基于词向量特征的循环神经网络语言模型*

张 剑 屈 丹 李 真

(中国人民解放军信息工程大学 信息系统工程学院 郑州 450001)

摘 要 循环神经网络语言模型能解决传统 N-gram 模型中存在的数据稀疏和维数灾难问题,但仍缺乏对长距离信息的描述能力. 为此文中提出一种基于词向量特征的循环神经网络语言模型改进方法. 该方法在输入层中增加特征层,改进模型结构. 在模型训练时,通过特征层加入上下文词向量,增强网络对长距离信息约束的学习能力. 实验表明,文中方法能有效提高语言模型的性能.

关键词 语音识别,语言模型,循环神经网络,词向量

中图法分类号 TP 391 **DOI** 10.16451/j.cnki.issn1003-6059.201504002

Recurrent Neural Network Language Model Based on Word Vector Features

ZHANG Jian, QU Dan, LI Zhen

(*Institute of Information System Engineering, The PLA Information Engineering University, Zhengzhou* 450001)

ABSTRACT

The recurrent neural network language model(RNNLM) solves the problems of data sparseness and dimensionality disaster in traditional N-gram models. However, the original RNNLM is still lack of long dependence due to the vanishing gradient problem. In this paper, an improved method based on contextual word vectors is proposed for RNNLM. To improve the structure of models, a feature layer is added into the input layer. Contextual word vectors are added into the model with feature layer to reinforce the ability of learning long-distance information during the training. Experimental results show that the proposed method effectively improves the performance of RNNLM.

Key Words Speech Recognition, Language Model, Recurrent Neural Network, Word Vector

* 国家 863 计划项目(No. 2012AA011603)、国家自然科学基金项目(No. 61175017)资助
收稿日期:2014-02-27;修回日期:2014-03-27
作者简介 张剑,男,1988 年生,硕士研究生,主要研究方向为语音识别、自然语言处理. E-mail: Crsmx_23@163.com. 屈丹,女,1974 年生,博士,副教授,主要研究方向为语音识别、智能信息处理. 李真,女,1982 年生,硕士,讲师,主要研究方向为语音识别、智能信息处理.

1 引 言

语言模型是自然语言处理中最重要的组成部分,在语音识别、机器翻译、信息检索等领域有着广泛的应用,其中最具代表性的是 N-gram 语言模型.但目前 N-gram 模型的发展逐渐陷入瓶颈,性能提高缓慢,且代价昂贵,因此神经网络语言模型(Neural Network Language Model, NNLM)等高级语言模型逐渐成为研究的热点[1].

随着深度学习(Deep Learning)理论的不断发展,神经网络在自然语言处理,尤其是语言模型中的应用开始受到关注. Bengio 等[2]率先提出利用神经网络构建语言模型的方法,通过对词的分布式表达(Distributed Representation)巧妙解决数据稀疏对统计建模的影响,同时克服模型参数的维数灾难问题.其训练的神经网络语言模型无需使用传统 N-gram 模型中复杂的平滑算法,具备较好的模型性能,在 Associated Press (AP) News 数据集上的对比实验表明,相比 Kneser-Ney 平滑的 N-gram 模型,NNLM 的性能有明显的提高.

但前馈神经网络语言模型在输入层上仍采用 $N-1$ 个词作为历史信息,并未解决长距离信息建模能力差的问题.为此,Mikolov 等[3-4]利用循环神经网络训练语言模型,通过隐含层的循环获得更多的上下文信息,同时降低模型的参数个数,称为循环神经网络语言模型(Recurrent NNLM, RNNLM),并进行一系列改进. RNNLM 的优势在于能利用更多的上下文信息进行词的预测,对语言具有更好的建模能力.但文献[5]的理论研究表明,循环神经网络中存在消失梯度(Vanishing Gradient)问题,使得基于梯度的训练算法对长距离信息的学习变得困难.模型性能分析表明 RNNLM 的性能类似于 N 为 8 或 9 的超长距离前馈神经网络模型[6].

因此,如何提高 RNNLM 对长距离上下文的学习能力需更进一步的研究.目前学者们进行的研究主要有 2 种:1)优化训练算法,利用高阶信息,避免消失梯度问题,使误差传递的层数更多,达到学习更长距离信息的目的,如 Hessian-free 优化方法[7];2)利用 Long Short-Term Memory 改进神经网络,采用门控神经元(Gating Neurons)锁存多重时序中的误差信号传递[8].

本文从模型结合的角度研究上述问题,提出基于词向量特征的循环神经网络语言模型.通过计算基于长距离历史的上下文词向量,并将其作为网络的一个输入特征进行神经网络语言模型的训练中,增强网络对长距离信息的学习,提高语言模型的精度.该方法的优势是可在神经网络中较方便地利用其他复杂的高级模型,这在传统的 N-gram 类模型中经常使用到,如引入语言学信息和主题特征等[9].本文利用连续词袋(Continuous Bags-of-Words, CBOW)模型和 Skip-gram 模型获取上下文相关的词向量,改进原有循环神经网络语言模型,并在 Penn Treebank 数据库上进行困惑度测试,在微软语料库和 Wall Street Journal 语料库上进行语音识别实验.实验表明,本文方法可提高神经网络语言模型性能,有效降低语言模型在测试集上的困惑度,对连续语音识别系统的性能也有较大幅度的提高.

2 神经网络模型结构

RNNLM 的网络结构包含输入层(Input Layer)、隐含层(Hidden Layer)和输出层(Output Layer)3 个部分,输入向量 $w(t)$ 代表 t 时刻时输入词,采用 One-hot 编码(也称为 1-of-N 编码),维数与词汇表大小相同.隐含层 $s(t)$ 为网络的状态,表示 t 时刻时上下文的历史信息,输出向量 $y(t)$ 表示待预测词在词汇表上的概率分布. U、V、W 分别为各层之间的权值矩阵.

在 t 时刻,网络的输入由当前词向量 $w(t)$ 和前一时刻的隐含层输出 $s(t-1)$ 构成,联合计算下一个隐含层,通过隐含层循环的方式,可利用更长的上下文信息,更好地表示自然语言,其性能优于传统的前馈神经网络语言模型,近年来被大量使用于语音识别、机器翻译、语言理解等领域[10-11].

在实际应用中,循环神经网络的消失梯度问题使网络不能较好学习较长距离的信息,其影响在传递过程中较易丢失.这是因为在循环网络中,时间上隔得较远的输入的微小变化对网络训练几乎不会产生影响,即使远距离输入发生较大变化时,其产生的影响仍存在不能被梯度检测到的问题,这就使得在一些特定情况下,基于梯度的训练算法对长距离依赖的学习变得困难.本文采用的是一种类似模型结合的方法,通过将上下文相关词向量直接加入到原有神经网络中,提高网络对长距离信息的学习能力,避免网络训练中的消失梯度问题.

因此,本文改进原有网络结构,在输入层中增加一个特征层(Feature Layer),并通过权值矩阵 F、G 分别与隐含层和输出层相连,网络结构如图 1 所示.此时网络的输入变为

$$x(t) = w(t) + s(t-1) + f(t).$$

在特征层中，输入向量 $f(t)$ 为词的上下文相关向量，其中包含更多的长距离信息，是对输入向量 $w(t)$ 的一个补充，可使词概率的计算更准确. 采用随机梯度下降法训练网络，此时输出向量 $y(t)$ 表示待预测词在给定当前词 $w(t)$ 、上下文 $s(t-1)$ 和特征向量 $f(t)$ 下的概率分布. 改进后的 RNNLM 计算公式为

$$s(t) = f(\boldsymbol{U}w(t) + \boldsymbol{W}s(t-1) + \boldsymbol{F}f(t)),$$

$$y(t) = g(\boldsymbol{V}s(t) + \boldsymbol{G}f(t)),$$

其中，$f(z)$ 为 sigmoid 激活函数:

$$f(z) = (1 + e^{-z})^{-1},$$

$g(z)$ 为 softmax 激活函数，

$$g(z_m) = \frac{e^{z_m}}{\sum_k e^{z_m}}.$$

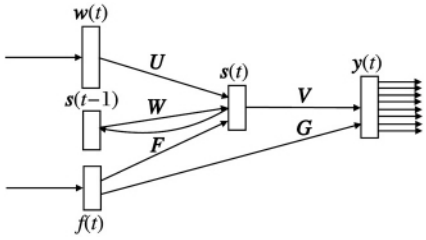


图 1　结合特征层的 RNNLM 结构

Fig. 1　Structure of RNNLM with feature layer

3　词向量特征

3.1　词向量定义

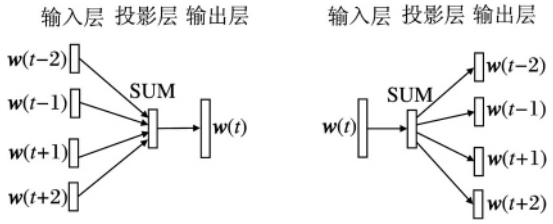
在自然语言处理中，要将自然语言理解的问题转化为机器学习的问题，就需将自然语言的符号数学化，其中最直观和常用的方法是 One-hot 表示法. 这种方法将每个词表示为一个很长的向量，其维数是词汇表大小，其中绝大多数元素为 0，只有一个维度的值为 1，这个维度就代表当前的词.

在自然语言处理中，常将 One-hot 表示采用稀疏的方式进行存储，即为每个词分配一个数字 ID. 该方法因其简单易用，广泛应用于各种自然语言处理任务中，如 N-gram 模型中就采用这种词向量表示法. 但这种表述方法也存在一定问题: 其表示的任意两个词之间是孤立的，无法表示这两个词之间的依赖关系，从词向量上看不出两个词是否相关; 采用稀疏表示法，在处理某些任务，如构建 N-gram 模型时，会引起维数灾难问题.

而在深度学习中，一般采用分布式表示(Distributed Representation)的方法表示词向量，这种表示法最早由 Hinton[12] 提出，通常称为 Word Representation. 这种方法将词用一种低维实数向量表示，优点在于相似的词在距离上更接近，能体现出不同词之间的相关性，从而反映词之间的依赖关系. 同时，较低的维度也使特征向量在应用时有一个可接受的复杂度. 因此，新近提出的许多语言模型，如潜在语义分析(Latent Semantic Analysis，LSA)模型和潜在狄利克雷分布(Latent Dirichlet Allocation，LDA)模型，及目前流行的神经网络模型等，都采用这种方法表示词向量.

3.2　连续词袋模型和 Skip-gram 模型

本文选择 CBOW 模型和 Skip-gram 模型获取上下文相关词向量. 这两种模型由 Mikolov 等[13] 提出，旨在以较小的计算复杂度获取词的分布式表示. 在传统神经网络模型的基础上，针对训练复杂度过大的问题进行改进，采用一种 Log-linear 模型结构，去除神经网络的非线性隐含层，同时将神经网络中 N-gram模型的训练和词向量的计算分开，提高训练效率，其模型结构如图 2 所示.



(a) CBOW　　(b) Skip-gram

图 2　CBOW 和 Skip-gram 模型结构

Fig. 2　Architectures of CBOW and Skip-gram model

从结构图中可看出，CBOW 的模型架构类似于前馈神经网络模型，但去除隐含层，只包括输入层、投影层(Projection Layer)和输出层. 输入层、输出层表示每个词的词向量，均采用分布式表示，维数一般为 50 或 100，投影层维数为 D，窗长 C 表示上下文长度. 在训练时与前馈神经网络模型的区别如下: 输入层的词向量在投影层上的投影不再按顺序排列，而是进行合并，采用均值表示单个词向量，达到降低计算量的目的. 词在历史信息中的顺序不影响其在投影层中的表示，因此这种结构称为连续空间中的词袋模型. 此外，由于无需进行语言模型概率的计算，模型可利用未来的信息 $w(t+1)$ 、$w(t+2)$ 等训练当前词 $w(t)$，真正实现根据上下文得到最优的词向量.

Skip-gram 模型的结构则与 CBOW 模型相反，其采用当前词 $w(t)$ 训练上下文的词向量，目的是得到有利于预期周围词的词向量，用以表征上下文信息，可理解为根据上下文分类当前词.

这两种词向量各有优势，CBOW 模型在语法测试中准确性更高，表明其通过对上下文的学习，获得更多的语法信息；Skip-gram 模型在语义测试中性能更好，说明其词向量的区分性更好，对单个词的信息描述更准确.

这两种模型的共同优点在于能从规模数据中得到高质量的词向量，Mikolov 等[14]的实验表明，在词向量维数为 1 000 时，可在十亿级的数据量下完成词向量的训练，这在以往是难以想象的. 对海量数据的有效利用使所得词向量具有更高的精度，能更好描述对不同词之间的相关性. 同时，这两个模型得到的词向量也能描述词和短语之间的相关性，对句子中存在的长距离的词之间的关系进行有效表达.

4 实验及结果分析

为验证本文语言模型的性能，通过一系列的实验进行验证，包括 Penn Treebank 语料库上的困惑度实验，在微软语料库和 Wall Street Journal 语料库上的连续语音识别实验. 实验采用的评价指标是语言模型的困惑度(Perplexity，PPL)和语音识别系统的词错误率(Word Error Rate，WER). 具体的实验设置及实验结果如下.

4.1 Penn Treebank 困惑度实验

4.1.1 实验设置

Penn Treebank 语料库[15]是国际上广泛采用的一个英文树库，由宾夕法尼亚大学建立，常用来进行词性标注、句法分析等自然语言处理任务，在语言模型领域也被广泛采用. 在实验时采用相同的实验设置(相同的训练、开发和测试数据，及相同的词汇限制)，这也有利于对比不同的语言模型技术.

实验中 Penn Treebank 语料划分如下：0～20 节为训练集，包含 93 万个 Tokens；21～22 节为确认数据，作为开发集，包含 7.4 万个 Tokens；23～24 节为测试集，包含 8.2 万个 Tokens；词汇表大小限制为 1 万，集外词采用标号〈unk〉表示.

4.1.2 实验结果

实验中采用 CBOW 模型和 Skip-gram 模型在训练集上训练上下文相关的词向量(维数为 50)在窗长范围为 10～50 的条件下，选取不同的窗长进行实验，最终选取窗长为 30 的词向量训练语言模型. 在这一窗长下，词向量不仅可完整表示一句话内的词关联信息，也能有效表示句子间的上下文信息，得到的语言模型性能最优.

实验中训练 3 组 RNNLM，隐含层神经元个数采用 10 或 100，类别层个数为 100，并在开发集和测试集上分别计算模型的困惑度，实验结果如表 1 所示. 为对比语言模型性能，还采用 Kneser-Ney 平滑的 5-gram 模型.

表 1 Penn Treebank 语料库上困惑度实验结果

Table 1 Experimental results of perplexity test on Penn Treebank corpus

语言模型	困惑度	
	开发集	测试集
5-gram (KN5)	148.0	141.2
RNN 10	252.8	239.3
RNN 10 + CBOW 50	202.9	192.2
RNN 10 + Skip-gram 50	187.1	179.1
RNN 100	158.9	151.1
RNN 100 + CBOW 50	140.9	134.6
RNN 100 + Skip-gram 50	133.2	127.0

表 1 中，KN5 表示采用 Kneser-Ney 平滑的 5-gram 模型，RNN 10 和 RNN 100 分别表示隐含层神经元个数采用 10 和 100 的循环神经网络模型，CBOW 50 和 Skip-gram 50 分别代表词向量维数为 50 的模型. RNN+CBOW、RNN+Skip-gram 分别表示采用两种模型进行词向量扩展的循环神经网络模型.

从表 1 中可看出，本文方法可提高神经网络语言模型的性能，其在开发集和测试集上的困惑度均有明显降低. 在隐含层神经元个数较小(隐含层大小为 10)时，结合词向量对模型性能的提高更明显. 在开发集上，RNN-CBOW 模型相比原有 RNN 模型，困惑度由 252.8 降至 202.9，下降 19.7%；采用 Skip-gram 模型时，困惑度降低到 187.1，下降幅度更大，达到 26%. 而当隐含层神经元个数增多时，随着原有模型训练更充分，结合词向量特征对性能的提高进一步减弱. 在开发集上，两种方法相比原有 RNN 模型，困惑度分别下降 11% 和 16.3%.

为进一步验证本文方法的性能，将本文方法与其他语言模型对比. 在相同的实验条件下，选取文献[16]中的结果对比，RNNLM 的隐含层大小为 300，词向量的维数为 50，窗长为 30. 实验中每个模型除单独测试困惑度外，还与 Kneser-Ney 平滑的 5-gram 模型进行结合，计算在测试集上的困惑度. 具体结果

如表 2 所示.

表 2　不同语言模型在测试集上的性能对比

Table 2　Performance comparison of different language models on test set

语言模型	困惑度	
	单个	+ KN5
5-gram (GT5)	165.2	–
5-gram (KN5)	141.2	–
平滑最大熵模型	142.1	138.7
随机森林模型	131.9	131.3
结构化模型	146.1	125.5
前馈经神经网络模型	140.2	116.7
句法神经网络模型	131.3	110.0
循环神经网络语言模型	124.7	105.7
RNN + CBOW LM	118.7	99.4
RNN + Skip-gram LM	**112.6**	**97.3**

表 2 中列出多种不同语言模型的实验结果. N-gram 模型由 SRILM 工具包训练得到,采用目前流行的两种平滑算法,其中,Kneser-Ney 平滑(KN5) 的性能优于 Good-Turing 平滑(GT5) . 最大熵模型、随机森林模型和结构化模型是 N-gram 模型的一些扩展,采用不同方法改进 N-gram 模型. 而神经网络类模型则由前馈神经网络模型(Feedforward Neural Network LM) 、句法神经网络模型(Syntactical Neural Network LM) 和循环神经网络语言模型组成.

通过实验对比发现,神经网络类语言模型的性能普遍优于 N-gram 类模型,其中句法神经网络模型和循环神经网络模型由于能表示更多的长距离依赖关系,具备更好的模型性能而被采用. 本文方法改进 RNN 模型,困惑度进一步下降. RNN+Skip-gram 模型在两项测试中的困惑度(112.6 和 97.3) 均为最低,具备最好的模型性能,进一步说明本文方法的有效性.

4.2　连续语音识别实验

在该章节中,选取两个国际上常用的语料库——微软语料库和 Wall Street Journal(WSJ) 语料库进行连续语音识别实验,验证本文语言模型在连续语音识别系统中的性能. 针对两个语料库的实验设置和实验结果如下.

4.2.1　基于微软语料库的实验

微软语料库 Speech Corpora (Version 1.0) 是一个汉语连续语音识别语料库,其训练集由 100 个年龄在 18 ~ 40 岁间的男性录音组成,共 19 688 句,约 33 h 的语音数据; 测试集为另外 25 个男性录音,共 500 句语音; 采样频率为 16 kHz,采用 16 bit 线性脉冲编码调制(Pulse Code Modulation, PCM) 量化.

实验中声学特征采用 13 维 MFCC 参数及其一阶和二阶差分,共 39 维特征矢量语音信号. 采用 Hamming 窗处理,帧长 25 ms,帧移 10 ms 声学模型为有调音节的声韵模型,共 187 个模型基元,采用三状态自左向右无跨越的隐马尔科夫模型(HMM) . 由于语音中存在协同发音现象,在声韵母结构上,上下文相关三音子(Triphone) 的声韵母结构语言模型的训练数据为语料库的语音标注文件,采用 SRILM 工具包训练得到 Kneser-Ney 平滑的 3-gram 和 4-gram 模型.

本实验采用 Kaldi 工具箱[17] 搭建连续语音识别系统,基线系统采用 3-gram,在基于加权有限状态机(Weighted Finite State Transducer, WFST) 的解码器中构建静态解码网络,进行一次解码,利用 Lattice-to-nbest 工具提取 N-best 列表,在二次解码中,采用 Kneser-Ney 平滑的 4-gram 和循环神经网络语言模型进行 N-best 重打分,得到最终的识别结果. 采用测试集上的词错误率(WER) 作为系统的评价指标.

在实验中,RNNLM 的隐含层个数为 100 和 200,词向量的维数为 50,窗长为 30,将模型送入解码器进行实验. 表 3 给出不同语言模型的系统识别结果.

表 3　不同语言模型在微软语料库上的词错误率

Table 3　WER of different language models on Microsoft corpus

语言模型	词错误率/%
基线系统	16.45
4-gram	16.31
RNN 100	15.91
RNN 100 + CBOW 50	15.90
RNN 100 + Skip-gram 50	**15.76**
RNN 200	15.64
RNN 200 + CBOW 50	15.63
RNN 200 + Skip-gram 50	**15.45**

从表 3 可看出,结合 Skip-gram 模型改进的 RNNLM 的系统具有最低的识别词错误率. 从识别结果看,在隐含层为 100 时,系统识别的词错误率为 15.76%,与原有 RNNLM 相比,性能提高 0.97%; 与基线系统相比,相对提高 4.19%. 在隐含层为 200 时, RNN-Skip-gram 模型的识别词错误率为 15.45%,在 RNN 语言模型基础上提高 1.21%; 与基线系统相比,相对提高 6.08%. 说明 Skip-gram 词向量的加入可使 RNNLM 的识别性能进一步提高,而 CBOW 模型特征对 RNN 语言模型的连续语音识别结果并未有明显改善,还需进一步的实验验证其性能.

4.2.2 基于 WSJ 语料库的实验

Wall Street Journal(WSJ) 语料库是一个国际上广泛使用的英文语料库,适用于进行大词汇量连续语音识别实验. 其训练集包含 WSJ0 和 WSJ1 两部分,共 81.3 h 语音数据,测试集由 Eval92 和 Dev93 组成.

实验设置与微软语料库类似,针对 WSJ-20K 语音识别任务进行实验,采用 SI-284 数据训练得到传统 HMM-GMM 模型,利用 Kaldi 自动生成的问题集进行三音子状态聚类,共 4 368 个不同的绑定状态. 语言模型训练数据包含 37 MB 的 Tokens.

在不同的隐含层个数下训练神经网络模型,观察词向量对模型性能的影响,采用的词向量维数为 50,窗长为 30. 训练选择的隐含层大小分别为 30、100、200 和 300,不同循环神经网络模型在开发集上的困惑度如图 3 所示.

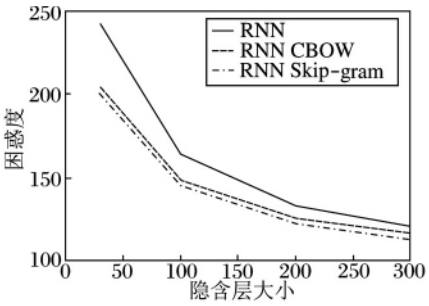


图 3 不同隐含层的 RNNLM 在有效数据上的困惑度

Fig. 3 Perplexity of RNNLM with different hidden layer sizes on valid data

从图 3 可看出,两种模型词向量的引入都能明显降低 RNNLM 的困惑度,其中 Skip-gram 模型的性能略优于 CBOW 模型. 当隐含层较小时,这种方法对模型困惑度的降低更明显,而随着神经元个数的增大,加入特征向量使模型性能的提高幅度减小. 这说明在隐含层较小时,特征向量的加入对网络的影响更大,可更好提高模型的性能. 随着网络本身训练更充分,影响开始减弱.

为进一步验证本文方法的性能,选择隐含层个数为 30 和 100 的语言模型进行语音识别实验,采用 100-Best 重打分进行二次解码,根据在 Eval92 和 Dev93 上的词错误率,对比不同模型对识别性能的影响. 实验结果如表 4 所示. 表 4 中,基线系统为一次解码的识别结果,4-gram 表示采用 4-gram 模型重打分的识别结果.

从表 4 可看出,本文方法训练的语言模型具有更好的识别效果,在 Eval92 和 Dev93 测试集上其识别性能在原有 RNNLM 基础上进一步提高. 在 Eval92 测试集上,隐含层个数为 30 时,结合 Skip-gram 词向量的 RNNLM 识别的词错误率为 7.67%,相比 RNN 和基线系统分别提高 1.41% 和 12.2%;隐含层个数为 100 时,词错误率为 7.21%,相比 RNNLM 和基线系统分别提高 2.2% 和 17.5%. 在 Dev93 测试集上可得到类似的实验结果.

表 4 不同语言模型在 WSJ 语料库上的词错误率

Table 4 WER of different language models on WST corpus

%

语言模型	Eval92	Dev93
基线系统	8.74	12.29
4-gram	8.35	11.57
RNN 30	7.78	11.39
RNN 30 + CBOW 50	7.64	11.28
RNN 30 + Skip-gram 50	7.67	11.23
RNN 100	7.37	11.26
RNN 100 + CBOW 50	7.30	10.77
RNN 100 + Skip-gram 50	7.21	10.93

RNN-CBOW 模型对系统的识别性能也有明显改善. 在 Eval92 测试集上,隐含层为 30 时,其词错误率为 7.64%,相比 RNN 和基线系统分别提高 1.8% 和 12.6%;隐含层为 100 时,词错误率相比 RNN 和基线系统分别提高 0.95% 和 16.5%.

与微软语料库的实验结果相比,本文方法在 WSJ 语料库上对系统识别性能的提升更明显. 这说明在语料库数据更大时,词向量的训练更充分,对长距离上下文信息的描述也更准确,对语言模型的性能提高也更大. 尤其是 RNN-CBOW 模型,在大词汇量语音识别系统中,其性能相比小词汇量识别系统有明显提高,证明此方法的有效性,也说明训练数据量的大小对词向量的性能有一定影响.

为更准确对比不同识别系统的性能,排除测试数据随机性对实验结果造成的干扰,得到更可靠的结论,在 WSJ 上对不同模型的识别结果进行统计假设检验. 利用 NIST 提供的 SCTK 工具包对词错误率进行显著性测试(Significance Test) ,包括: 配对句子分词错误率(Matched Pair Sentence Segment Word Error) 测试,简称 MP 测试; 符号成对比较说话人词准确率(Signed Paired Comparison Speaker Word Accuracy) 测试,简称 SI 测试; Wilcoxin 符号秩说话人词准确率(Wilcoxin Signed Rank Speaker Word Accuracy) 测试,简称 WI 测试. 测试结果如表 5 所示.

表 5 中的每项给出对应的两种语言模型的测试

结果,其中,“MP：RNN-CBOW”表示在 MP 测试中,RNN-CBOW 词错误率更低.“MP：RNN”表示在 MP 测试中,RNN 模型的词错误率更低,识别结果的不同是由于测试数据的随机性造成的. 从表 5 中可看出,3 种显著性测试结果均表明,在 5% 的显著性水平下,本文提出的两种词向量改进的 RNN 模型均优于传统的 N-gram 模型和原有 RNN 模型,肯定本文方法的有效性.

表 5　不同语言模型词错误率的显著性测试结果

Table 5　Significance test results for WER of different language models

	RNN-CBOW	RNN-Skip-gram
传统 N-gram	MP：RNN-CBOW SI：RNN-CBOW WI：RNN-CBOW	MP：RNN-Skip-gram SI：RNN-Skip-gram WI：RNN-Skip-gram
RNNLM	MP：RNN-CBOW SI：RNN-CBOW WI：RNN-CBOW	MP：RNN-Skip-gram SI：RNN-Skip-gram WI：RNN-Skip-gram

5　结 束 语

本文提出一种结合词向量特征的循环神经网络语言模型改进方法,通过引入词关联信息的方法解决模型中存在长距离信息的学习问题,改善语言模型的性能,并进行困惑度测试和连续语音识别实验. 实验表明本文方法可有效改善语言模型性能,降低识别系统的平均词错误率. 在今后的研究工作中,将重点关注如下方面: 提高词关联信息的准确率,考虑更长的关联信息,句子关联信息的有效利用.

参 考 文 献

[1] Schwenk H. Continuous Space Language Models. Computer Speech and Language, 2007, 21(3): 492-518

[2] Bengio Y, Ducharme R, Vincent P, *et al*. A Neural Probabilistic Language Model. Journal of Machine Learning Research, 2003, 3: 1137-1155

[3] Mikolov T, Karafiát M, Burget L, *et al*. Recurrent Neural Network Based Language Model // Proc of the 11th Annual Conference of the International Speech Communication Association. Makuhari, Japan, 2010: 1045-1048

[4] Mikolov T, Kombrink S, Burget L, *et al*. Extensions of Recurrent Neural Network Language Model // Proc of the IEEE International Conference on Acoustics, Speech and Signal Processing. Prague, Czech Republic, 2011: 5528-5531

[5] Bengio Y, Simard P, Frasconi P. Learning Long-Term Dependencies with Gradient Descent Is Difficult. IEEE Trans on Neural Networks, 1994, 5(2): 157-166

[6] Son L H, Allauzen A, Yvon F. Measuring the Influence of Long Range Dependencies with Neural Network Language Models // Proc of the NAACL-HLT Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT. Montreal, Canada, 2012: 1-10

[7] Martens J, Sutskever I. Learning Recurrent Neural Networks with Hessian-Free Optimization [EB/OL]. [2014-02-10]. http://www.icml-2011.org/papers/532_icmlpaper.pdf

[8] Sundermeyer M, Schlüter R, Ney H. LSTM Neural Networks for Language Modeling [EB/OL]. [2014-02-10]. http://www-i6.informatik.rwth-aachen.de/publications/download/820/Sundermeyer-2012.pdf

[9] Shi Y, Wiggers P, Jonker C M. Towards Recurrent Neural Networks Language Models with Linguistic and Contextual Features // Proc of the 13th Annual Conference of the International Speech Communication Association. Portland, USA, 2012: 1664-1667

[10] Auli M, Galley M, Quirk C, *et al*. Joint Language and Translation Modeling with Recurrent Neural Networks // Proc of the Conference on Empirical Methods in Natural Language Processing. Seattle, USA, 2013: 1044-1054

[11] Yao K, Zweig G, Hwang M Y, *et al*. Recurrent Neural Networks for Language Understanding [EB/OL]. [2014-02-10]. http://research.microsoft.com/pubs/200236/RNN4LU.pdf

[12] Hinton G E. Learning Distributed Representations of Concepts // Proc of the 8th Annual Conference of the Cognitive Science Society. Amherst, USA, 1986: 1-12

[13] Mikolov T, Chen K, Corrado G, *et al*. Efficient Estimation of Word Representations in Vector Space [EB/OL]. [2014-02-10]. http://arxiv.org/pdf/1301.3781.pdf

[14] Mikolov T, Sutskever I, Chen K, *et al*. Distributed Representations of Words and Phrases and Their Compositionality [EB/OL]. [2014-02-10]. http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf

[15] Marcus M P, Marcinkiewicz M A, Santorini B. Building a Large Annotated Corpus of English: the Penn Treebank. Computational Linguistics, 1993, 19(2): 313-330

[16] Mikolov T, Deoras A, Kombrink S, *et al*. Empirical Evaluation and Combination of Advanced Language Modeling Techniques [EB/OL]. [2014-02-14]. http://www.fit.vutbr.cz/~imikolov/~rnnlm/is2011_emp.pdf

[17] Povey D, Ghoshal A, Boulianne G, *et al*. The Kaldi Speech Recognition Toolkit [EB/OL]. [2014-02-10]. http://homepages.inf.ed.ac.uk/aghoshal/pubs/asru11-kaldi.pdf