基于标点信息和统计语言模型的语音停顿预测*

钱揖丽[1,2] 荀恩东[3]

[1](北京工业大学 计算机科学学院 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 语音停顿被认为是有声语言的标点符号.在语言交流中,说话人会在韵律短语的边界处插入长短不同的停顿.利用这一性质,在调查标点符号停顿作用的基础上,提出基于标点信息预测语音停顿的思想,阐述基于标点和统计模型的训练语料自动获取以及语音停顿预测方法,讨论训练语料规模对模型性能的影响,并比较基于标点信息的自动获取语料与人工标注语料的性能.实验结果显示,汉语的标点提供有价值的停顿信息,基于汉语标点信息能够有效预测语音停顿.

关键词 标点符号,语音停顿,统计语言模型,语料获取
中图法分类号 TP 391

Prediction of Speech Pauses Based on Punctuation Information and Statistical Language Model

QIAN Yi-Li[1,2], XUN En-Dong[3]

[1](*College of Computer Science, Beijing University of Technology, Beijing* 100022)
[2](*College of Computer and Information Technology, Shanxi University, Taiyuan* 030006)
[3](*College of Information Sciences, Beijing Language and Culture University, Beijing* 100083)

ABSTRACT

Speech pauses are considered as punctuation marks of spoken language. People always insert different pauses at the boundaries of rhythmic phrases when communicating by language. Based on this characteristic, the speech pause of punctuation marks is investigated and the concept of predicting speech pauses using punctuation information is proposed. The punctuation-based and SLM-based methods are introduced to obtain training corpus and predict speech pauses. The influence of training corpus size on the performance of model is discussed. And the performance of punctuation-based corpus and manually-labeled corpus is compared. Experimental results show that the Chinese punctuation supplies valuable information on pause, and the method based on punctuation information can predict the Chinese speech pauses effectively.

Key Words Punctuation Marks, Speech Pause, Statistic Language Model, Corpus Obtaining

* 国家自然科学基金资助项目(No. 60572159,60573184,60473139)
收稿日期:2007 - 06 - 21;修回日期:2008 - 01 - 21
作者简介 **钱揖丽**,女,1977 年生,讲师,博士研究生,主要研究方向为自然语言处理. E-mail: qyl@ sxu. edu. cn. **荀恩东**,男,1967 年生,博士,硕士生导师,主要研究方向为自然语言处理.

1 引 言

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.语音停顿体现了言语特有的节奏和韵律.适当的停顿,可以强化语言节奏,增强言语表达效果.因此,正确预测语音停顿,对于改善机器合成语音的自然度具有重要的意义.

目前计算语言学中对韵律信息的研究,主要是从语音合成的角度、基于汉语文本信息、应用统计或者知识推理的方法进行的.例如郑敏的概率频度方法[1],李剑锋的最大熵模型方法[2],曹剑芬的基于语法信息的方法[3],赵晟的规则学习方法[4],牛正雨的边界点词性特征统计方法[5],应宏的结构助词驱动方法[6],聂鑫的规则与统计相结合的方法[7]等.

上述研究的共同之处是,模型的训练语料都是采用人工标注的方法获得,由标注人员直接对文本进行韵律信息的标注.为了保证并提高模型的性能,研究者通常都希望基于较大规模的标注语料库开展工作.但是,人工标注的方法不仅费时费力,而且容易受到标注者主观因素的影响,降低标注结果的客观性和一致性.

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.表示停顿是标点符号的主要作用之一,而语音停顿也被认为是有声语言的标点符号.基于这一特性,本文提出将标点符号应用于语音停顿预测的思想.利用标点符号自动获取大规模训练语料,并利用基于标点信息建立的统计模型预测汉语句子的语音停顿位置.实验结果表明,汉语的标点提供非常有价值的停顿信息,基于汉语标点信息能够有效预测语音停顿.

2 汉语标点与语音停顿

2.1 标点的停顿作用

古时候并没有标点符号,写的文章不仅读起来很吃力,而且可能由于断句错误而产生误解.标点符号是在文字产生之后,随着书面交际的需要,陆续创造出来的.

根据《标点符号用法》,标点符号是用来表示停顿、语气和标明词语的性质、作用的符号.标点符号分为标号和点号两大类.标号的作用在于标明,主要标明语句的性质和作用.常用的标号有 9 种,即引号、括号、破折号、省略号、着重号、连接号、间隔号、书名号和专名号.点号的作用在于点断,主要表示说话时的停顿和语气.

点号又分为句末点号和句内点号.

1)句末点号用在句末,表示句末的停顿,同时表示句子的语气.句末点号有句号、问号、叹号 3 种,分别表示陈述句、疑问句和反问句、感叹句末尾的停顿.

2)句内点号用在句内,表示句内的各种不同性质的停顿.句内点号有逗号、顿号、分号、冒号 4 种.

(1)逗号表示一句话中间的停顿.

(2)顿号表示句子内部并列词语之间的停顿.

(3)分号表示复句内部并列分句之间的停顿.分号是介于逗号和句号之间的一种符号,它所停顿的时间比逗号长些、比句号短些.

(4)冒号表示提示性话语之后的停顿.

2.2 标点处的语音停顿调查

由于标号中的破折号和省略号常常有表示停顿的作用,所以将它们同点号一起列为考察对象.本文根据 3 000 多句的语音语料(语速较慢,为 2.59 音节/s;音节间的语音无声段平均时长为 114.21ms;每个音节的平均时长为 271.63ms),抽取并统计各个标点处出现的语音无声段的长度,如表 1 所示.

从表 1 中可以看到,各标点处的语音无声段平均时长都在 600ms 以上,它们明显大于语料所有音节间语音无声段的平均时长.因此,调查结果进一步验证了书面语中的标点符号能够表示停顿的性质,这为本文将标点符号应用于语音停顿的预测提供了理论基础和支持.

表 1 标点处的语音无声段

Table 1 Speech silence segments at punctuation marks

	句末点号			句内点号				部分标号	
	。	?	!	、	,	;	:	…	——
无声段时长范围(ms)	1315~2294	513~2800	433~3311	106~1481	165~2277	566~2764	404~2803	637~2371	303~1008
无声段平均时长(ms)	1794	1088	1132	628	738	1066	882	1205	711

3　基于标点和统计模型预测语音停顿

3.1　基于标点的语料获取

基于标点符号表示停顿的性质,用统一的符号▲(本文称为停顿符)自动替换文本语料中的上述7种点号和2种标号,然后对替换后语料进行自动分词,得到由词和停顿符▲构成的大规模训练语料,解决人工标注语料获取困难的问题. 这种基于标点的语料自动获取方法,方便快捷,对降低人力消耗、减少手工工作量等具有重要的意义.

本文的训练语料约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,包括20 815 504个句子,包含73 233 305个停顿符▲.

以下为训练语料的示例样本,其中,停顿符▲也被看作是一个词.

现在 成人 高校 本科 资格 的 审批 不够 严格 ▲ 缺乏 必要 的 资格 审批 程序 ▲ 没有 规范 的 标准 ▲ 从 明年 开始 将 对 所有 申报 的 本科 ▲ 专升本 专业 资格 进行 审定 ▲

3.2　建立统计模型

采用Tri-gram语言模型,考察前面2个历史信息,基于训练语料建立语言模型. 那么,任意一个词序列 $W = w_1 w_2 \cdots w_n$ 的先验概率 $P(W)$ 为

$$P(W) = P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_1 w_2)\cdots P(w_n \mid w_{n-2}w_{n-1}).$$

同时,针对模型的数据稀疏问题采用Good-Turing估计平滑. 即对于在样本中出现 r 次的事件,假设它的出现次数为

$$r^* = (r+1)\cdot\frac{n_{r+1}}{n_r},$$

其中 n_r 是Tri-gram的训练集中实际出现次数为 r 的事件的个数.

3.3　语音停顿预测算法

书面语中的标点能够表示停顿,而语音停顿被认为是有声语言中的标点符号,它是"说话时语音上的间歇"及"朗读语流中声音的中断". 因此本文认为句子中词语之间出现标点的可能性大小,能够用于估计该处出现语音停顿的可能性大小.

由于书面文本本身就包含标点符号,所以首先将出现标点的位置直接判断为语音停顿点. 其次,将标点位置作为切分点,把原句子切分为若干个词序列(即原句子中两个相邻标点之间的部分). 最后,采用下述算法预测这些词序列内部的语音停顿位置.

算法　语音停顿的预测算法

设　$length(W)$ 为词序列 W 中词的个数.

输入　任意词序列 $W_s = w_1 w_2 \cdots w_n$, $w_i (1 \leqslant i \leqslant n)$ 是第 i 个词,令 $W = W_s$

输出　预测并标注了语音停顿位置后的 W_s

step 1　For $i = 1$ to $length(W) - 1$ do

step 1.1　在 $w_i w_{i+1}$ 之间插入一个停顿符▲,形成新序列 $W'_i = w_1 w_2 \cdots w_i ▲ w_{i+1} \cdots w_n$.

step 1.2　利用Tri-gram语言模型,计算 W'_i 的概率:

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

step 2　令 $k = \underset{i}{\operatorname{argmax}} P(W'_i)$,并将 W 分裂为 $W_l = w_1 w_2 \cdots w_k$ 和 $W_r = w_{k+1} \cdots w_n$.

step 3　若 $length(W_l) \geqslant 2$ 且 $length(W_r) \geqslant 2$,则将 W 的 $w_k w_{k+1}$ 之间预测为一个语音停顿位置.

step 4　若 $length(W_l) \geqslant 4$,则令 $W = W_l$,跳转到step 1做递归处理.

step 5　若 $length(W_r) \geqslant 4$,则令 $W = W_r$,跳转到step 1做递归处理;否则,结束.

例如词序列"有效 地 杜绝 了 拍 脑袋 工程 花架子 项目"的语音停顿预测过程如图1所示.

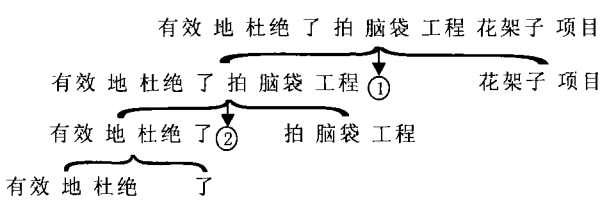

图1　语音停顿预测示意图1

Fig. 1　Sketchmap 1 for prediction of speech pauses

输出为"有效 地 杜绝 了 ② 拍 脑袋 工程 ① 花架子 项目",总共预测出2个语音停顿位置. 其中,① 表示是根据第一层分裂预测得到,同时也说明此处出现标点的可能性最大,那么存在语音停顿的概率也最大,依此类推.

上述算法中将某分裂点预测为语音停顿位置的约束条件是 $length(W_l) \geqslant 2$ 且 $length(W_r) \geqslant 2$,即要求分裂点左、右两边的词序列都至少包含2个词. 这是因为:

1)若对分裂点左、右两边都无约束,则任意两个相邻的词之间都将被预测为语音停顿位置,工作没有意义.

2)若只约束分裂点的左边(或右边),则图2中

3 个“★”所示的位置也被预测为语音停顿点. 对大量数据的统计表明,此时的召回率有小幅提高,但是准确率有大幅降低,总体上降低了模型的性能,负面作用大于正面作用.

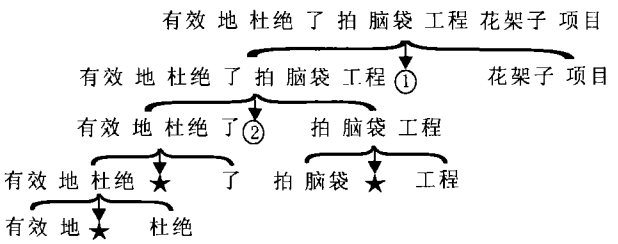

图 2 语音停顿预测示意图 2

Fig. 2 Sketchmap 2 for prediction of speech pauses

4 实验结果及分析

4.1 模型参数的选择

基于训练语料建立统计模型时,需要设置裁剪门限 α 和 β,舍弃出现次数分别低于 α、β 的 Tri-gram 和 Bi-gram. 通过测试和比较参数 α、β 的不同取值对模型性能的影响,本文取 $\alpha = 1, \beta = 8$.

4.2 开放测试结果

随机抽取 500 个包含 5 个以上词的序列 W(即 $length(W) > 5$)作为开放测试集. 词序列的平均含词数为 9.5 个,测试语料共包含 1 136 个语音停顿点(由标注人员结合录音语料标注). 开放测试结果如表 2 所示,其中

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

表 2 开放测试结果

Table 2 Results of open test

正确个数	识别个数	实际个数	召回率	准确率	F 值
946	1008	1136	83.27%	93.85%	88.24%

4.3 训练语料规模的影响

对于基于标点信息的语音停顿预测方法,训练语料的规模对模型的性能有何影响. 分别基于不同规模的训练语料建立语言模型并预测语音停顿位置,测试并观察训练语料规模对模型性能的影响,比较结果如图 3 所示.

依据图 3 可知,随着训练语料规模的加大,模型性能逐渐提高. 但是当训练语料达到一定规模后,性能曲线逐步趋于平缓,基本不再提高. 也就是说,当训练语料达到一定规模之后,模型的语音停顿预测能力趋于一种饱和的状态,规模的继续加大并不能带来性能的明显提高,此时性能的提高只能依靠方法上的改进.

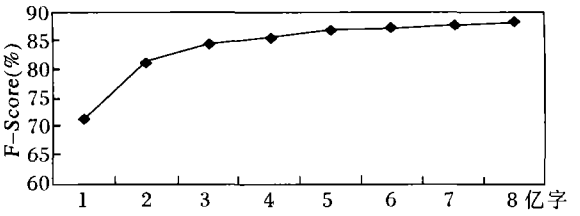

图 3 不同规模训练语料的性能比较

Fig. 3 Performance comparison of training corpus with different sizes

4.4 与人工标注语料的性能比较

本文进行了基于标点自动获取语料与人工标注语料的性能比较. 研究[1-2,4-5]等采用的人工标注训练语料的规模一般为几万个字,包含三五千个句子,一万多个停顿点. 所以,本文以人工标注的 20 万字语料(包含约 31 099 个停顿点)作为标准,比较其与基于标点自动获取语料的性能. 经开放测试得知,它与 250 万字的自动获取语料(包含 251 095 个停顿点)性能相当,性能比较结果如图 4 所示.

从图 4 可知,基于上述两种语料训练获得的模型性能相当. 但是,人工标注 20 万字语料费时费力,极其困难;而基于标点符号自动获取 250 万字语料则非常容易、快捷. 它们在实现的难易程度上存在显著的差别,以机器替代人工劳动具有重要的意义和价值.

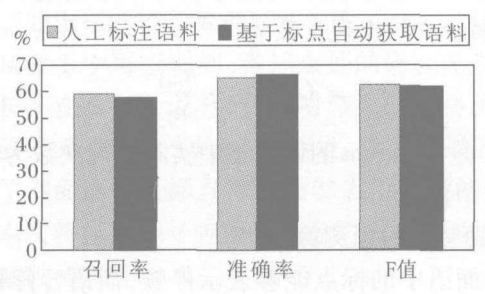

图 4 人工标注语料与自动获取语料的性能比较

Fig. 4 Performance comparison between manually-labeled corpus and automatically-obtained corpus

5 结 束 语

本文利用书面语中的标点符号表示停顿的性质,提出将标点符号应用于语音韵律停顿预测的思

想,讨论基于标点信息的语料获取和语音停顿预测方法并考察训练语料规模对模型性能的影响,同时进行标点语料与人工标注语料的性能比较. 实验结果显示,基于标点的方法不仅能够极大地减少人力消耗、降低手工工作量,而且能够取得较好的识别效果. 它对于语音韵律领域研究工作的开展,具有积极意义和实用价值.

参 考 文 献

[1] Zheng Min, Cai Lianhong. Statistical Model Based on Probability Frequency for Mandarin Prosodic Structure Prediction. Journal of Tsinghua University: Science and Technology, 2006, 46(1): 78 - 81 (in Chinese)
(郑 敏,蔡莲红. 基于概率频度的普通话韵律结构预测统计模型. 清华大学学报:自然科学版, 2006, 46(1): 78 - 81)

[2] Li Jianfeng, Hu Guoping, Wang Renhua. Prosody Phrase Break Prediction Based on Maximum Entropy Model. Journal of Chinese Information Processing, 2004, 18(5): 56 - 63 (in Chinese)
(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. 中文信息学报, 2004,18(5): 56 - 63)

[3] Cao Jianfen. Prediction of Prosodic Organization Based on Grammatical Information. Journal of Chinese Information Processing, 2003, 17(3): 41 - 46 (in Chinese)
(曹剑芬. 基于语法信息的汉语韵律结构预测. 中文信息学报, 2003, 17(3): 41 - 46)

[4] Zhao Sheng, Tao Jianhua, Cai Lianhong. Rule-Learning Based Prosodic Structure Prediction. Journal of Chinese Information Processing, 2002, 16(5): 30 - 37 (in Chinese)
(赵 晟,陶建华,蔡莲红. 基于规则学习的韵律结构预测. 中文信息学报, 2002,16(5): 30 - 37)

[5] Niu Zhengyu, Chai Peiqi. A Statistical Approach Based on Boundary POS Feature to Prosodic Phrasing. Journal of Chinese Information Processing, 2001, 15(5): 19 - 25 (in Chinese)
(牛正雨, 柴佩琪. 基于边界点词性特征统计的韵律短语切分. 中文信息学报, 2001,15(5): 19 - 25)

[6] Ying Hong, Cai Lianhong. Research on the Segmentation of the Prosodic Phrase Based on Driven by the Structural Auxiliary Word. Journal of Chinese Information Processing, 1999, 13(6): 41 - 46 (in Chinese)
(应 宏,蔡莲红. 基于结构助词驱动韵律短语界定的研究. 中文信息学报, 1999, 13(6): 41 - 46)

[7] Nie Xin, Wang Zuoying. Automatic Phrase Breaks Prediction in Chinese Sentences. Journal of Chinese Information Processing, 2003, 17(4): 39 - 44 (in Chinese)
(聂 鑫,王作英. 汉语语句中短语间停顿的自动预测方法. 中文信息学报, 2003,17(4): 39 - 44)

[8] Chu Min, Yao Qian. Locating Boundaries for Prosodic Constituents in Unrestricted Mandarin Texts. Computational Linguistics and Chinese Language Processing, 2001, 6(1): 61 - 82

[9] Yang Jinchen, Yang Yufang. Prosody Generation in Language Production. Advances in Psychological Science, 2004, 12(4): 481 - 488 (in Chinese)
(杨锦陈, 杨玉芳. 言语产生中的韵律生成. 心理科学进展, 2004,12(4): 481 - 488)

[10] Ostendorf M, Veilleux N. A Hierarchical Stochastic Model for Automatic Prediction of Prosodic Boundary Location. Computational Linguistics, 1994, 20(1): 27 - 54