语料对上面提出的方法进行评估[4] ，一组是 Big5 码，一组是 GB 码。表 2 是这两组语料的统计数据。按照 Bakeoff 规则，在每种组语料库上又分封闭和开放两种测试：封闭测试只允许从同组的训练语料中获取知识来从事分词；开放测试则不受此约束。由于开放测试涉及的方法和语言资源变化多样，并不专对分词技术本身做出有效评价。考虑到本文研究的性质，我们仅在封闭测试条件下进行实验比较。我们使用基于词的 *F* 值作为评估标准，它是准确率 *P* 和召回率 *R* 的调和平均值：*F=2RP/(R+P)*。为了和相关工作比较，我们也同时列出词表词的召回率 R_{iv} 和未登录词的召回率 R_{oov}。

表 2　评估语料的统计信息

	CityU2005（繁体 Big5 编码）	MSRA2005（简体 GB 编码）
训练语料词数	1.46M	2.37M
测试语料词数	41K	107K

4．1 子串标注的效果

我们首先抽取部分高频词构造有效的子串词典。为了比较不同选择的效果，分别抽取 1000，2000 和 3000 个最高频多字词作为初始的子串词典。通过迭代最大匹配过滤算法分别获得一个优化词典。相关的词数以及迭代次数信息如表 3 所示。

表 3 过滤后的子串词典的大小以及相应过滤算法的迭代次数

初始词典大小	CityU2005	MSRA2005
1000	637/3	604/3
2000	1312/4	1315/3
3000	2058/4	2027/3

在以下的实验比较中，为简化起见，双词典最大匹配算法的辅助词典也重复使用过滤后的子串词典。表 4 给出了不同的子串标注算法下的标记跨越现象的统计。在每个词边界标记上出现一个跨越，我们记标记跨越一次。从表 4 可以看到，初始子串词典越大，标记跨越现象会越严重。

考虑到一次标记跨越将会导致两个词被切分错误，使得系统评估的 *F* 值遭到加倍的惩罚性损失，也就是，如果直接使用最大匹配算法进行子串标注而不做专门处理，以初始词典为 3000 个词为例，按照表 4 中的数据，系统的 F 值将降低 0.4-1 个百分点。这是一个相当可观的性能损失。而依据所提出的算法，这一性能损失可以降低几十倍，说明所提出的算法具有很强的针对性。同时可以也注意到，仅使用双词典最大匹配算法能降低 1/4-1/3 的标记跨越，而仅使用过滤词典能将消解 95%以上的标记跨越。这说明，子串词典的选择在这一标注框架下具有更为重要的意义。

表 4 不同子串标注算法的标记跨越现象的统计

初始词典大小	子串标注算法	标记跨越的次数	
		CityU2005	MSRA2005
1000	原始词典+最大匹配	124	172
	过滤词典+最大匹配	6	6
	原始词典+双词典最大匹配	99	129
	过滤词典+双词典最大匹配	6	6
2000	原始词典+最大匹配	169	278
	过滤词典+最大匹配	7	10
	原始词典+双词典最大匹配	117	201
	过滤词典+双词典最大匹配	5	8
3000	原始词典+最大匹配	204	354
	过滤词典+最大匹配	9	12
	原始词典+双词典最大匹配	133	241
	过滤词典+双词典最大匹配	7	10

为了比较，我们也重新实现了基于 TMPT-R 模板集和三标注集的分词学习系统。在不同的初始子串词典的情形下，该系统和 TMPT-H 模板集与六标注集搭配的分词系统的效果对比如表 5 所示，表中特征标注集后的 o 和 w 分别代表不使用和使用我们所提出的子串词典生成和标注策略。从表 5 中可以看到，所提出的子串词典过滤以及标注策略，在不同规模的子串词典、不同的特征和标注集下都获得更为稳定的性能。而不做专门处理的基于子串的标注，在某些情形下其性能低于基于字的标注。

表 5 不同特征和标注集的性能比较（*F* 值）

语料	特征和标注集	字标注	子串标注（不同大小的初始词典）		
			1000	2000	3000
CityU2005	TMPT-R+3Tag/o	0.944	0.946	0.945	0.944
	TMPT-R+3Tag/w		0.948	0.949	0.949
	TMPT-H+6Tag/o	0.948	0.948	0.950	0.948
	TMPT-H+6Tag/w		0.950	0.951	0.952
MSRA2005	TMPT-R+3Tag/o	0.963	0.966	0.967	0.966
	TMPT-R+3Tag/w		0.966	0.967	0.967
	TMPT-H+6Tag/o	0.973	0.972	0.972	0.971
	TMPT-H+6Tag/w		0.974	0.974	0.974

4．2 与已有结果的比较

我们的结果和已有最佳结果的比较列在表 6 之中。这个表格中列出了 Bakeoff-2005 的最佳成绩[4]以及该届 Bakeoff 封闭测试的最佳参赛者 Tseng 的成绩[8]。同时，为了和已有工作作一个全面的比较，我们列出了 Zhang 完整系统的结果[10]以及[12]中报告的非正式 Bakeoff-2005 结果。

表 6 中的实验结果表明，我们基于子串标注的系统能够达到最佳的分词精度，基于子串标注的系统优于基于字标注的系统。同时，我们的结果也优于已有的类似的工作所获得结果。本文构建的是单一的基于子串标注的系统，但其性能优于[10]中相应的单一的子串标注系统，同时，也

优于其中使用了复杂集成技术的组合系统。在[10]中，作者认为基于字符或者子串标注的系统对未登录词的识别具有更大的贡献，而一个基于已知词典的概率分词方法能够更好地改进词表词的识别。因此，该文使用了一种加权方法来集成这两种分词系统（加权的权重是另外一个经验性参数，而非由自动学习获得）。我们的结果在此证明，仅使用单一的串标注学习方法，无需复杂的系统集成技术或者是自定义的后处理步骤，同样可以获得更好的性能。这再一次验证我们在[20]中的结论，即，强调词表词信息的复合系统不一定优于单一的标注系统。

表 6　实验结果比较(*F*/*Riv*/*Roov* 值)

（我们的系统均使用 TMPT-H 模板集和六字位标注集，以及 3000 个词的初始子串词典）

	CityU2005	MSRA2005
Tseng, 2005	0.943/0.961/0.698	0.964/0.968/0.717
Bakeoff-2005 最好成绩	0.943/0.961/0.698	0.964/0.968/0.717
Zhou, 2005	------/-------/------	0.966/**0.992**/0.387
Zhang, 2006/子串标注系统	0.930/0.967/0.736	0.952/0.972/0.716
Zhang, 2006/集成系统	0.951/0.969/**0.741**	0.971/0.976/0.712
基于字的标注系统	0.948/0.967/0.692	0.973/0.978/**0.750**
基于子串的标注系统	**0.952**/**0.972**/0.735	**0.974**/0.980/0.749

5 结论以及将来的工作

基于字标注的统计学习方法已经成为中文分词的主流技术。我们使用一种包含两个步骤的子串标注框架将这一方法推广到更为一般的情形。我们的研究目的是为了克服已有工作中的标记跨越问题，同时期望能够获得更高的分词性能。

在形式上，我们的直接目标是以一种更为可靠的算法寻找更长的标注单元来尽可能地捕捉有用的长串信息，从而改善中文分词的统计学习效能。我们提出了这种可靠的子串标注的一般化框架，它包括两个步骤，一是确定有效的子串词典，二是在给定文本上实现子串单元识别。对于前者，我们提出了一种迭代最大匹配过滤算法；对于后者，我们提出了一种双词典最大匹配算法。所提出方法的有效性在 Bakeoff-2005 的评估语料上获得了验证。实验结果表明，所提出的方法优于已有的最佳结果。同时，按照所提出的子串标注方法，所获得的基于子串的学习系统，也优于基于字标注的学习系统。

在未来的工作中，可以考虑有效地确定初始的子串词典。在本文的研究中，我们依然是依据经验性知识来确定初始的子串词典的规模，寻找更有效的判据来确定这个词典依然是一个很有意义的工作。其次，在子串的标注上，可以考虑使用概率算法来进一步改进标注的效果。

参考文献：

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

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

[3] 杨尔弘, 方莹, 刘冬明, 乔羽. 汉语自动分词和词性标注评测[J]. 中文信息学报, 20(1), pp. 46-51, 2006.

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

[5] John D. Lafferty, Andrew McCallum and Fernando C. N. Pereira. 2001. Conditional Random Field: Probabilistic

models for segmenting and labeling sequence data [A]. In: *ICML-18* [C], pp.282-289. June 28-July 01, 2001.

[6] 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 [C], pp.108-117, Sydney, July 2006.

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

[8] Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky and Christopher Manning. A conditional random field word segmenter for SIGHAN Bakeoff 2005 [A]. In: Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing [C], pp.168-171, Jeju Island, Korea: 2005.

[9] Hai Zhao, Chang-Ning Huang, Mu Li and Bao-Liang Lu. Effective tag set selection in Chinese word segmentation via conditional random field modeling [A]. In *PACLIC-20* [C], pp.87-94, Wuhan, China: November 1-3, 2006.

[10] Ruiqiang Zhang, Genichiro Kikui and Eiichiro Sumita. Subword-based tagging by Conditional Random Fields for Chinese word wegmentation [A]. In: *HLT/NAACL-2006* [C], pp.193-196. New York, 2006.

[11] Jorge Nocedal and Stephen J. Wright. Numerical Optimization [B]. Springer, 1999.

[12] Jun-Sheng Zhou, Xin-Yu Dai, Rui-Yu Ni and Jia-Jun Chen. A hybrid approach to Chinese word segmentation around CRFs [A]. In: Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing [C], pp.196-199, Jeju Island, Korea, 2005.

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

[14] Richard Sproat and Chilin Shih. A stochastic finite-state word segmentation algorithm for Chinese [J]. Computational Linguistics, 22(3): 377-404, 1996.

[15] 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[M]. 北京：中国标准出版社, 1993.

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

[17] 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, 161-164, 2005.

[18] Gina-Anne Levow. 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: 108-117, July 2006.

[19] Fuchun Peng, Fangfang Feng and Andrew McCallum. Chinese segmentation and new word detection using Conditional Random Fields [A]. In: *COLING 2004* [C], 562-568. Geneva, Switzerland, August 23-27, 2004.

[20] Chang-Ning Huang and Hai Zhao. Which is essential for Chinese word segmentation: Character versus word [A], In: *PACLIC 20* [C], pp.1-12, Wuhan, China, November 1-3, 2006.

[21] 黄昌宁，赵海. 中文分词十年回顾［J］, 中文信息学报，21（3），8-20，2007.

电子文献及载体类型标识:[DB/OL] --联机网上数据库 [DB/MT]--磁带数据库 [M/CD]--光盘图书
[CP/DK] --磁盘软件 [J/OL] --网上期刊 [EB/OL]--网上电子公告