基于马尔可夫间隔标注的中文分词算法*

姜文斌 王志洋 刘群 吕雅娟
中国科学院计算技术研究所 北京 100190
E-mail: {jiangwenbin, wangzhiyang, liuqun, lvyajuan}@ict.ac.cn

摘 要：典型的判别式方法通过标注每个字符在词中的相对位置，将分词看作字符标注问题。本文提出了一个形式化的标注策略——马尔可夫间隔标注，来对汉语进行分词。在每一步中，*N* 阶马尔可夫间隔标注对连续的 *N+1* 个字符间隔进行标注，并按照马尔可夫方式来处理这 *N+1* 个间隔。实验结果表明：在使用相似特征的前提下，当阶数由 0 渐变为 2 时，间隔标注方法的分词准确率也随之增加。

关键词：判别式方法；中文分词；特征模板选择；马尔可夫间隔标注

Word Segmentation by Markov Gap-Tagging

Jiang Wenbin , Wang Zhiyang, Liu Qun, Lü Yajuan
Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190
E-mail: {jiangwenbin, wangzhiyang, liuqun, lvyajuan}@ict.ac.cn

Abstract: Classical discriminative approaches treat word segmentation as a character-tagging problem. This paper presents an alternative, formal tagging strategy named markov gap-tagging for Chinese word segmentation. An *Nth* order markov gap-tagging tags *N+1* successive gap between characters at each step, and processes all these gap *N+1*-grams in a markov fashion. Experimental results show that, along with the markov order growing from 0 to 2, the accuracy of the gap-tagging-based segmentor increases continuously, although using similar features.

Keywords: discriminative approach; Chinese word segmentation; markov gap-tagging

1. 引言

对于没有明显的词定界符的语言，例如汉语，分词是很多自然语言处理（NLP）任务的基础。近年来，大量的基于语料库的机器学习方法被引入：例如，生成模型——隐马尔可夫模型（HMM）(Rabiner,1989)，还有判别模型——条件随机场（CRFs）(Lafferty et al.,2001)等。由于判别模型在表示特征时更为灵活，相比生成模型，它们往往能得到更好的结果。

根据字符在词中出现的规律，(Xue and Shen,2003) 将分词看作字符标注问题。通过从上下文中抽取特征，判别式分类器给每个字符标注一个表征它在词间相对位置的标签。当所有字符的位置标签都确定之后，我们便可从标注序列中得到分词结果。这种方法准确而且有效：一方面，很多判别式方法可以用来训练分类器，像最大熵 (Ratnaparkhi and Adwait,1996)；另一方面，如果仅使用局部特征，我们可以根据动态规划搜索到全局最优的结果。此外，若在标注时引入词性（POS）信息，这种方法也可以同时处理分词和词性标注问题 (Ng and Low,2004)。

在上述方法中，标注集的确定都是基于经验的，可能存在不能充分利用分类器能力的风险。本文提出马尔科夫间隔标注的标注策略来处理中文分词问题。给定一个汉语句子，在每一步中，

*本文承自然科学基金项目(项目号 60603095 和 60736014)和国家 863 重点项目(项目号No. 2006AA010108)的资助。

*N*阶马尔可夫间隔标注对连续的*N+1*个字符间隔进行标注，并按照马尔可夫方式来处理间隔。随着阶数的增大，本策略能更好的利用上下文特征，从而为词定界决策提供更多的信息。在中文树库*CTB5.0*和人民日报语料上的实验表明：在使用相似特征的前提下，当阶数由0变为2时，分词准确率随之不断的增加。此外，我们还提出了一个特征模板选择算法，用来选择适合N阶马尔可夫间隔标注的特征模板。

2. 基于间隔标注的分词

一个中文句子可以用字符序列$C_{1:n}=C_1C_2...C_n$表示。我们显式的将字符C_i和C_{i+1}间的间隔$G_i(i=1...n-1)$加入句子$C_{1:n}$中：$C_{1:n}\mid G_{1:n-1}=C_1G_1C_2G_2...G_{n-1}C_n$。这样带有间隔的切分结果可以表示如下：

$C_{1:e_1}\mid G_{1:e_1-1}$
G_{e_1}
$C_{e_1+1:e_2}\mid G_{e_1+1:e_2-1}$
G_{e_2}
...
$G_{e_{m-1}}$
$C_{e_{m-1}+1:e_m}\mid G_{e_{m-1}+1:e_m-1}$

$C_{l:r}\mid G_{l:r-1}$表示带有间隔序列$G_{l:r-1}$的字符序列$C_{l:r}$。如上所示，间隔被分为两组，一些出现在词中，像$G_1...G_{e_1-1}$；一些则出现在词间，像G_{e_1}。我们将出现在词中的间隔标记为 *adjoin*（表示为 *A*），出现在词间的间隔标记为 *split*（表示为 *S*）。这样候选切分结果是对应标记有特定的 *A/S* 的序列，分词问题可以转化为间隔标注问题。

2.1 *N*阶马尔可夫间隔标注

当在每一步同时考虑 *N*+1 个间隔时，将会产生2^{N+1}种决策选择。这些决策构成决策集，表示为$\{A,S\}^{\otimes N+1}$。这里操作符$\otimes$表示集合间的笛卡尔乘。*N* 阶马尔可夫间隔标注在位置G_i标注间隔序列$G_{i-N:i}$，在位置G_{i+1}标注$G_{i-N+1:i+1}$。通过线性规划可以对该马尔可夫链进行解码。

我们训练一个判别式分类器来指导间隔标注。给定一个字符/间隔序列$x=C_{1:n}\mid G_{1:n-1}$，我们的目标是找到一个间隔标注序列$F(x)$，它满足：

$$F(x)=\arg\max_{y\in\{A,S\}^{\otimes n-1}} Score(x,y) \qquad (1)$$

这里函数 *Score* 表示标注序列 *y* 经由分类器评估的结果：

$$Score(x,y)=\sum_i Eval(y_{i-N:i},\phi(x,i)) \qquad (2)$$

$\phi(x,i)$表示从字符/间隔序列 *x* 中的位置g_i所选择的特征。*Eval* 表示基于特征$\phi(x,i)$，分类器对标注选择$y_{i-N:i}$的评估分数。

2.2 对字符标注模型的解释

实际上，经典的字符标注模型可以用间隔标注理论来解释。(Ng and Low,2004) 将字符标注为以下四个标签之一：*b*，*m*，*e* 和 *s*，分别表示单词的起始字符、中间字符、结束字符以及单独

成词。(Xue and Shen,2003)中这四个标签叫做 *LL*，*MM*，*RR* 和 *LR*。一阶间隔标注等同于这种方式。例如给定一个字符/间隔序列片段$...G_{i-1}C_iG_i...$，将G_{i-1i}标记为 *AA* 等同于将C_i标记为 *m*，*AS* 等同于 *e*，*SA* 等同于 *b*，*SS* 则等同于 *s*。另外一种字符标注策略将每个字符分为为两种类型：*b* 和 *c*，分别表示一个词的开始和继续,对应于(Xue and Shen,2003)中的 *LL* 和 *RR*。这仍然可以用一阶马尔可夫间隔标注来解释：将 *SA* 和 *SS* 合在一起表示 *b*，*AA* 和 *AS* 合在一起表示 *c*。

(Xue and Shen,2003)的实验结果显示仅有两类标签的标注策略明显不如四类标签的标注策略。一个合理的解释是：四类标签的字符标注模型与一阶马尔可夫间隔标注是一一对应的，因此它能充分利用判别式模型的分类能力；而两类标签的字符标注模型则由于标签合并，损失了部分分类能力。

3. 特征模板选择

我们对(Berger et al.,1996)的特征选择算法进行了适当的修改，以从候选集*cand*中产生特征模板集。这一算法通过贪心方式增量式的生成特征模板集。算法(a)描述了这一贪心过程：在每一轮迭代中，我们从当前的候选模板集*cand*中选择带来准确率最大增益的模板 f^*。当候选模板集为空，或者准确率增益小于我们规定的阈值*min_increment*时，迭代停止。其中，*cand*是从由当前间隔周围大小为*K*的窗口中抽取的一元和二元特征集。

Input: candidate set $\mathcal{C}$, train set $\mathcal{T}$, develop set $\mathcal{D}$
$\mathcal{F} \leftarrow \emptyset$
$eval_{\mathcal{F}} \leftarrow 0$
repeat
  for $f \in \mathcal{C}$ **do**
    $m_{+f} \leftarrow BuildModel(\mathcal{T}, \mathcal{F} \cup \{f\})$
    $eval_{+f} \leftarrow TestModel(m_{+f}, \mathcal{D})$
  $f^* \leftarrow argmax_f eval_{+f}$
  if $eval_{+f^*} - eval_{\mathcal{F}} > min\_increment$ **then**
    $\mathcal{F} \leftarrow \mathcal{F} \cup \{f^*\}$
    $\mathcal{C} \leftarrow \mathcal{C} - \{f^*\}$
    $eval_{\mathcal{F}} \leftarrow eval_{+f^*}$
until $\mathcal{C} = \emptyset$ or $\mathcal{F}$ not enlarged
Output: template set: $\mathcal{F}$

(a)

Input: character/gap sequence $C_{1:n}|G_{1:n-1}$
for G_i in $C_{1:n}|G_{1:n-1}$ in ascending order **do**
  $\phi \leftarrow \Phi(C_{1:n}|G_{1:n-1}, i)$
  for each tagging choice $c \in \{A, S\}^{\otimes N+1}$ **do**
    $a^* \leftarrow \text{argmax}_{a \in ANC(c)} S[i-1, a] \cdot score$
    $S[i, c] \cdot score \leftarrow S[i-1, a^*] \cdot score + F(\phi, c)$
    $S[i, c] \cdot backp \leftarrow a^*$
$tail \leftarrow argmax_{c \in \{A,S\}^{\otimes N+1}} S[n-1, c] \cdot score$
$r^* \leftarrow Trace(tail)$
Output: best gap tag sequence r^*

(b)

算法　(a)特征模板选择算法；(b)N 阶马尔可夫间隔标注解码算法

4. 解码

基于间隔标注的分词模型的目标是找到得分最高的间隔标注序列。给定一个字符/间隔序列$C_{1:n}|G_{1:n-1}$，解码算法通过线性规划自左向右的进行处理。$C_{1:n}|G_{1:n-1}$中的每个间隔G_i，*N* 阶马尔可夫间隔标注对$G_{i-N:i}$的标注决策考虑所有的2^{N+1}种可能，并给每个标注赋以得分来表征该标注的置信程度。为了记录线性规划过程中子问题的最佳解决方案，在每个间隔G_i，对所有的2^{N+1}种候

选标注结果，我们保存以这种结果结尾的最好的$G_{1,i}$标注序列。因此，在间隔G_i，我们可以通过将当前候选标注的得分加上它的最好前驱标注序列得分，便可得到当前状态的最好得分。算法(b)描述了 N 阶马尔可夫间隔标注的解码算法。依据 2.1 部分的定义，（第 3 行）函数 Φ 从字符/间隔序列$C_{1:n} \mid G_{1:n-1}$中抽取当前考察的间隔序列$G_{i-N:i}$需要的特征。函数 F 基于分类器得到的间隔G_i上下文的特征集ϕ来生成评估结果。第 2-7 行通过线性规划搜索最佳的间隔标注序列，并使用函数 *Trace* 来追踪输出结果（第 9 行）。

5. 实验及结果分析

本部分记录了特征模板选择和 N 阶马尔可夫间隔标注的实验情况。不考虑训练语料领域和规模对特征模板选择过程的影响，我们在一个相对较小的语料——中文树库 *CTB5.0* 上抽取模板集。并利用这些模板集，来进行马尔可夫间隔标注阶数（0-3）变化的实验。

不论在特征模板选择，还是N阶马尔可夫间隔标注实验，我们均使用由张乐博士开发的最大熵工具包[†]来进行判别式分类，不使用高斯优先，只进行 100 轮迭代。

5.1 特征模板选择

依照统计句法分析的先例，我们将 *CTB5.0* 进行如下划分：1-260 章（18,074 个句子）作为训练集 T，301-325 章（350 个句子）作为开发集 D。零阶马尔可夫间隔标注的候选特征集包含了当前间隔上下文窗口大小为 4 的所有一元和二元字符特征。对于高阶的马尔可夫间隔标注，它们的候选特征集可以递归产生。为了充分利用 *Cand* 中的 n 元特征，阈值 *min_increment* 被赋以很小的值 0.0001。阶数从 0 到 3 所使用的特征模板见表(1)。

5.2 N 阶马尔可夫间隔标注

马尔可夫间隔标注阶数变化的实验是在两个不同规模的语料上进行的。第一个是用来做特征模板选择实验的中文树库 *CTB5.0*，除了上面提到的训练集和开发集外，它还有个测试集，由 271-300 章（348 个句子）组成。另外一个语料是人民日报，它的训练集有 100,344 个句子，测试集有 19,006 个句子，比中文树库 *CTB5.0* 大很多。为了更清楚地测试马尔可夫间隔标注的性能，我们只使用表(1)中提到的特征模板，不使用任何词典和标点信息。

表(2)是实验结果的详细记录，包括所耗内存、分词速度，以及对应的 F 值。所有的 8 个间隔标注模型都是在同一台 32 位 PC 机上测试的。由表(2)可以看出，随着阶数的变化，不同语料的两组实验结果表示出相似的变化趋势：阶数从 0 渐变到 2，F 值慢慢增加；当阶数变为 3 时，F 值下降。在 *CTB5.0* 上，阶数从 0 变为 1 时，错误率减少了 5.58%；从 1 变为 2 时减少了 4.37%；但当阶数由 2 变为 3 时，错误率反而增加了 16.66%。人民日报语料上的结果略好于 *CTB5.0*，阶数由 0 变为 1，错误率减少了 14.29%；由 1 变为 2，错误率减少了 8.33%；当阶数由 2 变为 3 时，错误率同样也增加了，尽管没有在 *CTB5.0* 上增加了那么明显。

在每组实验中，解码时间随着马尔可夫阶数的增加成倍增加。和等价于经典的字符标注的 1 阶马尔可夫间隔标注策略相比，0 阶马尔可夫间隔标注在保证可容许的准确率的前提下，分词速度快了一倍。仅仅依靠简单的线性扫描，0 阶马尔可夫间隔标注也能取得这样不错的结果。

[†] http://homepages.inf.ed.ac.uk/s0450736/maxent.html

Markov order	Feature template set
0	$C_{-1}C_0, C_0C_1, C_1C_2$
1	C_0C_1, C_1C_2, C_2C_3
2	$C_{-1}C_0, C_0C_1, C_1C_2$
3	$C_{-1}C_0, C_0C_1, C_1C_2, C_2C_3$

(1)

Markov Order	Memory	Time	F-measure
CTB 5.0			
0	49M	0.22s	0.9588
1	49M	0.44s	0.9611
2	49M	0.72s	0.9628
3	73M	1.35s	0.9566
People's Daily			
0	186M	34.1s	0.9580
1	190M	66.3s	0.9640
2	196M	120.0s	0.9670
3	303M	214.3s	0.9633

(2)

表 (1) 0-3 阶方法使用的特征模板集；(2) 马尔可夫间隔标注模型实验结果

5.3 结果分析

当阶数由 0 变为 2 时，间隔标注模型的准确率不断增加，解码时间消耗也随之增加，但模型大小基本不变（通过所使用的内存大小看出）。对于 N 阶马尔可夫间隔标注，每一步有 2^{N+1} 个决策选择。随着阶数的增加，更多的决策选择被引入。一方面，更多的决策选择能更充分的利用特征。首先，在某一步对更多的可能做出决策时，特征可以更有效地被利用。其次，随着阶数的增加，当前决策依赖于先前已有的决策结果，因此当前决策所使用的特征也在后续决策中会间接使用到。另一方面，由于自然语言本身的歧义性，以及机器学习算法能力的有限性，更多的决策选择可能导致糟糕的切分结果。因此，有必要对马尔可夫阶数有个折中。从表(2)的结果可以看出，在给定语料规模和算法的前提下，最好的马尔可夫阶数是 2，而不是等同于经典的字符标注模型的 1 阶模型。

我们还发现，阶数从 0 变为 3，人民日报语料上准确率增量（我们把 3 阶时的下降作为负增长）比中文树库 *CTB5.0* 更优。数据稀疏是准确率下降的一个主要原因，当阶数为 3 时，每一步的决策可能是 16 种！我们相信，随着训练语料的扩大，更高的阶数也适合于马尔可夫间隔标注。

6. 结论与展望

本文提出了一个新颖的方法——*N* 阶马尔可夫间隔标注，来对中文进行分词。这种策略考虑连续的 *N+1* 个字符间隔，并按照马尔可夫方式进行处理。当阶数在某一指定范围内增加时，*N* 阶马尔可夫间隔标注由于更好的利用了上下文特征和相邻解码步之间的关联性，分词准确率稳定的增加。在 *CTB5.0* 和人民日报上的语料上的实验表明，在使用相似特征的前提下，阶数由 0 增加到 2 时，分词准确率持续的增加。特别地，当阶数由 1 变为 2 时，分词效果比经典的判别式方法更好。这告诉我们一个有趣的事实：目前还没有充分挖掘基准模型的潜力，使用重排序技术(Jiang et al., 2008b)有点为时过早。

通过简单的分析，我们在马尔可夫间隔标注和经典的字符标注模型之间建立了联系：字符标注方法可以看作马尔可夫间隔标注的特例。此外，我们还给出了一个适应于马尔可夫间隔标注的特征模板选择算法。

尽管马尔可夫间隔标注模型充分利用了上下文特征来提高分词质量，但仍然还有一些问题。首先，本文提出的 N 阶马尔可夫间隔标注将标注集大小设为 2^N，如果将这些标注集再进行合理的分组，每组表示一个更泛化的结果，这样能否获得更好的结果？其次，最近联合了分词和词性标注的方法颇为流行(Jiang et al.,2008a; Zhang and Chark,2008)，我们有必要在间隔标注中也这样做吗？这些都是我们以后研究需要关注的。

参 考 文 献

[1] Lawrence. R. Rabiner, 1989. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition.Proceedings of IEEE, pages 257–286.

[2] John Lafferty and Andrew McCallum and Fernando Pereira, 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Proceedings of the 18th ICML, pages 282—289.

[3] Nianwen Xue and Libin Shen, 2003. Chinese word segmentation as LMR tagging. Proceedings of SIGHAN Workshop.

[4] Ratnaparkhi and Adwait, 1996. A maximum entropy part-of-speech tagger. Proceedings of the Empirical Methods in Natural Language Processing Conference.

[5] Hwee Tou Ng and Jin Kiat Low, 2004. Chinese Part-of-Speech Tagging: One-at-a-Time or All-at-Once? Word-Based or Character-Based? Proceedings of EMNLP.

[6] Adam L. Berger and Stephen A. Della Pietra and Vincent J. Della Pietra, 1996. A Maximum Entropy Approach to Natural Language Processing. Computational Linguistics.

[7] Wenbin Jiang and Liang Huang and Yajuan Lv and Qun Liu, 2008a. A Cascaded Linear Model for Joint Chinese Word Segmentation and Part-of-Speech Tagging. Proceedings of ACL.

[8] Wenbin Jiang and Haitao Mi and Liang Huang and Qun Liu, 2008b. Word Lattice Reranking for Chinese Word Segmentation and Part-of-Speech Tagging. Proceedings of COLING.

[9] Yue Zhang and Stephen Clark, 2008. Joint Word Segmentation and POS Tagging using a Single Perceptron. Proceedings of ACL.