第32卷 第19期 计 算 机 工 程 2006年10月
Vol.32 № 19 Computer Engineering October 2006
·博士论文· 文章编号：1000—3428(2006)19—0023—03 文献标识码：A 中图分类号：TP391

基于SLM的二叉树在语音停顿预测中的应用

钱揖丽[1,2]，荀恩东[3]，宋 柔[3]
(1. 北京工业大学计算机学院，北京 100022；2. 山西大学计算机与信息技术学院，太原 030006；
3. 北京语言大学信息科学学院，北京 100083)

摘 要：讨论基于统计语言模型SLM（Statistic Language Model）的二叉树在语音停顿预测中的应用。基于大规模语料，利用三元模型Trigram，建立统计语言模型；基于SLM为待处理句子生成相应的二叉树；将生成的二叉树所包含的信息，从不同角度应用于语音停顿的预测。实验结果表明，基于SLM生成的二叉树能够较好地为语音停顿的预测做出贡献。

关键词：统计语言模型；二叉树；语音停顿；预测

Application of Bintree Based on SLM in Speech Pauses' Prediction

QIAN Yili[1,2], XUN Endong[3], SONG Rou[3]
(1. College of Computer, Beijing University of Technology, Beijing 100022; 2. College of Computer and Information Technology, Shanxi University, Taiyuan 030006; 3. College of Information Science, Beijing Language and Culture University, Beijing 100083)

【Abstract】This paper discusses the application of bintree based on SLM (Statistic Language Model) in speech pauses' prediction. It constructs Trigram statistic language model based on large-scale corpus, and builds corresponding bintree for the sentence waiting disposal; and then it predicts speech pauses at two different angle using information provided by tree. The results of experiments show that the bintree based on SLM can make contribution to speech pauses' prediction effectively.

【Key words】Statistic language model; Bintree; Speech pause; Prediction

文语转换系统（Text-to-Speech）的目的是将文字的输入自动地转换成语音的输出。它在信息发布系统、语音应答系统、电子邮件中的语音服务、文稿校对系统以及残疾人语音辅助等许多方面有很大的应用前景。合成高可懂度、高自然度的语音，一直是语音合成所追求的目标。经过十几年的研究，现阶段合成语音的可懂度已经达到相当高的水平，但自然度还不够高，输出语音的质量与实际应用的要求还有一段距离[5]。

影响合成语音自然度的原因之一，就是语音停顿的预测问题。人们在正常发音时，并不会把一个较长的句子一口气念出，而会把它分隔成若干个短语，并在短语的边界处插入长短不同的停顿，正确预语音停顿，从而改善合成语音的自然度，具有重要的意义。国内外已经提出了许多韵律短语自动切分的方法，并取得了一定的进展。

本文讨论基于统计语言模型SLM的二叉树在语音停顿预测中的应用。实验结果表明，基于SLM生成的二叉树能够较好地为语音停顿的预测做出贡献。

1 建立统计语言模型SLM

1.1 停顿与标点符号

关于停顿，现在能被普遍认可的说法是：“说话时语音上的间歇”；“朗读语流中声音的中断”；它是有声语言的标点符号[7]。

标点符号是辅助文字记录语言的符号，是书面语言的有机组成部分。英语中，标点隔开的一定是一个完整的句法成分，汉语则不然。汉语的标点是用来表示停顿、语气以及词语性质的标记。不同的标点符号，表示的停顿也不一样。它们的作用分别为[9]：

(1) 顿号表示句子内部并列词语之间的停顿。如：桃树、杏树和梨树，都开满了花。

(2) 逗号用于句子内部主语与谓语之间、句子内部动词与宾语之间、句子内部定语或状语后边需要的停顿，复句内各分句之间的停顿。如：王诚，我们的校长，到现在我还不认识。今天早晨小王在上学的路上等公共汽车的时候，捡到了一个内有巨款的包。

(3) 分号表示复句内部并列分句之间的停顿。如：先生还是写一点罢；刘和珍生前就很爱看先生的文章。

(4) 冒号用在称呼语等之后表示提起下文，用在总说性话语之后表示引起下文的分说，用在需要解释的词语之后表示引出解释或说明，也可以用在总括性话语的前边以总结上文。如：大家注意：嫌疑人出现了！

(5) 句号用于陈述句末尾的停顿。

(6) 问号用于疑问句和反问句末尾的停顿。

(7) 叹号用于感叹句末尾的停顿。

综上所述，在汉语书面语中，表示停顿，是标点符号的主要作用之一。因此，在一个无标点长句的各个词语之间，有停顿的可能性大小可以用该处出现标点的可能性大小来估计。而出现标点的可能性大小，则可以用N-gram的统计语言模型来估计，本文N-gram的语言单位是词和标点符号。

1.2 训练语料获取

建立语言模型需要大规模的训练语料，但获取大规模的

基金项目：国家自然科学基金资助项目(60272055)；国家“863”计划基金资助项目(2001AA114111)；教育部科学技术研究重点基金资助项目(00128)

作者简介：钱揖丽(1977－)，女，讲师、博士生，主研方向：自然语言处理；荀恩东，副教授；宋 柔，教授、博导

收稿日期：2005-12-26 **E-mail：**qyl@blcu.edu.cn

标注了语音停顿的语料，是一件非常困难的事情。利用停顿与标点符号之间存在的上述关系，就能够比较方便地解决这个问题。

本文所用训练语料约 8.25 亿字，来源于人民日报、科技日报、Web、求是杂志、计算机杂志、南方周末等，共包括 20 815 504 个句子。本文首先用单一的分隔符(下文称为停顿符，用△表示)替换语料中的以上 7 类标点符号，然后对语料进行分词。将处理过的语料作为训练语料，建立以词为单位的 N-gram 统计语言模型，其中停顿符△也被看作一个词。

1.3 N-gram 语言模型

统计语言模型是对语言信息源进行统计意义上的描述而得到的模型。近年来，被广泛应用到语音识别、文本校错、机器翻译、词性标注等各个应用领域。

本文采用目前最流行的 N-gram 语言模型。令 $W = w_1 w_2 \cdots w_n$ 为任意一个词序列， W 的先验概率 P(W)为：

$P(W) = \prod_{i=1}^{n} P(w_i \mid w_1 \cdots w_{i-1})$，其中，

$$P(w_i \mid w_1 \cdots w_{i-1}) \approx \frac{Count(w_1 \cdots w_{i-1} w_i)}{Count(w_1 \cdots w_{i-1})}$$

这种统计语言模型依赖于上下文环境，包含了前N-1 个词所能提供的全部信息，这些词对于当前词的出现具有很强的约束力[6]。根据前人的研究经验，认为三元模型（Trigram）是实际应用中表现最佳的模型[6]，对于估计P(W)是非常有效的。所以，本文采用了三元模型（Trigram），即考虑前面 2 个历史词。

针对 Trigram 的数据稀疏问题，本文采用了 Good-Turing 估计。对于在样本中出现 r 次的事件，假设它的出现次数为 r^*，$r^* = (r+1)\frac{n_{r+1}}{n_r}$。其中，$n_r$ 是 N-gram 的训练集中实际出现次数为 r 的事件的个数。

2 构建二叉树

2.1 基本思想

对于任意输入的句子（已分词） $W = w_1 w_2 \ldots w_n$，$w_i (1 \le i \le n)$ 是句子中的第 i 个词。本文认为，每个词对 $w_{i-1} w_i$ 之间都存在一个潜在的停顿点。所以，包含 n 个词的句子共存在 n-1 个潜在停顿点，从左到右记潜在停顿点为 pos，$pos \in [1 \cdots n-1]$。

分别在每一个潜在停顿点插入一个停顿符，形成新的句子如 $W' = w_1 w_2 \ldots w_{i-1} \Delta w_i \ldots w_n$，并基于训练生成的语言模型，计算插入停顿符后整个句子的概率：

$$P(W') = P(w_1)P(w_2 \mid w_1) \cdots P(w_i \mid w_{i-1}\Delta) \cdots P(w_n \mid w_{n-2} w_{n-1})$$

找出使得句子概率最大的停顿点，即

$$\arg\max_{pos} P(W')$$

其中，$pos \in [1 \cdots n-1]$。在这个停顿点位置，将句子分裂左、右子句，同时也形成二叉树的左、右子树。分别对子句重复以上工作，直到所有潜在停顿点处理完毕，即全部都是叶子结点为止。这样，就为任意一个输入的句子，生成了一棵对应的二叉树。

2.2 生成算法

输入句子 W 对应二叉树的生成算法 *Tree(W)*可以描述如下：

(1)对每一个潜在停顿点，从 pos=1到 $wordnum(W)-1$，分别计算在 pos 处加入停顿符后构成的新句子 W' 的概率 $P(W')$；

(2)找到使句子概率最大的停顿点 $\arg\max_{pos} P(W')$；

(3)为停顿点左边的子句 $leftsent = w_1 w_2 \ldots w_{pos}$，构建对应的二叉树 $Tree(leftsent)$；

(4)为停顿点右边的子句 $rightsent = w_{pos+1} \ldots w_{wordnum(W)}$ 构建对应的二叉树 $Tree(rightsent)$；

(5)若句子 W 所包含词的个数 $wordnum(W) = 1$，则结束。

2.3 二叉树表示

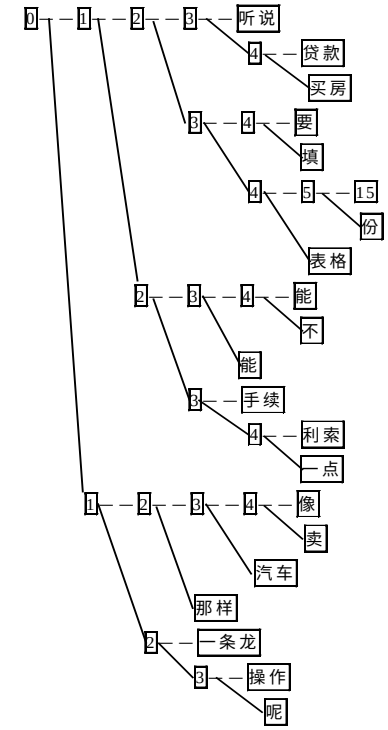

图 1 为句子(1)生成的二叉树

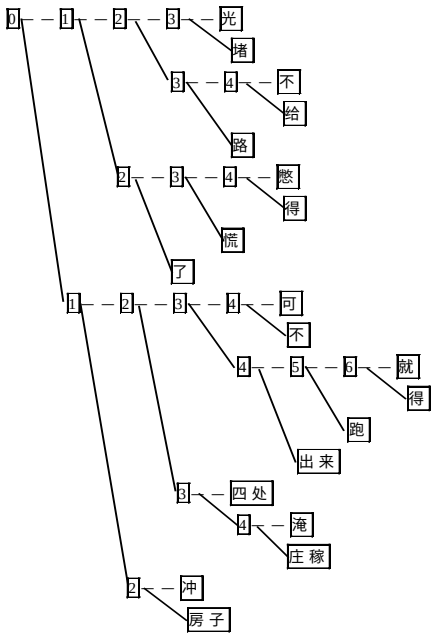

图 2 为句子(2)生成的二叉树

通过两个实例，了解二叉树的表示形式。例如，输入以下两个分词后的句子：

(1) 听说/贷款/买房/要/填/15/份/表格/能/不/能/手续/利索/一点/像/卖/汽车/那样/一条龙/操作/呢

(2) 光/堵/不/给/路/憋/得/慌/了/可/不/就/得/跑/出来/四处/淹/庄稼/冲/房子

这两个句子经过以上二叉树生成算法 $Tree(W)$ 的处理，并将输出结果表示成二叉树的形式，分别如图 1、图 2 所示。

3 二叉树在语音停顿预测中的应用

3.1 语音停顿预测

任意输入待处理的句子，上述算法为其生成相应的二叉树，并根据二叉树中的层次关系，在同一层的子树之间插入停顿符 Δ，预测句子的语音停顿情况。例如：

(1) 根据图 1 中树的第 1 层，为句子(1)标注出一个停顿位置：

听说/贷款/买房/要/填/15/份/表格/能/不/能/手续/利索/一点/ Δ 像/卖/汽车/那样/一条龙/操作/呢

(2) 根据图 1 中树的第 2 层，又为句子(1)标注出两个语音停顿位置：

听说/贷款/买房/要/填/15/份/表格/ Δ 能/不/能/手续/利索/一点/ Δ 像/卖/汽车/那样/ Δ 一条龙/操作/呢

本文随机抽取 500 个句子（共包含 13 096 个词，平均句长为 26.192 个词）做开放测试。分别根据二叉树的第 $i(1 \le i \le 7)$ 层标注语音停顿，实验结果如表 1 所示。

表 1 分层实验结果

	识别个数	正确个数	正确率	召回率
第 1 层	494	455	92.11%	13.96%
第 2 层	742	636	85.71%	19.52%
第 3 层	797	530	66.50%	16.26%
第 4 层	762	434	56.96%	13.32%
第 5 层	580	273	47.07%	8.38%
第 6 层	277	106	38.27%	3.25%
第 7 层	55	35	63.64%	1.07%

表 1 反映了二叉树各个层次的语音停顿预测能力。从工表 1 可以看出，根据二叉树的第 1 层进行标注，能够正确识别出 13.96%的停顿点，且正确率达 92.11%；依此类推。随着层次的深入，正确率呈现下降的趋势，这与二叉树的构建方法是相吻合的。因为依据 SLM 生成二叉树时，随着树层次的深入，停顿点的可信度逐渐降低，也就使得语音停顿预测的正确率逐渐下降。

表 2 考察二叉树前 $n(1 \le n \le 7)$ 层的预测能力，其中

$$F-Score=\frac{2\times 正确率\times 召回率}{正确率+召回率}\times 100\%$$

表 2 前 n 层实验结果

	识别个数	正确个数	正确率	召回率	F-Score
前 1 层	494	455	92.11%	13.96%	24.25%
前 2 层	1236	1091	88.27%	33.48%	48.55%
前 3 层	2033	1621	79.73%	49.74%	61.26%
前 4 层	2795	2055	73.52%	63.06%	67.89%
前 5 层	3375	2328	68.98%	71.43%	70.18%
前 6 层	3652	2434	66.65%	74.69%	70.44%
前 7 层	3707	2469	66.60%	75.76%	70.89%

3.2 语音停顿检验

前面图 1、图 2 所示两个句子的语音停顿正确标注结果分别为：

(1)听说/贷款/买房/ Δ 要/填/15/份/表格/ Δ 能/不/能/手续/利索/一点/ Δ 像/卖/汽车/那样/ Δ 一条龙/操作/呢。

(2)光/堵/不/给/路/ Δ 憋/得/慌/了/ Δ 可/不/就/得/跑/出来/ Δ 四处/淹/庄稼/ Δ 冲/房子。

对比正确标注结果和二叉树表示结果，本文发现一种现象：在句子中位于两个停顿点之间的部分，往往恰好可以构成一棵完整的子树。例如：在句子(1)中，两个停顿点之间的子句“能/不/能/手续/利索/一点/”，恰好就是二叉树中一棵完整的子树，如图 3 所示。

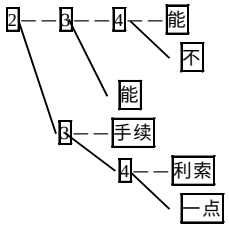

图 3 图 1 中的一棵子树

因此本文认为，考察某个子句是否能够在句子对应的二叉树中构成一棵完整的子树，就可以为句子语音停顿的预测作出贡献。

仍然使用以上随机抽取的 500 个句子做开放测试，将相应的二叉树与正确标注结果进行对比，考察停顿点之间的部分构成二叉树中完整子树的情况。实验结果如表 3 所示。

表 3 语音停顿检测实验结果

500 句共包含停顿	满足条件	不满足条件	正确率
3 259 个	2 791 个	468 个	85.64%

从表 3 可以看出，在 500 个开放测试句子所包含的 3 259 个停顿点中，有 2 791 个都满足本文提出的条件，即：句子中两个停顿点之间的部分，85.64%都恰好可以构成整个句子的对应二叉树中的一棵完整子树。我们可以将这一现象应用于汉语语音停顿的预测。

4 结束语

本文讨论基于统计语言模型 SLM 的二叉树在语音停顿预测中的应用。充分利用汉语中标点符号的停顿作用，本文建立了 8 亿多字的大规模语料，并利用三元文法 Trigram，建立了统计语言模型 SLM。基于 SLM 为任意待处理句子生成相应的二叉树，并进而将生成的二叉树所包含的信息应用于语音停顿的预测。实验表明，基于 SLM 的二叉树可以较好地为语音停顿的预测作出贡献，具有一定的价值。

同时，分析实验过程，本文存在的不足主要有：

(1) 训练语料方面

停顿是用标点来表示的，但并不是话语中的每一处停顿在书面上都要用标点表示出来。事实上，句中的有些停顿有时是不用标点的[8]。如：马克思列宁主义是从客观实际产生出来又在客观实际中获得了证明的最正确最科学最革命的真理。(毛泽东《整顿党的作风》)。这个句子比较长，句中需要有些停顿，可是在书面上，这些需要停顿的地方并没有用标点符号表示出来。因此，标点符号只能提供有限的停顿信息，还有很多信息在训练语料中并没有表现出来；而获得包含丰富停顿信息的大规模语料，是比较困难的。这样，训练语料所包含信息的缺失带来了一定的负面影响。

(2) 语言模型方面

本文采用了 Trigram 语言模型，只考虑了 2 个历史词的信息。解决语音停顿预测问题，Trigram 是否已经足够，是否

（下转第 28 页）