一种自适应概率语言模型的训练方法及其应用于中文分词*

徐志明#* 揭春雨* Jonathan Webster#

#香港城市大学 中文、翻译及语言学系

*哈尔滨工业大学 计算机学院

E-mail: {ctxuzm, ctckit, enjjw}@cityu.edu.hk

摘 要：本文提出一种自适应的概率语言模型的训练方法，采用 EM 迭代优化算法在未切分的语料库上训练概率语言模型的参数。本文用该算法训练了中文的基于词的 N-gram 模型，并应用于概率分词。实验结果显示，该算法能显著地改善汉语分词的精度。

关键词：语言模型，词网格，EM 算法，概率分词

An Adaptive Training Algorithm for Probabilistic Language Model and its Application to Chinese Word Segmentation

Zhiming Xu#* Chunyu Kit# Jonathan Webster#

#Department of Chinese, Translation & Linguistics, City University of Hong Kong

* School of Computer Science & Technology, Harbin Institute of Technology

E-mail: {ctxuzm, ctckit, enjjw}@cityu.edu.hk

Abstract: This paper presents an adaptive training algorithm for probabilistic language models using unsegmented corpus based on the EM algorithm. It is applied to train an N-gram model for probabilistic Chinese word segmentation. Experimental results show that it can improve word segmentation performance significantly.

Keywords: Language model, word lattice, EM algorithm, probabilistic word segmentation

1 引言

概率语言模型一般用来估计一个词序列的出现概率，广泛应用于自然语言处理各领域[1]，例如语音识别[6]。本文重点讨论基于词的 N-gram 模型参数的训练方法，及其应用于中文概

*本文受香港大学教育资助委员会（UGC）研究资助局(RGC)角逐研究用途补助金(CERG)项目 "Example-based Machine Translation (EBMT) for Legal Texts (基于实例的法律文本的机器翻译)"（项目号#9040482）的资助。

率分词。一般来说，基于词的 N-gram 模型需要在切分好的大规模语料库上进行训练，以获得可靠的概率参数。但是，大规模的切分好的语料库涉及大量的人工劳动，不易获得。若是没有切分好的语料库，概率语言模型的训练和概率分词，就不可避免地形成一个典型的“鸡蛋问题”（“chick-and-eggs”）：构造基于词的 N-gram 模型要求有切分好的训练语料库；而概率分词，需要基于词的 N-gram 模型提供可靠的概率参数来处理切分歧义。

为了解决这一问题，研究人员为概率语言模型提出了一些自适应的训练方法，并应用于概率分词，新词识别等任务[2][3] [7][8]。大多数的研究人员采用了 EM 迭代优化算法[4]在未切分的语料库上训练分词用的语言模型参数，也有些学者使用 EM 算法研究无词典的自动分词方法[5]。一般很少涉及高元（$N \geq 2$）N-gram 的自动训练问题。根据信息理论，随着 N 的增长，N-gram 模型的熵值逐渐下降，使得语言模型的描述能力不断增强。与低元的 N-gram 相比，高元的 N-gram 可以提供更可靠的概率估计，有助于提高概率分词的精确度。

本文提出了一种自适应的汉语概率语言模型的训练方法，采用 EM 算法在未切分的汉语生语料上训练语言模型的参数，用来指导汉语的概率分词问题。实验结果表明，这种训练方法显著提高概率分词的精度。

本文以下各节组织如下：第 2 节讨论概率分词和 N-gram 模型；第 3 节详细描述概率分词的状态空间——词网格；第 4 节讨论自适应的语言模型的训练方法；第 5 节介绍词典和概率参数库的结构，以及试验结果；最后一节总结研究成果。

2 汉语概率分词和 N-gram 模型

给定一个字串 $c = c_1 c_2 \cdots c_n = c_1^n$，将它切分成一个词序列 $s = w_1 w_2 \cdots w_m = w_1^m$，概率分词的任务是在众多的切分候选中选择一个最佳的切分，即概率最大的切分：

$$\hat{s} = \arg\max_{s} p(w_1^m) = \prod_{i=1}^{m} p(w_i \mid w_1^{i-1}) \tag{1}$$

由于数据稀疏问题，一般限制上下文 w_1^{i-1} 为固定小个数的 N-1 个词，并由相关词序列频度估算(1)中的条件概率，定义如下：

$$p(w_i \mid w_1^{i-1}) = \frac{C(w_1^i)}{C(w_1^{i-1})} \cong \frac{C(w_{i-N+1}^i)}{C(w_{i-N+1}^{i-1})} \tag{2}$$

$C(x)$为词序列 x 在训练集中出现的频度，一般记入一个参数库，用于计算条件概率，而该词序列则称为一个 N-gram 项。

3 词网格和概率分词

给定一个未切分的汉语句子 $c_1 c_2 \ldots c_n$，通过词典查询，我们可以找出其中所有的词候选（词形），为每个词候选生成一个节点（词候选节点）。根据节点中词候选的长度，节点可分成单字词节点和多字词节点。所有的词候选节点分成列，每个列号定义为该词候选的尾

字在句中的位置。词候选的上下文可根据其前向邻接关系确定。例如，由单字词 c_i 构成的节点被挂在第 i 列，它的前邻接列为 i-1。多字词 $c_i...c_j$ 构成的节点被挂在 j 列，它的前邻接列为 j-1。将所有具有前邻接关系的节点用有向边连接起来，我们可得到一个有向图，称为成为词网格，词网格可以被看作概率分词问题的状态空间。词网格中的每条路经（自左向右），是句子的一个切分候选。例如，给定字序列"香港特别行政区成立"，其词网格如图 1 所示。

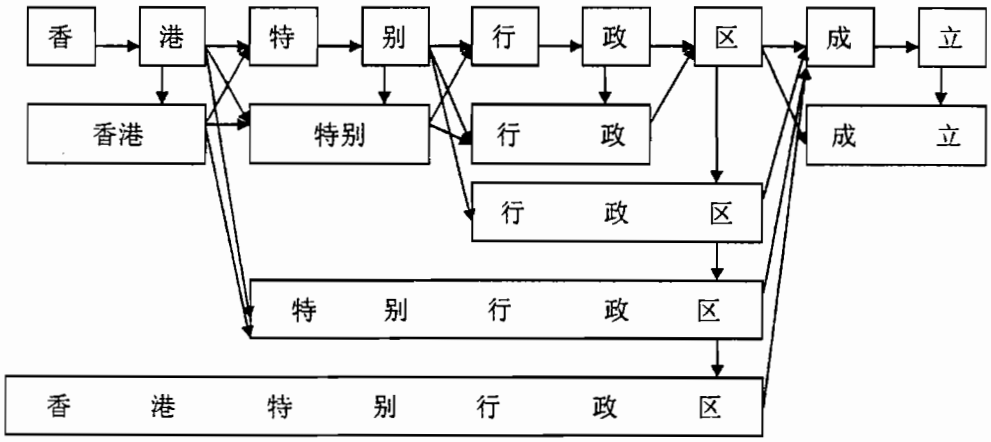

图 1：词网格例示

4 自适应的 N-gram 模型的训练算法

根据 EM 算法的定义[4]，我们可以得到 N-gram 模型参数的迭代训练公式如下：

$$p^{k+1}(w_{i-N+1}^{i}) = \frac{\sum_{T}\sum_{s} C_s(w_{i-N+1}^{i}) \cdot p^{k}(s)}{\sum_{w_i}\sum_{T}\sum_{s} C_s(w_{i-N+1}^{i}) \cdot p^{k}(s)} \tag{3}$$

等式左边是新一轮的 N-gram 模型参数，根据上一轮的模型参数计算出来，其中，k 是循环次数，T 是训练语料库，s 是当前句子的一个切分候选，$C_s(x)$是一个词序列 x 在 s 中出现的次数。本文采用 EM 算法作为自适应的 N-gram 模型的训练算法，分成两个阶段：（1）模型参数的初始化（2）模型参数的迭代训练。

4.1 模型参数初始化

在模型参数的初始化阶段，给定一个输入句子 $c_1c_2...c_n$，一般可以假设任意一个切分候选 $w_1w_2...w_m$ 是等概率出现的。根据这种观点，如果输入句子一共有 K 个切分候选，则每个的出现概率为 $1/K$，词序列的记数应该为

$$\frac{1}{K}\sum_{s} C_s(w_{i-N+1}^{i}) \tag{4}$$

但是，本文发现上述初始化阶段的参数估计存在下列问题：（1）该方法需要枚举词网格中的所有的切分候选，其时间复杂度为 $O(k_1k_2...k_n)$，其中 k_i 为第 i 列词候选节点的数目；（2）该方法对词网格中一个词序列的记数过度繁复。我们的初始化方法是在词网格中设置一个尺寸为 N 的滑动的“窗口”，对应着词网格的 N 列，我们对窗口内每个词序列记数，然后向前滑动窗口并计数，直到句子结束。这样的初始化参数算法的时间复杂度为 $O(k_1k_2...k_N+k_2k_3...k_{N+1}+...+k_{n-N+1}k_{n-1}k_n)$，远小于前一方法。对于某一词序列，假设其出现在第 l 列到第 l+N-1 列之间的窗口内的次数为 C，如果用前一方法对它直接计数，则要做 C $k_1k_2k_3...k_{l-1}k_{l+N}...k_n$ 次计数，因为有这么多条路径经过它，复杂度非常高。与此相比，我们的计数更近似于普通的 N-gram 语言模型的构造过程，更加自然。试验中，参数训练过程的快速收敛性和概率分词的精度，也验证了这种初始化参数方法的合理性。

4.2 模型参数的迭代训练

对于输入的每个句子，我们采用 Viterbi 算法在词网格中搜索概率最大的切分候选，一旦得到概率最大的切分候选，我们就沿着这条路径回溯，并对出现在该路径上所有的相关词序列重新计数，记入到新的参数库中。该迭代训练算法的时间复杂度等于 Viterbi 算法的时间复杂度。在参数初始化阶段，我们相当于得到了一个近似的 N-gram 模型，本文称之为基于词形的 N-gram 模型。该模型参数为词形的 N-gram 在切分候选中的出现次数，而不是在正确的切分候选中出现的次数，经过不断的迭代训练，使得模型逐渐逼近于真实的基于词的 N-gram 模型。

4.3 模型参数训练算法的其他策略

在模型参数初始化阶段，我们采用了递进式统计策略和归并策略来加快参数的初始化参数过程。我们将训练文本切割成固定大小的一组子文件，对它们分别进行初始化参数统计，把结果归并到一个参数库中。这种方法明显加快模型参数的初始化参数过程。在模型参数训练阶段，我们采用的递进式统计策略，也同样降低了算法的内存损耗，加快了训练过程。

5 实验结果

我们将自适应训练的基于词的 Bigram 模型算法，应用到中文的概率分词问题中，通过概率分词的准确率来验证模型的优劣和训练算法的性能。我们采用 1994 年的人民日报作为训练语料库（40M 字节）。我们自有的中文词典收录 54961 词，称之为自有词典。我们采用北京大学标注语料库（8.6M 字节）作为正确的切分标本，然后去掉所有的切分标记，生成

本文的测试语料库，应用我们的概率分词模型进行切分，其切分结果和正确的切分标本自动对照，以得到准确率结果。另外，本文根据北京大学标注语料库，单独抽取其词典，这里称之为在线词典。

表 1：应用自适应训练的基于词的 N-gram 模型进行中文概率分词的实验结果

词典使用	测试方式	切分正确率
在线词典	封闭式	97.96%
在线词典+自有词典	封闭式	94.27%
自有词典	开放式	87.07%

试验观察发现，作为切分标本的北京大学标注语料库存在着许多不一致的切分。由于我们假设北京大学标注语料的切分结果是唯一正确的，大量的正确结果被误判，因此，表 1 中的准确率实际上是明显低估的。按经验估计，表 1 中的三个试验的真正准确率大概会比所显示的数据高出 2%-6%左右，但是，仍需要进一步探讨确认。另外，使用混合词典（在线词典+自有词典）的切分结果低于单独使用在线词典，这是因为自有词典中的很多词汇超出了在线词典的内容，因此加大了切分结果的差异，使得被误判的切分结果更多。而在开放式试验中的大部分错误来自大量的未登录词（特别是专用名词，如人名，地名，时间，数字等等）。由于本文的概率分词算法尚未对时间表达式和数字表达式，以及人名和地名进行识别，因此错误率较高，我们将在后续研究中继续探讨上述专有名词的处理。

实验结果表明，我们提出的训练算法有如下特点：（1）针对交叉型歧义(overlapping ambiguities)，具有很强的切分消歧能力；（2）针对组合型歧义（Combinational ambiguities），它会自动合并过度切分的词串；（3）一旦切分错误被纠正之后，不会再改错；（4）快速收敛，和达到相当高的切分精度；（5）一旦初始化参数之后，即使在比较小的训练集上，该算法也能取得很高的切分精度。这也验证了初始化方法的合理性和该算法的可拆卸性（portability）。

6 结论

本文提出了一种自适应的基于词的 N-gram 模型的参数训练方法，该算法采用 EM 迭代优化算法进行模型参数训练，并使用词网格作为训练算法的状态空间。本文提出了一种新的模型参数初始化方法。同惯常的方法相比，它提供了更合理的初始值，同时也降低了算法的计算法时间复杂度。另外，在模型参数初始化阶段，我们采用了递进式统计和归并策略，则进一步降低了算法的时间复杂度。在模型的参数训练过程，我们也采用了递进式统计策略，同样加快了参数训练过程。训练后的模型参数应用于汉语的概率分词，取得了很高的切分精度。除了快速收敛和极高的切分精度之外，该算法还具有自适应，可以适用于不同的语言。

參 考 文 獻

[1] Chen, Stanley F. 1996. Building Probabilistic Models for Natural Language, Ph.D. thesis, Harvard University.

[2] Chang, J.-S., and K-Y. Su. 1997. An unsupervised iterative method for Chinese new lexicon extraction. *International Journal of Computational Linguistics & Chinese Language Processing*, **1**(1):101-157.

[3] Chen, K.-J., and M.-H. Bai. 1998. Unknown word detection for Chinese by a corpus-based learning method. *Computational Linguistics and Chinese Language Processing*, **3**(1):27-44.

[4] Dempster, A., N. Laird, and D. Rubin. 1997. Maximum-likelihood from incomplete data via the EM algorithm. *J. Royal Statistical Society, Series B*, **39**:1-38.

[5] Ge, X., W. Pratt, and P. Smyth. 1999. Discovering Chinese words from unsegmented text. SIGIR-99, pp.271-272,

[6] Jelinek, Frederick. 1997. *Statistical Methods for Speech Recognition*. The MIT Press.

[7] Sproat, R., C. Shih, W. Gale, and N. Chang. 1996. A stochastic finite-state word segmentation algorithm for Chinese. *Computational Linguistics* **22**(3):377-404.

[8] Sproat, R., and C. Shih. 1990. A Statistical Method for Finding Word Boundaries in Chinese Text. *Computer Processing of Chinese and Oriental Languages*, **4**(4):336-351.