# 基于改进主题分布特征的神经网络语言模型

刘 畅[①②] 张一珂[①②] 张鹏远*[①②] 颜永红[①②③]
[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)
[②](中国科学院大学 北京 100049)
[③](中国科学院新疆理化技术研究所新疆民族语音语言信息处理实验室 乌鲁木齐 830011)

**摘 要**：在递归神经网络(RNN)语言模型输入中增加表示当前词所对应主题的特征向量是一种有效利用长时间跨度历史信息的方法。由于在不同文档中各主题的概率分布通常差别很大，该文提出一种使用文档主题概率改进当前词主题特征的方法，并将改进后的特征应用于基于长短时记忆(LSTM)单元的递归神经网络语言模型中。实验表明，在 PTB 数据集上该文提出的方法使语言模型的困惑度相对于基线系统下降 11.8%。在 SWBD 数据集多候选重估实验中，该文提出的特征使 LSTM 模型相对于基线模型词错误率(WER)相对下降 6.0%；在 WSJ 数据集上的实验中，该特征使 LSTM 模型相对于基线模型词错误率(WER)相对下降 6.8%，并且在 eval92 测试集上，改进隐含狄利克雷分布(LDA)特征使 RNN 效果与 LSTM 相当。

**关键词**：语音识别；语言模型；隐含狄利克雷分布；长短时记忆
**中图分类号**： TP391.42 **文献标识码**： A **文章编号**：1009-5896(2018)01-0219-07
**DOI**: 10.11999/JEIT170219

# Neural Network Language Modeling Using an Improved Topic Distribution Feature

LIU Chang[①②] ZHANG Yike[①②] ZHANG Pengyuan[①②] YAN Yonghong[①②③]
[①](*Key Laboratory of Speech Acoustics and Content Understanding*, *Institute of Acoustics*, *Chinese Acadamy of Sciences*, *Beijing* 100190, *China*)
[②](*University of Chinese Academy of Sciences*, *Beijing* 100049, *China*)
[③](*Xinjiang Laboratory of Minority Speech and Language Information Processing*, *Xinjiang Technical Institute of Physics and Chemistry*, *Chinese Academy of Sciences*, *Urumqi* 830011, *China*)

**Abstract**: Attaching topic features to the input of Recurrent Neural Network (RNN) models is an efficient method to leverage distant contextual information. To cope with the problem that the topic distributions may vary greatly among different documents, this paper proposes an improved topic feature using the topic distributions of documents and applies it to a recurrent Long Short-Term Memory (LSTM) language model. Experiments show that the proposed feature achieved an 11.8% relatively perplexity reduction on the Penn TreeBank (PTB) dataset, and reached 6.0% and 6.8% relative Word Error Rate (WER) reduction on the SWitch BoarD (SWBD) and Wall Street Journal (WSJ) speech recognition task respectively. On WSJ speech recognition task, RNN with this feature can reach the effect of LSTM on eval92 testset.

**Key words**: Speech recognition; Language model; Latent Dirichlet Allocation (LDA); Long Short-Term Memory (LSTM)

## 1 引言

近年来，递归神经网络(RNN)语言模型(LM)[1]已经成为语言模型常见的实现方法之一，它使语言模型在困惑度和语音识别整体错误率方面都有显著改善[2]。与典型前向反馈神经网络相比，RNN 通过隐含层的节点来得到历史信息，从而提高模型对历史信息的利用率。然而随时间距离当前时刻越远，梯度逐层变小，模型在时间维度上会出现梯度消失问题，因此通过 RNN 学习到长时间跨度信息并不容易。另一种常用的语言模型实现方式为长短时记忆(Long Short-Term Memory, LSTM)神经网络[3]，它通过在 RNN 结构中引入 LSTM 单元来改善前面提到的梯度消失问题。

收稿日期：2017-03-17；改回日期：2017-10-06；网络出版：2017-10-27
*通信作者：张鹏远 pzhang@hccl.ioa.ac.cn
基金项目：国家自然科学基金(11590770-4, U1536117, 11504406, 11461141004)，国家重点研发计划重点专项(2016YFB0801203, 2016YFB0801200)，新疆维吾尔自治区科技重大专项(2016A03007-1)
Foundation Items: The National Natural Science Foundation of China (11590770-4, U1536117, 11504406, 11461141004), The National Key Research and Development Plan (2016YFB0801203, 2016YFB0801200), The Key Science and Technology Project of the Xinjiang Uygur Autonomous Region (2016A03007-1)

另一方面，一种简单的增加历史信息影响的技巧是增加缓存信息[4]。该方法的一种常见形式为在通用语言模型中，根据词在固定长度历史内的出现情况动态地调整语言模型中的概率分布。除此之外还有利用隐含语义分析(LSA)来改进语言模型的方法[5]。这类模型将当前词和长时间历史信息分别表示为隐含语义空间中的向量，并通过两个向量在隐含语义空间中向量的余弦相似度得到预测词的概率，最后将该概率与 $n$ 元文法($n$-gram)语言模型概率的插值结果作为最终结果。但是由于 LSA 算法本身对一词多义及一义多词问题上的处理能力较弱，该模型性能对历史信息的利用并不充分。

Khudanpur 等人[6]提出在最大熵框架中使用主题特征，并且提出在语音搜索任务中使用表示当前词是否在用户的输入历史中出现过的特征来改进的最大熵模型。最后，Lau 等人[7]提出一种基于句子的模型，它通过判断词对是否在在一句话中同时出现来产生一个触发特征，进而达到使用长上下文信息的目的。这几种方法只利用到历史中同一词多次出现的信息，没有利用到更深层的语义信息。

而对于主题信息来说，通常的主题语言模型[8]把数据分成若干子集，使每个子集中的数据都只包含一个主题，再使用不同子集训练不同的小语言模型，解码时使用最相关的小语言模型与采用所有数据训练的通用语言模型插值得到最终结果。这类方法需要对数据集进行拆分，因而导致各小语言模型的训练数据都相对较少，加重了训练语料的稀疏性。

Mikolov 等人[9]提出在 RNN 上增加当前时刻的主题信息，并把该信息作为特征分别连到 RNN 的输入层和输出层来增加上下文对模型的影响。该方法在使用主题信息的同时避免了文献[8]中数据的拆分问题，并且考虑到了历史信息的影响会随离当前词距离增大而衰弱的现象。但该方法使用在计算词的主题分布时假设所有主题均匀分布，即仅使用该词在所有训练语料上的主题分布，没有考虑到不同语境的影响，因此对当前词的主题特征估计不够准确。除该方法外，也有使用当前词的词向量[10]或当前词词性标注[11]作为特征加入到 RNN 模型中工作。其中词向量为词的一种低维实数向量表示，可以使词义相似的词在词向量空间上更接近，并体现出相邻词之间的关系，因此该向量可以表示一定的历史信息；词性则可以在一定程度上表示一些语法信息，从而也是一种对历史信息的利用。但这两种方法都只考虑到了当前词该特征所携带的信息，而缺少文献[9]中对历史词特征的利用。

除文献[10]外，国内在语言模型上的工作还有左玲云等人[12]在中文电话交谈语音任务中使用基于 LSTM 结构的深度神经网络语言模型对多候选进行重评估来增加历史信息的影响，并讨论了重评估过程中使用不同长度历史的影响；王龙等人[13]提出一种基于批处理(mini-batch) 的并行优化训练算法.，该算法利用 GPU 的计算能力来提高网络训练时的矩阵及向量运算速度，优化后的网络能同时并行处理多个数据流即训练多个句子样本，加速训练过程等。

本文的工作在文献[9]的基础上进行，首先提出一种特征计算方法的改进方法，再将几种特征应用在 LSTM 语言模型上，最后在 PTB, SWBD 和 WSJ 数据集上通过实验测试新模型的性能。

## 2 模型结构

最简单的 RNN 语言模型[1]由输入层、包含循环连接的隐含层、输出层和连接他们的权重矩阵组成。通常情况下，模型会在隐含层前增加一个投影层来降低输入维度，目前主流做法是将稀疏向量换为词矢量(word embedding)[14,15]，使相邻近的词在词矢量空间中的位置相近。

LSTM 模型通过增加“门”来控制将信息加载到节点的能力。简单的 LSTM 在 RNN 的基础上增加了输入门、遗忘门和输出门，来分别控制输入、前一时刻节点的历史信息和输出信息的流通。

本文在 RNN 及 LSTM 模型中分别加入一个特征层 $\boldsymbol{g}_t$，并将它分别连向 RNN 层和输出层，如图 1 所示。特征层的输入为当前时刻的主题信息。

本文 RNN 模型隐含层及输出层节点的计算公式如式(1)~式(3)：

$$\boldsymbol{x}_t = \boldsymbol{P}\boldsymbol{w}_t + \boldsymbol{F}\boldsymbol{g}_t \tag{1}$$

$$\boldsymbol{s}_t = \sigma\left(\boldsymbol{U}\boldsymbol{x}_t + \boldsymbol{W}\boldsymbol{s}_{t-1}\right) \tag{2}$$

$$\boldsymbol{y}_t = \varphi\left(\boldsymbol{V}\boldsymbol{s}_t + \boldsymbol{G}\boldsymbol{g}_t\right) \tag{3}$$

式中，$\sigma$，$\varphi$ 分别表示 sigmoid、softmax 激活函数，

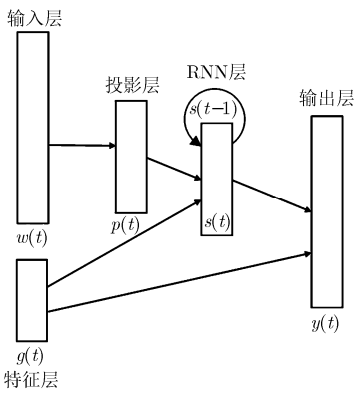


图 1 含有特征层的 RNN 或 LSTM 模型

$\boldsymbol{x}_t$ 为神经网络的输入向量，由当前词的 one-hot 向量 $\boldsymbol{w}_t$ 的投影和使用 $F$ 矩阵加权后的当前时刻的特征 $\boldsymbol{g}_t$ 相加得到。$\boldsymbol{s}_t$ , $\boldsymbol{s}_{t-1}$ 分别表示当前时刻、前一时刻神经网络单元的值。$\boldsymbol{y}_t$ 为输出向量，即预测下一个词的概率。$\boldsymbol{U}$ , $\boldsymbol{W}$ , $\boldsymbol{V}$ 和 $\boldsymbol{G}$ 分别为对应的权重矩阵。

LSTM 模型计算方法与 RNN 类似：

$$\boldsymbol{x}_t = \boldsymbol{P}\boldsymbol{w}_t + \boldsymbol{F}\boldsymbol{g}_t \tag{4}$$

$$\boldsymbol{i}_t = \sigma\left(\boldsymbol{W}_{xi}\boldsymbol{x}_t + \boldsymbol{W}_{mi}\boldsymbol{m}_t + \boldsymbol{W}_{ci}\boldsymbol{c}_{t-1} + \boldsymbol{b}_i\right) \tag{5}$$

$$\boldsymbol{f}_t = \sigma\left(\boldsymbol{W}_{xf}\boldsymbol{x}_t + \boldsymbol{W}_{mf}\boldsymbol{m}_t + \boldsymbol{W}_{cf}\boldsymbol{c}_{t-1} + \boldsymbol{b}_f\right) \tag{6}$$

$$\boldsymbol{c}_t = \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \odot \tanh\left(\boldsymbol{W}_{xc}\boldsymbol{x}_t + \boldsymbol{W}_{mc}\boldsymbol{m}_{t-1} + \boldsymbol{b}_c\right) \tag{7}$$

$$\boldsymbol{o}_t = \sigma\left(\boldsymbol{W}_{xo}\boldsymbol{x}_t + \boldsymbol{W}_{mo}\boldsymbol{m}_t + \boldsymbol{W}_{co}\boldsymbol{c}_{t-1} + \boldsymbol{b}_o\right) \tag{8}$$

$$\boldsymbol{y}_t = \varphi\left(\boldsymbol{o}_t \odot \tanh\left(\boldsymbol{c}_t\right) + \boldsymbol{G}\boldsymbol{g}_t\right) \tag{9}$$

其中式(4)输入向量 $\boldsymbol{x}_t$ 的计算方式与RNN模型相同，式(5)–式(9)分别为输入门、遗忘门、输出门、记忆单元和隐含层输出的计算方法。式(10)中输出向量 $\boldsymbol{y}_t$ 与RNN中相同，为所求的预测概率。模型在训练过程中通过梯度下降算法得到最优的权重矩阵，使训练数据上由历史得到下一个词的平均对数似然值最大[16]。

由于 RNN 语言模型需要较多历史，在识别过程中无法将其应用到一遍解码。通常的做法是将 RNN 与 $n$-gram 模型进行融合，即以一定比例对两种模型得到的概率线性加权后对第 1 次识别的前 $n$ 个最佳候选($n$-best)重新打分。模型融合的公式为

$$\text{score}_{\text{lm}}(s) = \sum_{i=1}^{m} \lg\left(p\left(w_i \middle| w_1 \cdots w_{i-1}\right)\right) \tag{11}$$

$$\begin{aligned}\text{score}(s) = {} & \lambda \cdot \text{score}_{\text{ac}}(s) + \alpha \cdot \text{score}_{n\text{-gram}}(s) \\ & + (1-\alpha)\text{score}_{\text{RNN}}(s)\end{aligned} \tag{12}$$

式中，$s$ 表示当前句子，由该语句的原始文本和句子开始符号(<bos>)，句子结束符号(<eos>)组成，设 $s$ 长度为 $m$，则其可表示为序列 $w_1, w_2, \cdots, w_m$，其中 $w_1$，$w_m$ 分别为<bos>, <eos>符号。式(14)中，$\text{score}_{\text{lm}}(s)$ 表示 $s$ 的 RNN 语言模型得分。重打分后 $s$ 的语言模型得分由 $n$-gram 与 RNN 语言模型得分的线性加权得到，总得分为语言模型与原声学模型的加权和。最后对每条语音取总得分最高的候选句子作为系统输出。

## 3　主题特征提取方法

### 3.1 LDA 模型

本实验使用潜在狄利克雷分配(LDA)[17,18]来得到主题分布信息。LDA 模型假设文档的生成过程如下：

(1)由泊松分布采样得到文档的长度 $N$，即 $N \sim \text{Poisson}(\xi)$；

(2)对于文档 $d$，从参数为 $\alpha$ 的狄利克雷分布中采样得到这篇文档主题分布的多项分布参数 $\boldsymbol{\Theta} \sim \text{Dir}(\alpha)$；

(3)对于每个主题，从参数为 $\boldsymbol{\beta}$ 的狄利克雷分布中采样得到的各个词的多项式分布参数 $\Gamma \sim \text{Dir}(\boldsymbol{\beta})$；

(4)对文档 $N$ 个词中的第 $n$ 个词，首先选择这个词的主题分布 $\boldsymbol{z}_n \sim \text{Multinomial}(\boldsymbol{\Theta})$，然后得到这个词在主题分布和 $\boldsymbol{\beta}$ 参数下的概率 $p(\boldsymbol{w}_n|\boldsymbol{z}_n, \boldsymbol{\beta})$。

可见，超参数 $\alpha$ 为其中一个重要的参数。通常认为不同文档 $\alpha$ 是相同的。当 $\alpha$ 小于 1 时，对每篇文章主题分布的采样都是一个少数主题概率较大，其他主题概率很小的一个尖锐的分布；$\alpha$ 等于 1 时，采样结果为各个主题分布的概率相同；$\alpha$ 大于 1 时，主题分布越平均的分布概率越大。在本文涉及的实验中，我们使用由 Blei 等人开发的工具包 [1)]来实现 LDA 算法。给定训练数据、主题个数和初始 $\alpha$ 后，工具包通过迭代得到最优的 $\alpha$、每个主题下各个词出现的概率分布参数 $\boldsymbol{\beta}$，以及每个文档的主题分布参数。

### 3.2 LDA 特征提取

**3.2.1 直接 LDA 特征**　一种直接获得当前主题信息的方法为取该词前窗长为 $L$ 的窗内单词作为一篇文档，使用 LDA 工具包对文档进行推测(inference)操作得到该文档的主题分布，并将该文档主题概率分布作为当前主题特征[9]。然而，对每个词的历史信息进行推测操作非常费时，需要找到一个更快捷便利的当前主题信息估计方法。

**3.2.2 快速 LDA 特征**　为避免使用推测操作直接得到历史信息 LDA 特征，一种可行的方法为使用当前词的主题信息与历史主题信息来估计当前的主题信息[9]。其中每个词的主题概率密度 $p(z \mid w(t))$ 可由各主题下每个词出现的概率分布参数 $\boldsymbol{\beta}$ 与各主题出现的概率得到：

$$p\left(z \mid w(t)\right) = \frac{p\left(w(t) \mid z\right)p(z)}{\sum_k p\left(w(t) \mid k\right)p(k)} \tag{12}$$

一般在 LDA 训练中，在全部训练文本中各主题出现概率相同。即，若 $K$ 为主题个数，对所有的 $j = 1,2,\cdots,K$，有 $p(z) = 1/K$。因此在训练文本集合上每个词的主题概率密度计算可简化为

$$p\left(z \mid w(t)\right) = \frac{p\left(w(t) \mid z\right)}{\sum_k p\left(w(t) \mid k\right)} \tag{13}$$

式(14)可以通过对 $\boldsymbol{\beta}$ 矩阵的列向量归一化实

[1)] http://www.cs.princeton.edu/blei/lda-c/

现。可以使用当前词与它之前 $L$ 个单词按对应主题的主题概率密度相乘的结果估计当前特征。具体计算公式为

$$\boldsymbol{g}_t = \frac{1}{Z}\prod_{i=0}^{L-1} p\left(z \mid w(t-i)\right) \tag{15}$$

其中，$Z$ 为归一化系数，其目的为使 $\boldsymbol{g}_t$ 向量所有元素和为 1。该特征可以通过依次减少较远历史主题分布权重来改善，即可以通过前一时刻的特征与当前词的主题概率在对数域上线性加权得到当前时刻特征。其计算公式为

$$\boldsymbol{g}_t = \frac{1}{Z}\boldsymbol{g}_{t-1}^{\gamma} p\left(z \mid w(t)\right)^{1-\gamma} \tag{16}$$

式中，$\gamma$ 为加权权重，取值范围为 0 到 1。$\gamma$ 越小表示历史词的主题信息衰落的越快，当前词主题信息的影响越大。

**3.2.3 基于文档信息的词 LDA 特征** 上一节快速 LDA 特征的计算中使用了基于所有训练文本的词的主题概率分布 $p\left(z \mid w(t)\right)$。实际上，在不同语段或句子中，不仅主题的分布差别很大，而且相同单词在不同语言环境下对主题分布的贡献也不相同。基于上述想法，本文提出了一种使用基于文档的词主题概率分布 $p\left(z \mid w(t), d\right)$ 来计算 LDA 特征的方法。本方法首先计算出当前文档的主题分布 $p\left(z \mid w(t)\right)$，然后利用该信息计算当前词在该文档下的主题分布 $p\left(z \mid w(t), d\right)$，并用该分布代替式(14)中基于训练集合整体的词的主题分布，其完整的计算公式为

$$p\left(z \mid w(t), d\right) = \frac{p\left(w(t) \mid z\right) p\left(z \mid d\right)}{\sum_k p\left(w(t) \mid k\right) p\left(k \mid d\right)} \tag{17}$$

其中，$d$ 表示 $w(t)$所属的文档，$p\left(z \mid w(t)\right)$ 与快速 LDA 特征中相同，$p\left(z \mid d\right)$ 通过 LDA 算法得到。当前词所对应的特征的计算方法与快速 LDA 特征类似，其计算公式为

$$\boldsymbol{g}_t = \frac{1}{Z}\boldsymbol{g}_{t-1}^{\gamma} p\left(z \mid w(t), d\right)^{1-\gamma} \tag{18}$$

## 4 实验结果及分析

### 4.1 PTB 实验结果

Penn TreeBank(PTB)语料库[19]由宾夕法尼亚大学建立，是目前国际上广泛使用的测试语料之一。它使用的词表大小为 $10^4$，词表外的词映射为<unk>。其中第 0–20 节作为训练集($9.3\times10^5$ 万词)，第 21, 22 节作为验证集($7.4\times10^4$ 万词)，第 23 节、第 24 节作为测试集($8.2\times10^4$ 万词)。

训练 LDA 模型时，按照前人的工作[9]将每 10 句话作为一篇文档，并将主题个数 $K$ 设为 40。由于 PTB 数据集将语段截开划分成句子，所以前一个句子的主题信息可能会对后面单词主题信息预测有影响。因此在训练语言模型及提取 LDA 特征过程中均采用把所有句子连在一起的策略。本文在计算快速 LDA 特征时，对概率做了平滑处理：$\boldsymbol{\beta}$ 矩阵列归一化后，若某个词的某个主题概率小于 0.001，则设为 0.001，然后重新进行归一化操作。在计算基于文档信息的词 LDA 特征时，将每个词在主题上分布的概率增加词表大小的倒数再进行归一化。特征计算中将所有乘法操作都使用对数加法进行代替，避免数据溢出问题。实验使用 google 公司的开源工具 tensorflow 实现，实验中所有 RNN 及 LSTM 层的节点数均设为 100。

困惑度是评价语言模型的一个重要指标，其基本评价方式为，对测试集赋予高概率值的语言模型评价更高，其具体计算公式为

$$\mathrm{PPL}(s) = \sqrt[L]{p(w_1)\prod_{i=2}^{L} p\left(w_i \mid w_1 \cdots w_{i-1}\right)} \tag{19}$$

易知，PPL 越小，该指标下，模型性能越好。

在 RNN 与 LSTM 语言模型上分别加入不同特征的结果如表 1 所示。

表中 RNN+POS, RNN+word2vec 分别为特征层使用词性标注和词向量的结果，其中词性特征标注使用 nltk 工具包得到，特征向量为各词性对应的 one-hot 向量；word2vec 特征向量为由 word2vec-master 工具训练得到的每个词的 40 维词向量。可见，这两种方法性能与快速 LDA 相近，但不如直接 LDA 及改进 LDA 特征。另一方面几种 LDA 特征相对基线模型均有不同程度的改进，并且在 RNN，LSTM 模型中都是快速 LDA 特征对基线系统困惑

**表 1 RNN 和 LSTM 语言模型加入不同特征时模型性能比较**

| 模型 | 验证集困惑度 | 测试集困惑度 |
|---|---|---|
| RNN | 149.9 | 142.6 |
| RNN+POS | 141.7 | 135.9 |
| RNN+word2vec | 136.8 | 130.9 |
| RNN+快速 LDA 特征($\gamma$=0.1) | 140.4 | 134.2 |
| RNN+直接 LDA 特征($L$=50) | 129.9 | 124.2 |
| RNN+改进 LDA 特征($\gamma$=0.2) | 116.4 | 111.4 |
| LSTM | 120.4 | 115.0 |
| LSTM+快速 LDA 特征($\gamma$=0.1) | 116.3 | 111.2 |
| LSTM+直接 LDA 特征($L$=50) | 115.2 | 110.7 |
| LSTM+改进 LDA 特征($\gamma$=0.4) | 105.9 | 101.4 |

度减小最小(测试集上分别为 5.9%, 3.3%)，改进 LDA 特征的减小最大(测试集上分别为 21.9%, 11.8%)。但是加入长时间历史信息的影响后 RNN 与 LSTM 模型之间差距要远小于不加特征的情况。首先比较特征层的连接方式对模型性能的影响，实验使用 RNN 单元及快速 LDA 特征，结果如表 2 所示。

**表 2 使用快速 LDA 特征时特征层连接方式效果比较**

| 特征层连接方式 | 验证集困惑度 | 测试集困惑度 |
|---|---|---|
| 不使用特征层 | 149.9 | 142.6 |
| 只连接到 RNN 层 | 141.6 | 135.6 |
| 只连接到输出层 | 147.0 | 140.0 |
| 同时连接到 RNN 层与输出层 | 140.4 | 134.2 |

由表 2 结果可以看出，特征只连接到 RNN 层或输出层时，测试集上模型性能相比不使用特征层分别有 1.8%和 4.9%的提升，均不如特征同时连接到 RNN 层与输出层的情况。可见特征层到 RNN 层与输出层的连接分别能引入一定信息，均不能被另一连接取代。在直接 LDA 特征的实验中，首先研究了使用不同历史长度信息推测主题概率的效果。结果如表 3 所示。

从表 3 结果可以看出，增加直接 LDA 特征可以使 RNN 模型性能有 10%以上的提高，并且提高程度随历史长度变化不大。其中在使用历史长度 20 到 50 的数据得到 LDA 特征时，由于历史提供的信息增多，模型困惑度会随历史信息的增多而减少，但超过 50 词后历史信息增多反而使模型性能有微弱下降。这种现象产生的原因可能是过长历史的主题与当前主题有一定不同，因此过长历史可能对当前特征产生一定干扰。最后一行中“两句”是指使用当前词所在句子的历史单词以及前面两句话的信息作为历史。这种做法可以考虑到句子边界信息的影响。经统计 PTB 数据集中平均每句话长度为 22 词，考虑句子边界的结果略好于长度相当(即长度为 40, 50)时的结果。在计算改进 LDA 特征的过程中，文档规模的选取对模型性能具有很大影响，具体实验结果如表 4 所示。

**表 3 计算直接 LDA 特征时使用不同长度历史信息效果比较**

| 模型及特征使用历史长度 | 验证集困惑度 | 测试集困惑度 |
|---|---|---|
| RNN, $L$=20 词 | 133.4 | 127.5 |
| RNN, $L$ =40 词 | 132.9 | 126.4 |
| RNN, $L$ =50 词 | 129.9 | 124.2 |
| RNN, $L$ =60 词 | 131.7 | 125.2 |
| RNN, $L$ =80 词 | 131.8 | 125.4 |
| RNN, 2 句 | 131.0 | 124.0 |

**表 4 计算改进 LDA 特征时使用不同长度历史信息效果比较**

| 模型及计算特征时文档长度 | 验证集困惑度 | 测试集困惑度 |
|---|---|---|
| RNN-快速LDA特征 | 140.4 | 134.2 |
| RNN-10句 | 127.6 | 121.4 |
| RNN-1句 | 116.4 | 111.4 |

由表 4 易见，选择过长数据作为一篇文档的效果不如短文档。这个现象也从侧面印证了前面关于直接 LDA 特征历史长度选择实验的分析结果。因此，本文所有改进 LDA 特征的实验中文档长度都选为 1 句话。在改进 LDA 特征的实验中，参数 $\gamma$ 的影响如图 2 所示。

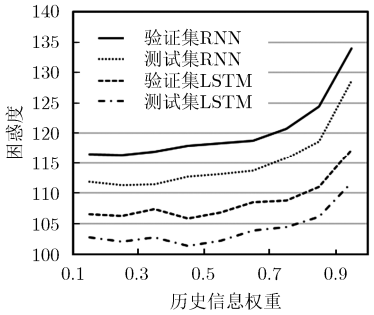


图 2 改进特征中历史信息权重对模型困惑度影响

实验表明，使 RNN 和 LSTM 模型性能最好的 $\gamma$ 参数分别为 0.2 和 0.4。并且当 $\gamma$ 较小时，即当前词主题概率占比重较大时，模型性能较好，且受 $\gamma$ 参数影响不大；当 $\gamma$ 较大(大于 0.6)时，模型性能随 $\gamma$ 增大急速下降。可见，历史词的信息对于在 RNN, LSTM 模型性能的改善都有一定作用。我们将表 1 中 RNN 和 LSTM 加上特征后效果最好的模型分别于 5-gram-cache 语言模型(KN5-cache)融合。其中 5-gram-cache 语言模型使用 srilm 工具训练，使用 Kneser-Ney 平滑方法，缓存历史设为 180。实验结果如表 5 所示。

由实验结果可以看出，在测试集上融合后的 RNN, LSTM 模型的困惑度相对 5-gram-cache 分别有 35.2, 39.0 的下降。

### 4.2 多候选重估实验

本阶段实验为使用新语言模型对语音识别结果

表 5 加改进 LDA 特征的模型与 5-gram 语言模型融合实验结果

| 模型 | 验证集困惑度 | 测试集困惑度 |
|---|---|---|
| KN5-cache | 131.6 | 128.3 |
| KN5-cache+RNN-LDA | 94.6 | 93.1 |
| KN5-cache+LSTM-LDA | 90.9 | 89.3 |

的前 100 候选答案重新打分。基线识别系统使用 Kaldi 语音识别工具搭建，声学模型使用采用交叉熵(CE)准则训练的时延神经网络(TDNN)模型，语言模型使用 KN 平滑的 3-gram 模型(KN3)。

**4.2.1 SWBD 实验** 本节实验中语言模型训练数据为训练集语音所对应的标注文本，词典为训练集中所有出现过的单词集合，词表大小为 29671 词。测试数据使用 eval2000 测试集。测试数据使用 eval2000 测试集。实验结果如表 6 所示。

表 6 改进语言模型在 SWBD 重估实验错误率比较

| 语言模型 | eval2000 词错误率(%) |
|---|---|
| KN3 | 20.0 |
| RNN | 19.6 |
| RNN-改进 LDA 特征 | 19.3 |
| LSTM | 18.9 |
| LSTM-改进 LDA 特征 | 18.8 |

由表 6 结果可知，在 RNN、LSTM 中使用改进 LDA 特征可以使词错误率相对 3-gram 语言模型分别降低 3.5%, 6%。其中该特征对 RNN 的改进效果大于对 LSTM 改进，其变化规律与 4.1 节中 PTB 实验结果基本相同。

**4.2.2 WSJ 实验** 本节实验中语言模型训练数据为 WSJ 的新闻文本数据。由于训练数据总量较大，并且经验证，使用全部训练数据和使用 1/5 训练数据训练的效果基本相同，因此本阶段实验训练集选为全部数据的 1/5。该子集共 $3.224\times10^6$ 词，词典为所有出现次数大于 1 的单词的集合，词表大小为 51857 词。实验结果如表 7 所示。

可见，加入改进 LDA 特征后，整个系统的错误率在 dev93, eval92 上与基线系统相比分别有相对 6.8%, 14.2%的下降。并且在 WSJ 数据集上改进 LDA 特征同样对 RNN 的改进效果要远大于对 LSTM。其中在 eval92 测试集上该特征可使 RNN 达到与 LSTM 相同的效果。这也从另一方面说明了该特征的加入使神经网络学到了更多历史信息，并且该信息与 LSTM 通过门机制学到的内容基本类似，因此 LSTM 上该特征并没有太多作用。而 RNN 增加特征的方法相对于 LSTM，模型复杂度更小，并且训练时间、预测时间也都更短。

表 7 改进语言模型在 WSJ 重估实验错误率比较

| 语言模型 | dev93 词错误率(%) | eval92 词错误率(%) |
|---|---|---|
| KN3 | 9.6 | 7.0 |
| RNN | 9.4 | 6.6 |
| RNN-改进 LDA 特征 | 9.0 | 6.0 |
| LSTM | 9.0 | 6.1 |
| LSTM-改进 LDA 特征 | 8.9 | 6.0 |

## 5 结论

为了缓解语言模型不能有效利用长时间跨度信息的问题，本文提出一种基于文档信息的词 LDA 特征计算方法，并将该特征用在神经网络语言模型中。实验表明本文提出的特征性能明显优于直接 LDA 特征和快速 LDA 特征，在 PTB 数据集上使 RNN, LSTM 模型困惑度分别有相对 21.9%, 11.8%的降低；在 SWBD 语音识别任务的多候选重估实验中，在 RNN 上使用本文提出的方法可在 eval2000 测试集上，使 WER 相对基线模型降低 3.5%；在 WSJ 的 dev93, eval92 测试集上，该特征分别使 WER 相对基线模型下降 6.8%, 14.2%。

## 参 考 文 献

[1] MIKOLOV T, KARAFIÁT M, BURGET L, et al. Recurrent neural network based language model[C]. INTERSPEECH, Makuhari, Chiba, Japan, 2010: 1045–1048.

[2] MIKOLOV T, JOULIN A, CHOPRA S, *et al*. Learning longer memory in recurrent neural networks[OL]. https://arxiv.org/abs/1412.7753v22014.

[3] MEDENNIKOV I and BULUSHEVA A. LSTM-based language models for spontaneous speech recognition[C]. International Conference on Speech and Computer, Athens, Greece, 2016: 469–475.

[4] HUANG Z, ZWEIG G, and DUMOULIN B. Cache based recurrent neural network language model inference for first pass speech recognition[C]. IEEE International Conference on Acoustics, Speech and Signal Processing, Florence, Italy, 2014: 6354–6358.

[5] COCCARO N and JURAFSKY D. Towards better integration of semantic predictors in statistical language modeling[C]. International Conference on Spoken Language Processing, Sydney, Australia, 1998: 2403–2406.

[6] KHUDANPUR S and WU J. Maximum entropy techniques for exploiting syntactic, semantic and collocational

dependencies in language modeling[J]. *Computer Speech & Language*, 2000, 14(4): 355–372.

[7] LAU R, ROSENFELD R, and ROUKOS S. Trigger-based language models: A maximum entropy approach[C]. IEEE International Conference on Acoustics, Speech, and Signal Processing, Orlando, Florida, USA, 2002: 45–48.

[8] ECHEVERRY-CORREA J D, FERREIROS-LÓPEZ J, COUCHEIRO-LIMERES A, *et al*. Topic identification techniques applied to dynamic language model adaptation for automatic speech recognition[J]. *Expert Systems with Applications*, 2015, 42(1): 101–112.

[9] MIKOLOV T and ZWEIG G. Context dependent recurrent neural network language model[C]. Spoken Language Technology Workshop, Miami, Florida, USA, 2012: 234–239.

[10] 张剑, 屈丹, 李真. 基于词向量特征的循环神经网络语言模型[J]. 模式识别与人工智能, 2015, (4): 299–305. doi: 10.16451/j.cnki.issn1003-6059.201504002.
ZHANG Jian, QU Dan, and LI Zhen. Recurrent neural network language model based on word vector features[J]. *Pattern Recognition and Artificial Intelligence*, 2015, (4): 299–305. doi: 10.16451/j.cnki.issn1003-6059.201504002.

[11] GONG C, LI X, and WU X. Recurrent neural network language model with part-of-speech for Mandarin speech recognition[C]. International Symposium on Chinese Spoken Language Processing, Singapore, 2014: 459–463.

[12] 左玲云, 张晴晴, 黎塔, 等. 电话交谈语音识别中基于LSTM-DNN 语言模型的重评估方法研究[J]. 重庆邮电大学学报(自然科学版), 2016, 28(2): 180–186. doi: 10.3979/j.issn.1673-825X.2016.02.007.
ZUO Lingyun, ZHANG Qingqing, LI Ta, *et al*. Revaluation based on LSTM –DNN language model in telephone conversation sqeech recognition[J]. *Journal of Chongqing University of Post and Telecomunications*, 2016, 28(2): 180–186. doi: 10.3979/j.issn.1673-825X.2016.02.007.

[13] 王龙, 杨俊安, 陈雷, 等. 基于循环神经网络的汉语语言模型并行优化算法[J]. 应用科学学报, 2015, 33(3): 253–261. doi: 10.3969/j.issn.0255-8297.2015.03.004.
WANG Long, YANG Junan, CHEN Lei, *et al*. Parallel optimization of chinese language model based on recurrent neural network[J]. *Journal of Applied Sciences*, 2015, 33(3): 253–261. doi: 10.3969/j.issn.0255-8297.2015.03.004.

[14] PIOTR Bojanowski, EDOUARD Grave, ARMAND Joulin, *et al*. Enriching word vectors with subword information[OL]. https://arxiv.org/abs/1607.04606v2.

[15] GANGULY D, ROY D, MITRA M, *et al*. Word embedding based generalized language model for information retrieval[C]. The International ACM SIGIR Conference, Santiago, Chile, 2015: 795–798.

[16] LI X. Recurrent neural network training with preconditioned stochastic gradient descent[OL]. https://arxiv.org/abs/1606.04449v2, 2016.

[17] BLEI D M, NG A Y, and JORDAN M I. Latent dirichlet allocation[J]. *Journal of Machine Learning Research*, 2003, 3: 993–1022.

[18] BHUTADA S, BALARAM V V S S S, and BULUSU V V. Semantic latent dirichlet allocation for automatic topic extraction[J]. *Journal of Information & Optimization Sciences*, 2016, 37(3): 449–469.

[19] MARCUS M P, MARCINKIEWICZ M A, and SANTORINI B. Building a large annotated corpus of English: the penn treebank[J]. *Computational Linguistics*, 1993, 19(2): 313–330.

刘　畅：女，1992 年生，博士生，研究方向为语音信号处理、语音识别、语言模型等.

张一珂：男，1991 年生，博士生，研究方向为语音信号处理、语音识别、语言模型、自然语言理解等.

张鹏远：男，1978 年生，研究员，硕士生导师，研究方向为大词表非特定人连续语音识别、关键词检索、声学模型、鲁棒语音识别等.

颜永红：男，1967 年生，研究员，博士生导师，研究方向为语音信号处理、语音识别、口语系统及多模系统、人机界面技术等.