特定领域的汉语语言模型平滑算法比较研究

杨　琳，张建平，颜永红
（中科院声学所 中科信利语音实验室，北京 100080）
E-mail: lyang@hccl.ioa.ac.cn

摘　要：为了完成特定领域的语音识别任务，利用有限的语料建立高性能的语言模型成为提高系统性能的关键。针对此问题，对特定领域的语言模型进行了研究。提出了利用高频新词来加强模型的领域特征的方法，采取了两种方案：一种是将高频新词直接加入原有字典，并在训练过程中增加这些新词的权重，使模型更能表达与领域相关的特征；一种是基于高频新词统计出一个和领域相关的小词表，并对这两种方案进行了比较研究。通过实验研究了适合汉语语言的平滑策略。最后，实验结果表明，对于特定领域问题，语言模型平滑算法对模型性能影响较大；采用适合汉语的 Witten-Bell 插值平滑，可以使识别率达到 88.4%，比通用模型性能相对提高了 18.18%。

关键词：语言模型；特定领域；语音识别；平滑；字典

文章编号：1002-8331(2006)32-0014-03　**文献标识码**：A　**中图分类号**：TP391.4

Comparative Study on Smoothing Algorithms for Domain-Specific Chinese Language Models

YANG Lin, ZHANG Jian-ping, YAN Yong-hong
(ThinkIT Laboratory, Institute of Acoustics, Chinese Academy of Sciences, Beijing 100080, China)

Abstract: It is important to build a powerful language model by using limited corpora in the field of speech recognition for a specific domain. To deal with this problem, two methods concerning how to process new words with high frequencies in a specific domain are presented. One way is to add the new words to the dictionary directly and then give them a high weight in the procedure of training. The other is to work out a new dictionary according to the new words. And based on some comparative experiments, these two methods and various smoothing algorithms are studied in detail. At last, it can be concluded that the performance of language model is affected by the smoothing algorithm greatly, and the Witten-Bell interpolation method could improve the recognition rate to 88.4%, which is 18.18% higher than the general language model.

Key words: language model, specific domain, speech recognition, smoothing algorithm, dictionary

1　引言

语言模型自上个世纪 80 年代提出以来，已经广泛地应用于语音识别、印刷体文字识别、手写体文字识别和文字校对等领域中。根据贝叶斯理论，语音识别任务可以表示为：$W^* = \arg\max_W P(W|O) = \arg\max_W P(O|W)P(W)$。其中，$O$ 表示观察序列，W 是从声音信号提取的特征向量；$P(O|W)$ 是待识别的汉字串；为声学模型，表示在词序列 W 情况下产生特征向量 O 的条件概率；$P(W)$ 预测在某种语言中词序列出现的概率，是语言模型的主要任务。可以看出，语言模型在语音识别中以先验概率的形式发挥着重要作用。

多年来尽管有一些新的语言模型建模方法被提出，如 class-based 模型、最大熵模型、结构化语言模型以及潜在词义分析模型等，但是目前在语音识别任务中，对统计语言模型的研究还局限于 N 元文法模型，即 Ngram。Ngram 模型面临的最主要问题就是数据稀疏问题和对领域的强依赖性。利用大量语料训练的通用语言模型，即使在通常情况下取得了比较好的结果，但是当用到特定领域的识别任务时，由于与领域中词语的概率分布不一致，效果也会不尽人意。

本文利用特定领域的语料研究了 Ngram 语言模型的训练方法，为了解决数据稀疏问题，通过实验比较了各种平滑方法在小语料特定领域情况下的性能。结果表明，恰当的处理与领域相关的高频新词，选择合适的平滑方法可以增强模型的性能，提高识别率。

2　Ngram 模型及其平滑算法

Ngram 模型采用了 Markov 假设，认为每个预测变量只与长度为 N-1 的上文有关，即给定句子 $W=w_1w_2\cdots w_L$ 的情况下，W 出现的概率可以表示为

$$P(W)=\prod_{i=1}^{L}P(w_i|w_{i-n+1},w_{i-n+2},\cdots,w_{i-1}) \quad (1)$$

其中 n 为模型的阶数，当 n 分别取 1、2、3 时，就构成了常用的 Unigram，Bigram，Trigram。n 的取值决定了模型的精度和复杂度，理论和实验表明，n 值越大，对单词间依赖关系的描述越准确，但是随之带来的是模型复杂度增加，模型参数急剧膨胀，因

基金项目：国家自然科学基金资助项目（编号：60535030）。
作者简介：杨琳，博士研究生；张建平，副研究员；颜永红，主任研究员。

此，合适的 n 值是在模型精度和复杂度上的一个折中选择。在语音识别任务中，一般选择 Trigram 语言模型。在基于训练语料建立的 Ngram 模型中，模型参数的计算，一般采用最大似然(Maximum Likelihood)法，即

$$P_{ML}(w_i|w_{i-1}^{i-n+1})=\frac{C(w_i|w_{i-1}^{i-n+1})}{C(w_{i-1}^{i-n+1})} \tag{2}$$

其中 $C(w_{i-1}^{i-n+1})$ 表示单词串 $w_{i-n+1},\cdots,w_{i-1}$ 在训练语料集中出现的频率。

显然，当某个词串在训练语料中没有出现时，采用最大似然法得出的概率为 0。当此词串在测试语料中出现时，就会造成被测试语句整句话概率为 0 的结果，使得模型性能很差，根本无法应用到实际系统中，这个问题就是统计语言模型中的关键问题——数据稀疏问题。有实验[2]表明，对一个包含约 240 M 单词的语料库采用最大似然方法建立词典数为 60 000 个单词的 Trigram 模型，当该模型用于实际测试语料时，发现测试集中只有 69%的 Trigram 包含在模型中。因此，数据稀疏问题是 Ngram 语言模型必须解决的一个问题。

目前解决 Ngram 模型的数据稀疏问题主要依靠各种平滑(smoothing)方法，包括折扣(discounting)和预测两部分。折扣算法主要是对测试语料中出现的 Ngram 的词频按照某种方式打一个折扣，从概率总体中分出一部分概率，再以一定的规则分配给未出现的情况，主要包括的算法有：Jelinek-Mercer, Additive, Good-Turing, Witten-Bell, Absolute, Kneser-Ney 以及 Modified Kneser-Ney 等[3]。而预测算法，是利用低阶的信息来预测高阶的概率，其重要实现方法有回退(backoff)和插值(interpolate)两种。回退算法以 Katz backoff 用的最多，可以表示为

$$P_{smooth}(w_i|w_{i-n+1}^{i-1})=\begin{cases}P_{smooth}(w_i|w_{i-n+1}^{i-1}) & \text{if } C(w_{i-n+1}^{i-1})>0\\ \gamma(w_{i-n+1}^{i-1})P_{smooth}(w_i|w_{i-n+2}^{i-1}) & \text{if } C(w_{i-n+1}^{i-1})=0\end{cases} \tag{3}$$

其中，C_{i-n+1}^{i-1} 表示词序列 $w_{i-n+1},w_{i-n+2},\cdots,w_{i-1}$ 出现的次数，权重系数 $\gamma(w_{i-n+1}^{i-1})$ 的求取依据条件概率和为 1 的原则。

插值的方法可以表示为

$$P_{smooth}(w_i|w_{i-n+1}^{i-1})=(1-\lambda_{i-n+1}^{i-1})P_{ML}(w_i|w_{i-n+1}^{i-1})+\lambda_{i-n+1}^{i-1}P_{smooth}(w_i|w_{i-n+2}^{i-1}) \tag{4}$$

比较两种方法，可以看出，对于出现次数为 0 的 Ngram，两种方法都利用了低阶的概率分布；但是对于出现次数大于 0 的 Ngram，插值方法利用了来自低阶的信息，而回退方法并不包含低阶信息。

使用不同的打折方法和预测方法，都会影响模型的性能，尤其是在小语料的情况下。对模型性能的衡量一般采用模型针对某一测试集的复杂度(Perplexity, PP 值)为标准，这是一种基于平均概率的评判准则，定义为

$$PP=\exp\left\{-\frac{1}{L}\log P(w_1w_2\cdots w_L)\right\} \tag{5}$$

其中，L 是测试集中所包含的词数。复杂度反映了测试文本相对于当前模型的不确定性，不确定性越大，PP 值越大，说明模型表示的分布与测试集相差越远，也给语音识别任务造成更大的困难。因此，对于特定领域的语音识别，任务就是建立一个模型，使其针对这一领域选出的测试语料有更小的 PP 值。

在语音识别领域，在相同的声学模型和解码算法的基础上，利用识别率和词误识率(Word Error Rate, WER)也可以达到验证语言模型性能的目的。

3 特定领域的问题

特定领域的语言模型面临的主要问题，一是数据稀疏问题，二是对该领域出现的高频新词的处理问题。对于第一个问题，可以采用平滑方法给予一定的解决。对于领域内出现的高频新词如果处理不当，会增加更多的词典外词汇(Out Of Vocabulary, OOV)，不但无法提高模型的性能，甚至还会造成模型性能的下降，从而影响识别效果。

为了尽量多地利用语料中高频新词的信息，使模型更具备领域特征，对于这些高频新词的处理，提出了两种方案，并进行了实验比较。

(1)把通过半自动半手工从语料中提取的高频新词直接加入到原有的 43 000 词的词表中，并且在模型训练过程中适当增加这些高频词的权重，以此来增强模型的领域特征。

(2)根据利用原始词典从语料中统计出的该领域的高频词，再加上新词，以及 1 000 多个常用的单字词组成一个规模只有几千词的较小的新词典。以新词典为基础，处理语料和训练模型。这种方案可以使模型结构更加紧凑，从而提高识别效率，但是，也可能由于词表的大规模减小，造成语料信息利用不充分，以及 OOV 的增加，从而影响模型的性能。

4 实验与结果

4.1 实验数据

实验是根据用户需要，在身体健康和老年保健这一特定领域进行。实验所用语料来自互联网与此领域相关的网页，经过处理后，得到的可用文本 4.5 M，包含 130 多万词条。测试语料也是来自互联网的领域相关的句子以及部分人工整理的此领域常见的短语，组成两份测试语料：Test1 由较多的短语组成；Test2 由短句和长句共同组成。测试用的语音是从测试语料中整理出的句子和短语，也分为两部分，一部分包含 50 个短语，另一部分包含 50 个短语和 30 个长句。

实验所采用的一套 LVCSR 系统，在使用通用语言模型的基础上，对 431 个说话人的通用测试集进行测试的识别率为 86.5%。然而对特定领域的测试集 1 和 2 的识别率分别为 69%和 74.8%，足以说明语言模型对领域的强依赖性和研究特定领域语言模型的必要性。

4.2 两种高频新词处理方案比较

经过半自动半手工的方法，挑选出与此领域相关的 120 个高频新词。采用第一种方案，将这些新词添加到词典中，在训练过程中给予这些新词一个较大的权重，以此来突出模型的领域特征；采用第二种方案，创建一个新的词表，包括依据老词表统计的高频词，新词以及常用单字词，新词表大小约为 6 000 词条。采用 PP 值进行比较，结果如表 1 所示。

表 1 两种高频新词处理方案的实验比较结果(PP 值)

	Test1			Test2		
打折方法	GT	KN	MKN	GT	KN	MKN
方案 1	236.793	229.108	272.252	49.845	30.305	48.959
方案 2	244.879	224.58	266.495	200.57	92.209	110.318

表中 GT 表示 Good-Turning 打折方法，KN 表示 Knesser-Ney 打折方法，MKN 表示 Modified KN 打折方法。全部采用 backoff 预测策略实现平滑。通过比较实验结果可以看出，对于

以短语为主体的测试语料，虽然对于不同的平滑方法，优劣性能略有差异，但是从 PP 值来看，两种方案的性能相差不多，因为对于短语的测试主要利用了模型的低阶特性，两种方案在 Unigram, Bigram 的分布上相似。而对于包含长句和短语的测试集而言，第一种方法明显优于第二种方法。正如预期的那样，方案二虽然提高了解码速度，但是由于词表规模的大幅度减小，不仅丧失了语料中的信息，而且对模型的高阶分布产生了较大的影响，降低了模型的性能。因此，从整体来看，第一种方案优于第二种方案，以下进行的平滑方法的比较实验都采用第一种方案。

4.3 各种平滑方法对模型性能的影响

为了解决语言模型的关键问题——数据稀疏问题，对平滑方法进行了比较研究，主要包括 Good-Turing 打折，Witten-Bell 打折的回退和插值实现，Knesser-Ney 打折，Modified Knesser-Ney 打折的回退和插值实现，分别表示为：GT, WB_b, WB_i, KN, MKN_b, MKN_i。

对测试语料 1 和测试语料 2 进行 PP 值比较，结果如表 2 所示。

表 2 不同平滑算法的比较结果(PP 值)

	GT	WB_b	WB_i	KN	MKN_b	MKN_i
Test1	236.793	233.714	198.648	229.108	272.252	218.031
Test2	49.845	32.389 2	29.321 2	30.305 8	48.959 5	41.565 4

从两个表都可以看出，Test2 的 PP 值明显低于 Test1 的 PP 值，这是由于 Test2 中大部分是完整的长句，里面的语言现象与训练语料相符，即测试集和训练集的词语分布比较一致。而 Test1 中大部分是此领域中经常出现的短语和较短的句子，其词语概率分布和训练集不是非常一致，因此，相比之下模型更好地反映了第二个测试集。通过在同一测试集上采用不同的平滑算法，观察其 PP 值的变化趋势，来比较各种平滑算法的优劣。

理论上，PP 值越小模型性能越好，对应于语音识别任务中的识别率越高。对应于两组语音测试数据的识别结果如图 1 和图 2 所示。

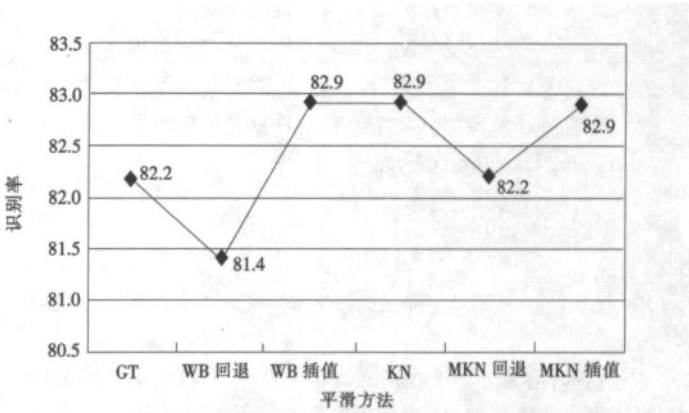

图 1 测试语音集 1 的识别结果比较

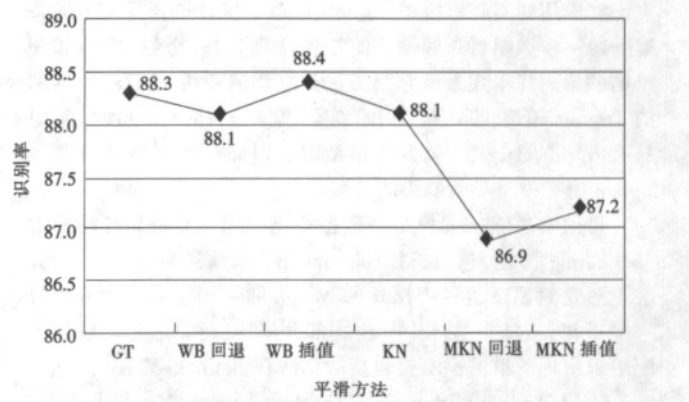

图 2 测试语音集 2 的识别结果比较

实验结果表明，在大部分情况下，语言模型的 PP 值与识别率有一定的对应关系，PP 值越小，识别率越高；对于特定领域的问题，各种平滑算法对模型的性能和识别结果有较大的影响，Witten-Bell 的插值方法相对其他平滑方法性能更优。虽然有文献[3]表明，在英文中 MKN 插值方法在不同规模的训练集上都优于其他方法，但是由于特定领域词语的概率分布对领域有一定的依赖性，而且英语和汉语也存在一定的差异性，如汉语由字构成词，大部分常用搭配都作为一个独立的词出现；而英语中的词就是一个个独立的单词，单词之间的依赖关系差异很大，类似 great deal, San Francisco 的常用搭配词组比较多，而 KN[4]和 MKN 方法就是为解决这类语言现象而提出的，因此考虑到两种语言的构词差异，实验结果与前人的研究并不冲突。从实验结果还可以看出，在测试集 1 上，插值方法比回退方法的识别率最多可绝对提高 0.3%，在测试集 2 上，最多可绝对提高 1.5%，因此采用插值方法确实比回退方法更好，这与文献[3]的结论一致。而且，与通用模型相比，在测试集 2 上识别率从 74.8%提高到了 88.4%，相对识别率提高了 18.18%。

5 结论

针对特定领域建立语言模型能够更好地完成与主题相关的语音识别、文本检索以及文本分类等任务，如何解决语言模型共有的数据稀疏问题和如何处理特定领域出现的高频特征词成了提高系统性能的关键问题。本文针对以上问题，通过实验比较了通用模型和特定领域模型的差异，两种处理高频新词的方法以及各种平滑算法的性能，结果表明：语言模型对领域有很强的依赖性，通用模型的性能在文中特定领域的情况下，使识别率相对下降了 13.5%；适当的添加与领域相关的新词，增强语言模型的领域特征，以及选择适合汉语语言的 Witten_Bell 插值平滑，可以在相同的识别框架下，使识别率从通用模型的 74.8%提高到特定领域模型的 88.4%。在今后，针对汉语语言的特点，寻找适合汉语的平滑策略或者以新的方法建模都是值得研究的方向。(收稿日期：2006 年 9 月)

参考文献：

[1] ROSENFELD R.Two decades of statistical language modeling: where do we go from here? [C]//Proceedings of the IEEE, 2000, 88: 1270~1278.

[2] Rosenfeld R.A maximum entropy approach to adaptive statistical language modeling[J].Computer Speech and Language, 1996, 10: 187~228.

[3] CHEN S F, GOODMAN J.An empirical study of smoothing techniques for language modeling[J].Computer speech and language, 1999, 13: 359~394.

[4] Kneser R, Ney H.Improved backing-off for m-gram language modeling[C]//Proceedings of the IEEE International Conference on Acoustics: Speech and Signal Processing, 1995: 181~184.

[5] CHEN S, ROSENFELD R.A survey of smoothing techniques for ME models[J].IEEE Trans Speech and Audio Processing, 2000, 8: 37~50.