**基于字的神经网络语言模型训练方法及其系统**

1. **背景**

语言模型描述语言内部的词语搭配关系，被广泛应用在自然语言处理和语音识别等领域。一个高质量的语言模型对连续语音识别和机器翻译等任务都具有重要意义。

1. **统计语言模型**

当前主流的语言模型为基于概率的统计语言模型，即用概率来描述词与词之间的搭配关系和句子产生能力。统计语言模型被广泛应用于各种自然语言处理问题，如中文分词、词性标注、机器翻译、自动摘要等。在语音识别领域，统计语言模型对减小搜索空间，提高识别准确度也具有重要意义[7]。

举例而言，假如输入的拼音串为“nixianzaiganshenme”，对应的汉字输出可以有多种形式，如“你现在在干什么”、“你西安在干什么”，等等，那么到底哪个才是正确的结果呢？利用统计语言模型，我们可以计算出前者的概率大于后者，因此转换成前者在多数情况下更为合理。

具体而言，统计语言模型计算由一串词) 组成句子s的概率：

 (1.1)

由概率的乘法准则可知：

 (1.2)

(1.2)式意味着每一个词的出现概率依赖于之前出现的所有词，即(。这一子串定义为的历史子串，即:

　 　(1.3)

则1.2式写成:

 (1.4)

随着句子长度的增加，将线性增长，导致模型复杂度过高而无法建模。通常的方法是假设只包含前面的(n-1)个词，即:

 (1.5)

当前主流的统计语言模型是n-gram语言模型。这一模型通过统计词与其历史子串的共现概率来计算，即：

 (1.6)

其中为某一子串在训练语料中出现的次数。利用(1.6)式的计算方法和(1.5)式的近似假设，配合一系列平滑化算法[2]，n-gram语言模型可以得到相当好的建模效果。实际应用中，(1.5)式中的n一般取为n=3或n=4，即三元和四元n-gram语言模型。

1. **基于神经网络的语言模型**

N-gram语言模型的一个显著缺陷是：基于(1.6)式，新词和低频词难以得到有效的概率统计。基于此，人们发明了各种平滑算法，如discount, back-off, interpolation等。这些方法在一定程度上改善了n-gram在低频词上的性能，但基于模型本身的缺陷，这一困难始终无法从根本上解决。

随着神经网络的兴起，人们开始尝试利用神经网络构造语言模型。与n-gram不同，神经网络对参数进行高度共享，因此对低频词具有天然的平滑能力。神经网络语言模型(Neural Network Language Model, NNLM) 的最早由Bengio等人在2001年提出[1]，近年来一些学者开始展开这方面的研究，并取得一系列成果，如[3,4,5,6,8]，但总体而言, 对NNLM的研究还处在起步阶段。

具体而言，NNLM通过一个多层感知网络(MultiLayer Perceptron, MLP)来计算:

 (1.7)

(1.7) 式中为模型所接受的词表V中的第 个词，为神经网络的映射函数， 为输入为时神经网络输出的第个值，对应取的概率值。

图 1.1 给出一个典型的NNLM语言模型。该模型由四层结构组成，分别为输入层、映射层、隐藏层和输出层。输入层对应词的历史子串。将中的每个词表示为一个维度为|V|的向量，其中 |V|表示模型词表的大小。该向量只包含一个非零值，对应于该词在词表V中的索引值。的长度为(n-1)，因此网络的输入层结点数为(n-1)|V|。



图 1.1 词神经网络语言模型

输入层中的每个词对应的|V|维向量经过一个线性变换投影到一个维度为m的向量，称为的词向量。中所有词的词向量首尾相接拼起来，组成一个的向量b，即图1.1中的映射层。

网络的隐藏层由映射层经过线性变换M附加一个非线性激活函数tanh()得到。记隐藏层第i个节点为:

 (1.8)

其中为线性变换矩阵M的第(i,j)个元素。

网络的输出层由个节点组成，记其中第i个节点的输出值为，对应wk为词表V中第i个词的概率。该概率由隐藏层输出经过线性变换R后附加非线性激活函数softmax()得到，计算公式为：

 (1.9)

 (1.10)

其中为线性变换矩阵R的第(i,j)个元素。

1. **问题描述**

相比当前主流的n-gram语言模型，NNLM对模型参数的共享更直接有效，因而对低频词具有天然的光滑性，在建模能力上具有显著优势。另一方面，当前基于词的NNLM语言模型也具有明显的弱点，包括：

1. 训练效率低，耗时长。当词表增大到超过10w以上时，训练时间变得难以接受；

2 解码时间较长，不能满足实际要求；

3 对未登录词难以解决。中文是动态语言，新词不断涌现，当前基于词的NNLM无法处理这些新词，同时不断增加的新词使得模型越来越复杂，即使要重新训练模型也极为困难。

为了解决上述问题，本发明提出基于汉字来对中文进行NNLM语言模型建模。汉语中汉字的数量是固定的，常用的字数量大约在6千左右，远远小于词语的数量，因而基于字来建模可以极大降低模型的复杂度，提高训练和解码的效率。同时，因为新出现的词语也可以汉字串的形式表示，因此完全可以被基于字的神经网络所描述，这省去了现有NNLM系统对新词进行重新训练的困难。

然而，由于汉字本身具有多义性，基于字的NNLM语言模型很难对字概率进行精确建模，因此需要一系列新方法对其进行改进，才能满足实际应用的需要。

1. **发明要点**

为了解决上述大词表NNLM语言模型学习中的低效率问题和处理未登录词上的困难，本发明提出基于字的NNLM语言模型建模方法;　为解决基于字的NNLM模型在刻画汉字多义性上的缺点，本发明提出对字向量进行预训练的学习方法。

**1. 基于字的NNLM语言模型**

为了解决大词表NNLM语言模型训练的低效率问题和处理未登录词上的困难，本发明提出基于汉字进行NNLM语言模型建模。因为在中文里汉字的数量是一定的，常用汉字数量大约在6千左右，远远的小于词的数量，因此基于字的NNLM模型要远小于基于词模型。同时，因为任何新词都可以被拆分成字串，基于字的语言模型可以不经过重新训练而对新词具有较好的描述能力。

基于字的NNLM语言模型如图3.1所示。与基于词的NNLM语言模型相比，基于字的NNLM模型的输入层由原来的词变成了字，如“电脑 手机”变为“电 脑 手 机”。为了获得更多的上下文信息，基于字的NNLM语言模型中的历史子串长度（即输入层的字数）要大于基于词的NNLM语言模型的历史子串长度，以保证模型能学习足够的上下文信息。图3.1给出了本发明提出的基于字的NNLM模型的网络图。



图3.1 基于字的神经网络语言模型

如图3.1所示，该神经网络输入层对应字的历史子串。将中的每个字表示为一个维度为|V|的向量，其中 V为所有汉字集合， |V|表示这个集合的大小，即汉字总数。输入层中的每个字对应的|V|维向量经过一个线性变换投影到一个维度为m的向量，称为的字向量。中所有字的字向量首尾相接拼起来，即组成一个的向量b，即图3.1中的映射层。

网络的隐藏层由映射层经过线性变换M附加一个非线性激活函数tanh()得到。记第隐藏层的第i个节点为:

 (3.1)

其中为线性变换矩阵M的第(i,j)个元素。

网络的输出层由个节点组成，其中第i个节点的输出值记为，对应ek为字表V中第i个字的概率。该概率由隐藏层输出经过线性变换R后附加非线性激活函数softmax()得到，计算公式为：

 (3.2)

 (3.3)

其中为线性变换矩阵R的第(i,j)个元素。

1. **基于预训练字向量的NNLM建模**

上述基于字的NNLM语言模型解决了大词NNLM语言模型的训练效率问题，解决了NNLM处理未登录的困难。但是，由于中文的汉字具有多义性，用传统方法训练基于字的NNLM语言模型很难对这些歧义进行有效学习。为了解决此问题，本发明提出了基于预处理字向量的NNLM语言模型。

基于预处理字向量的NNLM语言模型是对基于字的NNLM语言模型的改进和增强。事实上，NNLM语言模型建模包括两个步骤：（1）字向量映射，即通过学习，将输入层中的每个汉字投影为映射空间中的一个字向量; (2) 字概率预测，即在字向量的基础上对句当前汉字进行概率计算。现有的NNLM语言模型训练方法将这两个步骤统一到一个网络中，因此字向量映射与字概率预测学习是混淆在一起的，不仅效率低，而且字向量难以得到优化。本发明提出对字向量进行单独预训练，使其对语义信息具有更强的区分性。由于预处理的目的是获得更具有区分性和表达性的字向量，而不是优化语言模型，因此在预处理中所用的模型更加简单高效，训练所得模型对字的多义性具有更强的区分能力。本发明中，我们采用简单神经网络的办法来学习字向量模型，详细描述如下：



图3.2 预训练字向量模型

图3.2为我们采用的字向量学习模型。和图3.1中的NNLM语言模型相对照，字向量模型可以看作是NNLM语言模型的输入层和映射层，这两层结构的目的也是学习字向量。然而，与NNLM的字向量学习网络不同的是，图3.2中的预训练字向量模型的输出不仅包含当前字的前向上下文，还包含其后向上下文，这意味着该模型的训练目标除了对历史信息进行优化外，还对未来信息进行优化。具体而言，预训练字向量模型把当前汉字作为输入，在输出层预测出当前汉字前后上下文中一定窗长范围内的字序列。即：给定一个字序列，模型的优化的目标是使下式取值最大化：

 (3.5)

其中，是上下文窗长。c的值越大，模型所刻画的预测信息就越复杂，得到的向量性能越好，代价是训练时间越长。模型中，由下式得到：

 (3.6)

其中，为字e的字向量，由模型中的映射层得到。公式分母中的e对所有汉字进行加和。通过对映射层的连接权重（映射矩阵）进行优化，我们得到独立的字向量模型。与传统NNLM语言模型相比，预训练字向量模型使得产生的字向量对汉字的内在意义具有更强的代表性和区分性。



图3.3 基于预训练字向量的NNLM语言模型

依照这个思路，我们的NNLM语言模型训练系统如图3.3所示：在预训练阶段，我们首先训练出一个独立的字向量模型; 在NNLM语言模型建模时，我们利用前期得到的预训练字向量模型作为映射层并保持不变，对剩余的隐藏层和输出层进行优化。由于只对两层网络进行优化，这一方法要比传统NNLM三层优化方法更高效。

1. **方案优势**
2. 本方案利用中文汉字的有限闭集特性，以汉字为单位代替词进行神经网络语言模型建模，避免了大词表NNLM模型效率低下的问题，解决了基于词的NNLM在处理未登录词上的困难。
3. 我们将字向量模型从神经网络语言模型中独立出来，提高了训练效率的同时，也增强了字向量的表征能力和语义区分能力。
4. **实现流程**

****

图5.1 基于预训练字NNLM语言模型建模流程

六 参考文献

[1] Bengio Y, Schwenk H, Senécal J S, et al. Neural probabilistic language models[M] Innovations in Machine Learning. Springer Berlin Heidelberg, 2006: 137-186.

[2] Katz S. Estimation of probabilities from sparse data for the language model component of a speech recognizer[J]. Acoustics, Speech and Signal Processing, IEEE Transactions on, 1987, 35(3): 400-401.

[3] Tomas Mikolov and Geoffrey Zweig, [Context Dependent Recurrent Neural Network Language Model](http://research.microsoft.com/apps/pubs/default.aspx?id=176926), in *Spoken Language Technologies*, IEEE, 2012

[4] Zamora-Martínez, F., Castro-Bleda, M., España-Boquera, S.: [Fast evaluation of connectionist language models](http://dx.doi.org/10.1007/978-3-642-02478-8_5). In: 10th International Work-Conference on Artificial Neural Networks. LNCS. Springer (2009) 144--151.

[5] Bengio, Y. and Senecal, J.-S. (2008) [Adaptive Importance Sampling to Accelerate Training of a Neural Probabilistic Language Model](http://www.iro.umontreal.ca/~lisa/pointeurs/importance_samplingIEEEtnn.pdf). IEEE Transactions on Neural Networks, to appear.

[6] Schwenk, H. and Gauvain, J.-L. (2004) Training Neural Network Language Models On Very Large Corpora. In Joint Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 201-208.

[7] Manning, C. and Schutze H. (1999) [Foundations of Statistical Natural Language Processing](http://nlp.stanford.edu/fsnlp/), MIT Press.

[8] Schwenk, H., Gauvain, J.-L. (2002) Connectionist Language Modeling for Large Vocabulary Continuous Speech Recognition. In ICASSP 2002, pages I:765-768.