中 文 信 息 学 报
第20卷 第3期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol. 20 No. 3

文章编号:1003-0077(2006)03-0001-05

应用二叉树剪枝识别韵律短语边界

荀恩东[1],钱揖丽[1],郭 庆[2],宋 柔[1]

(1. 北京语言大学 语言信息处理研究所,北京 100083; 2. 富士通研究开发中心,北京 100016)

摘要:句子的韵律短语识别是语音合成的重要研究内容。本文提出了应用统计语言模型生成的二叉树,结合最大熵方法识别待合成汉语句子的语音停顿点。文中给出了二叉树相关的模型训练和生成算法;二叉树与语音停顿点之间的关系;在最大熵方法中应用二叉树剪枝识别句子的韵律短语。实验结果表明,在搜索算法中,利用二叉树进行剪枝,可以很大程度上提高语音停顿预测的正确率和召回率,基于试验数据的f-Score提高了近35%。

关键词:人工智能;自然语言处理;统计语言模型;二叉树;韵律短语;最大熵

中图分类号:TP391 **文献标识码**:A

Using Binary Tree as Pruning Strategy to Identify Prosodic Phrase Breaks

XUN En-dong[1], QIAN Yi-li[1], GUO Qing[2], SONG Rou[1]

(1. Center for Language and Information Processing, Beijng Language and Culture University, Beijing 100083, China;
2. Fujistu Research Center, Beijing 100016, China)

Abstract: It is important to recognize the prosodic phrase breaks in text-to-speech. In this paper, a new method is introduced for this purpose, which uses binary tree as pruning strategy in the Maximal Entropy Model (MaxEnt) framework. First of all, the concept of binary tree generated from a statistical language model is given. Then the process of generating the binary tree is discussed. In the process of applying MaxEnt to seeking optimal prosodic phrases, the binary tree is exploited so as to narrow the search space and improve the performance. Experimental results show that the F-score of predicating prosodic phrase breaks is about 35% better than the previous system, in which the binary tree strategy is not adopted.

Key words: artificial intelligence; natural language processing; statistical language model; binary tree; prosodic phrase; Maximal Entropy Model

1 引言

可懂度和自然度是衡量文语转换系统(Text-to-Speech)水平的两个主要指标。目前TTS系统普遍采用基于语音语料库方法,系统可懂度已经达到相当高的水平,但自然度还不够高,提高合成语音的自然度是TTS的主要研究内容之一。

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。我们把停顿之间的短语称为韵律短语,在合成时,如果能够正确预测句子的语音短语,可以提高合成语音的自然度。

收稿日期:2005-07-20 **定稿日期**:2005-10-23

基金项目:国家自然科学基金资助项目(60573184)

作者简介:荀恩东(1967—),男,博士,副教授,主要研究方向为自然语言处理.

国内外已经提出了许多韵律短语自动切分的方法。例如:基于规则学习的方法[1],基于边界点词性特征统计的方法[2],基于结构助词驱动的方法[3],基于语法信息的方法[4],以及基于最大熵模型的方法[5]等,并取得了一定的进展。

本文给出了基于统计语言模型生成的二叉树方法,这种二叉树结构与韵律短语有非常好的同构性,在应用最大熵模型搜索最优韵律短语时,依照二叉树结构进行剪枝,可以大大提高最大熵方法识别正确率。本文首先给出二叉树的训练模型和生成算法;论述了二叉树与语音停顿点之间的关系;在最大熵方法中如何利用二叉树信息识别输入语音的停顿点。实验结果表明,利用基于统计语言模型生成的二叉树,可以很大程度上提高语音停顿的预测的正确率和召回率,基于测试数据的 F-Score 提高了近 35%。

2 应用统计语言模型建立二叉树

2.1 停顿与标点符号

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,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,可以用统计语言模型来估计。

2.2 生成二叉树的统计语言模型

建立语言模型,需要大规模的训练语料,而获取大规模的标注了语音停顿的语料,是一件非常困难的事情。这里采用了两个语言模型线性组合的方法:一个模型是利用语音停顿与标点符号之间存在的这种关系,从大规模生语料中训练语言模型;另一个是从带有停顿标识的训练语料中训练得到。

设 $W = w_1 w_2 \cdots w_n$ 为任意一个已经分词处理的词序列,W 的概率 $P(W)$ 为:

$$P(W) = \lambda \times P_G(W) + (1 - \lambda) \times P_D(W) \tag{1}$$

这里:$P_G(W)$ 为通用三元语言模型,从大规模汉语生语料中训练得到。在训练时,原语料中所有标点符号统一替换成一个标识停顿位置的符号 Δ。

$$P_G(W) = P(w_1) \times P(w_2) \prod_{i=3}^{n} P(w_i \mid w_{i-2}, w_{i-1})$$

$P_D(W)$ 为从训练语料中计算得到的二元语言模型。在训练时,替换训练语料中已标注的停顿标识为符号 Δ。

$$P_D(W) = P'(w_1) \times \prod_{i=2}^{n} P'(w_i \mid w_{i-1})$$

训练 $P(w_i \mid w_{i-2}, w_{i-1})$ 和 $P'(w_i \mid w_{i-1})$ 时,采用了 Good-Turing 数据平滑方法。

2.3 利用语言模型生成二叉树

对于任意已分词输入的句子 $W = w_1 w_2 \cdots w_n$,$w_i (1 \leqslant i \leqslant n)$,是句子中的第 i 个词。假设每个词对 $w_{i-1} w_i$ 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 $W' = w_1 w_2 \cdots w_{i-1} \Delta w_i \cdots w_n$,应用公式 1 计算 $P(W')$,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,将句子分裂为左、右子句,同时也形成二叉树的左、右子树。分别对子树重复以上工作,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。

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 $leftsent=w_1w_2\cdots w_{pos}$,构建对应的二叉树 $Tree(leftsent)$;

(4)为停顿点右边的子句 $rightsent=w_{pos+1}\cdots w_{wordnum(W)}$,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。

例如,输入两个已分词处理的句子:

句子1:他　在　粮食　企业　摸爬滚打　30　多　年

句子2:一　张　大红　请帖　搅　得　你　心神不定

经过以上二叉树生成算法 $Tree(W)$ 的处理,将输出结果表示成二叉树的形式,结果为图1、图2所示。

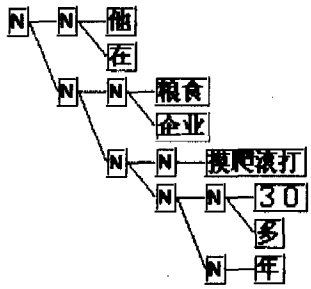

图1　句子1生成的二叉树

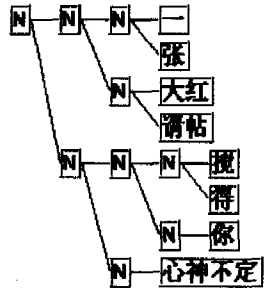

图2　句子2生成的二叉树

以上两个图的树结构和该树对应的韵律短语有很好的同构性,韵律短语往往对应二叉树的一个完整子树,试验结果表明这种对应性可以高达96%左右。

3　在最大熵模型中应用二叉树约束

3.1　问题表示

设输入已经分词和词性标注的句子:$W=w_1w_2\cdots w_n$,$T=t_1t_2\cdots t_nw_i$ 为第 i 个词,t_i 为 w_i 的词性($1\leqslant i\leqslant n$)。句子的语音停顿结果可以表示为,$P=p_1p_2\cdots p_m$,($m\leqslant n$),p_i 称为韵律短语,它对应句子中一个或者多个词的序列 $w_j\cdots w_{j+k}$,对应词性序列 $t_j\cdots t_{j+k}$,($1\leqslant j\leqslant n$,$0\leqslant k\leqslant n$);当 $k=0$ 时,韵律短语 p_i 的标识定义为其对应词的词性,当 $k>0$ 时,韵律短语 p_i 的标识定义为 SP,这样,最后的识别结果为词性和 SP 组成的序列,例如,标注语料中一个句子:

表1　问题表示示例

类型 / 韵律短语	词序列	词性序列	标识
一张	一+张	M+q	SP
大红请帖	大红+请帖	B+n	SP
搅得你	搅+得+你	V+u+r	SP
心情不定	心情不定	I	I

一/m 张/q | 大红/b 请帖/n | 搅/v 得/u 你/r | 心神不定/i

识别结果就是(m q)/SP (b n)/SP (v u r)/SP (i)/I,对应的表示如表1。

3.2 模型特征

最大熵模型的解具有以下形式：

$$P(y \mid x) = \frac{1}{Z(x)}\exp\left(\sum_i \lambda_i f_i(x,y)\right)$$

这里 y 表示预测的韵律短语边界的结果，x 表示输入句子的上下文，$f_i(x,y)$ 满足 x 上下文的一个特征，λ_i 表示特征 $f_i(x,y)$ 对应的权重，$Z(x)$ 为一个归一化因子。

在应用最大熵模型时，需要提取模型特征 $f_i(x,y)$，训练特征的统计参数 λ_i，我们取了以下韵律短语的上下文作为模型特征。这里综合利用了词信息、词性信息、韵律短语的词数目和字数目。对训练语料中，所有的韵律短语按照表 2 中的特征模板生成模型特征，统计语料中特征出现次数，把特征数大于 5 的作为最后的特征，参与模型训练和参数计算。

表 2 特征模型

当前的韵律短语的词性序列	当前韵律短语的末词
前一个韵律短语的词性序列	当前韵律短语的词的个数
当前的韵律短语标识	当前韵律短语的字的个数
前一个韵律短语的标识	前一个韵律短语的词的个数
前两个韵律短语的标识	前一个韵律短语的字的个数
前一个韵律短语的末词与当前韵律短语的首词	

3.3 搜索算法

搜索过程就是利用训练得到的最大熵模型参数，求解全局概率最大的韵律短语，这里采用了 Beam-Search 的方法对输入的带有分词和词性标注结果的句子进行韵律短语识别。在这里韵律短语的识别结果可以表示为词性和 SP 组合形式，搜索阶段的任务就是在所有可能的词性和 SP 组合中求解概率最大的作为最终识别结果，在建立搜索路径时，这里采用了从训练语料中抽取得到的、韵律短语的词性序列的各种组合来构建搜索空间。

例如，输入待语音停顿边界识别的句子：

一/m 张/q 大红/b 请帖/n 搅/v 得/u 你/r 心神不定/I

依照从训练语料中抽取的韵律短语的所有可能词性序列，依次从左到右构造弧，构成以下有向图，如图 3 所示，这样从左到右，任一个路径都有可能构成识别结果。考虑到输入句子对应的二叉树结构图 2，具有实线的弧对应二叉树中一个完整子树，而画有虚线的弧不对应在二叉树中完整子树，我们依据弧要对应输入句子二叉树中一个完整子树的原则，可以对待搜索空间进行剪枝。在此例子中滤掉所有虚线的弧，然后对余下的搜索图，应用最大熵计算公式，求解概率最大的一条路径，本例中，图 3 粗实线对应的路径为韵律短语识别结果，即对应以下标注：

一/m 张/q | 大红/b 请帖/n | 搅/v 得/u 你/r | 心神不定/I

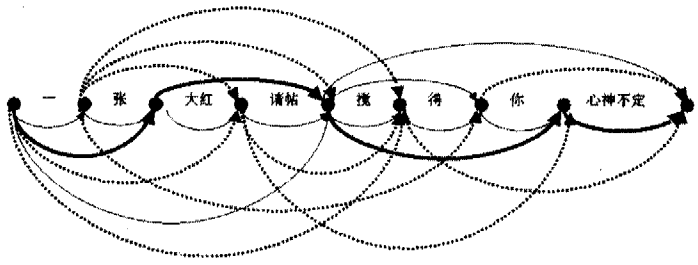

图 3 特征模型

4 实验

实验语料包括已经带有分词和词性标注的句子3,000个,该语料从富士通标注的1998年《人民日报》中随机抽取,平均每句包含9.17个词,每句平均包含3.68个语音停顿点标记。为了保证实验有更好的说明性,由程序随机生成10组实验数据,每组包括2700句作训练语料,300句做为测试语料。测试时考察标注的正确率、召回率和F-Score。

输入分词句子,应用统计语言模型生成二叉树时,采用了线形插值方法,即组合从一般语料训练的语言模型和带有语音停顿点标记的训练语料,图4给出了λ的不同取值,测试语料的韵律短语覆盖二叉树完整子树的正确率。依据曲线,实际中设置λ=0.075,在该设定下,正确率平均达到了96.1%。由实验数据可知二叉树与韵律短语具有同构性非常好。从而为二叉树作为剪枝条件提供了可能。

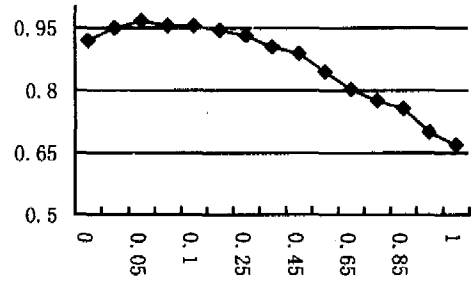

图4 λ对子数覆盖正确率影响

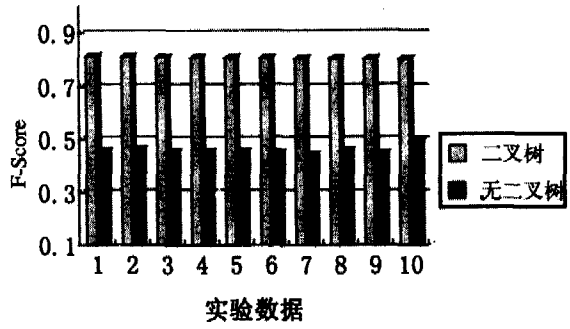

图5 是否应用二叉树的F-Score的比较

在最大熵方法特征提取时,在训练语料中,应用特征模板生成特征,选取的次数大于5的特征为有效特征,图5中给出不同组测试数据,是否应用二叉树剪枝识别结果的F-Score比较,实验表明:在搜索算法中,依照韵律短语要覆盖二叉树完整子数原则,对搜索空间进行剪枝,F-score提高了近35%。

表3 10组测试数据韵律短语的识别结果

	1	2	3	4	5	6	7	8	9	10	平均
正确率	0.835	0.827	0.827	0.832	0.823	0.821	0.821	0.821	0.825	0.829	0.826
找回率	0.791	0.798	0.787	0.793	0.794	0.792	0.791	0.791	0.786	0.781	0.79
F-Score	0.813	0.812	0.809	0.809	0.808	0.807	0.806	0.806	0.805	0.804	0.808

图5和表3给出了不同10组测试数据的韵律短语识别的正确率、召回率和F-Score,平均正确率为0.825,平均召回率为0.79,平均F-Score为0.808。

表4 不同方法韵律短语的识别对照结果

方 法	正确率	召回率	F-Score	训练语料规模(句子数)
文献[10]	0.811	0.787	0.797	9,000
文献[5]	—	—	0.65	41,000
文献[9]	—	—	0.829	12,000
本方法	0.825	0.79	0.808	3,000

表4给出了前人做的类似工作和本实验结果对比,因为每个实验采用的训练和测试语料都不同,对韵律短语的的把握标准也不同,该表仅供参考。从表中可看到,本方法的应

(下转第28页)

195.
[5] 王萌,何婷婷,姬东鸿,王晓荣. 基于 HowNet 概念获取的中文自动文摘系统[J]. 中文信息学报,2005,19(3):440-446.
[6] 钱铁云,王元珍,冯小年. 结合类频率的关联中文文本分类[J]. 中文信息学报,2004,18(6):30-36.
[7] Dong Zhengdong, Dong Qiang the download of Hownet[EB/OL], http://www.keenage.com.
[8] Yang Yimin, and Pedersen J O. A comparative study on feature selection in text categorization[A]. In: proceedings of the 14th International Conference on Mahine Learning(ICML-97) [C],1997.
[9] 代六玲,黄河燕,陈肇雄. 中文文本分类中特征抽取方法的比较研究[J]. 中文信息学报,2004,18(1):26-32.
[10] 李凡,鲁明羽,陆玉昌. 关于文本特征抽取新方法的研究[J]. 清华大学学报(自然科学版),2001,(7):99-102.
[11] FABRIZIO SEBASTIANI machine learning in automated text categorization[C]. ACM computing surveys, Vol. 34, No 1, March 2002, P1.

(上接第 5 页)

用的训练语料规模,要比其他方法少的多。同时应用了二叉树剪枝方法,也提高了系统的运行效率。

5 结束语

本文采用了最大熵方法识别输入句子的韵律短语边界,在搜索最佳识别结果时,引入了二叉树方法作为剪枝策略,对不满足韵律短语不完全覆盖二叉树子树的韵律短语对应的弧进行剪枝,大大缩小了搜索空间,提高了识别 F-Score 近 35%,在小训练语料上,平均 F-Score 达到了 80.8%。

参 考 文 献:

[1] 赵晟,陶建华,蔡莲红. 基于规则学习的韵律结构预测[J]. 中文信息学报,2002,16(5):30-37.
[2] 牛正雨,柴佩琪. 基于边界点词性特征统计的韵律短语切分[J]. 中文信息学报,2001,15(5):19-25.
[3] 应宏,蔡莲红. 基于结构助词驱动韵律短语界定的研究[J]. 中文信息学报,1999,13(6):41-46.
[4] 曹剑芬. 基于语法信息的汉语韵律结构预测[J]. 中文信息学报,2003,17(3):41-46.
[5] 李剑锋,胡国平,王仁华. 基于最大熵模型的韵律短语边界预测[J]. 中文信息学报,2004,18(5):56-63.
[6] 叶竹钧. 朗读中的停顿探析[J]. 语文教学通讯,1995,(Z1):78-79,1995,(7):30-31.
[7] 汪国胜. 标点符号概说[J]. 高等函授学报(哲学社会科学版),1996,(6):19-23.
[8] 中华人民共和国国标《标点符号用法》,1996,6.
[9] Min Chu, Yao Qian, Locating Boundaries for Prosodic Constituents in Unrestricted Mandarin Texts[J], 2001, Computational Linguistics and Chinese Language processing, Vol 6, No. 1, 61-83.
[10] 赵永贞,刘挺,王志伟,陈惠鹏,邵艳秋. 汉语文语转换系统中停顿指数的自动标注[J]. 中文信息学报,2004,18(5):48-55.
[11] 聂鑫,王作英. 汉语语句中短语间停顿的自动预测方法[J]. 中文信息学报,2003,17(4):39-44.
[12] 吴志勇,蔡莲红. 语音合成中韵律关联模型[J]. 中文信息学报,2004,18(2):44-50.