归一化的邻接类别方法在基于条件随机场的中文分词中的应用*

何赛克[1] 王小捷[2] 董远[1,3] 张韬政[2] 白雪[2]

北京邮电大学 信息与通信工程学院 北京 1000876；2. 北京邮电大学 计算机科学与技术学院 北京 1000876；
3. 法国电信北京研发中心 北京 100080

E-mail: hsk000@gmail.com ; xjwang@bupt.edu.cn; yuandong@orange-ft.com
zhangtaozheng@gmail.com; bc003@sina.com

摘 要：在自然语言处理中，中文分词系统的性能在很大程度上受制于其对未登录词（unknown words）的处理能力。本文提出了一种无监督和有监督相结合的中文分词方法。即：将邻接类别方法引入基于条件随机场的中文分词系统中。并针对邻接类别方法（Accessor Variety，AV）在处理较少的训练数据（training data）时存在的缺陷，提出了一种归一化的改进方法，以减轻计算 AV 值时产生的波动。此外，其它的一些后处理方法，如：一致性检测和基于转换的错误学习方法（TBL）也被用于提升中文分词系统的性能。

关键词：无监督分词，条件随机场，归一化的邻接类别方法，基于转换的错误学习方法

Normalized Accessor Variety in Chinese Word Segmentation Based on Conditional Random Fields

He Saike[1]，Wang Xiaojie[2]，Dong Yuan[1,3]，Zhang Taozheng[2]，Bai Xue[2]

1.School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 1000876, China;

2. School of Computer Science and Technology, Beijing University of Posts and Telecommunications, Beijing 1000876, China;

3. France Telecom R&D (Beijing), Beijing 100080, China

E-mail: hsk000@gmail.com ; xjwang@bupt.edu.cn; yuandong@orange-ft.com
zhangtaozheng@gmail.com; bc003@sina.com

Abstract: In the field of natural language processing (NLP), the performance of Chinese word segmentation (CWS) system is greatly limitedly by its competence on dealing with unknown words. This paper proposes a method combining supervised learning with unsupervised method to conduct CWS, which incorporates unsupervised segmentation into Conditional Random Fields (CRFs). Based on the flaw inherent in Accessor Variety (AV) when dealing with limited training data, normalization is involved in order to alleviate the fluctuation in the calculation of access variety value in the phrase of unsupervised segmentation. Some other post-processing measures such as consistency checking and transformation-based error-driven learning (TBL) are also employed to improve word segmentation performance.

Keywords: Unsupervised Segmentation, CRFs, Normalized Accessor Variety, TBL.

1 前言

词是自然语言处理（NLP）中的基本单元，因为它为进一步进行处理提供了词汇层面的基础。由于中文文本中缺少定界符，如空格，这使得中文分词（CWS）变成一项有趣而又具有挑

*本文得到了高等学校学科创新引智计划（项目编号：B08004）、国家支撑计划项目(项目编号：2007BAHo5B02-04)的支持。

战性的任务。

目前，机器学习方法已被成功应用于自然语言处理的各个领域。由于中文分词可被看作一个简单而有效的序列标注问题，条件随机场（CRFs）[1]成为中文分词的一个主流方法。虽然，条件随机场对已登录词（known words，在测试语料和训练语料中均出现的词）有着较高的准确率，但是，中文分词系统的性能在很大程度上受制于其对未登录词（unknown words，在测试语料中出现，而未在训练语料中出现）的处理能力。因此，如何对未登录词进行切分，成为中文分词中的一个亟待解决的问题。

本文基于我们在第四届国际中文自然语言处理（Bakeoof-4）[1]参赛的中文分词系统，引入了一种无监督的学习方法（unsupervised learning）。这种方法最初源于对语音流的的切分[2]，并且已被成功的用于分词[3]及词语提取[4]任务。针对这种无监督方法在训练数据过少时的不足，本文提出了一种改进的方法。最后给出实验数据及结果分析。

2 中文分词系统（CWS）的架构

在自然语言处理领域，CRFs 是用于中文分词的一项主流技术。根据前人的工作[5]，在此，使用一阶线性链式 CRFs 作为中文分词系统的基本架构。

2．1 条件随机场（CRFs）

对于序列标注问题，条件随机场较产生式模型（如 HMMs）及判别式模型（如 MEMM）有着明显的优势。CRFs 基于一个无向图：G = (V, E)，其中 V 是随机变量集合：$Y = \{Y_i | 1 \le i \le n\}$，代表输入文本序列中的每个字符。E 是边集合：$E = \{(Y_{i-1}, Y_i) | 1 \le i \le n\}$，这些边组成一条线性链。在给定输入序列：$(o_1, o_2 \ldots o_n)$ 的条件下，状态序列：$(s_1, s_2 \ldots s_n)$ 的条件概率定义如下：

$$P_\Lambda(s \mid o) = \frac{1}{Z_o} \prod_{c \in C(s,o)} \exp\left(\sum_{t=1}^{T} \sum_{k=1}^{K} \lambda_k f_k(s_{t-1}, s_t, o, t)\right) \quad (1)$$

其中，f_k 为特征函数；λ_k 为其对应特征函数的权重；Z_0 是个归一化因子，用于保证概率的归一性。现在，原问题可被看作是一个最优化问题：使用迭代算法，如 L-BFGS[6]，使得整个句子的条件概率 $P_\Lambda(s \mid o)$ 最大化。

2．2 标记集

前人的工作[3-4]表明，6-tag 标记集能够使 CRFs 达到更好的性能。因此，本文中的 CWS 系统也采用此种标记集，即：B，B2，B3，M，E 以及 S。

表 1 6-tag 范例

词长	对一个词的标记序列
1	S
2	BE
3	BB2E
4	BB2B3E
5	BB2B3ME
≥6	BB2B3M ··· ME

2．3 特征模板

[1]第一届中国中文信息学会汉语处理评测（CIPS-CLPE）暨第四届国际中文自然语言处理 Bakeoff（Bakeoff-4）http://www.china-language.gov.cn/bakeoff08/

表 2 CWS 系统中使用的特征集

特征类型	特征
一元语法	C_n (n = -2,-1,0,1,2)
二元语法	C_nC_{n+1} (n = -2,-1,0,1)
跳跃特征	$C_{-1}C_1$
标点特征	$Pu(C_0)$
数字，日期/时间，英文字符	$T_{-1}T_0T_1$

表 2 描述了 CWS 系统中使用的特征，其中 C 代表字符，下标 n 代表特征相对于当前字符的位置。Pun 指示当前字符是否为标点符号。T 当前字符的类别属性：数字为第一类；时间和日期为第二类；英文字符为第三类；标点符号为第 4 类；其它字符为第五类。

3 无监督的分词

除了上面的 N 元语法特征外，本文还将使用一种称为邻接类别的无监督分词（unsupervised segmentation）结果作为辅助特征，这样可以从统计上捕获更多的关于词的边界信息，从而进一步提升系统的性能。

3．1 邻接类别方法（Accessor Variety）

在中文文本中，句子中的每一个字串都可以成为一个潜在的词，然而只有那些能够表达特定含义的子串可以形成一个真正意义上的词。邻接类别方法（AV）[4]反映子串在上下文语境中的灵活程度。如果子串的灵活性越强，那么它就更加有可能是一个词。一个字串 s 的邻接类别定义如下：

$$AV(s) = \min\{Lav(s), Rav(s)\} \tag{2}$$

其中， Lav(s) 是子串 s 的左邻接类别，它被定义为子串 s 的不同前驱字符的数目，加上它在不同句首出现的次数； Rav(s) 为子串 s 右邻接类别，它被定义为子串 s 的不同后继字符的数目，加上它在不同句尾出现的次数。

在本文中，针对邻接类别方法在训练数据过少时表现出的不足（Bakeoof-4 中的数据就是如此），提出了一种改进的方法。在此将采用一种改进版本的邻接类别方法 – 归一化的邻接类（normalized accessor variety, NAV）来作为无监督的分词标准，这在下面将做详细介绍。

3．2 目标函数

给定句子中某个特定字串的邻接类别计算公式，分词问题就可被看作是一个优化问题，即：使得定义在一个句子上所有子串 AV 值的目标函数（target function）最大化。一个由 n 个中文字符构成的句子 s，其切分结果 S 可表示为由 m 个词构成的词串：

$$s = c1c2c3\ldots ci\ldots cn$$

$$S = w1w2w3\ldots wi\ldots wm \tag{3}$$

其中ci代表字符，wi代表切分后的一个词。此时，目标函数F的定义为：

$$F(S) = \sum_{i=1}^{m} f(w_i) \tag{4}$$

其中， $f(w_i)$ 为 AV 函数，其定义为：

$$f(w) = |w|^{\varepsilon} \times AV^{d}(w) \tag{5}$$

|w|为子串 w 的字符数，参数 c 取值为整数。尽管 AV 函数有很多种定义方式，但是，在此我们采用多项式形式的 AV 函数，这在前人的工作[4]中被证实十分的有效。通常情况下，可获得的训练数据十分的有限，因此对于那些具有较多字符的字串和具有极少字符的字串，在使用公式（5）计算子串 w 的 AV 值时，就会存在一些波动，进而导致 AV 值不具有很强的统计意义。比如：仅由单个汉字构成的子串，如助词‘的’，或疑问词‘吗’往往会被赋予较低的 AV 值，从而使得它们与其临近词造成‘粘连’，尽管它们被当做单个词来看可能会更好一些；而对于那些由较多汉字构成的子串，虽然它们不能表达任何的实际意义，仅仅由于它们的共现频率很高，往往会被赋予一个较高的 AV 值。因此，我们有必要对上述两种子串进行区别对待。然而，公式（5）中隐含的这个潜在问题，在过去的研究中均被忽略了[3-4]（至少没有被详细的阐述过）。为了解决这个问题，我们提出了一种归一化的 AV 函数 f（normalized accessor variety, NAV）来减弱在 AV 值计算时产生的波动，归一化的 AV 函数 f*的定义如下：

$$f^*(w)=\frac{|w|^c \times AV^d(w)}{1+\left(\frac{w}{Norm}\right)} \tag{6}$$

在公式（5）的基础上，我们引入了一个实值的归一化因子 Norm。公式（6）是基于这样的考虑：一方面，当子串 w 中包含的子串足够多时，除非其按（5）计算出的 AV 值很高，否则我们将不认为 w 为一个词，在公式（6）的计算标准下，w 将会得到一个较低的 AV 值；另一方面，当子串 w 中包含的字符过少时，除非其按公式（5）计算出的 AV 值也很低，否则，在公式（6）的计算标准下，w 仍然会得到一个较高的 AV 值。这个度量标准不需要对单字词和多字词进行分别处理，并且在一定程度上缓和了由于训练数据过少而造成的 AV 值计算上的波动，因此能够更加真实的反映子串在上下文中的灵活程度。

有了上述定义，就可以计算一个句子 s 的目标函数 F(S)了。由于切分后句子中的每个词 w_i 可以被独立的计算，因此 F(S)可以通过动态规划的方法来计算求得，这种方法的时间复杂度与句子长度呈线性关系。令 f_i 为每个子句 $c_1c_2\ldots c_i$ 的目标函数值，$w_{j\ldots i}$ 代表词 $c_{j+1}c_{j+2}\ldots c_i$ ($j \leq i$)，那么可以得到下面的动态规划公式：

$$\begin{aligned} &f_0 = 0; \\ &f_1 = f(w_{1\ldots1} = c_1); \\ &f_i = \max_{0<i-j<7} f_j + f(w_{j\ldots i}), \text{ for } i > 1; \\ &f(S) = f_n. \end{aligned} \tag{7}$$

值得注意的是，在每次迭代过程中，最多有 N 个可能的选择词（我们的实验中最大词长 N=6）。

3．3 AV 特征

在确定了 AV 值的计算公式及目标函数的形式后，在如何利用 AV 值时，我们有两种选择：<1>将基于 AV 值的无监督的分词结果（即 6-tag 标记集），作为 CRFs 模型的辅助特征。<2>直接将每个字串的 AV 值作为其特征。在后一种方式中，我们需要定义一个特征函数来对 AV 值的连续取值空间进行量化，以得到离散的 AV 值，从而避免数据稀疏（data sparsity）的问题。在这里，我们采用如下定义的特征函数[3]：

$$f_n(s) = t, \text{if } 2^t \leq AV(s) < 2^{t+1} \tag{8}$$

其中，t 是一个正整数，用它来对 AV 值取对数。因为前人研究工作中没有任何证据表明上

述 2 种使用特征的方式中，哪种效果更好一些，因此我们在实验部分做了一组实验来对此给出一个明确的回答。

4 后处理（Post-Processing）

最后，为了进一步提升分词系统的性能，我们还采用了一些后处理手段，包括：一致性检测（consistency checking）[7] 和基于转换的错误学习方法（TBL）[8]。一致性检测对每一个潜在词的每次出现，检查其边界。TBL 从 CWS 系统的输出结果的错误中学习一些规则，从而修正一些局部的标记错误。基于之前的实验结果，本文选择了一套对 CWS 系统是最为有效的规则模板[9]来学习 TBL。

值得注意的是：TBL 仅仅被用于 CWS 的开放测试集上，因为在 Bakeoff-4 中的评测结果[5]表明：TBL 将会使 CWS 系统在封闭测试集上的性能下降。

5 实验结果

本文的实验基于 Bakeoff-4 中提供的中文分词数据，共来自 5 个数据集：CITYU, CKIP, CTB, NCC and SXU。其中 CITYU 的数据是繁体中文，而其它数据则是简体中文。

在 Baseline CWS 系统（仅仅使用有监督的学习）架构中，首先使用三重交叉验证来训练初始的 CRFs 模型，并用这 3 个训练好的模型来测试剩下的一份语料，之后 TBL 通过比较训练语料和三个初始的 CRFs 的测试结果来完成训练。表 3 列出了我们在开放测试中分别用于训练 CRFs 和 TBL 的语料。

表 3 分别用于训练 CRFs 和 TBL 的语料

Run ID	CRFs	TBL
CityU	2005,2006,2007	2003
CKIP	2007	2006
CTB	2006,2007	2007

在这个基本架构的基础上，我们引入了 NAV 来进一步提升分词系统的性能。为了深入的分析 NAV 对分词系统性能的影响，我们做了一组对比试验，这将在后续的实验中见到。

在后面的实验中，无监督方法中最大的分词长度设为 6，因为中文文本中长度大于 6 的词比较少见；公式（6）中的 2 个参数 c 和 d 分别取值为 1 和 2（这是沿用了在实验中的最佳参数设置）。值得一提的是，所有子串的 AV 值都是基于全部语料计算得到的，即：全部训练语料 和 测试语料。

表 4 两种利用 AV 值方式的对比结果[2]

Run ID	F-Score	
	Auxiliary Seg	NAV value
CITYU	94.50	94.93
CKIP	93.21	94.04
CTB	94.89	95. 39
NCC	92.41	93.93
SXU	95.63	96.19

为了得到一种更好的 AV 值的使用方式，表 4 中给出一组对比试验的实验结果。其中“Auxiliary Seg”是指将基于 AV 值的无监督的分词结果，即标记集：B，B2，B3，M，E 以及 S，作为 CRFs 模型的辅助特征。“NAV value” 是指直接将每个字串的 AV 值（通过特征函数离散

[2]评测工具可以从以下网址下载：http://www.china-language.gov.cn/bakeoff08/

化后）作为其特征。在NAV方式下，公式（6）中的参数Norm取值为2.5，这在本文之前的实验中取得了最好的效果。表 4 中的实验结果表明：采用‘NAV’方式的效果较好。这个结论可以解释为 “Auxiliary Seg”中的标记错误在一定程度上干扰了CRFs的学习过程。因此，在后面的实验中，我们将采用‘NAV’值赋值的方式来使用AV特征。

表5 四个系统在CWS封闭测试集上的性能比较(未使用TBL)[3]

Run ID	F-Score			
	BaseLine	+AV	+NAV	Best
CITYU	94.43	94.78	94.93	95.10
CKIP	93.17	93.90	94.04	94.70
CTB	94.86	95. 45	**95.39**	95.89
NCC	92.99	93.00	93.93	94.05
SXU	95.46	96.15	96.19	96.23

为了证实 NAV 在封闭集上的有效性，表 5 列出了四种系统在 CWS 的性能比较。其中‘BaseLine’代表我们参加Bakeoff-4的基准系统，它仅仅利用了Table2中定义的特征，在后处理阶段，仅仅使用了一致性检测；‘+AV’表示在‘BaseLine’系统的基础上加入了 AV 特征；‘+NAV’表示在‘BaseLine’系统的基础上加入了NAV特征；‘Best’代表在相应的封闭测试集上当时最好系统的性能。实验结果表明邻接类别方法（AV）能够提升基于CRFs架构的CWS系统的性能，而基于NAV的CRFs性能最好。这正是由于NAV缓和了AV值的波动，从而带了性能的进一步提升。然而，值得指出的是，NAV在CTB数据集上的性能较AV方式略有下降，但仍然高于‘BaseLine’系统的性能。一个合理的解释是Norm取2.5对于CTB数据来说可能并不是最佳的参数设置。

表6 四个系统在CWS开放测试集上的性能比较(使用了TBL)

Run ID	F-Score			
	BaseLine	+AV	+NAV	Best
CITYU	**96.97**	97.00	96.99	96.97
CKIP	93.64	94.48	94.53	95.63
CTB	97.93	97.94	97.96	99.20
NCC	-	-	-	
SXU	-	-	-	

为了进一步证实在开放测试中，当在后处理阶段除了一致性检测，还引入了额外信息训练了TBL的情况下，此时再引入NAV是否还会带来性能的一步提升，表 7 列出了四种系统在开放测试集合上的性能比较（Norm=2.5）。在开放测试中，由于没有额外的资源，因此没有对NCC和SXU进行实验。通过比较表5和表6中‘BaseLine’系统的性能可以发现，在表6中引入的TBL大幅度的提升了CWS系统的性能，这是因为大量的额外信息被加入到了系统中，从而为系统提供更加丰富的语言学知识。与此同时，这导致AV和NAV能够挖掘的额外的上下信息变得更加少了，从而使得在开放测试中的性能提升不如封闭测试中那么的明显了。

从表7中可以看出，当公式（6）中的参数Norm设为2.5时，NAV的性能较好。另外值得注意的是，当Norm参数取值在[2，3]之间的时候，NAV的性能良好。但是，在本文的实验过程中，当Norm取值超过此范围时，性能会有不同程度的恶化，这是个需要进一步改进的地方。

[3]官方结果可以从以下网址下载：

http://www.china-language.gov.cn/bakeoff08/bakeoff-08_basic_chs.html

表 7 不同 Norm 设置下，NAV 在 CWS 封闭测试任务中的性能

Run ID	F-Score		
	Norm=2	Norm =2.5	Norm =3
CITYU	94.92	94.93	94.87
CKIP	93.94	94.04	94.05
CTB	95.50	95.39	95.35
NCC	93.91	93.93	93.94
SXU	96.15	96.19	96.08

6 结语

在自然语言处理中，中文分词系统的性能在很大程度上受制于其对未登录词的处理能力，而仅仅依靠有监督的学习是无法解决这个问题的。本文提出了一种无监督和有监督相结合的中文分词方法，即：将邻接类别方法引入基于条件随机场的中文分词系统中，并对邻接类别方法固有的缺陷进行了改进。实验结果表明：NAV 的性能在原有 AV 方法的基础上，又有了进一步的提升。与此同时，本文还简要介绍了系统在后处理中所才采用的两种关键技术：一致性检测和基于转换的错误学习方法，之前的实验表明这对于中文分词系统的整体性能也起着关键作用。最后，AV 函数的选取及参数的设置对无监督学习部分的性能有着直接影响，这将是 NAV 方法需要进一步改进的地方。

参 考 文 献

[1] J. Lafferty, A. McCallum, and F. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proceedings of the 18th ICML, 2001: 282–289, San Francisco, CA.

[2] Zellig Sabbetai Harris. Morpheme within words. In Papers in Structural and boundaries Transformational Linguistics, 1970: page 68 - 77.

[3] Hai Zhao and Chunyu Kit, Unsupervised Segmentation Helps Supervised Learning of Character Tagging for Word Segmentation and Named Entity Recognition, The Sixth SIGHAN Workshop on Chinese Language Processing (SIGHAN-6), 2008: pp.106-111, Hyderabad, India, January 11-12.

[4] Haodi Feng, Kang Chen, Chunyu Kit, and Xiaotie Deng. Unsupervised segmentation of Chinese corpus using accessor variety. In K.-Y. Su, J. Tsujii, J. H. Lee, and O. Y. Kwong, editors, Natural Language Processing-IJCNLP 2004, volume 3248 of Lecture Notes in Computer Science, 2005: pages 694–703, Sanya, Hainan Island, China. Springer Berlin / Heidelberg.

[5] Xinnian Mao, Yuan Dong and Saike He, Sencheng Bao and Haila Wang, Chinese Word Segmentation and Name Entity Recognition Based on Condition Random Fields, The Sixth SIGHAN Workshop on Chinese Language Processing (SIGHAN-6), 2008, Hyderabad, India

[6] R.H. Byrd, J. Nocedal and R.B. Schnabel. Representations of quasi-Newton matrices and their use in limited memory methods. Mathematical Programming, (63), 1994:129-156.

[7] Jin Kiat Low, Hwee Tou Ng and Wenyuan Guo, A Maximum Entropy Approach to Chinese Word Segmentation, 2005.

[8] David D Palmer. A trainable rule-based algorithm for word segmentation[C]// ACL. Proc. of the 35th annual meeting on ACL. Madrid, Spain: Morgan Kaufmann Publishers, July, 1997: p.321-328.

[9] Nan He, Xinnian Mao, Yuan Dong, Haila Wang. Transformation-based Error-driven Learning as Post-precessing for Chinese Word Segmentation, In Proceedings of the 7th International Conference on Chinese Computing, 2007: 46-51, Wuhan, China.