基于频率的语言模型在OCR中的应用限制

雷·史密斯Goo
gle Inc
美国山景城
Rays@google.com

摘要—尽管在语音识别和机器翻译应用程序中使用了大语言模型，但OCR系统在使用语言模型时却“远远落后”。造成这种情况的原因不是OCR社区的落后，而是基于频率的语言模型在不经过精心应用的情况下，造成的损害多于好处的事实。本文借助Google图书n-gram语料库对这种差异进行了分析，并得出结论，需要使用噪声通道模型来密切建模基础分类器和细分错误。

关键词-OCR;语言模型。

一，引言

语言模型在OCR应用中至关重要。在英语，仅I / l / 1和O / 0的歧义就证明了这一点，因此OCR系统在过去的二十年中一直具有单词表和字符n-gram。然而，在OCR上超过二十年里，向作者提出的最常见的问题是：“为什么不使用字典？”用户继续看到“愚蠢”错误，其中错误地识别了明显的词典单词。如果拼写检查器可以找到正确的单词，那么将更好的语言模型正确地与形状分类器集成在一起肯定会做得更好吗？

答案是语言模型的权重是针对平均准确度进行优化，并增加其权重将解决一些明显的错误，但会增加不正确的词典单词的幻觉率。尽管这对于任何最佳系统都应该是正确的，但它并没有说明增加语言模型的复杂性的效果，也没有提及其修正的选择性（应提高最佳精度）。

语音识别系统已经使用了更多复杂的语言模型比OCR系统要几十年[2] [3]，但是OCR系统并未遵循。这种方法二分法有几种可能的原因：

- 起源于1980年代，软件OCR应用程序的内存占用量很小，很难找到增加较大语言模型所需的100％或更多内存增加量的理由。
- OCR研究人员尝试了更大的语言模型，但发现它们没有成功。由于很少报告有负面结果，因此这一直是OCR社区维护良好的“秘密”。
- OCR研究人员是落后的，他们没有做出必要的努力以充分利用更好的语言模型。

具有语言模型的语音系统的成功[4]，而带有阿拉伯语OCR的BBN Byblos系统[5]“不可能”以上的第二个原因。在本文中，我们着手研究影响使用语言模型的OCR系统中最佳精度的因素，以及语言模型对OCR的作用可能不如对语音识别系统的作用。该调查将OCR形状分类器的简化模型和不同的语言模型（在第III部分中定义）应用于1011个单词的大型Google图书n-gram语料库[6]。

二。背景和ZIPF定律

Zipf的定律[7]指出当一种语言中的单词是如果按降序排列，则排名为n的单词的频率与1 / n成正比。因此，英语中最常见的100个单词占语料库所有单词的三分之一以上。频繁的单词是如此一致，以致它们构成了攻击简单密码的基础[8]，并且类似的方法已经在未经初始分类器训练的情况下多次应用于OCR文本[9] [10]。

1 / n分布的另一端是“长尾巴”，表示大型语言模型永远不会完整，无论它多么庞大，因此总会有一些单词或短语会因不可能而愚弄语言模型。

例如，考虑“arıd”，它可以与“ and”'n'折断，或'arid'，'i'中缺少点。由于“和”比“干旱”要频繁得多，因此语言模型需要确定的上下文才能接受“干旱”。然而，在：“ Capability Brown称条件……”，大多数人类和语言模型仍会选择“ and”，而不会继续使用“ ...，因为年降雨量少于...”，或者知道Capability Brown是18世纪的英国景观园丁。

前面的例子是炮制的，但现实甚至更离奇。在其发展的某一时刻，Tesseract [11]仅仅通过从字符分类器的n个最佳列表中提取最佳选择的词典词，并映射错误映射8-> o，1即可将“8½”识别为“烤箱”。/-> v，2-> e，“-> n。

轶事的例子很有趣，但并不令人信服。第四节中使用的数据集是Google图书英语n-gram语料库[6]，它由10个总单词集合中的大约4×106个唯一单词组成。公共语料库已被过滤[13]，仅包括至少40本书中出现的单词，但未经过滤的版本包含超过108个唯一单词。我们将语料库分为词典D和测试语料库C。W□C是m个符号或形状的真值词，Ŵ□D是候选结果词。

三，方法

为了提供合理简单的分析，我们在本文中考虑孤立的形状分类器的情况，并结合具有单词n-gram频率的语言模型

（1≤n≤3）模型或二进制n-gram字典模型。在每种情况下，形状分类器都提供一个形状列表，每个形状都有一个“概率”。不考虑允许形状由多个模型状态组成的基于HMM的模型，例如BBN系统[5]。简化假设的选择是在夸大最终准确性的方面进行的，因此为使用语言模型可以进行的改进提供了一个上限。

A.分类器模型对于每个分类的形状，（形状可以是

字形到字素簇的字形），形状分类器返回概率为p（stc）的顶级选择形状stc，还返回形状集合si中的所有其他形状：0≤i

此分类器模型是一个简单的简化。无论如何

在OCR系统中，人们可能希望形状分类器偶尔返回具有几乎相等或什至完全相等概率的多个形状，例如l / l / 1。人们还希望，如果不是正确的答案，那么正确的答案至少将接近最佳选择。虽然通常是正确的，但也确实是错误的选择通常接近正确的选择（例如，使用l / l / 1），并且经常出现在最佳答案中找不到正确答案的情况。形状分类器提供的列表。语言模型必须从此类灾难性错误中恢复的唯一机会是允许使用通配符。此处使用的简化模型实质上允许语言模型在每种情况下都使用通配符，并且权重控制通配符的应用程度。

为了进一步简化，我们假设形状为

公认的细分。因此，我们在每个位置j□[1，m]上将首选分类器形状sTC，j的串联称为首选分类器词WTC。尽管在实际的OCR系统中也需要进行大幅度的简化和分段，但是错误模型不包含任何分段错误，因此无需校正分段错误，添加这些分段错误只会增加语言模型产生幻觉的范围。不正确的字典单词。

另一个简化的假设是形状分类器

错误在统计上是独立的。实际上，由于图像质量问题，错误通常会在突发中发生，但统计独立性的假设允许对形状与字错误率的以下简单分析：如果首选形状分类器错误的概率为pe，且错误为在统计上是独立的，则长度为W的W中的一个最佳选择错误的概率为mpe（1-pe）m-1，而总错误概率为1-（1-pe）m。B.词n-gram语言模型语言模型选择通过词的最佳路径

使用对数线性模型[12]对形状分类器加权的语言模型来优化组合概率的分割图。对于n>

1，通常会使用波束搜索来选择单词的总体最佳顺序，但是由于C是n-gram的集合，而不是连续文本，因此对n>

1的处理简化为持有n-1

常数（上下文）一词，并仅考虑单个词进行错误模拟/纠正。因此，上下文结合字长m将C和D划分为许多单独的情况。

由于我们没有细分错误，因此该语言

模型选择通过晶格的最佳路径，从而：

成本= -wLM lnp□W□-wSM∑

$$\text{cost} = -w_{LM} \ln p(\hat{W}) - w_{SM} \sum_{i=1}^{n} \ln p(s_i)$$

对于最佳单词最小化。对于每个选择i，这样si，j拼写Wt和p（Ŵ）是语言模型概率。wLM，wSM分别是语言模型和形状模型的权重，最好为模型的每个字长使用不同的wLM值。

如果让r = p（W）/ p（WTC），并假设一个形状

WTC中的错误，则语言模型可以通过用单词WW中较低的选择sLM，k替换单词WTC中位置j的单个首选形状TC，k来纠正它，前提是：

在线服务

$$\ln r > \frac{w_{SM}}{w_{LM}} \left(\ln p_{\square sLM_{jk}\square\square\sum j \neq k} - \ln p_{\square sTC} \right) + \sum_{j \neq k} \ln p(s_{ij})$$

该模型表明，

C.二进制n元语法字典语言模型使用二进制字典，语言模型可以判断是否

单词包含在词典（IID）中，或不在词典中（OOD）。没有频率信息。为了将语言模型与形状分类器结合起来，OCR系统（例如Tesseract [11]）将搜索分类器结果，以找到分类器可能性最大的IID词。权重wIID控制WTC和最可能的IID字之间的平衡。使用上述简化的分类器模型，如果wIID

该模型表明，

二进制字典将取决于字典的大小，因此我们将字典的纠正能力作为字典大小的函数进行探讨。

IV。实验内容

实验是在Google图书n-

克英语语料库，每隔几十年就分裂一次。奇数年出版的书籍用于创建D，而数十年出版的书籍用于C。

A.词n-gram频率模型对于基于频率的语言模型，

单词W的位置j□[1，m]的可校正比为：

r j =min□p□W□/p□W□: □W∈Ω□W，j□，□W≠W，

$$r_j = min\left(p(\hat{W}) / p(W) \right) : \hat{W} \in \Omega(W, j), \hat{W} \neq W,$$

r j =p□W□/□如果Ω□W，j□= {W}□□□

$$r_j = p(W) / \varepsilon \text{ if } \Omega(W, j) = \{W\}$$

其中Ω（W，j）是在单词W中的位置j处由D进行通配符替换产生的单词集合，而ε是单词在D中出现一次的概率。在单词中，W的频率至少是出现频率的rj倍作为其最接近的通配符竞争对手。

类似地，损坏率

rc =max□p□W□/p□W□□: j∈[1，m]，W□∈W，j□□□□

定义了不正确但更有可能的单词击败正确单词的边距。对于每个具有频率w的语料库词W，针对形状错误率的最终单词错误率直方图将表I所示的值累加为log
r的函数，并显示了可纠正性与可破坏性之间的权衡。

该分析忽略了形状的可能性
分类器在W中的一个位置处产生错误，而语言模型通配符在另一位置处发生错误，与校正原始形状分类器错误相比，产生的单词频率更高。这是一个相对低概率的事件，但是会使最终误差被低估。

B.二进制语言模型在二进制模型中，仅尝试更正

当WTC为OOD时。在长度为m个形状的单词W的位置j处，我们估计IID幻觉的可能性，其中形状分类器错误创建了不正确的词典单词，因为h = d / Ns其中d = |Ω（W，j）|为可以产生IID字的形状集的分数。假设W□Ω（W，j），并假设语言模型在Ω（W，j）中随机选择，则具有单个形状错误的单词可以用概率pc =（1-h）/ d校正。

如果□□Ŵ□Ω（W，j），ŴWWp（W）
当字典包含Ŵ但不包含W时损坏，即大小阈值位于p（W）和p（Ŵ）之间。因此，我们要求任何不正确的通配符单词的最大频率：

xc =max□logp□W□: j∈[1，m]，W□∈W，j□，□W≠W□□□

图2中直方图的x轴是
字典的大小，但是形状错误率pe的单词错误率直方图累积了表II中显示的值，作为x = log p（Ŵ）：ŴΩΩ（W，j）的函数作为字典大小的直方图代理，并且实际的字典大小是使用相同的存储区来计算的。

C.覆盖问题过滤后的语料库仅包含在

至少40本不同的书籍[6] [13]，以及D的覆盖范围

表I.基于频率的模型的字错误率直方图

值范围说明		
w（1−（1−pe）n）0≤r≤∞一个或多个形状分类器误差的概率。		
w（1-pe）n0≤r≤rcTop-choice分类器正确，但被语言模型破坏。		
-wpe（1-pe）n-10≤r≤rj发生位置j的单个错误并已得到纠正。应用于每个j□[1，n]。		∈

表II。二进制模型的字错误率直方图

值范围说明		
w（1−（1−pe）n）-∞≤x≤0一个或多个形状分类器错误的概率。		
w（1-pe）nlog p（W）		首选分类器 正确，但由于语言模型而损坏。
-w (1-h) pe (1-pe) n-1 /d-∞≤x≤log p (W)		位置j处发生单个错误并已纠正。应用于每个j□[1，n]。 ∈

即使超过3克，C也超过99.99％。这在实际的OCR数据中是不现实的，因此使用原始语料库重新运行了实验。这将1克，2克和3克的覆盖率分别降低到99.76％，97.55％和88.66％。

D.二进制-频率混合-基于频率的语言模型的清晰结果

是正确的单词被太频繁地破坏了。随着形状分类器的错误率下降，除非语言模型的权重降低得如此之低以至于几乎没有影响，否则这种破坏效果将开始占主导地位。（参见图1。）相反，随着错误率的降低，二进制模型从更大的字典中受益更多。这建议了一个简单的二进制混合模型，其中语言模型不尝试更改IID WTC（无论是幻觉还是正确的），但是当它确实纠正了单词时，便使用频率来选择结果。由于要更正的单词始终是OOD，因此可破损率略有不同：

rc =max□p□W□/□□: j∈[1，n]，W∈□□W，j□□

$$r_c = max\left(p(\hat{W}) / \varepsilon \right) : j \in [1, n], \hat{W} \in \Omega(W, j)$$

其中ε是与OOD单词相关的频率-对应于训练语料库中的计数1。与基于频率的模型一样，直方图累积的值是log r的函数，如表III所示。

五，结果

图1显示了n-gram的最终单词错误率
频率模型，形状分类器的错误率分别为（2）中log r的2％，1％和0.2％。图1（a），（b），（c）在过滤后的语料库上分别使用了1、2和3词的词频模型，图3显示了原始语料库的结果。在y轴上，如果一个通配符替换中有一个可用的语言模型，则语言模型将以形状错误的比率替换一个更频繁的单词。随着r的增加，损坏率降低，校正率也降低，直到r =∞时，输出错误率才是输入形状分类器错误率。在低输入错误率的情况下，最佳点几乎不比输入错误率好，因为由于Zipf分布的长尾，损坏率渐近地下降到非零值：Zipf的墙。

表III。二进制混合模型的字错误率直方图

值范围说明		
w（1−（1−pe）n）0≤r≤∞一个或多个形状分类器误差的概率。		
w（1-pe）n0≤r≤rcTop-choice分类器正确，但被语言模型破坏。		
-w (1-h) pe (1-pe) n-10≤r≤rj		位置j的单个错误为OOD，并已纠正。应用于每个j□[1，n] ∈

图2显示了二进制文件的最终字错误率

形状分类器的误码率分别为In（字典大小）的2%，1%和0.2%。图2（a），（b），（c）分别使用单词1-gram，2-gram和3-gram二进制模型，类似地，图4显示了原始语料库的结果。字典越大，损坏率越低，因为更多的单词被视为IID且未更改，但更正率也降低了，因为可以找到更多的通配符。对于2克和3克模型，仅在最大字典大小下才能获得最佳错误率。图4（c）中的突然下降是由于在字典语料库D中添加了计数为1的单词引起的。

图5显示了混合词的最终字错误率

模型，形状分类器的误差率为2%，1%和0.2%，是log r的函数。图5（a），（b），（c）分别使用单词1-gram，2-gram和3-gram模型。表IV列出了所有模型的最小误差点。表IV和图5显示了简单的非线性规则的优点，即不应触摸IID字。

VI。结论：语音VS OCR和进一步的工作

频率模型的基本限制是

假设单词在语言中出现的概率（给定当前语境中n> 1的n-gram的上下文）与单词正确的概率之间存在密切的关系。如果分类器（用于OCR的形状，或用于语音的声音）较弱，则最常见的单词是最佳猜测，因为它平均获胜。弱分类器也会经常使不正确的词典单词产生幻觉，因此根据先前的n-1个单词预测单词的功能要强大得多。当分类器从根本上更加准确时，平衡发生变化，即使在前n-1个单词的上下文中，最可能的单词也不再是最佳猜测。如果分类器的首选是IID单词，那么它比编辑距离1内最可能出现的单词更可能是正确的。这解释了语音和OCR之间的方法二分法。

复杂的语言模型在

语音分类器看到的语音错误率很高，但是OCR形状分类器通常更准确（对于基于拉丁语的语言（例如英语）），因此语言模型的效果不佳。对于OCR仍具有较高错误率的语言，例如阿拉伯语和北印度语，强大的语言模型仍然非常有用。

基于频率的语言模型和对数

对于OCR，线性模型可以声明为无效；的

表IV。字错误率最小值的摘要

语料库过滤原始							
形状误差率2.00%1.00%0.20%2.00%1.00%0.20%							
输入字错误率8.40%4.20%0.84%9.10%4.60%0.91%							
型号n-gram最小输出字错误率（%）							
频率1 4.54 2.50 0.65 5.01 2.76 0.73							
频率2 3.90 2.22 0.61 3.82 2.13 0.57							
频率3 2.71 1.65 0.52 3.23 1.72 0.43							
二进制1 4.73 2.58 0.60 5.77 3.17 0.77							
二进制2 3.69 1.80 0.35 5.08 2.56 0.51							
二进制3 1.78 0.82 0.15 3.83 1.89 0.37							
混合动力1 2.87 1.37 0.26 4.64 2.28 0.45							
混合动力2 1.28 0.56 0.10 2.00 0.94 0.18							
混合动力3 0.65 0.24 0.03 1.82 0.86 0.16							

二进制频率混合显示，当WTC为OOD时，最常见的通配符词仍然是一个非常好的猜测，并且可以使用非线性特征函数将其表示为对数线性组合：

cost = -wLM lnp□W□-wTCf □W□-wSM∑ j lnp□si，j□□□□

$$cost = -w_{LM} \ln p(W) - w_{TC} f(W) - w_{SM} \sum_i \ln p(s_{i,j})$$

如果f = WTCWD，则f（Ẁ）= 0，否则为-1。

该实验的一个总体简化是

分类器错误模型，其中语言模型必须从整个字符集中进行猜测。展望未来，专门针对特定OCR形状分类器的替换和分割错误进行训练的噪声通道模型[14]似乎比盲通配符更合适，例如e-> c比o-> x更合适。那应该减少不正确的字典词的幻觉频率，使语言模型相对更强大。

Google图书n-gram语料库可用于

除了英语以外的其他语言，比较其他语言的结果将是有用且有趣的。

致谢

作者要感谢David Rika Antonova Eger，Oded Fuhrmann，Dar-Shyang Lee和Ranjith Unnikrishnan为本文做出了宝贵贡献。

参考资料

[1] M. Bokser，“ Omnidocument Technologies”，Proc.Natl.A则。 IEEE卷1992年7月，第80（7）页，第1066-1078页，IEEE。

[2] L. R. Bahl，P。F. Brown，P。V. de Souza，R。L. Mercer，“用于自然语言语音识别的基于树的统计语言模型，” IEEE trans。声学，语音和信号处理，第37卷，1989年7月，pp1001-1008。

[3] R. Kuhn，R。De Mori，“用于语音识别的基于缓存的自然语言模型”，IEEE trans。模式分析和机器智能第12卷，1990年6月，pp570-583。

[4] A. Lee，T。Kawahara，K。Shikano，“ Julius —一个开源实时大型词汇识别引擎”，Proc.Natl.Acad.Sci。 Eurospeech-2001，pp1691-1694。

[5] L. Zhidong，R。Schwartz，P Natarajan，I。Bazzi，J。Makhoul，“ BBN Byblos OCR系统的先进技术”，Proc.Natl.Acad.Sci.USA，87：3877-5。第五届ICDAR，pp337-340，IEEE Sep 1999。

[6] Ngrams.googlelabs.com/info

[7] L. A. Adamic，“ Zipf，幂律和pareto –排名教程”，http：//www.hpl.hp.com/research/idl/papers/ranking/ranking.html。

[8] S. Singh，《密码本》，1989年，Doubleday。

[9] T。 K. Ho，G. Nagy，“未经形状训练的OCR”，Proc。第15届ICPR，第27-30页，IEEE 2000。

[10] A. Kae，E。Learned-Miller，“动态学习：无字体方法难以解决的OCR问题”，Proc。第10届ICDAR，pp571-575，IEEE 2009。

[11] http://code.google.com/p/tesseract-ocr.[12] F. J. Och，“统计机中的最小错误率训练翻译”，Proc。 ACL 2003，第41届计算语言学协会年会，第160-167页。

[13] M. Jean-Baptiste等人，“使用数百万本数字化图书”，《科学》杂志331，p176，2011，doi：10.1126 / science.11□□99644。

[14] E. Brill，R。C. Moore，“噪声通道的改进错误模型拼写更正。” 2000年第38届计算语言学协会年会，pp286-293，ACL

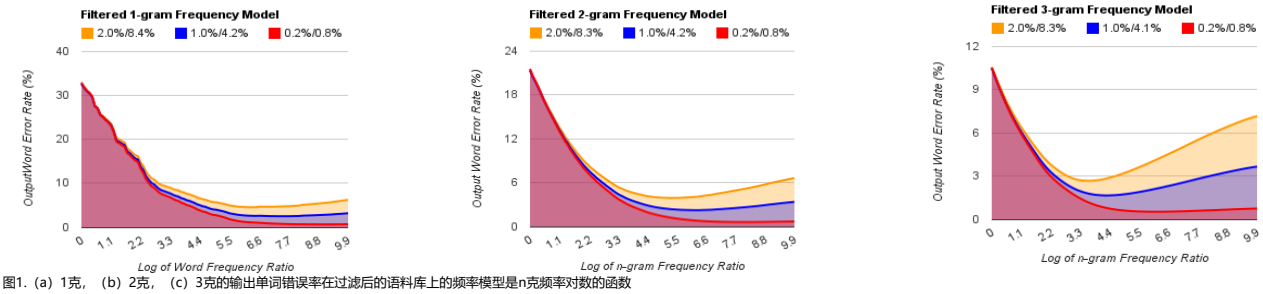

图1. (a) 1克， (b) 2克， (c) 3克的输出单词错误率在过滤后的语料库上的频率模型是n克频率对数的函数

输入形状分类器/单词错误率分别为2%/ 8.4%，1%/ 4.2%和0.2%/ 0.8%。

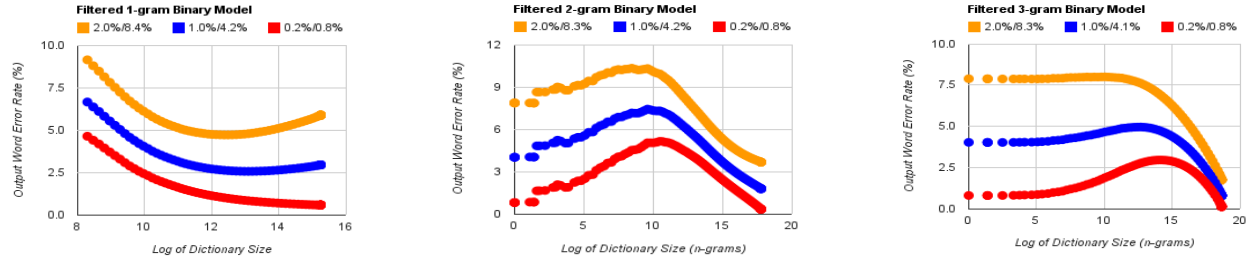

图2.输出字错误率（a）1克， （b）2克， （c）3克过滤后的语料库上的二元模型与n克字典大小的对数的函数，

输入形状分类器/单词错误率分别为2%/ 8.4%，1%/ 4.2%和0.2%/ 0.8%。

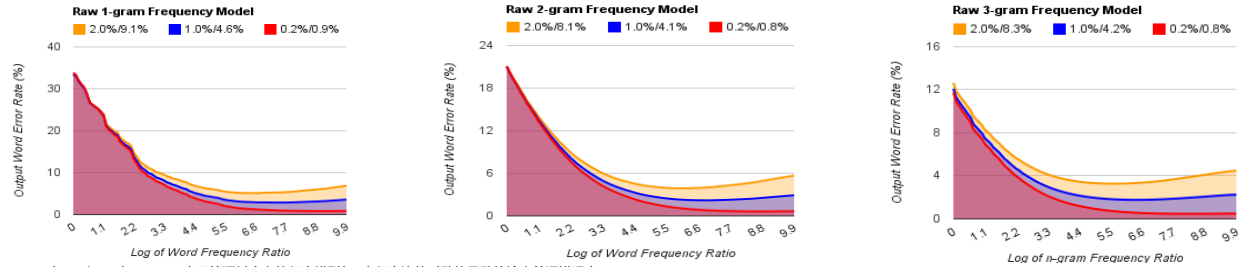

图3. (a) 1克， (b) 2克， (c) 3克原始语料库上的频率模型与n克频率比的对数的函数的输出单词错误率，

输入形状分类/单词错误率分别为2%/ 9.1%，1%/ 4.6%和0.2%/ 0.9%。

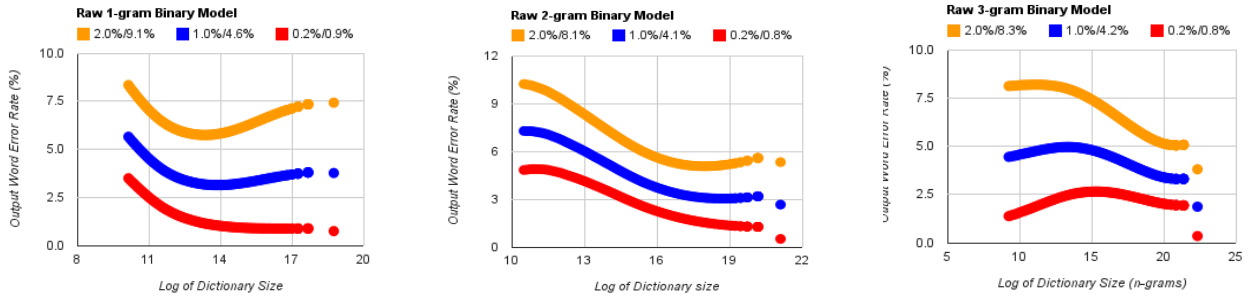

图4. (a) 1克， (b) 2克， (c) 3克的原始单词的输出单词错误率是n语法字典大小的对数的函数，其中

输入形状分类器/单词错误率分别为2%/ 9.1%，1%/ 4.6%和0.2%/ 0.9%。

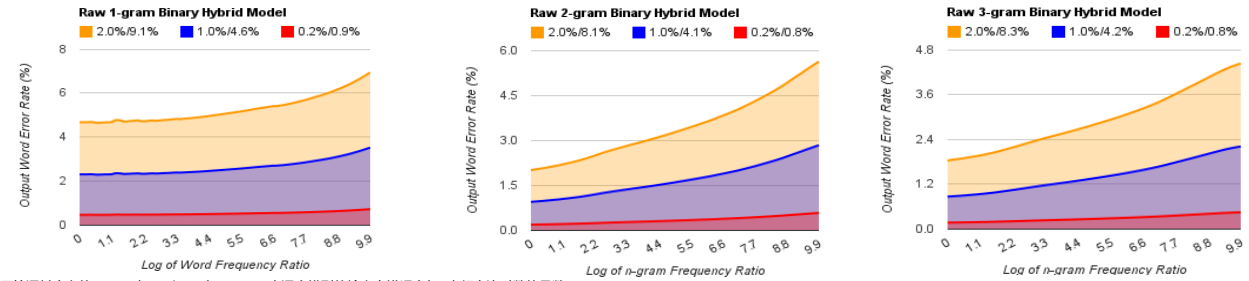

图5.原始语料库上的（a）1克， （b）2克， （c）3克混合模型的输出字错误率与n克频率比对数的函数，

输入形状分类器/单词错误率分别为2%/ 9.1%，1%/ 4.6%和0.2%/ 0.9%