文章编号：1003-0077(2009)05-0003-06

基于 CRFs 边缘概率的中文分词

罗彦彦，黄德根

（大连理工大学 计算机科学与工程系，辽宁 大连 116024）

摘　要：将分词问题转化为序列标注问题，使用 CRFs 标注器进行序列标注是近年来广泛采用的分词方法。针对这一方法中 CRFs 的标记错误问题，该文提出基于 CRFs 边缘概率的分词方法。该方法从标注结果中发掘边缘概率高的候选词，重组边缘概率低的候选词，提出 FMM 的奖励机制修正重组后的子串。在第四届 SIGHAN Bakeoff 中文简体语料 SXU 和 NCC 上进行闭式测试，分别在 F-1 值上达到了 96.41％和 94.30％的精度。

关键词：计算机应用；中文信息处理；中文分词；条件随机场(CRFs)；边缘概率；最大向前匹配(FMM)；全局特征

中图分类号：TP391　　**文献标识码**：A

Chinese Word Segmentation Based on the Marginal Probabilities Generated by CRFs

LUO Yanyan, HUANG Degen

(Department of Computer Science and Engineering, Dalian University of Technology, Dalian, Liaoning 116024, China)

Abstract: The method of treating the word segmentation issue as a sequence tagging problem and using CRFs has been widely applied recently. However, in this method, some wrong tags are produced by CRFs. To reduce the number of wrong tags, we propose a new method based on the marginal probabilities generated by CRFs for Chinese word segmentation. Firstly, the candidate words with high marginal probabilities are extracted from the tagging results. Then, the candidate words of low marginal probabilities in the tagging results are recombined. Finally, a mechanism of premium that is built on FMM is introduced to complement the sub-strings produced by the recombinant procedure. Evalued by the closed track of SXU and NCC corpora in the fourth SIGHAN Chinese Word Segmentation Bakeoff, this method produces an F-score of 96.41％ and 94.30％, respectively.

Key words: computer application; Chinese information processing; Chinese word segmentation; Conditional Random Fields(CRFs); Marginal probability; Forward Maximum Matching(FMM); global feature

1　引言

中文分词经过多年的研究取得了一定的成果。特别是 SIGHAN 举办的国际中文分词评测活动 Bakeoff 促进了分词的比较和发展。文献[1]运用标注的思想将中文分词问题转换为序列标注问题，然后使用最大熵模型进行字的标注，获得良好的分词效果。自此这种基于字标注的机器学习方法获得了广泛地关注，并在随后两届的 Bakeoff 上获得了很大的成功，性能领先的系统[2]几乎都采用了这种类似的标注方法，成为分词研究领域中的主流技术。CRFs 能够克服最大熵模型的标记偏置问题，在基于字的分词系统上要优于最大熵模型，成为基于字标注的首选模型[3-4]。

文献[5-6]在基于字标注的基础上，采用了基于子词的分词策略，使用 CRFs 进行标注。基于子词的模型仍然采用序列标注的思想，把子词作为最小的切分单元来看待。然而，由于所用的子词词典以

收稿日期：2008-10-17　**定稿日期**：2009-01-13

基金项目：国家 863 高技术资助项目(2006AA012140)；国家自然科学基金资助项目(60673039)

作者简介：罗彦彦(1985—)，女，硕士生，研究方向为自然语言理解与机器翻译；黄德根(1965—)，男，教授，主要研究领域为自然语言理解与机器翻译。

及标注的不确定性，导致了大量的标记跨越现象，影响了分词的最终性能。

为提升CRFs的分词性能，一些研究者尝试使用多层模型：文献[7]基于双层CRFs模型的中文分词与词性标注，第一层基于CRFs进行分词获得N个分词结果，第二层用CRFs进行词性标注从这N个结果中选择一个最佳的结果；文献[8]围绕CRFs用多种方法进行中文分词，第一层进行基于CRFs的基本分词，第二层使用CRFs进行命名实体的识别，最后使用统计和规则的方法对结果进行修正；文献[9-10]基于词典和子词标注模型的中文分词，两者均采用基于子词的CRFs模型，但文献[9]采用一种少数服从多数的选举算法，在多个标注结果中选择“票数”最多的结果。而文献[10]结合HMM采用一种置信度的方式来修正CRFs的分词结果，提高系统对词典词的识别能力。近来文献[11-12]提出用无指导的分词结果指导CRFs的学习——将无指导的分词结果转化为CRFs的特征，可以加强CRFs的学习能力，提高CRFs的分词表现。

在对CRFs的分词结果进行细致地分析后，我们发现凡是被CRFs标错的字符，其边缘概率都比较低。鉴于此，本文提出使用一种基于边缘概率来解决CRFs标注错误的方案：如果边缘概率低于一个限值，则对其进行优化，根据优化后的边缘概率重新判定标注的结果。与以往只关注CRFs标注结果的方案相比，本方法更关注于CRFs标注的决策过程。

2 CRFs模型

CRFs是一种基于无向图的条件概率模型，其核心思想是利用无向图理论使序列标注的结果在整个观察序列上达到全局最优。CRFs能够使用复杂的、重叠性的和非独立的特征进行训练和推理，同时能够在某种程度上克服最大熵等模型中出现的节点偏置问题，是性能比较好的标注器。在实验中我们采用一阶线性CRFs① 作为分词的基本框架。

2.1 CRFs模型算法简介

对于给定参数 $\Lambda=(\lambda_1,\lambda_2,\cdots,\lambda_n)$ 的一阶线性链CRFs，在给定输入序列 $X=x_1\cdots x_T$ 上其对应的状态序列 $Y=y_1\cdots y_T$ 的条件可能性为：

$$P_\Lambda(Y\mid X)=\frac{1}{Z_X}\exp\Big\{\sum_{t=1}^{T}\sum_{k}\lambda_k f_k(y_{t-1},y_t,X,t)\Big\}$$

其中 Z_X 是使整个状态序列概率之和为1的规一化因子；$f_k(y_{t-1},y_t,X,t)$ 是一个二值的特征函数，λ_k 是它的权重。t 是输入序列的索引号。

输入序列 $X=x_1\cdots x_T$ 的最大可能的标注序列为

$$Y^*=\arg\max_{Y}P_\Lambda(Y\mid X)$$

可以考虑使用Viterbi算法进行解码。

2.2 标注集

文献[13]的实验结果表明：使用6标记(S,B,B_2, B_3,M,E)的基于字的CRFs，其分词表现要优于使用其他标记。所以我们使用6标记的标注集。表1给出了不同词长的标注序列。

表1 不同词长的标注序列

字长	标注序列	描 述
1	S	S表示该词为单字
2	BE	B,E分别表示一个词的开头与结尾
3	BB_2E	B_2 表示词头第二个汉字
4	BB_2B_3E	B_3 表示词头第三个汉字
5	BB_2B_3ME	M表示词中(词的第四个汉字至倒数第二个汉字)
≥6	$BB_2B_3M\cdots\cdots ME$	表示同上

2.3 局部特征

字符的n元特征在基于机器学习的中文分词中由于其高效性而被经常运用。本文使用6个(字的)上下文特征，它们分别是 C_{-1}、C_0、C_1、$C_{-1}C_0$、C_0C_1、$C_{-1}C_1$，字母C代表一个字符，其下标－1、0、1分别代表前一个字符、当前字符、后一个字符。同时采用字的分类特征 $T_{-1}T_0T_1$，T代表预先定义的类：汉字(C)，标点(P)，数字(N)，字母(L)，这个特征模板是从文献[1]中改进而得到的。

2.4 词的全局特征

除了上述的局部特征，将从无分割标记的训练和测试语料中获得的AV(Accessor Variety)[14-15]统计信息做为全局特征加入到CRFs学习中。AV全局特征是Hai Zhao等[11-12]探索出来的可以加强CRFs学习的有效特征。AV的基本思想是：一个

① http://crfpp.sourceforge.net/

子串若在多种语境下出现，那么该子串成为词的可能性就比较高。AV 定义如下：

$$AV(s) = \min\{L_{av}(s), R_{av}(s)\}$$

$L_{av}(s)$，$R_{av}(s)$分别表示子串 s 的不同前驱和后继的字符数。文献[14]中使用启发式规则去除包含停用词的子串，但实验中发现这种处理对指导 CRFs 学习的效果并不理想，原因是去除这些子串的同时，也去除了一些有用的信息。

用多特征模板来表示按 AV 标准获得的不同词长的候选词，为了避免数据稀疏问题，定义特征值为[12]：

$$f_n(s) = t \quad 2^t \leqslant AV(s) < 2^{t+1}$$

文献[13]对连续两届 Bakeoff 语料进行统计，发现 6 字长以上的词所占比例不到 0.1%，故为了提高效率，我们只考虑 5 字长以下的且 AV 值大于 1 的候选词。

2.5　字的全局特征

词的全局特征值不能直接用于 CRFs 学习，需要将获取的候选词的全局特征值转化为基于字的全局特征值。如句子“乒坛常青树老瓦”，假设当前字符为“树”且候选词 s 的长度为 5 字长，首先考虑候选词“乒坛常青树”，统计“树”做为词尾的 AV 值，记为 $f_n(c)$-E，接着考虑候选词“坛常青树老”，统计“树”被标为“M”的 AV 值，记为 $f_n(c)$-M。依此类推，分别统计“树”被标为“B_3”，“B_2”，“B”的相应的 $f_n(c)$-B_3，$f_n(c)$-B_2，$f_n(c)$-B 值。然后取这些值中最大者作为字符“树”最终的 5 字特征。同理可以统计出“树”在其他不同词长候选词中所处不同位置的全局特征值。表 2 给出了实验所用的特征模板。

表 2　实验所用的特征模板

代号	类　型	特　　征	描　　述
a	一元	C_n，$n=-1,0,1$	前一个(当前，后一个)字符
b	二元	C_nC_{n+1}，$n=-1,0$	前一个(后一个)，当前字符
		$C_{-1}C_1$	前一个和后一个字符
c	字符分类	$T_{-1}T_0T_1$	前一个，当前，后一个字符的分类
d	单字	A_0	当前字作为单字的 log(AV(s))值
e	双字		当前字在不同词长词中所处不同位置的 $f_n(c)$最大值，其值为以下形式：$f_n(c)$-B，$f_n(c)$-E，……
f	三字		
g	四字		
h	五字		

注：其中代号 a，b，c 为局部特征，代号 d，e，f，g，h 为全局特征。

3　边缘概率的利用与优化

这一过程主要涉及两个问题：第一，除用融合 AV 全局和局部特征的 CRFs 来提高标注结果边缘概率的置信度外，如何获得置信度更高的边缘概率。第二，如何利用原始语料构建 FMM 所用的词典。

3.1　字符边缘概率的首次优化

用 6 标记的 CRFs 进行中文分词，可以获得每个字符被分别标为(S，B，B_2，B_3，M，E)的边缘概率，其中 S，B 都可表示一个词的开始，这样可以只用一个标记 B 来取代 S，B。用 I 来取代其他的标记。之所以这样转化，是因为 B，I 两标记就可以区分不同的词，而 B、I 两标记在 CRFs 学习中效果不好。通过转化可以得到关于每个词边界的更可靠的边缘概率。新标记的边缘概率的计算如公式(1)(2)(3)所示。

$$P_B = \frac{(P_S + P_B)}{\sum P_t} \tag{1}$$

$$P_I = \frac{(P_{B_2} + P_{B_3} + P_M + P_E)}{\sum P_t} \tag{2}$$

$$P_t = \frac{\sum_{Y=t_0t_1\cdots\cdots t_M, t_i=t} P(Y \mid X, x_i)}{\sum_{Y=t_0t_1\cdots\cdots t_M} P(Y \mid X)} \tag{3}$$

其中 $t \in \{S, B, B_2, B_3, M, E\}$。

3.2　用于 FMM 的词典的构建

在基于 FMM 的方法中，单字词所提供的信息

量有限,故在构建词典时只选择训练语料中两字长以上的词,而对于标注后测试语料中的候选词,利用转换后的边缘概率进行判断,若一个候选词的所有组成字符的边缘概率都高于下限值 α,则认为该候选词是一个词并将其加入到词典中。即使一个候选词中只有一个字符的边缘概率低于 α,该候选词也不能加入到词典中。α 是一个经验值,实验中取0.87。如:

例 1:"正 B/0.999 799 如 I/0.968 899 赛 B/0.810 304 前 I/0.734 964 所 B/0.670 241 预 B/0.924 975料的那样……"

"正 B/0.999 799 如 I/0.968 899"该候选词中每个字符其边缘概率都高于值 α,故将"正如"这个候选词加入到词典中,而"赛前"中因"赛","前"的边缘概率都低于值 α,不能加入到词典中。重复上述过程直到标注的测试语料的末尾,这样便得到了用于FMM的词典。

3.3 子串低边缘概率的再优化

这一过程同样要利用转化后的边缘概率,若标注的测试语料中,候选词中任一字符的边缘概率低于一个上限值 β(β 也是一个经验值,实验中取 0.8),则将该候选词提取出来用于合并,得到一子串,合并过程直到一个新候选词出现的概率大于该上限值为止,此时得到的子串即为要进行边缘概率优化的子串,上述例 1 中:候选词"正如"的所有字符其边缘概率都大于 β,"赛"被标为 B 的边缘概率为 0.810 304$>\beta$,故"正如"被认为是一个词不需要进行后处理,其边缘概率维持不变。"赛前"中"前"的边缘概率低于 β,将"赛前"提取出来。其后的"所"作为一个词开头的概率低于 β,提取该词进行合并得到"赛前所",由于"预"的边缘概率高于 β,此时表明出现一个新候选词,合并过程结束,得到最终要进行优化的子串"赛前所"。若"预"的边缘概率仍低于 β 则需继续合并,直到出现一个字符被标为 B,且其边缘概率高于 β,合并过程才能终止,才能得到最终的后处理子串。

对后处理子串,依照上文构造的词典进行FMM处理,选择FMM是基于FMM对子串的处理要比对句子的处理精确率高,且实现简单。修正后字符边缘概率的定义见公式(4)。

$$P_C = P_{C_o} + \lambda(1 - P_{C_o}) \tag{4}$$

P_{C_o} 为字符的初始边缘概率,P_C 为修正后字符的边缘概率,λ 为奖励因子,其值根据 FMM 对子串标注的结果而有所不同:

$$\lambda = \begin{cases} 0.8 & \text{如果 FMM 标注结果与 CRFs 标注结果一致} \\ 0.3 & \text{CRFs 标注结果为 I 而 FMM 标注结果为 B} \end{cases}$$

通过公式(4)得到后处理子串中每个字符新的边缘概率,若其值低于 β,则采用相反的标注结果(I变B,B变I),否则接受现有的标注结果。另外一种特殊情况就是当FMM标注结果为I,而CRFs标注结果为B时,完全接受FMM的标注结果(词典词是有限的,一旦匹配,就应对其进行大的奖励)。

4 实验与分析

训练语料与测试语料源自 2007 年 SIGHAN Bakeoff 的语料库,由于编码方式的不同,实验仅在中文简体语料库上进行了测试。表 3 显示了 SXU 和 NCC 两种语料库的信息。

表 3 SXU 和 NCC 语料库信息

语料库名称	训练大小	训练单词数	测试大小	测试单词数
SXU	2.66MB	528 238	367K	113 527
NCC	4.55MB	917 255	456K	152 354

4.1 结果评测

为了便于结果的比较,我们遵循 Bakeoff 闭式规则,不使用任何额外资源与外部信息,只利用训练和测试语料自身的信息,以最终的 F,P, R 值作为评价的标准(F=2PR/(R+P))。表 4、表 5 分别列出在 SXU,NCC 语料上经过边缘概率优化前后的F,P,R 值。

表 4 SXU 语料的评测

	单纯 CRFs 标注	采用边缘概率优化
F	96.22%	96.41%
P	96.20%	96.24%
R	96.23%	96.58%

表 5 NCC 语料的评测

	单纯 CRFs 标注	采用边缘概率优化
F	93.88%	94.30%
P	93.69%	93.87%
R	94.07%	94.74%

从上述结果可以看出本文提出的边缘概率优化的方法能够有效地提高分词的精确率与召回率，最终提高分词系统的 F 值。

我们将该实验的结果与 2007 年 SIGHAN Bakeoff 上取得前三名的结果进行了比较，见表 6。

表 6　与 2007 年 Bakeoff 的结果比较

语料库	参赛者 ID	F	P	R
SXU	2	96.23%	96.25%	96.22%
	26	95.88%	95.54%	96.23%
	28	95.80%	96.11%	95.49%
	采用本文的方法	**96.41%**	**96.24%**	**96.58%**
NCC	2	94.05%	94.07%	94.02%
	26	93.86%	93.2%	94.52%
	5	93.65%	93.65%	93.65%
	采用本文的方法	**94.30%**	**93.87%**	**94.74%**

表 6 说明上文提出的边缘概率优化的方法能够使中文分词获得良好的效果，而在我们的实验中仅进行了一次 CRFs 的学习与训练，避免了多次 CRFs 训练所带来的时间与空间的巨大开销[9-10]。

4.2　结果分析

给定观察序列 X，$Y_t=y$ 时的边缘概率 $p(Y_t=y|X)$，在一定程序上反映了该标记的置信度。因而对边缘概率低的候选词进行修正具有很强的针对性。另外，对于相同的词，由于上下文环境的不同，CRFs 得出的结果也会出现很大差异。如：

例 2a　“因 B 秉 B 性 I 忠 B/0.986 672 厚 I/0.969 763……”

例 2b　“这 B 个 B 人 B 的 B/0.900 856 确 B/0.591 197忠 I/0.576 859 厚 B/0.457 475 老 B 实 I……”

在上述例 2a 中候选词“忠厚”在训练语料中没有出现，但其边缘概率非常高，表明 CRFs 对该候选词成为一个词的判定较肯定，而例 2b 中候选词“的”“确忠”“厚”分别成为词的置信度较低。重组这些置信度较低的候选词得到子串“的确忠厚”，然后利用置信度较高的候选词“忠厚”和训练语料中的词“的确”对子串“的确忠厚”进行修正，便能得到正确的标注结果。

同时对系统中的分词错误进行分析，发现其主要原因可以归结为以下几个方面：

1）训练语料中词的划分标准不一致。例如：在 SXU 训练语料中“不是”有时作为一个词出现，有时作为“不”“是”出现。这种训练语料的不一致性会降低系统的性能，当然，这个问题会影响到所有的分词系统。

2）条件随机场对绝大多数的非汉字词[16]都能准确识别。主要是 CRFs 使用了字符分类特征，但同样由于训练语料中分词标准的不一致及小部分未登录词的影响如“长城 2007”，使非汉字词的分词错误也成为不可忽视的一部分。

3）一些前后缀词是否从属于词与语料中的标准存在差异。例如“高”“科技”就被划分为“高科技”，利用 AV 全局特征这个问题就更加突出一“高科技”的 AV 值比较高，那么它提供给 CRFs 的信息就是“高科技”可能是一个词，且成为一个词的可能性也比较大，从而促使 CRFs 得出一个关于该词的一个较高的边缘概率。

4）歧义问题在错误中所占的比重也比较大。“立法会 议员”由于边缘概率较高，很难被修正，因而被错误切分成“立法 会议员”。而“无线 电视”虽然其边缘概率较低，但采用 FMM 的方法又被错误切分为“无线电 视”。

5）由于训练语料大小的限制，未登录词识别依然是最严重的问题，通过窗口特征 C_n 和 C_nC_{n+1} 虽然能够识别出部分未登录词，但对于多字符（字符长度超过 4）的其他未登录词，CRFs 的识别还存在困难，对于成语绝大多数都被拆分，而人名，机构名的错误则没有什么特殊规律，有的被拆，有的被合。

5　结论

本文在已有的基于字标注的 CRFs 分词的基础上，提出基于 CRFs 边缘概率的中文分词方法：优化 CRFs 标注的边缘概率，利用 CRFs 边缘概率高的候选词，选取 FMM 修正边缘概率低的子串的标注。通过 SIGHAN Bakeoff 2007 年的语料上的实验表明，该方法能够有效地减少 CRFs 的标注错误，提高 CRFs 的分词水平。此外，该方法对其他使用 CRFs 进行标注的任务也具有参考价值。

基于 CRFs 边缘概率的中文分词还有诸多问题，例如如何使 CRFs 获得置信度更高的边缘概率，如何对子串的低边缘概率进行更有效地优化，是否存在比 FMM 更好的方法可以修正低边缘概率子串的标注等，这些问题均有待进一步的深入研究。

参考文献：

[1] Nianwen Xue. Chinese Word Segmentation as Character Tagging[J]. Computational Linguistics and Chinese Language Processing, 2003, 8(1): 29-48.

[2] Hai Zhao, Chang-Ning Huang and Mu Li. An Improved Chinese Word Segmentation System with Conditional Random Field[C]//Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing. Sydney, Australia: 2006: 108-117.

[3] John Lafferty, Andrew McCallum and Fernando Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data[C]// Proc. of ICML-18 . Williams College, USA, 2001: 282-289.

[4] Fuchun Peng, Fangfang Feng and Andrew McCallum. Chinese Segmentation and New Word Detection using Conditional Random Fields[C]//COLING 2004. Geneva, Switzerland, 2004: 562-568.

[5] 赵海，揭春雨. 基于有效字串标注的中文分词[J]. 中文信息学报，2007，21(5)：8-13.

[6] Ruiqiang Zhang, Genichiro Kitkui and Eiichiro Sumita. Subword-based Tagging by Conditional Random Fields for Chinese Word Segmentation[C]//HLT/NAACL-2006. New York, USA: 2006, 193-196.

[7] Yanxin Shi, Mengqiu Wang. A Dual-layer CRFs Based Joint Decoding Method for Cascaded Segmentation and Labeling Tasks[C]//Proc. of International Joint Conferences on Artificial Intelligence. Hyderabad, India, 2007: 1707-1712.

[8] Zhou Jun-sheng, Dai Xin-yu, Ni Rui-yu and Chen jiajun. A Hybrid Approach to Chinese Word Segmentation around CRFs[C]//Proceedings of the Fouth SIGHAN Workshop on Chinese Language Processing. JejuIsland, Korea, 2005: 196-199.

[9] Dong Song and Anoop Sarkar. Voting between Dictionaray-based and Subword Tagging Models for Chinese Word Segmentation[C]//Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing. Sydney, Australia, 2006: 126-129.

[10] Ruiqiang Zhang, Genichiro Kikui and Eiichiro Sumita. Subword-based tagging for confidence-dependent Chinese word segmentation[C]//Proc. of the COLING/ACL on Main conference poster sessions. Sydney, Australia, 2006: 961-968.

[11] Hai Zhao and Chuyu Kit. Incorporating global information into supervised learning for Chinese word segmentation[C]//PACLING-2007. Melbourne, Australia, 2007: 66-74.

[12] Hai Zhao and Chunyu Kit. Unsupervised Segmentation Helps Supervised Learning of Character Tagging for Word Segmentation and Named Entity Recognition[C]//Proceedings of the Six SIGHAN Workshop on Chinese Language Processing. Hyderabad, India: 2008, 106-111.

[13] Hai Zhao, Chang-Ning Huang, Mu Li and Bao-Liang Lu. Effective tag set selection in Chinese word segmentation via conditional random field modeling [C]//PACLIC-20. Wuhan, China, 2006: 87-94.

[14] Haodi Feng, Kang Chen, Xiaotie Deng and Weimin Zheng. Accessor variety criteria for Chinese word extraction[J]. Computational Linguistics, 2004, 30 (1): 75-93.

[15] Haodi Feng, Kang Chen, Chuyu Kit and Xiaotie Deng. Unsupervised segmentation of Chinese corpus using accessor variety[C]//Natural Language Processing-IJCNLP 2004. Sanya, China, 2004: 694-703.