HDP 与互信息相结合的中文无指导分词□

曹自强，李素建

北京大学计算语言学教育部重点实验室 北京 100871

E-mail: {ziqiangyeah, lisujian}@pku.edu.cn

摘要：本文探讨了无指导条件下的中文分词，这对构建语言无关的健壮分词系统大有裨益。互信息与 HDP（Hierarchical Dirichlet Process）是无指导情况下常用的分词模型，本文将两者结合，并改进了采样算法。不考虑标点符号，在两份大小不同的测试语料上获得的 F 值为 0.693 与 0.741，相比 baseline 的 HDP 分别提升了 5.8%和 3.9%。本文还用该模型进行了半指导分词，实验结果比常用的 CRF 有指导分词提升了 2.6%。

关键词：HDP；互信息；无指导分词

Unsupervised Chinese Word Segmentation Based on HDP and Mutual Information Getting together

Abstract: This paper explores Chinese word segmentation without training data, which greatly benefits the foundation of language-independent word segmentation system. Mutual information and HDP are both widely used methods for unsupervised segmentation task. We combine these two models and improve the sampling algorithm. Without regard to punctuations, the f-scores of two test corpus with different sizes are 0.693 and 0.741. Compared to HDP baseline, the scores rise 5.8% and 3.9%, respectively. Finally, our model is applied to semi-supervised word segmentation. The f-score is 2.6% larger than the common supervised CRF model.

Key words: HDP; mutual information; unsupervised word segmentation

1 引言

分词是中文、日文等没有明确词分隔符的亚洲语言文本处理的最初环节，效果直接影响后续步骤。当前绝大多数的分词方法是有指导的。相比无指导方法，有指导分词往往能取得更高准确率。但是，构建训练语料的代价非常大，并且对新词的识别没有太好办法。而无指导分词提供了一种近似语言无关的分词方法，对一些我们不了解的语言尤为重要。

无指导分词的最初研究主要来自互信息[1]。随后，其他一些统计量，包括χ^2值[2]，t-测试差[3]，直接串频统计[4]等等，也逐渐被引入。也有学者提出将这些统计量相结合的方法[5]。近期，非参数贝叶斯模型在无指导分词中的应用越来越广泛。比如 Pitman-Yor Process（PYP）[6]，Dirichlet Process（DP）与 HDP[7, 8]，Hierarchical PYP（HPYP）[9]，Hierarchical HPYP（HHPYP）[10]等等。

基于统计量和基于贝叶斯模型的分词方法有其各自的优缺点，如下表所示。

表 1.1：两种无指导分词方法各自的特点

方法	优点	缺点
统计量	实现简单；切分速度快	通过阈值调节，容易过拟合
贝叶斯模型	模型完备；全局最优切分，通用性好；可结合训练集。	速度慢

针对两者优缺点间的互补性，本文将这两种方法结合。HDP 的先验设定是个重要问题，会直接影响到后续实验的速度和性能，而统计量互信息是通过概率表示的，对语言现象给出了较好的描述，因此本文提出互

□本项目受到国家自然科学基金项目（编号： 61273278）、国家社会科学项目（编号：12&ZD227），国家科技支撑计划子课题项目（编号：2011BAH10B04-03）和国家 863 计划（编号：2012AA011101）的资助。

信息与 HDP 相结合的分词方法，用互信息表示 HDP 的先验。同时，本文改进了 HDP 的采样过程，使其能够处理文献[8]中容易出现的多错误长字符串问题。最后，利用本模型可以结合已标注语料的特点，进行了半指导分词实验。

本文其他章节的内容组织如下：第 2 节简单介绍了互信息和 HDP 的相关知识。第 3 节详述了 HDP 与互信息的结合方法，并对采样过程进行了改进。第 4 节是实验部分，包括改进前后的无指导分词和改进后的半指导分词。最后，第 5 节是对全文的总结并提出了进一步工作。

2 背景知识介绍

2.1 互信息简介

互信息可以衡量两个对象的结合能力。假设给定两个对象x,y，它们之间的互信息定义为：

$$\mathrm{mi}(\mathrm{x},\mathrm{y})=\log\frac{p(x,y)}{p(x)p(y)} \quad (1)$$

$\mathrm{mi}(\mathrm{x},\mathrm{y})$越大，$\mathrm{x},\mathrm{y}$就越倾向于以一个整体的形式出现。

应用于分词过程，文献[1]中x,y代表任意相邻的两个字符。通过设定阈值就能决定两者间是否需要切分。为了与 HDP 相结合，本文将x,y扩展为任意两个相邻字符串。

2.2 HDP简介

对于无监督参数模型，类的总数确定往往是中一项很艰巨的挑战，极易造成过拟合或欠拟合。而 HDP 是一种非参数贝叶斯模型。其参数数目随样本数的增加而自适应，因此用于聚类过程，不需要事先指定好类的总数。

2.2.1 DP 模型

在介绍 DP 之前，必须先了解 Dirichlet 分布。Dirichlet 分布是一种分布的分布，即给定 K 种可能的离散事件，对应度量参数为$\vec{\alpha}$的情况下求后验分布为$\vec{p}$的概率：

$$\mathrm{p}(\vec{p}|\vec{\alpha})\triangleq\frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K}p_k^{\alpha_k} \quad (2)$$

回到 DP，我们定义：在一个划分成 K 部分，每一部分T_i拥有先验概率$\mathrm{H}(T_i)$的空间Θ中，如果一随机过程 G 的概率分布满足：$\left(\mathrm{G}(T_1),\mathrm{G}(T_2),\ldots,\mathrm{G}(T_K)\right)\sim Dir(\alpha\mathrm{H}(T_1),\alpha\mathrm{H}(T_2),\ldots,\alpha\mathrm{H}(T_K))$，则称 G 满足 DP 过程。可以证明，对于给定的α,H，符合条件的G是唯一的，记为$\mathrm{G}\sim\mathrm{DP}(\alpha,\mathrm{H})$，其中的$\alpha$是集中参数，其大小反映了$\mathrm{G}$与先验分布$\mathrm{H}$的相似程度。

DP 最重要的应用是分类，对$\mathrm{G}\sim\mathrm{DP}(\alpha,\mathrm{H})$，在给定 N 个样本$\bar{\theta}_i\sim G$的情况下，G 的后验概率的期望为：

$$\mathrm{E}(\mathrm{G}(\mathrm{t})|\bar{\theta}_i;\alpha,\mathrm{H})=\frac{\alpha H(t)+\sum_{i=1}^{N}I(\bar{\theta}_i=t)}{\alpha+N} \quad (3)$$

如果$\alpha=0$，等式就变为最大似然估计。因此，DP 很大程度上可以理解为数据平滑[9]。

应用于分词过程， DP 可以视为一元模型——文本中一个词产生的概率只和它自身的性质有关，与其所处的上下文环境无关。即：

$$\mathrm{p}_1(w_i=c_1c_2,\ldots,c_K|h)=\frac{t_{w_i}+\alpha_0p_0(w_i)}{t+\alpha_0} \quad (4)$$

$$\mathrm{p}(w_i|\mathrm{text})\approx\mathrm{p}_1(w_i|h)$$

其中$p_0(w_i)$表示由字符串$c_1c_2,\ldots,c_K$组成词w_i的先验概率，p_1为一元模型对应后验概率，h 代表当前切分状况，t_{w_i}表示切分出词w_i的数目，t表示总词数，α_0为集中参数。

2.2.2 HDP 模型

HDP 是对 DP 的改进，它包含多层 DP，前一层 DP 的结果作为后一层 DP 的先验。

应用于分词过程，两层 HDP 可以被视为二元模型，它考虑了前一个词的影响，具体概率计算公式：

$$\mathrm{p}_1(w_i|h)=\frac{t_{w_i}+\alpha_0 p_0(w_i)}{t+\alpha_0}$$

$$p_2(w_i|w_{i-1},h)=\frac{n_{<w_{i-1},w_i>}+\alpha_1 p_1(w_i|h)}{n_{w_{i-1}}+\alpha_1} \quad (5)$$

$$\mathrm{p}(w_i|\mathrm{text})\approx p_2(w_i|w_{i-1},h)$$

其中p_2为二元模型对应后验概率，$n_{w_{i-1}}$代表以w_{i-1}开头的二元组数目，$n_{<w_{i-1},w_i>}$代表二元组$< w_{i-1},w_i >$的数目，α_1为第二层 DP 参数的集中参数。

2.2.3 基于 Gibbs Sampling 的潜变量估计

Gibbs Sampling 是蒙特卡洛方法。对应于分词任务，文献[8]提出的 Gibbs Sampling 具体步骤为：

一、 对语料进行随机切分，求得(4)(5)两式所需的各二元组数目；

二、 对语料中的每一句重新进行切分。先扣除这一句本来切分的二元组数目，再考虑每个位置是切分还是合并。

假设当前词串为$\alpha w_L w_1 w_R w_{RR}\beta$，α、β为运算中不需要考虑的相同部分。对于$w_1$是否需要分割成$w_2$、$w_3$，我们有两个假设$H_1, H_2$：

$$H_1(w_1|h)=p(w_1|w_L)\times p(w_R|w_1) \quad (6)$$

$$H_2(w_2w_3|h)=p(w_2|w_L)\times p(w_3|w_2)\times p(w_R|w_3) \quad (7)$$

式中每个条件概率都通过(5)求得。其中，字符串$c_1c_2,\dots,c_K$成词 w 的先验概率为：

$$p_0(w=c_1c_2,\dots,c_K)=\frac{p_\$}{1-p_\$}\prod_{k=1}^{K}(1-p_\$)p(c_k) \quad (8)$$

$p_\$$表示产生切分符号的概率（常用一个自定义值表示），$p(c_k)$表示每个字符在文档中出现的概率。

类似的，对于w_1w_R是否需要合并为w_4，我们也有两个假设H_3, H_4：

$$H_3(w_1w_R|h)=p(w_1|w_L)\times p(w_R|w_1)\times p(w_{RR}|w_R) \quad (9)$$

$$H_4(w_4|h)=p(w_4|w_L)\times p(w_{RR}|w_4) \quad (10)$$

根据两个假设我们得到了切分和合并的概率比。然后用随机数模拟的方式决定每个位置是否切分。

3 模型改进

3.1 结合互信息的先验

(8)式给出的先验保证了高频词的具有较高的成词概率，但是带来了两个问题：

一、 缺乏字符间的顺序考量。对于给定文本，究竟有哪些潜在词汇是固定的。比如短语“我爱我家”，可以切分出的 bigram 只有“我爱”、“爱我”、“我家”。但按照(9)式，还统计了“爱家”、“家爱”、“家我”、“我我”这些不可能出现的 bigram 的概率。特别是“我我”计算出的概率还是最大的。这显然不合理。

二、 本先验并不能提供是否需要划分和在什么位置切分的依据。仅考虑先验，我们有词w需要切分为w_1, w_2的概率比：

$$\frac{p_0(w=c_1c_2,\dots,c_K)}{p_0(w_1=c_1c_2,\dots,c_l)p_0(w_2=c_{l+1}c_{l+2},\dots,c_K)}$$

$$=\frac{\frac{p_\$}{1-p_\$}\prod_{k=1}^{K}(1-p_\$)p(c_k)}{(\frac{p_\$}{1-p_\$}\prod_{k=1}^{l}(1-p_\$)p(c_k))\times(\frac{p_\$}{1-p_\$}\prod_{k=l+1}^{K}(1-p_\$)p(c_k))}$$

$$=\frac{1-p_\$}{p_\$} \quad (11)$$

是一个固定值，**和划分位置无关**。大大弱化了先验的作用。

利用互信息的思想，本文直接计算字符串出现的概率表示该字符串成词的先验概率：

$$p_0(w = c_1c_2,\dots,c_K) = \frac{n(c_1c_2,\dots,c_K)}{n(all\ strings)} \quad (12)$$

则w是否要切分成w_1, w_2的先验决策就取决于比例：

$$\frac{p_0(w)}{p_0(w_1)p_0(w_2)} = 2^{\mathrm{mi}(w_1,w_2)} \quad (13)$$

恰好是互信息的等价形式，可以较好指导分词。

同时，本文还考虑了汉语中的词长分布特性：长度为 1 或 2 的词远多于其他，因此在先验中引入一个词长概率分布$\mathrm{p(length = w.length)}$。为简单起见，本文假设汉语的词长满足泊松分布。所以，最终的先验概率计算公式为：

$$p_0(w = c_1c_2,\dots,c_K) = \frac{n(c_1c_2,\dots,c_K)}{n(all\ strings)} \times \mathrm{Possion(K,\lambda)} \quad (14)$$

3.2 采样的改进

文献[8]的采样过程中仅考虑当前切分的字符串w_1是否需要二分为w_2、w_3，这是带有缺陷的。比如$w_1 =$ "学习和掌握"，是由 3 个词组成，不论如何二分，w_2、w_3中最多有一个划分正确。不妨令$w_2 =$ "学习"、$w_3 =$ "和掌握"。由于切分有误的词汇在语料中一般很少出现，我们有：

$$p(w_1|w_L), p(w_R|w_1), p(w_3|w_2), p(w_R|w_3) \approx 0$$
$$0 \ll p(w_2|w_L) < 1$$

计算(6)(7)两式之比：

$$\frac{H_1(w_1|h)}{H_2(w_2w_3|h)} = \frac{p(w_1|w_L) \times p(w_R|w_1)}{p(w_2|w_L) \times p(w_3|w_2) \times p(w_R|w_3)} \approx \frac{1}{p(w_2|w_L)}$$

即不切分的概率反而高于要切分，即使w_2是正确的。

经过实验统计，绝大多数的错误就来源此类需要三分的问题。本文决定在二分完毕后，再对每个词长超过一定阈值（设为 3）且H_1接近 0 的w_1考察是否需要三分为w_2、w_3、w_4。即新假设：

$$H_5(w_2w_3w_4|h) = p(w_2|w_L) \times p(w_3|w_2) \times p(w_4|w_3) \times p(w_R|w_4) \quad (15)$$

为了计算快捷，仅将最大的H_5与H_1比较，衡量是否需要三分。

4 实验部分

本文准备了两份语料——A：2000 年 12 月《人民日报》部分（总词数：88435）；B：1998 年 1 月《人民日报》（以 A 为参照，已登录词数：785056，未登录词数：163387）。A 和 B 的分词效果差异用于衡量文本大小对无指导分词的影响。A 还在半指导分词实验中作为已标注训练集。

在 4.1 节中进行了 3 种实验：

一、 进行文献[8]的 HDP 无指导分词作为 baseline；

二、 进行 3.1 和 3.2 节的改进后的 HDP 无指导分词；

三、 改进后的 HDP 半指导分词，并与常用有指导分词对比。

实验三的有指导分词基于开源工具 CRF++，并直接使用工具包内示例的分词模板。

由于标点符号必然需要切分，并且可以轻易识别，本文决定在评价时**删除标点**。用准确率、召回率和 F 值衡量分词效果，计算公式如下：

$$\text{准确率} = \frac{\text{正确分词总数}}{\text{系统分词总数}}$$
$$\text{召回率} = \frac{\text{正确分词总数}}{\text{答案分词总数}}$$

$$\text{F 值} = \frac{2 \times \text{准确率} \times \text{召回率}}{\text{准确率} + \text{召回率}}$$

实验结果表明经过改进后的 HDP 无指导分词在执行时间以及 F 值上相比原方法均有明显改善。加入较小训练语料后对已登录词的识别还有很大提升。

4.1.1 原方法

分别对语料 A、B 进行文献[8]的 HDP 分词。实验结果如下：

表 4.1：原方法 Gibbs Sampling 收敛结果表

语料	收敛迭代次数	迭代速度/ $s \cdot 轮^{-1}$	总耗时/h	准确率	召回率	F 值
A	5000	0.8	1.1	0.615	0.701	0.655
B	12000	9.2	30.7	0.709	0.717	0.713

语料 B 收敛总耗时超过 30h，说明本 Gibbs Sampling 的自相关作用（autocorrelation）很大，并不是一种高效的算法。

4.1.2 改进后的无指导和半指导方法

进行 3.1 和 3.2 节的改进后的实验结果如下表所示。其中“A+B”代表以 A 为已标注语料对 B 进行的半指导分词，用于同无指导分词对比。

表 4.2：改进方法 Gibbs Sampling 收敛结果表

语料	收敛迭代次数	迭代速度/ $s \cdot 轮^{-1}$	总耗时/h	准确率	召回率	F 值
A	1000	1.0	0.3	0.663	0.726	0.693
B	3000	12.3	10.3	0.740	0.742	0.741
A+B	3000	12.3	10.3	0.808	0.810	0.809

可见，通过改进，分词的 F 值和切分速度均有明显改进。语料 B 收敛总耗时 10h，仅为原来三分之一。加入较小的已标注语料后，总体分词效果还有较大提升。

用 A 作为训练语料，对 B 测试的 CRF 分词结果如下表：

表 4.3：CRF 有指导分词结果表

训练耗时/min	测试耗时/min	准确率	召回率	F 值
12	1	0.779	0.788	0.783

同半指导的 HDP 不同，CRF 的时间主要消耗在训练模型上，在测试时不需要自举迭代。由于本实验训练语料规模远小于测试语料，因而 CRF 的耗时自然远小于 HDP。但也正因为训练语料的不足，纯有指导分词又无法利用测试语料的信息，CRF 分词结果不如 HDP。

分析半指导分词的数据，结果见下表：

表 4.3：半指导分词结果分析表

类别	已登录词数	未登录词数	已登录正确数	未登录正确数
标准答案	785056	162818	/	/
无指导结果	713812	236282	620274	83087
半指导结果	779627	170249	685791	81949

从表中可知，虽然 A 的语料规模不到 B 的 10%，但 B 相对于 A 的未登录词仅有 17%，说明两个文档相似程度很高，半指导方法的优越性并未很好体现。从无指导结果看，这些已登录词大部分也是高频词，能够被正确识别。对比无指导和半指导结果。我们发现：已登录词数之差 ≈ 已登录正确数之差。F 值的差别就体现在这 6.6 万个被正确识别切分的已登录词。数据显示半指导分词对未登录词的切分改进效果有限。

对半指导分词，$\frac{\text{已登录词正确数}}{\text{已登录词总数}}$可以衡量歧义切分能力。

4.2 误差分析

实验求得的无指导分词 F 值 0.741 与半指导分词 F 值 0.809 均不是特别高。造成误差的可能原因有：

1、分词的规则。人工分词的结果和机器分词的结果有一些内在的不同。人往往会考虑词性的区分，比如“了的”必然不会划分成一个词。而机器只看统计结果，由于“了的”往往共同出现，因此易认为是一个词。这类情况普遍存在，比如人名。标准答案中人名姓和名是分开的但对机器而言常见人名被分成一个词的概率很高。

2、长词的切分错误。（14）的词长先验分布和（15）的三分假设都基于“汉语倾向于生成长度为 1 或 2 的词”该假设。从实际统计看，该假设是合理的。不考虑标点，两者之和能占总词数的 90%。但这两点必然造成分词程序输出长词的概率大大降低。语料 B 标准答案、无指导分词、半指导分词的具体词长分布情况如下：

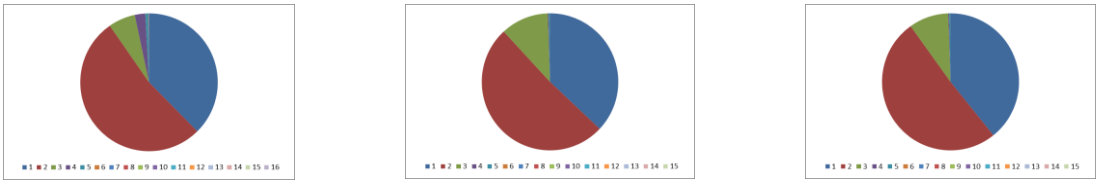

图 4.3：语料 B 的词长分布图。左：标准答案。中：无指导。右：半指导

从图上可以看出，由于 3.2 节设阈值为 3，采样过程中长度大于 3 的词基本完全被切分为较短的词。虽然有利于整体 F 值的提升，但长词错误率增大了不少。可能需要更加合理的权衡方法。

3、低频词。由于本实验是无监督的，字串的切分与否完全取决于两种状态的出现概率比。但低频词成为一个词的概率很低，因此会倾向于划分成单字。语料 A 的分词结果，一元词比例比答案多了 12%，二元词比例比答案少了 11%。可以看到由于 A 的语料规模远小于 B，其中的低频词比例要高得多。相当部分的二元词被错误切分成一元词，影响了准确率。因此，要使无指导分词准确，语料规模必须足够大。

5 结论与进一步工作

本文提出了 HDP 与互信息相结合的分词方法，并改进了采样算法。实验表明：本方法的切分速度、切分准确率同 baseline 的 HDP 相比均有较大改进。加入较小的训练语料后，总体 F 值还有较大提升。

我们的下一步工作是引入演化的 HDP 模型[11, 12]，同时实现不同日期不同领域文本的新词、领域词与新搭配发现。

参考文献:

[1] SPROAT, RICHARD, SHIH C. A statistical method for finding word boundaries in Chinese text [J]. Computer Processing of Chinese and Oriental Languages, 1990, 4(336-51.

[2] 黄萱菁, 吴立德. 基于机器学习的无需人工编制词典的切词系统 [J]. 模式识别与人工智能, 1996, 9(4): 297-303.

[3] MAOSONG S, DAYANG S, TSOU B K. Chinese word segmentation without using lexicon and hand-crafted training data; proceedings of the Proceedings of the 17th international conference on Computational linguistics-Volume 2, F, 1998 [C]. Association for Computational Linguistics.

[4] 刘挺, 吴岩, 王开铸. 串频统计和词形匹配相结合的汉语自动分词系统 ① [J]. 1998,

[5] 孙茂松, 肖明, 邹嘉彦. 基于无指导学习策略的无词表条件下的汉语自动分词 [J]. 计算机学报, 2004, 27(6): 736-42.

[6] PITMAN J, YOR M. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator [J]. The Annals of Probability, 1997, 25(2): 855-900.

[7] GOLDWATER S, GRIFFITHS T L, JOHNSON M. Contextual dependencies in unsupervised word segmentation; proceedings of the Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, F, 2006 [C]. Association for Computational Linguistics.

[8] GOLDWATER S, GRIFFITHS T L, JOHNSON M. A Bayesian framework for word segmentation: Exploring the effects of context [J]. Cognition, 2009, 112(1): 21-54.

[9] TEH Y W. A hierarchical Bayesian language model based on Pitman-Yor processes; proceedings of the Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, F, 2006 [C]. Association for Computational Linguistics.

[10] WOOD F, TEH Y W. A hierarchical, hierarchical Pitman-Yor process language model; proceedings of the ICML 2008 Workshop on Nonparametric Bayes, F, 2008 [C].

[11] XU T, ZHANG Z, YU P S, et al. Evolutionary clustering by hierarchical dirichlet process with hidden markov state; proceedings of the Data Mining, 2008 ICDM'08 Eighth IEEE International Conference on, F, 2008 [C]. IEEE.

[12] ZHANG J, SONG Y, ZHANG C, et al. Evolutionary hierarchical dirichlet processes for multiple correlated time-varying corpora; proceedings of the Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, F, 2010 [C]. ACM.