面向口语统计语言模型建模的自动语料生成算法

司玉景 [1] 肖业鸣 [1] 徐 及 [1] 潘接林 [1] 颜永红 [1]

摘 要 在资源相对匮乏的自动语音识别 (Automatic speech recognition, ASR) 领域, 如面向电话交谈的语音识别系统中, 统计语言模型 (Language model, LM) 存在着严重的数据稀疏问题. 本文提出了一种基于等概率事件的采样语料生成算法, 自动生成领域相关的语料, 用来强化统计语言模型建模. 实验结果表明, 加入本算法生成的采样语料可以缓解语言模型的稀疏性, 从而提升整个语音识别系统的性能. 在开发集上语言模型的困惑度相对降低 7.5 %, 字错误率 (Character error rate, CER) 绝对降低 0.2 个点; 在测试集上语言模型的困惑度相对降低 6 %, 字错误率绝对降低 0.4 点.

关键词 自动语音识别, 资源匮乏, 语言模型, 等概率事件, 语料生成算法

引用格式 司玉景, 肖业鸣, 徐及, 潘接林, 颜永红. 面向口语统计语言模型建模的自动语料生成算法. 自动化学报, 2014, **40**(12): 2808−2814

DOI 10.3724/SP.J.1004.2014.02808

Automatic Text Corpus Generation Algorithm towards Oral Statistical Language Modeling

SI Yu-Jing[1] XIAO Ye-Ming[1] XU Ji[1] PAN Jie-Lin[1] YAN Yong-Hong[1]

Abstract Data sparseness is a serious issue for language model (LM) in automatic speech recognition (ASR) towards resource-lack domains, e.g. the telephone conversation speech recognition task. In this paper, an event of equal probability based text corpus generation algorithm is proposed in order to alleviate the sparseness of language model. Experimental results show that 7.5 % relative reduction in perplexity and a 0.2 % absolute reduction in character error rate (CER) can be obtained on the develop set. And, a 6 % relative reduction in perplexity and a 0.4 % absolute reduction in CER can be obtained on the test set.

Key words Automatic speech recognition (ASR), resource-lack, language model (LM), equality probability event, text corpus generation

Citation Si Yu-Jing, Xiao Ye-Ming, Xu Ji, Pan Jie-Lin, Yan Yong-Hong. Automatic text corpus generation algorithm towards oral statistical language modeling. *Acta Automatica Sinica*, 2014, **40**(12): 2808−2814

自动语音识别技术 (Automatic speech recognition, ASR) 的研究目标是将人类的语音逐词 (字) 逐句的转化为相应的书面语言 (即文字). 近年来, 随着软硬件技术的不断发展, 自动语音技术发展迅速, 在工业、家电、通信、汽车电子、医疗、家庭服务、消费电子产品、国防等各个领域得到了广泛的应用. 大词汇连续语音识别系统处理的对象是大规模的音频数据, 广泛应用在电话监听、通信、嵌入式设备智能控制等领域. 如日本和欧洲有很多研究机构计划实现两种语言之间的直接通信, 即通过 "语音识别 — 机器翻译 — 语音合成" 将一种语言直接转成另一种语言, 用来消除不同语种间人们之间沟通的语言障碍[1].

语言模型 (Language model, LM) 在语音识别系统中的作用是非常显著的. 由于声学信号的动态时变、瞬时和随机性, 单靠声学模式的匹配与判断不可能完成语音无误的识别和理解. 一些较高层次的语言知识的利用可以在声学识别的层次上减少模式匹配的模糊性, 从而提高识别的准确性. 目前, 比较流行的语言模型是 N 元文法语言模型. N 元文法语言模型构造直接简单、计算复杂度低, 可以利用大规模语料充分训练模型的参数, 但是当训练语料不

收稿日期 2013-12-18 录用日期 2014-06-03
Manuscript received December 18, 2013; accepted June 3, 2014
国家高技术研究发展计划 (863 计划) (2012AA012503), 国家自然科学基金 (10925419, 90920302, 61072124, 11074275, 11161140319, 91120001, 61271426), 中国科学院战略性先导科技专项 (XDA06030100, XDA06030500), 中国科学院重点部署项目 (KGZD-EW-103-2) 资助
Supported by National High Technology Research and Development Program of China (863 Program) (2012AA012503), National Natural Science Foundation of China (10925419, 90920302, 61072124, 11074275, 11161140319, 91120001, 61271426), the Strategic Priority Research Program of Chinese Academy of Sciences (XDA06030100, XDA06030500), and the Chinese Academy of Sciences Priority Deployment Project (KGZD-EW-103-2)
本文责任编委 吴玺宏
Recommended by Associate Editor WU Xi-Hong
1. 中国科学院声学研究所语言声学与内容理解重点实验室 北京 100190
1. The Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190

足的时候, N 元文法语言模型面临着严重的数据稀疏问题[2]. 即便采用一些平滑算法, 数据稀疏问题依然很严重. 如面向电话交谈任务的语音识别领域, 特定领域的语料主要是通过人工标注获取的, 而人工标注的成本高并且效率低下, 获取的语料规模有限. 因此, 面向电话交谈任务的 N 元文法语言模型面临着严重的数据稀疏问题.

目前主要存在三种途径可以在一定程度上缓解语言模型的数据稀疏性. 1) 借助于互联网. 由于互联网上存在大量的文本语料, 我们可以利用这些文本语料构建语言模型, 然后和领域内的语言模型插值. 然而, 互联网上的文本语料大多都是 "书写式" 的语料, 有时并不能满足特定领域的需求. 比如, 面向电话交谈的语音识别任务需要的是 "交谈式" 的语料, 它风格上和网络语料有很大的差异, 因此, 网络语料对于提升面向电话交谈任务语音识别系统性能的能力是有限的. 2) 采用泛化能力和表达能力更强的语言模型建模技术来提升语言模型对未知数据的鲁棒性, 比如基于贝叶斯的语言模型[3]、最大熵语言模型[4]、前向神经网络语言模型 (Feed-forward neural network language model, FFNNLM)[5]、递归神经网络语言模型 (Recurrent neural network language model, RNNLM)[6] 等. 尤其是递归神经网络语言模型, 它在很多标准集子上, 比如 WSJ、TIMIT 等, 相比 N 元文法语言模型, 都有很大的提升[7]. 虽然这些语言模型建模技术的性能优于 N 元文法语言模型, 但是它们的计算复杂度远远高于 N 元文法语言模型, 在解码阶段直接应用这些高级语言模型建模技术会严重影响解码速度. 因此, 这些计算复杂度较高的语言模型建模技术普遍应用在后处理阶段, 如词图重估阶段[8]. 3) 通过生成语料的方式来解决数据稀疏问题. 比如, 可以通过半自动扩展方法来扩充语言模型的训练语料[9], 通过扩展原始语料中的同义词和词类来达到扩展训练语料的目的. 然而, 这种方法需要对扩展结果进行相应的人工删减, 并且其效果与同义词和词类的质量息息相关.

本文提出了一种基于等概率事件的采样语料生成算法, 可以自动生成特定领域风格的文本语料. 算法的基本思想是: 首先, 利用递归神经网络技术 (Recurrent neural network, RNN) 对领域内原始语料进行建模; 然后, 再利用基于等概率事件的语料生成算法生成服从 RNNLM 概率分布的采样语料, 利用这些采样语料构建 N 元文法语言模型; 最后, 再和网络语料语言模型和标注语料语言模型进行插值. 由于递归神经网络语言模型的训练数据是特定领域内的语料, 因此, 采样语料也是符合特定领域风格的语料. 这样, 加入这些采样语料可以缓解资源匮乏领域中 N 元文法语言模型面临的数据稀疏问题, 从而进一步提升语音识别系统的性能.

本文按照下列结构组织: 第 1 节介绍递归神经网络语言模型的建模技术; 第 2 节详细介绍本文提出的基于等概率事件的语料生成算法以及在 RNNLM 框架下自动语料生成算法的具体流程; 之后在第 3 节介绍基于多语言模型融合的采样语料生成算法; 第 4 节是具体的实验结果和分析; 最后, 第 5 节给出结论.

1 递归神经网络语言模型

基于递归神经网络的语言模型首先将词历史信息投射到连续空间上, 然后将相似的词历史在连续空间上进行聚类, 以提高处理未知数据的泛化能力[6]; 另一方面, 递归神经网络语言模型可以利用长历史词信息来估计当前词的概率, 同时训练参数不会急剧增长. 借助递归神经网络架构的短时记忆功能, 可以利用无限长的词历史信息来估计当前词的概率[10−11].

基于递归神经网络的语言模型的架构图如图 1 所示. 它主要包含三个部分: 输入层、隐藏层、输出层. 各部分的数学表达式为

$$\text{输入层: } x(t) = w(t) + s(t-1) \tag{1}$$

$$\text{隐藏层: } s_j(t) = f\left(\sum_i x_i(t) w_{ji}\right) \tag{2}$$

$$\text{输出层: } y_k(t) = g\left(\sum_j s_j(t) w_{kj}\right) \tag{3}$$

其中, $f(z)$ 为 Sigmoid 激活函数: $f(z) = 1/(1+\mathrm{e}^{-z})$. $g(z)$ 为概率规整函数: $g(z_m) = \mathrm{e}^{z_m}/\sum_k \mathrm{e}^{z_k}$

输入层 $x(t)$ 包括前一个词编号 $w(t)$ 和词历史信息 $s(t-1)$, 如式 (1) 所示. 词编号 $x(t)$ 采用 "N 中选一" 编码模式, 采用长度为 V (字典中的词个数) 的单位矢量表示, 其中, 词在字典中的编号位为 1, 其他位为 0; 词历史信息 $s(t-1)$ 是神经网络前向计算隐藏层的输出值. 隐藏层 $s(t)$ 是词历史信息在连续空间上的表示, 可以回馈给输入层, 作为下一个词的输入, 计算公式如式 (2) 所示. 输出层节点为给定词历史, 下一个词出现的概率, 计算公式如式 (3) 所示. 本文采用基于类的递归神经网络语言模型建模技术[11], 包含词部分和类部分. 词的类别可以根据训练语料中的词频得到[11], 也可以采用更精确的词聚类算法, 比如布朗算法[12]. 定义 $P(c_i)$ 为类 c_i 的概率; $P(w_i|c_i)$ 为词

w_i 在类 c_i 的概率. 那么, 给定词历史 $w_{i-1},\cdots,w_1$, 当前词 w_i 的概率为: $P(w_i|w_{i-1},\cdots,w_1) = P(c_i|w_{i-1},\cdots,w_1)P(w_i|c_i)$.

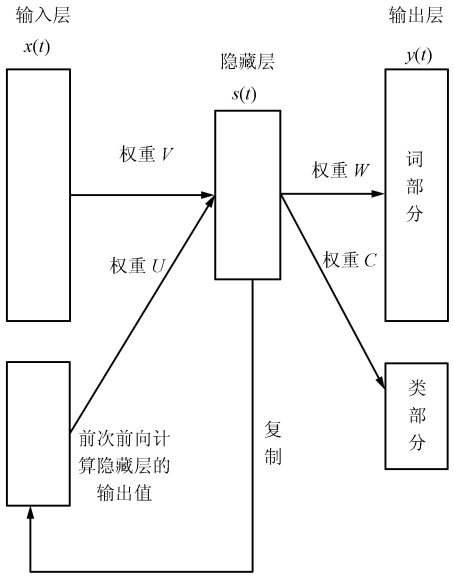

图 1 递归神经网络语言模型框架

Fig. 1 The framework of RNNLM

2 基于等概率事件的采样语料生成算法

如图 2 所示, 我们定义两个事件.

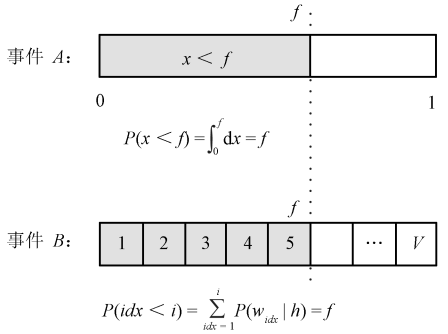

图 2 基于等概率事件的采样语料生成算法原理图

Fig. 2 Sampling text generation algorithm based on equality probability event

1) 事件 A: 随机变量 X 小于 f, 其中 f 为 0 到 1 的小数, X 服从均匀分布, 那么事件 A 发生的概率为: $P(A) = P(x < f) = \int_0^f \mathrm{d}x = f$

2) 事件 B: V 为语音识别系统中字典的大小, 给定词历史 h, 可以通过统计的方法赋予每个词一个概率 p, 即 $p = p(w_{idx}|h)$, 其中 idx 为该词在字典中的索引, 事件 B 定义为词索引 idx 小于词索引 i. 则事件 B 的发生概率为

$$P(idx < i) = \sum_{idx=1}^{i} p(w_{idx}|h) = f \tag{4}$$

首先假设: 事件 A 和事件 B 的发生概率如果相同, 那么两者便会同时发生. 因此, 首先按照均匀分布生成随机数 f, 产生事件 A, 然后去找词累加概率大于等于 f 的词标号, 生成事件 B, 那么, 该词标号便为给定词历史 h 后的一个输出采样.

RNNLM 框架下基于等概率事件的采样语料生成算法流程如图 3 所示. 其中, Num 表示当前生成的词数目, targetNum 表示期望生成的词数目. 函数 random(0, 1) 的功能是生成服从均匀概率分布的随机数, 范围是从 0 到 1, 用来生成图 3 中的事件 A 的发生概率. 得到事件 A 的发生概率以后, 利用式 (6) 寻找使得事件 B 的发生概率等于事件 A 的发生概率时所对应的词标号. RNNLM 框架下基于等概率事件的采样语料生成算法主要包含三个步骤:

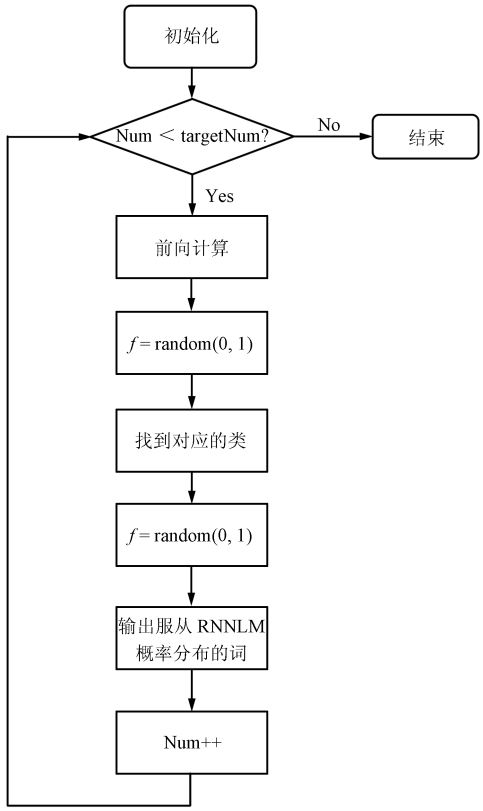

图 3 基于递归神经网络语言模型的自动语料生成算法

Fig. 3 Automatic text generation algorithm under the framework of RNNLM

1) 如果当前词数量 (Num) 小于目标词数量 (targetNum), 则进行递归神经网络语言模型的前向计算, 然后按照均匀分布生成 0 到 1 之间的随机数

f, 计算类的累加概率 g, 直到 $g \geq f$, 得到当前的类 c;

2) 重新按照均匀概率分布生成随机数 f, 在当前的类 c 中计算词的累加概率 g, 直到 $g \geq f$, 然后将当前词作为输出;

3) 转到步骤 1), 将当前词作为下一次前向计算的输入.

3 基于多语言模型融合的采样语料生成算法

文献 [7] 显示, 通过多个语言模型融合, 可以提升语言模型的性能. 受此启发, 本文期望通过融合多个语言模型产生的采样语料, 进一步提升面向电话交谈任务的语言模型. 算法流程如图 4 所示. 首先利用特定领域内语料训练多个性能较好、但计算复杂度较高的语言模型; 然后采用本文提出的基于等概率事件的采样语料生成算法分别生成各个语言模型对应的采样语料; 之后分别训练采样语料对应的 N 元文法语言模型; 最后, 在开发集上计算最优的插值系数, 得到最终的模型. 由于各个语言模型的建模方法不同, 它们生成的采样语料可以相互补充, 因此, 其效果要优于只利用单个语言模型的采样语料.

4 实验结果和分析

4.1 实验配置

本文在资源相对匮乏的面向电话交谈的语音识别任务上评测基于等概率事件的语料生成算法. 测试环境采用实验室自主开发的大词汇连续语音识别系统, 该系统是采用基于紧致状态网络的单遍解码器, 其他参数详见文献 [13]. 开发集采用在线录制的电话交谈风格的语音, 共 2 个小时, 用来确定 N 元文法语言模型的插值系数以及递归神经网络语言模型训练中的迭代次数等. 另外, 采用时长 3 个小时的电话交谈风格的语音作为 “测试集”.

基线语言模型训练语料包括网络语料和电话交谈领域语料两部分. 其中, 网络语料是从万维网上下载的语料, 总共包含 20 亿个词. 电话交谈领域语料为人工标注的电话交谈风格的语料, 该语料为纯粹的自然交谈风格, 主题不限, 内容和说话方式也没有任何限制, 总共 1 300 万个词. 本文使用 RNNLM Toolkit 训练递归神经网络语言模型. 由于训练递归神经网络语言模型的计算复杂度过高, 我们只利用电话交谈领域内语料训练递归神经网络语言模型. 关于递归神经网络语言模型的训练参数设置方法可以参考文献 [14].

4.2 实验结果

我们分别训练了隐层为 100, 200, 300, 400, 500, 600 的基于词的递归神经网络语言模型以及一个隐层为 600 的基于字的递归神经网络语言模型. 基于等概率事件的采样语料生成算法, 得到电话交谈风格的采样语料, 附录 A 和附录 B 为生成的采样语料实例. 从采样语料实例中可以看出, 有些句子并不通顺, 但是考虑到 N 元文法语言模型建模中只利用前 $N-1$ 个词历史, 因此并不会影响最终模型的性能. 用采样语料训练得到的 3 元文法语言模型, 在开发集上的困惑度如表 1 所示. 可以看出, 各个模型产生的采样语料是有互补性的. 下面我们的实验分别从命中率指标、困惑度指标以及字错误率指标出发来评估基于等概率事件的采样语料算法的性能表现.

4.2.1 命中率指标[15]

表 2 是各个语言模型在开发集和测试集上命中率的表现. 以开发集为例, 基于特定领域内语料的语言模型的 3 元文法命中率只有 38 %. 和网页语言模型插值以后, 可以大大地提高 3 元文法的命中率 (64.6 %). 最后, 利用采样语料, 可以进一步提高 3 元文法的命中率 (70.2 %). 为了验证算法的鲁棒性, 我们在测试集上也做了相同的实验, 实验结果如表 2 所示, 加入本算法自动生成的语料以后, 3 元文法的命中率从 68.2 % 提升到了 74.3 %.

4.2.2 语言模型困惑度指标

如表 3 所示, 利用采样语料以后, 语言模型在开发集和测试集上的困惑度分别相对降低了 7.5 % 和 6 %.

4.2.3 字错误率指标

以字错误率为衡量指标, 测试了算法生成的语料在面向电话交谈任务的语音识别系统中的表现, 实验结果如表 4 所示. 可见, 在网络语料和标注语料的基础上增加本算法生成的采样语料, 可以进一步提升语音识别系统的性能. 相比基线系统, 在开发集上字错误率绝对降低了 0.2 个点, 在测试集上字错误率绝对降低了 0.4 个点. 虽然提升的幅度不是很大, 但是字错误率在测试集和开发集上都有一致的降低.

5 结论

本文提出了一种基于等概率事件的语料生成算法, 用来进一步缓解文本语料匮乏领域中语言模型的稀疏问题. 假设: 事件 A 和事件 B 的发生概率相同, 那么事件 A 发生了, 事件 B 也会发生. 在递归

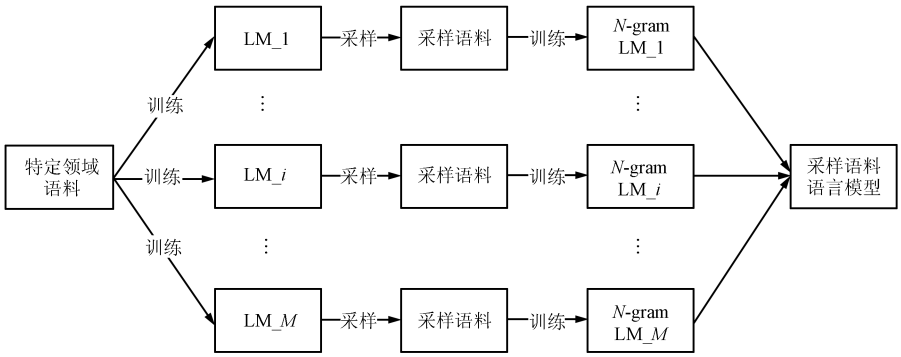

图 4 基于多语言模型融合的采样语料生成算法

Fig. 4 Sampling text generation algorithm based on multiple language model combination

表 1 各个采样语料语言模型在开发集上困惑度的表现

Table 1 The performance of sampling language model on the development set in term of perplexity

类型	模型	规模 (词总数)	语言模型文法总数	困惑度
基于词的递归神经网络语言模型	RNNLM.h100. 采样语料	77 980 138	35 235 024	257.37
	RNNLM.h200. 采样语料	78 143 300	33 957 611	224.37
	RNNLM.h300. 采样语料	76 573 019	33 326 799	239.20
	RNNLM.h400. 采样语料	78 152 499	33 965 205	218.23
	RNNLM.h500. 采样语料	76 360 827	33 781 465	237
	RNNLM.h600. 采样语料	78 081 537	33 721 658	216.67
基于字的递归神经网络语言模型	RNNLM.h600. 采样语料	54 998 697	26 963 296	233.12
	以上采样语料插值			203

[1] 首先将字合并, 然后用前向最大匹配分词, 最后训练 N 元文法语言模型.

表 2 自动生成的语料对命中率的影响 (%)

Table 2 The performance of sampling corpus in terms of hit-rate (%)

模型	开发集			测试集		
	1-gram	2-gram	3-gram	1-gram	2-gram	3-gram
标注语言模型	10.1	51.9	38.0	8.4	49.8	41.8
+ 网页语言模型	0.7	34.7	64.6	0.5	31.2	68.2
+ 自动生成的语料语言模型	0.4	27	72.6	0.3	25.4	74.3

表 3 语言模型困惑度表现

Table 3 The performance of sampling model in terms of perplexity

语言模型	开发集	测试集
标注语言模型	232	195
+ 网页语言模型	198	151
+ 采样语料语言模型	183	142

表 4 字错误率表现

Table 4 The performance of sampling model in terms of character error rate

	开发集	测试集
基线系统	48.6	54.1
+ 采样语料语言模型	48.4	53.7

神经网络语言模型的框架下利用等概率事件算法生成采样语料. 首先, 对电话交谈任务的标注语料进行建模; 然后, 利用基于等概率事件的采样语料生成算法获取采样语料, 利用这些采样语料构建 N 元文法语言模型; 最后, 再和网络语料训练的语言模型和标注语料训练的语言模型进行插值. 实验结果表明, 基于等概率事件的采样语料生成算法不仅可以降低语言模型的困惑度, 而且可以提升语音识别系统的整体性能. 另外, 本算法也同样适用于其他高级语言模型建模技术, 比如最大熵语言模型.

附录 A

想 干 妈 一 直 说 我 出 生 什 么 意 见
我 不 念 那 时 候 给 补 给 他 呢
那 干 啥 的 你 你 儿 子 警 要 是 还 跟
我 打 电 话 叫 她 不 介
位 子 放 在 家 里 呢 也 得
闲 到 地 里 我 怕
不 让 邮 汇 那 钱
告 诉 我 把 他 误 解 浪 费
好 像 哦 哦 我
是 吗
恩 很 少
据 说 是 二 十 磅 呢
一 个 是
那 你 不 是 那 种 你 那 个 琳 达 还 告 诉 我
对 呀 我 现 在 就 犹 豫 的 话

附录 B

过生日 了
要 找 工作 赶紧 退 呀
那 不成 就 办了 很高 第一
才 啥 时候 呢
反正 是 几 号 走 那 在 学校 那 事 那
他 也不 是 也不 用 提 一些 事情 的 问 吧
嗯 没有 问题 呀
现在 就是 都有 钱 了 看 环境 速度 太 慢 了
钱 要 上课 呀
嗯 呐 看看 那个 还想 不想 回家 明天 行了
啊 那 你 那天 把 水果 乔 盛 那两 六十
啊 啊 我 说 我 是 不爱 吃
啊 看这 个 东西 基本上 是
你们 要 那个 有时候 就 就 开始 累 得 很
他 是 租 的 是不是
人家 都是 去 温州 人 嘛 嗯 嗯
妈妈 的 我 一个 棒
对 哎哟 他 要不 要 问 一下 你 啊 有时候 你 自己 去
之前 你 这 有 这么 急 吗 刚
那个 老 姨 姐 幸好 没 给我 打了 几个 电话
那 什么 材料 呢 一般 也不 能 付 呗 写信 给 这个 人

References

1 Yang Xing-Jun, Chi Hui-Sheng. *Digital Processing of Speech Signals*. Beijing: Electronic Industry Press, 1995. 330−331
(杨行竣, 迟惠生. 语音信号数字处理. 北京: 电子工业出版牡, 1995. 330−331)

2 Chen S F, Goodman J. An empirical study of smoothing techniques for language modeling. In: Proceedings of the 34th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics. Santa Cruz, CA, 1996. 310−318

3 Allauzen C, Riley M. Bayesian language model interpolation for mobile speech input. In: Proceedings of the 2011 Interspeech. Italy, 2011. 1429−1432

4 Khudanpur S, Wu J. A maximum entropy language model integrating n-grams and topic dependencies for conversational speech recognition. In: Proceedings of the 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Phoenix, AZ: IEEE, 1999. 553−556

5 Schwenk H. CSLM — a modular open-source continuous space language modeling toolkit. In: Proceedings of the 2013 Interspeech. Lyyon, France, 2013. 1198−1202

6 Mikolov T, Karafiát M, Burget L, Černocký J H, Khudanpur S. Recurrent neural network based language model. In: Proceedings of the 2010 INTERSPEECH. Lyon, France: ISCA, 2010. 1045−1048

7 Mikolov T, Deoras A, Kombrink S, Burget L, Cernocky J H. Empirical evaluation and combination of advanced language modeling techniques. In: Proceedings of the 2011 Interspeech. Italy, 2011. 605−608

8 Liu X, Wang Y, Chen X, Gales M J F, Woodland P C. Efficient lattice rescoring using recurrent neural network language models. In: Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). FLORENCE, ITALY, 2014. 4941−4945

9 Huang Yun-Zhu, Wei Wei, Luo Yang-Yu, Li Cheng-Rong. Word-class expansion method about training corpus of language modal in restrcited domain. *Application of Computer System*, 2011, **20**(11): 55−58
(黄韵竹, 韦玮, 罗杨宇, 李成荣. 限定领域语言模型训练语料的词类扩展方法. 计算机系统应用, 2011, **20**(11): 55−58)

10 Bengio Y, Boulanger-Lewandowski N, Pascanu R. Advances in optimizing recurrent networks. In: Proceedings of the 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Vancouver, Canada: IEEE, 2013. 8624−8628

11 Sutskever Ilya. Training Recurrent Neural Networks [Ph. D. dissertation], University of Toronto, Canada, 2013.

12 Si Y J, Zhang Z, Li T, Pan J, Yan Y. Enhanced word classing for recurrent neural network language model. *Journal of Information & Computational Science*, 2013, **10**(12): 3595−3604

13 Shao J, Li T, Zhang Q Q, Zhao Q W, Yan Y H. A one-pass real-time decoder using memory-efficient state network. *IEICE Transactions on Information and Systems*, 2008, **1**(91): 529−537

14 Mikolov T, Kombrink S, Deoras A, Burget L, Cernocky J H. RNNLM-Recurrent neural network language modeling toolkit. In: Proceedings of the 2011 IEEE Workshop on Automatic Speech Recognition and Understanding, UK, 2011. 16−19

15 Shao Jian. Chinese Spoken Term Detection towards Large-Scale Telephone Conversational Speech [Ph. D. dissertation]. Institute of Acoustics, Chinese Academy of Sciences, China, 2008.
(邵建. 面向大规模电话交谈语音的汉语语音检索 [博士学位论文], 中国科学院声学研究所, 中国, 2008.)

司玉景 中国科学院声学研究所博士研究生. 2009 年获得吉林大学通信工程学院信息工程系学士学位. 主要研究方向为统计语言模型建模, 语音识别解码技术, 机器学习, 深度神经网络技术, 自动语音文本同步技术. 本文通信作者.
E-mail: siyujinglj@126.com
(**SI Yu-Jing** Ph. D. candidate at the Institute of Acoustics, Chinese Academy of Sciences. He received his bachelor degree from Ji Lin University in 2009. His research interest covers statistical language modeling, speech recognition, machine learning, deep neural network, and automatic speech text synchronization technology. Corresponding author of this paper.)

肖业鸣 中国科学院声学研究所博士研究生. 2008 年获得北京航空航天大学学士学位. 主要研究方向为大词汇量连续语音识别, 深度学习和神经网络技术.
E-mail: xiaoyeming@hccl.ioa.ac.cn
(**XIAO Ye-Ming** Ph. D. candidate at the Institute of Acoustics, Chinese Academy of Sciences. He received his bachelor degree from Beihang University in 2008. His research interest covers automatic speech recognition, deep learning, and neural network.)

徐 及 中国科学院声学研究所博士研究生. 2008 年和 2011 年分别在清华大学电子工程系获得学士和硕士学位. 主要研究方向为大词汇量连续语音识别, 深度学习和语音合成.
E-mail: xuji@hccl.ioa.ac.cn
(**XU Ji** Ph. D. candidate at the Institute of Acoustics, Chinese Academy of Sciences. He received his bachelor and master degrees from the Department of Electronic Engineering, Tsinghua University in 2008 and 2011, respectively. His research interest covers large vocabulary continuous speech recognition, deep learning, and speech synthesis.)

潘接林 教授. 1986 获得北京大学的学士学位, 1989 年获得清华大学电子工程的硕士学士. 从 2000 年 1 月至 2001 年 7 月, 他担任英特尔中国研究中心高级研究员和语音识别项目组经理. 2002 年 12 月加入中国科学院声学研究所语音声学和内容理解重点实验室. 主要研究方向为语音识别, 语音分析, 声学模型和搜索算法. E-mail: panjielin@hccl.ioa.ac.cn
(**PAN Jie-Lin** Professor. He received his bachelor degree from Peking University in 1986 and his master degree in electronic engineering from Tsinghua University in 1989. From 2000 to 2001 he was working in Intel China Research Center as a senior researcher and speech recognition group manager. In 2002 he joined the Key Laboratory of Speech Acoustics and Content Understanding, Chinese Academy of Sciences. His research interest covers speech recognition,, speech analysis, acoustic model, and search algorithm.)

颜永红 中国科学院声学研究所语言声学与内容理解重点实验室教授. 1990 年在清华大学获得学士学位, 1995 年 8 月于美国俄勒冈研究院 (Oregon Graduate Institute, OGI) 获计算机科学和工程博士学位. 他曾在 OGI 担任助理教授 (1995), 副教授 (1998) 和副主任 (1997). 从 1998 年到 2001 年, 在英特尔担任人机界面研究委员会主席, 微处理器研究实验室主任和英特尔中国研究中心的首席工程师. 主要研究方向为语音处理和识别, 语言/说话人识别, 人机界面.
E-mail: yanyonghong@hccl.ioa.ac.cn
(**YAN Yong-Hong** Professor at the Key Laboratory of Speech Acoustics and Content Understanding, Chinese Academy of Sciences. He received his bachelor degree from Tsinghua University in 1990, and his Ph. D. degree from Oregon Graduate Institute (OGI). He worked in OGI as assistant professor (1995), associate professor (1998) and associate director (1997) of Center for Spoken Language Understanding. He worked in Intel from 1998 ~ 2001, chaired Human Computer Interface Research Council, worked as Principal Engineer of Microprocessor Research Laboratory and Director of Intel China Research Center. His research interest covers speech processing and recognition, language/speaker recognition, and human computer interface.)