第 23 卷 第 11 期 电子与信息学报 Vol.23 No.11
2001 年 11 月 JOURNAL OF ELECTRONICS AND INFORMATION TECHNOLOGY Nov. 2001

一种基于 N-gram 模型和机器学习的汉语分词算法 [1]

吴应良 韦 岗* 李海洲*

(华南理工大学工商管理学院 广州 510641)
*(华南理工大学电子与通信工程系 广州 510641)

摘 要 汉语的自动分词，是计算机中文信息处理领域中一个基础而困难的课题．该文提出了一种将汉语文本句子切分成词的新方法，这种方法以 N-gram 模型为基础，并结合有效的 Viterbi 搜索算法来实现汉语句子的切词．由于采用了基于机器学习的自组词算法，无需人工编制领域词典．该文还讨论了评价分词算法的两个定量指标，即查准率和查全率的定义，在此基础上，用封闭语料库和开放语料库对该文提出的汉语分词模型进行了实验测试，表明该模型和算法具有较高的查准率和查全率．

关键词 汉语分词，N-gram 模型，机器学习，查准率，查全率

中图号 TN-051.1

1 引 言

汉语与英语不同，英语文本是小字符集上的词串，而汉语文本是大字符集上的字串．将汉语字串切分成词串，就是汉语分词 [1] 所要完成的任务．汉语分词又是各种中文信息处理应用系统中共同的、基础性的工作，例如：语音识别、字符识别、语音合成、文本较对、信息检索和机器翻译等。

汉语中的汉字是一种形、声、义相结合的方块字，由汉字组成的词或词组没有形态的变化。在文本中，词或词组间没有分隔标记，词的界定缺乏自然标准，人们是靠语义从一个汉字字符串中辨别出词和词组的 [2] 。另外，不同的应用目标对词需求不同．这就决定了分词规范的两难问题。而分词算法还存在切分歧义排除和未登录词的识别两方面的困难。大多数传统的分词方法采用词典和基于规则的启发式搜索 (Heuristic Searches) 算法．但是，基于传统分词方法的系统词典覆盖率有限，而且对具体的应用领域变化的适应性较差．另外，生词 (包括未知词、超出词汇表的词) 的处理，对汉语的语法分析是一个实质性的问题。

在英语语法分析方面，建立了采用标注语料库的概率性技术 [3] ，这些技术采用基于词方法，因为英语自然文本中的词由空格分隔开。因此，未知词的存在对分词结果没有严重影响．而在汉语的处理中，如果采用基于词的方法，则存在两方面的问题：一方面，未知词的存在会对分词结果产生严重的影响；另一方面，汉语的切分存在较大的歧义性 (Segmentation Ambiguity)，即对同一切分对象，可能具有多种切分结果，即候选词．而这些候选词的长度不同，需要我们对它们进行比较和判别．如果采用基于词 (Word-Based) 的模型，我们必须规整这些候选词的概率。

在英语词法分析方面，已有许多使用统计语言模型、具有高精度的方法报道 [3,4] 。近来提出了一些基于统计的词法分析方法，其中，采用的标注模型有二元文法 (Bigram)[5] ，基于字符的隐马尔可夫模型 (HMM)，词三元文法 (Word Trigram) 和字符三元文法 [6] 等．相比之下，运用随机方法的汉语词法分析成果报道则较少．大多数传统的包括分词的汉语词法分析使用基于规则的启发式搜索方法，这些方法使用一种连接矩阵作为语言模型，例如最大匹配法 LMM 或者最少子句法 LBNM(Least Bunsetu's Number Method) 等。

为了克服传统的基于规则的分词方法的不足，同时充分利用汉语的特质，我们提出一个具有高查准率的分词算法，该算法采用一种基于字符的统计语言模型—— N-gram 模型来构造汉语

[1] 1999-09-29 收到，2000-04-06 定稿
广东省自然科学基金资助课题

分词模型，同时结合基于机器学习的自组词算法 [7,8] 来实现汉语文本的自动分词。它具有如下优点：

(1) 汉字集是一个闭合的系统，其数量有限。汉语字符集中字符数较小，虽然汉字总数超过 6 万，但 GB2312-80 中的 6763 个汉字，覆盖率在 99.99% 以上，而中文词则以千万计。采用基于字符的 N-gram 模型，系统开销较小，处理较简单、速度快，且易于实现。

(2) 平均词长大约为两个字符。于是，一个字符可认为具有与某个词的关联信息。

(3) 基于字符方法无需人工编制领域词典 (人名和地名除外)，因此，我们不必担心未知词问题。

(4) 对于一个给定的句子，它包含的字符数是恒定的，我们在比较分词候选词时无需规一化其概率。

(5) 机器可读语料丰富，容易获取，因此可处理大规模的真实语料，并可较充分地获得和利用反映语言特点的统计信息，从而提高分词系统的性能。

2 基于字符的汉语 N-gram 模型

本文采用基于字符的 N-gram 作为语言模型，即将语言中字符的发生近似为 $(n-1)$ 阶 Markov 模型。也就是说，设有 l 个字符的汉字字符串 $c_1, c_2, \cdots, c_l$，在其上下文关系中，只有前 $n-1$ 个字符对下一个字符即第 n 个字符出现的概率有影响，用概率表示就是：

$$P(c_l|c_1, \cdots, c_{l-1}) \approx P(c_l|c_{l-n+1}, \cdots, c_{l-1}) \tag{1}$$

根据概率乘法定理和 N-gram 模型，汉字字符串 $c_1, \cdots, c_l$ 的概率可表示为组成该字符串的字符的概率的乘积：

$$P(c_l, c_2, \cdots, c_l) = \prod_{i=1}^{l} P(c_i|c_1, \cdots, c_{i-1}) \approx \prod_{i=1}^{l} P(c_i|c_{i-n+1}, \cdots, c_{i-1}) \tag{2}$$

N-gram 模型的参数可以根据字符串在训练语料库出现的频率来估计得到：

$$P(c_l|c_{l-n+1}, \cdots, c_{l-1}) = \frac{C(c_{l-n+1}, \cdots, c_l)}{C(c_{l-n+1}, \cdots, c_{l-1})} \tag{3}$$

这里函数 $C(c_1, \cdots, c_l)$ 用于计算其自变量中字符串的数量。

一般地说，高阶模型能更好地刻划语言的结构，但同时也有不足。因为从给定的语料库所能获得的有关数据是有限的，而且很多字符串在语料库中出现次数很少，或者根本不会出现。于是就会出现这样一种结果：高阶模型仅能给出那些与训练用数据近似的字符串的合理概率，而不能得到那些与训练用数据不相近的字符串的合理概率，即所谓的数据稀疏问题 [4,9]。另外，高阶模型需要的计算量、存储空间都较大，实现较困难。因此在实际处理中，取 n=2 或 n=3 的模型即可满足许多实际应用，这时，N-gram 模型分别被称为二元文法 (Bigram) 模型和三元文法 (Trigram) 模型。

为了克服数据稀疏问题，我们采用基于 Good-Turing 估计法的补偿平滑技术 [6,9]。根据 Good-Turing 估计法，对任何发生 r 次的 N-gram 文法，我们假定它会发生 r^* 次，这里

$$r^* = (r+1)\frac{n_{r+1}}{n_r} \tag{4}$$

式中 n_r 为在训练数据中精确出现 r 次 N-gram 的次数。为将该数转换成概率，我们只需作如下的规一化：设字符 c 在语料库中总共出现 r 次，我们取

$$P(c) = r^*/N \tag{5}$$

其中 N 为语料库所包含的字符的总数，定义比率 $d_r = r^*/r$ 为贴现系数 (Discount Coefficient)。

特别地，对于 Bigram 模型，如果字符串 c_1c_2 出现在语料库中，运用 Good-Turing 估计法，条件概率 $P(c_2|c_1)$ 可以表示如下：

$$P(c_2|c_1) = \frac{C^*(c_1c_2)}{C(c_1)} = \frac{C^*(c_1c_2)}{C(c_1c_2)}\frac{C(c_1c_2)}{C(c_1)} = d_{C(c_1c_2)}\frac{C(c_1c_2)}{C(c_1)} \tag{6}$$

3 分词算法

我们提出的分词算法由两个步骤组成，即首先通过对语料库进行基于字符统计 Bigram 统计分析和计算，生成单字同现频度库，并由该库自组织生成分词词典库；然后用 Viterbi 算法进行汉语文本的切分。

3.1 自组词算法

采用基于机器学习的自组词算法的目的，是从语料库中自动、动态地生成分词词典，这样的分词词典能较好地反映分词对象领域的特有词汇；另一方面，词典的生成过程独立于领域，生成过程中产生的详细统计信息可供后处理利用。

本文中的自组词算法采用基于字符 (对汉字来说就是单字) 统计频度自组词算法 [7,8]，这种方法只需统计单字同现频度，系统可根据单字同现频度库来自组织生成所对应的词，进而生成分词词库。因此，它具有占用系统存储空间小、处理效率高、实现比较容易，同时通过利用频度库中词的信息所隐含的汉语语言知识 (如语法、句法等) 进一步提高查准率。

3.2 分词算法

在 N-gram 模型中，在考虑某个汉字串 $c_1c_2\cdots c_n$ 时，第 n 个字符 c_n 的出现只与其前面的 $n-1(n \geq 1)$ 个字符有关。如果它前面的字符串包含切分定界符 <d>，则它在该位置之前已被切分。下面将 <s> 和 < /s> 分别用作句子的开始标记与结束标记，图 1 给出了有关切分实例。

<s> 自然语言处理 <d> 是 <d> 知识 <d> 信息处理 <d> 中 <d> 的 <d> 核心课题 <d>

图 1 训练语料切分举例

与每个字符位置相关联的状态有两种，记为 S 和 NO-S， S 表示某个切分的开始位置，而 NO-S 则相反。一种前向计算给出在字符串中每个位置的可能性 (其中 $k \geq 2$)：

$$\begin{aligned} P_{\text{NO-S}}(c_1\cdots c_k) = \max(&P_{\text{NO-S}}(c_1\cdots c_{k-1}) \times p(c_k|c_{k-2}c_{k-1}) \\ &\times P_{\text{S}}(c_1\cdots c_{k-1}) \times p(c_k| < \text{d} > c_{k-1})) \end{aligned} \tag{7}$$

$$\begin{aligned} P_{\text{S}}(c_1\cdots c_k) = \max(&P_{\text{NO-S}}(c_1\cdots c_{k-1}) \times p(< \text{d} > |c_{k-2}c_{k-1})p(c_k|c_{k-1} < \text{d} >), \\ &P_{\text{S}}(c_1\cdots c_{k-1}) \times p(< \text{d} > | < \text{d} > c_{k-1})p(c_k|c_{k-1} < \text{d} >)) \end{aligned} \tag{8}$$

其中

$$P_{\text{NO-S}}(c_1) = p(c_1 | < \text{s} >) \tag{9}$$

$$P_{\text{S}}(c_1) = 0 \tag{10}$$

例如，给定语句 $S = c_1 \cdots c_m$(如图 2)，用 Viterbi 算法进行最优路径搜索，即发现由 S 和 NO-S 组成的序列。因为在句子结束符 < /s> 前没有定界符 <d>，所以最大似然路径从 NO-S 开始。

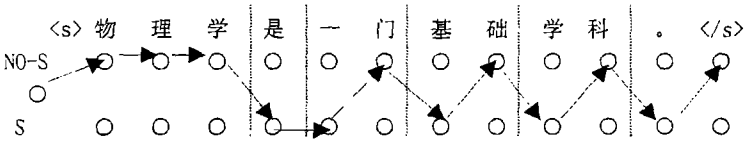

图 2　采用 Viterbi 算法对句子的切分实例

4 算法性能分析和讨论

因为汉语是一种粘结性 (Agglutinative) 的语言，因此要一致性地确定词的切分界限是困难的。为了定量地分析和表达系统切分性能，我们将英语语法分析的有关评价标准应用于汉语的词法分析，并进行了初步的实验。

下面给出分词性能的两个指标——查全率 R(Recall) 和查准率 P(Precision) 的定义。为此，设语料库中待切分词总数为 S_{cop}，系统切分的词总数为 S_{sum}，系统切分的匹配词数为 M，则查全率和查准率分别定义为：

$$\text{R} = M/S_{\text{cop}} \tag{11}$$

$$\text{P} = M/S_{\text{sum}} \tag{12}$$

我们采用 ATR 对话数据库 (ADD) 对我们提出的分词模型进行训练和测试，该语料库是用于会议注册领域键盘对话所产生资源库的一部分。作为训练数据，我们使用了具有 10000 条语句的 ADD 语料库，用 <d> 作为切分标志符，<s> 和 <s/ > 分别作为句子的开始标志符和结束标志符。训练语料的 N-gram 概率采用基于补偿平滑方法的 Carnegie Mellon 统计语言模型 (CMU SLM) 工具包进行估计 [10]。

我们对训练语料进行了封闭和开放两种测试。每种测试集的句子、词和字符的数量如表 1 所示。

表 1　测试数据情况表

组成	种类	
	封闭语料	开放语料
句子数	10,000	2679
词数	141,658	31742
字符数	573,456	135923

表 2 列出了该分词模型和算法在封闭和开放两种情况下语句的切分性能，即查全率和查准率。

表 2 分词模型性能

语料	无自组词算法		本文算法	
	查全率 R	查准率 P	查全率 R	查准率 P
封闭文本	94.43%	93.32%	97.68%	96.76%
开放文本	93.72%	92.12%	98.30%	96.79%

从表 2 可以看出，本文提出的分词模型和算法在上述分词领域中，测试语料为开放文本时的切分准确性高于测试语料为封闭文本的情况，并具有较高的查全率和查准率。这是因为基于 N-gram 模型的自组词汉语分词方法首先通过机器对生语料库的训练学习自动生成分词词典，能大幅度地减少词互扰和高频锐化现象所造成的干扰。

由于自组词分词算法生成了词，所以下一步的分词可以完全利用所有针对基于词库的分词算法，因此本文提出的分词算法具有很好的开放性，并且使得分词算法仅仅使用了单词频度库的空间而可同时获得基于词库的分词效果。

在实验中，我们发现大部分的切分错误来源于大词库所带来的切分歧义。可以通过改进 Viterbi 算法来减少切分错误。

对于人名、地名等特定领域的词切分问题，可以利用针对这些领域的专用词典与我们的自组词方法相结合，可以大幅度提高对真实文本中人名、地名等专有词汇的切分的准确率。

5 结语和以后的工作

本文提出了一种新的汉语分词方法，这种方法采用统计语言模型和一种 Viterbi 算法，通过机器学习的自组词算法构造分词领域词典。用 ADD 语料库的模拟实验证明了这种方法的有效性。为了提高模型的鲁棒性，今后将进一步研究基于聚类算法 (Clustering Algorithm) 的分词模型，这样可进一步利用词的有关知识，将字和词分成类，有利于提高分词系统的性能。另外，如何将人工神经网络 (ANN)、粗集 (Rough Sets)、模糊集 (Fuzzy Sets) 等智能信息处理理论与方法应用于分词统计语言模型的构造、优化和实现，都是有待探索的领域。

参 考 文 献

[1] 梁南元，汉语计算机自动分词知识，中文信息学报，1989, 4(2), 29–33.

[2] 王德春，应用语言学概论，上海，上海外语教育出版社，1997 年 12 月第 1 版，88–120.

[3] E. Charniak, C. Hendrickson, N. Jacoboson, M. Perkowitz, Equations for part-of speech tagging, AAAI-93, 1993, 784–789.

[4] K. Church, A stochastic parts program and noun phrase parser for unrestricted text, ANLP-88, 1998, 136–143.

[5] S. Sakai, Morphological category bigram: A single language model for both spoken language and text, ISSD-93, 1993, 97–90.

[6] M. Yamamoto, A re-estimation method for stochastic language modeling from ambigous observations, in Proceeding of WVLC-96, California, 1996, 155–167.

[7] 赵以宝，孙圣和，一种基于单字统计二元文法的自组词音字转换算法，电子学报，1998, 26(10), 55–58.

[8] F. Jelinek, Self-Organized Language Modeling for Speech Recognition, IBM Research Report, IBM T, J. Watson Research Center, 1985. Reprinted in Reading in Speech Recognition, Waibel, A., and Lee, K-F. (Eds.), Morgan Kaufann Publishers, 1990, 450–506.

[9] S. M. Katz, Estimation of probailities from sparse data for the language model component of speech recognizer, IEEE Trans. on Acoustics, Speech, and Signal Processing, 1987, ASSP-35(3), 400–401.

[10] R. Rosenfeld, The CMU statistical language modeling toolkit and its use in the 1994 ARPA CSR evaluation, In the Proc. of ARPA Spoken Language Systems Technology Workshop, Washington, 1995, 47–50.

A WORD SEGMENTATION ALGORITHM FOR CHINESE LANGUAGE BASED ON N-GRAM MODELS AND MACHINE LEARNING

Wu Yingliang Wei Gang* Li Haizhou*

(*School of Business Administration, South China Univ. of Tech., Guangzhou* 510641 China)

*(*Dept. of Electron. and Info. Eng., South China Univ. of Tech., Guangzhou* 510641 China)

Abstract Automatic word segmentation for the Chinese language is a fundamental and difficult problem in the field of computer Chinese language information processing. This paper presents a new method for segmenting the input Chinese language text sentence into words, which consists of a character-based N-gram model and an efficient Viterbi search algorithm. In addition, two performance evaluation ration targets, i.e. Recall and Precision for word segmentation algorithm are discussed, The effectiveness has been confirmed by evaluation experiments using the closed texts and open texts corpus.

Key words Chinese language word segmentation, N-gram model, Machine learning, Precision, Recall

吴应良: 男, 1963 年生, 博士, 副教授, 主要从事智能信息处理、计算机网络与信息安全、电子商务等方面的教学和科研工作.

韦 岗: 男, 1963 年生, 博士, 教授, 博士生导师, 主要从事现代通信理论与技术、多媒体信息处理、模式识别、神经网络等方面的教学和科研工作.

李海洲: 男, 1964 年生, 博士, 教授, 博士生导师, 主要从事语音识别、神经网络等方面的教学和科研工作.