第4章 语言模型

北京市海淀区中关村东路95号
邮编：100190

电话：+86-10-8254 4588
邮件：jjzhang@nlpr.ia.ac.cn

4.1 基本概念

4.1 基本概念

大规模语料库的出现为自然语言统计处理方法的实现提供了可能，统计方法的成功使用推动了语料库语言学的发展。

基于大规模语料库和统计方法，可以

- 发现语言使用的普遍规律
- 进行机器学习、自动获取语言知识
- 对未知语言现象进行推测

4.1 基本概念

如何计算一段文字(句子)的概率？

阳春三月春意盎然，少先队员脸上荡漾着喜悦的笑容，鲜艳的红领巾在他们的胸前迎风飘扬。

- 以一段文字(句子)为单位统计相对频率？
- 根据句子构成单位的概率计算联合概率?

$$p(w_1)\times p(w_2)\times \ldots \times p(w_n)$$

4.1 基本概念

语句 $s = w_1 w_2 \ldots w_m$ 的先验概率：

$$p(s) = p(w_1) \times p(w_2|w_1) \times p(w_3|w_1 w_2) \times \ldots \times p(w_m|w_1 \ldots w_{m-1})$$

$$= \prod_{i=1}^{m} p(w_i \mid w_1 \cdots w_{i-1}) \qquad \ldots (5\text{-}1)$$

当 i=1 时，$p(w_1|w_0) = p(w_1)$。

$\boxed{\text{语言模型}}$

4.1 基本概念

说明:

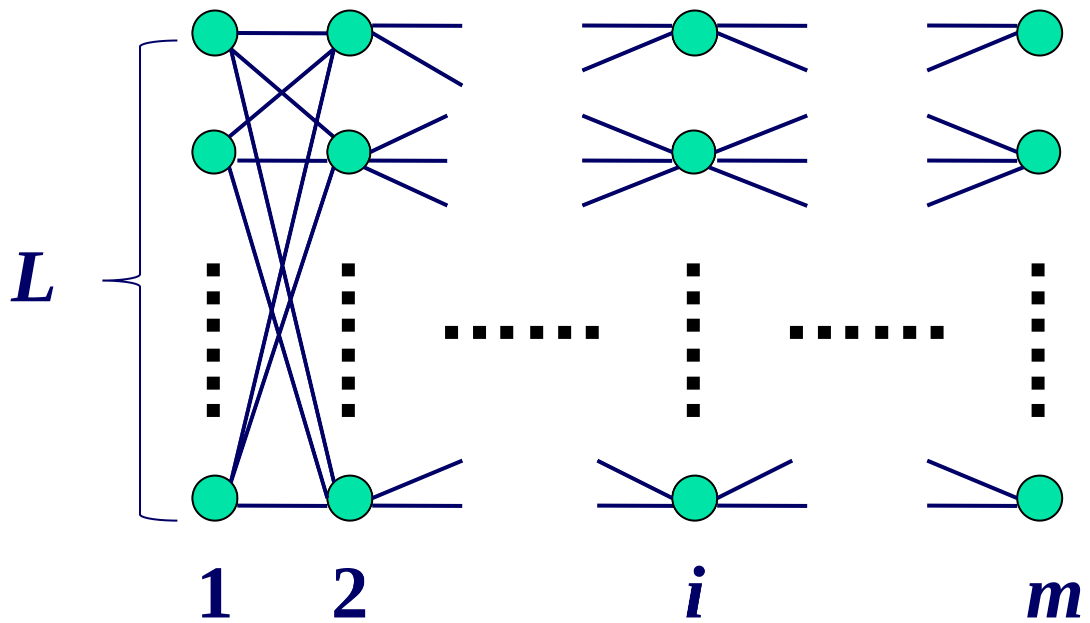
(1) w_i 可以是字、词、短语或词类等等，称为统计基元。通常以“词”代之。

(2) w_i 的概率由 $w_1, \ldots, w_{i-1}$ 决定，由特定的一组$w_1, \ldots, w_{i-1}$ 构成的一个序列，称为 w_i 的历史（history）。

4.1 基本概念

问题：随着历史基元数量的增加，不同的“历史”(路径)按指数级增长。对于第 i （$i>1$）个统计基元，历史基元的个数为 i-1，如果共有 L 个不同的基元，如词汇表，理论上每一个单词都有可能出现在1到 i-1的每一个位置上，那么，i 基元就有 L^{i-1} 种不同的历史情况。我们必须考虑在所有的 L^{i-1} 种不同历史情况下产生第 i 个基元的概率。那么，模型中有L^m 个自由参数 $p(w_m|w_1\ldots w_{m-1})$。

4.1 基本概念



如果 L=5000, m = 3, 自由参数的数目为 1250 亿！

4.1 基本概念

◆问题解决方法

设法减少历史基元的个数，将 $w_1\ w_2\ \ldots\ w_{i\text{-}1}$ 映射到等价类 S($w_1\ w_2\ \ldots\ w_{i\text{-}1}$)，使等价类的数目远远小于原来不同历史基元的数目。则有：

$$p(w_i \mid w_1, \ldots, w_{i-1}) = p(w_i \mid S(w_1, \ldots, w_{i-1}))$$

… (5-2)

4.1 基本概念

◆如何划分等价类

将两个历史映射到同一个等价类，当且仅当这两个历史中的最近 ***n*-1** 个基元相同，即：

$$H_1:\ w_1\ w_2 \ldots\ \ldots\ \underbrace{w_{i\text{-}n+1}\ w_{i\text{-}n+2} \ldots w_{i\text{-}1}}_{n\text{-}1}\ w_i \ldots\ldots$$

$$H_2:\ v_1\ v_2 \ldots\ \ldots\ \overbrace{v_{k\text{-}n+1}\ v_{k\text{-}n+2} \ldots v_{k\text{-}1}}\ v_k \ldots\ldots$$

$$\boxed{\begin{array}{l} S(w_1, w_2, \ldots, w_i) = S(v_1, v_2 \ldots, v_k) \\ \quad iff\ \ H_1:(w_{i-n+1}, \ldots, w_i) = H_2:(v_{k-n+1}, \ldots, v_k) \end{array}} \qquad \ldots (5\text{-}3)$$

4.1 基本概念

这种情况下的语言模型称为 *n* 元文法(*n*-gram)模型。

通常地，

❖ 当 n=1 时，即出现在第 i 位上的基元 w_i 独立于历史。一元文法也被写为 uni-gram 或 monogram；

❖ 当 n=2 时, **2-gram** (bi-gram) 被称为1阶马尔可夫链；

❖ 当 n=3 时, **3-gram**(tri-gram)被称为2阶马尔可夫链，依次类推。

4.1 基本概念

为了保证条件概率在 i=1 时有意义，同时为了保证句子内所有字符串的概率和为 1，即 $\sum_s p(s)=1$，可以在句子首尾两端增加两个标志：<BOS> $w_1\ w_2 \ldots w_m$ <EOS>。不失一般性，对于n>2 的 n-gram，$p(s)$ 可以分解为：

$$p(s)=\prod_{i=1}^{m+1} p(w_i \mid w_{i-n+1}^{i-1}) \qquad \ldots (5\text{-}4)$$

其中，w_i^j 表示词序列 $w_i \ldots w_j$，w_{i-n+1} 从 w_0 开始，w_0 为 <BOS>，w_{m+1} 为 <EOS>。

4.1 基本概念

◆举例：

给定句子：John read a book

增加标记：<BOS> John read a book <EOS>

Unigram：<BOS>, John, read, a, book, <EOS>

Bigram：(<BOS>John), (John read), (read a), (a book), (book <EOS>)

Trigram：(<BOS>John read), (John read a), (read a book), (a book <EOS>)

4.1 基本概念

<BOS> John read a book <EOS>

基于2元文法的概率为：

$$p(\text{John read a book}) = p(\text{John}|\text{<BOS>}) \times p(\text{read}|\text{John}) \times p(\text{a}|\text{read}) \times p(\text{book}|\text{a}) \times p(\text{<EOS>}|\text{book})$$

4.1 基本概念

◆应用-1：音字转换问题

给定拼音串：ta shi yan jiu sheng wu de

可能的汉字串：踏实研究生物的

他实验救生物的

他使烟酒生物的

他是研究生物的

… …

4.1 基本概念

$$\hat{CString} = \arg\max_{CString} p(CString \mid Pinyin)$$

$$= \arg\max_{CString} \frac{p(Pinyin \mid CString) \times p(CString)}{p(Pingyin)}$$

$$= \arg\max_{CString} p(Pinyin \mid CString) \times p(CString)$$

$$= \arg\max_{CString} p(CString)$$

4.1 基本概念

CString = {踏实研究生物的， 他实验救生物的， 他是研究生物的， 他使烟酒生雾的, … …}

如果使用 2-gram：

$p(CString_1)$ =*p*(踏实|<BOS>)×*p*(研究|踏实)×

p(生物|研究)×*p*(的|生物)×*p*(<EOS>|的)

$p(CString_2)$ =*p*(他|<BOS>)×*p*(实验|他)×*p*(救|实验)×

p(生物|救)×*p*(的|生物)×*p*(<EOS>|的)

……

4.1 基本概念

如果汉字的总数为：N

- 一元语法：1）样本空间为 N
  2）只选择使用频率最高的汉字
- 2元语法：1）样本空间为 N^2
  2）效果比一元语法明显提高
- 估计对汉字而言四元语法效果会好一些
- 智能狂拼、微软拼音输入法基于 *n*-gram.

4.1 基本概念

◆应用-2：汉语分词问题

给定汉字串：他是研究生物的。

可能的汉字串：

1）他|是|研究生|物|的

2）他|是|研究|生物|的

4.1 基本概念

$$\hat{Seg} = \arg\max_{Seg} p(Seg \mid Text)$$

$$= \arg\max_{Seg} \frac{p(Text \mid Seg) \times p(Seg)}{p(Text)}$$

$$= \arg\max_{Seg} p(Text \mid Seg) \times p(Seg)$$

$$= \arg\max_{Seg} p(Seg)$$

4.1 基本概念

如果采用2元文法：

$$p(\text{Seg1}) = p(\text{他}|\text{<BOS>}) \times p(\text{是}|\text{他}) \times p(\text{研究生}|\text{是}) \times p(\text{物}|\text{研究生}) \times p(\text{的}|\text{物}) \times p(\text{的}|\text{<EOS>})$$

$$p(\text{Seg2}) = p(\text{他}|\text{<BOS>}) \times p(\text{是}|\text{他}) \times p(\text{研究}|\text{是}) \times p(\text{生物}|\text{研究}) \times p(\text{的}|\text{生物}) \times p(\text{的}|\text{<EOS>})$$

问题：如何获得 *n* 元语法模型？

4.2 参数估计

4.2 参数估计

◆两个重要概念：

➢ 训练语料(*training data*)：用于建立模型，确定模型参数的已知语料。

➢ 最大似然估计(*maximum likelihood Evaluation*, **MLE**)：用相对频率计算概率的方法。

4.2 参数估计

对于 ***n*-gram**，参数 $p(w_i \mid w_{i-n+1}^{i-1})$ 可由最大似然估计求得:

$$p(w_i \mid w_{i-n+1}^{i-1}) = f(w_i \mid w_{i-n+1}^{i-1}) = \frac{c(w_{i-n+1}^{i})}{\sum_{w_i} c(w_{i-n+1}^{i})} \qquad \ldots(5\text{-}5)$$

其中，$\sum_{w_i} c(w_{i-n+1}^{i})$ 是历史串 w_{i-n+1}^{i-1} 在给定语料中出现的次数，即 $c(w_{i-n+1}^{i-1})$，不管 $\boldsymbol{w_i}$ 是什么。

$f(w_i \mid w_{i-n+1}^{i-1})$ 是在给定 w_{i-n+1}^{i-1} 的条件下 w_i 出现的相对频度，分子为 w_{i-n+1}^{i-1} 与 $\boldsymbol{w_i}$ 同现的次数。

4.2 参数估计

例如，给定训练语料：

"John read Moby Dick",

"Mary read a different book",

"She read a book by Cher"

根据 2 元文法求句子的概率？

4.2 参数估计

$$p(John \mid <BOS>) = \frac{c(<BOS>\ John)}{\sum_w c(<BOS>\ w)} = \frac{1}{3} \quad p(a \mid read) = \frac{c(read \quad a)}{\sum_w c(read \quad w)} = \frac{2}{3}$$

$$p(read \mid John) = \frac{c(John \quad read)}{\sum_w c(John \quad w)} = \frac{1}{1} \qquad p(book \mid a) = \frac{c(a \quad book)}{\sum_w c(a \quad w)} = \frac{1}{2}$$

$$p(<EOS> \mid book) = \frac{c(book \quad <EOS>)}{\sum_w c(book \quad w)} = \frac{1}{2}$$

$$p(John\ read\ a\ book) = \frac{1}{3} \times 1 \times \frac{2}{3} \times \frac{1}{2} \times \frac{1}{2} \approx 0.06$$

<BOS>John read Moby Dick<EOS>

<BOS>Mary read a different book<EOS>

<BOS>She read a book by Cher<EOS>

4.2 参数估计

$p(\textbf{Cher read a book}) = ?$

$$=p(Cher|<BOS>) \times p(read|Cher) \times p(a|read) \times p(book|a) \times p(<EOS>|book)$$

$$p(Cher \mid <BOS>) = \frac{c(<BOS>\ Cher)}{\sum_w c(<BOS>\ w)} = \frac{0}{3}$$

$$p(read \mid Cher) = \frac{c(Cher \quad read)}{\sum_w c(Cher \quad w)} = \frac{0}{1}$$

于是，$p(Cher\ read\ a\ book) = 0$

<BOS>John read Moby Dick<EOS>

<BOS>Mary read a different book<EOS>

<BOS>She read a book by Cher<EOS>

4.2 参数估计

问题：

数据匮乏(稀疏) (*Sparse Data*) 引起零概率问题，如何解决？

数据平滑(data smoothing)

4.3 数据平滑

4.3 数据平滑

◆**数据平滑的基本思想：**

调整最大似然估计的概率值，使零概率增值，使非零概率下调，“**劫富济贫**”，消除零概率，改进模型的整体正确率。

◆**基本目标：**测试样本的语言模型**困惑度越小越好**。

◆**基本约束：** $\sum_{w_i} p(w_i \mid w_1, w_2, \cdots, w_{i-1}) = 1$

4.3 数据平滑

- **回顾—困惑度的定义：**

对于一个平滑的 ***n*-gram**，其概率为 $p(w_i \mid w_{i-n+1}^{i-1})$，可以计算句子的概率：

$$p(s) = \prod_{i=1}^{m+1} p(w_i \mid w_{i-n+1}^{i-1})$$

假定测试语料 ***T*** 由 ***l_T*** 个句子构成 **(t_1, …, t_{lT})**，则整个测试集的概率为：

$$p(T) = \prod_{i=1}^{l_T} p(t_i)$$

4.3 数据平滑

模型 $p(w_i \mid w_{i-n+1}^{i-1})$ 对于测试语料的交叉熵：

$$H_p(T) = -\frac{1}{W_T}\log_2 p(T)$$

其中，W_T 是测试文本 T 的词数。

模型 p 的困惑度 $PP_p(T)$ 定义为： $PP_p(T) = 2^{H_p(T)}$

***n*-gram 对于英语文本的困惑度范围一般为 50～1000，对应于交叉熵范围为 6～10 bits/word。**

4.3 数据平滑

◆数据平滑方法

(1) 加1法(Additive smoothing)

基本思想：每一种情况出现的次数加1。

例如，对于 ***uni*-gram**，设 w_1, w_2, w_3 三个词，概率分别为：**1/3, 0, 2/3**，加**1**后情况？

2/6，1/6，3/6

4.3 数据平滑

对于2-gram 有：

$$p(w_i \mid w_{i-1}) = \frac{1 + c(w_{i-1}w_i)}{\sum_{w_i}[1 + c(w_{i-1}w_i)]}$$

$$= \frac{1 + c(w_{i-1}w_i)}{|V| + \sum_{w_i} c(w_{i-1}w_i)}$$

其中，*V* 为被考虑语料的词汇量（全部可能的基元数）。

4.3 数据平滑

在前面 3 个句子的例子中,

$$p(Cher\ read\ a\ book) = p(Cher|\text{<BOS>}) \times p(read|Cher) \times p(a|read) \times p(book|a) \times p(\text{<EOS>}|book)$$

<BOS>John read Moby Dick<EOS>

<BOS>Mary read a different book<EOS>

<BOS>She read a book by Cher<EOS>

原来:

$\boldsymbol{p(Cher|\text{<BOS>}) = 0/3}$

$\boldsymbol{p(read|Cher) = 0/1}$

$p(a|read) = 2/3$

$p(book|a) = 1/2$

$p(\text{<EOS>}|book) = 1/2$

4.3 数据平滑

词汇量：$|V| = 11$

<BOS>John read Moby Dick<EOS>

<BOS>Mary read a different book<EOS>

<BOS>She read a book by Cher<EOS>

平滑以后：

$p(Cher|\langle BOS\rangle) = (0+1)/(11+3) = 1/14$

$p(read|Cher) = (0+1)/(11+1) = 1/12$

$p(a|read) = (1+2)/(11+3) = 3/14$

$p(book|a) = (1+1)/(11+2) = 2/13$

$p(\langle EOS\rangle|book) = (1+1)/(11+2) = 2/13$

$$p(Cher\ read\ a\ book) = \frac{1}{14} \times \frac{1}{12} \times \frac{3}{14} \times \frac{2}{13} \times \frac{2}{13} \approx 0.00003$$

4.3 数据平滑

同理，对于句子*John read a book* 数据平滑后：

$p(John|\text{<BOS>}) = 2/14$, $p(read|John) = 2/12$,

$p(a|read) = 3/14$, $p(book|a) = 2/13$, $p(\text{<EOS>}|book) = 2/13$

于是，$p(John\ read\ a\ book) = p(John|\text{<BOS>}) \times p(read|John) \times p(a|read) \times p(book|a) \times p(\text{<EOS>}|book)$

$$= \frac{2}{14} \times \frac{2}{12} \times \frac{3}{14} \times \frac{2}{13} \times \frac{2}{13} \approx 0.0001$$

<BOS>John read Moby Dick<EOS>
<BOS>Mary read a different book<EOS>
<BOS>She read a book by Cher<EOS>

4.5 语言模型应用举例

4.5 语言模型应用举例

◆汉语分词问题：

句子：这篇文章写得太平淡了。

这/ 篇/ 文章/ 写/ 得/ 太/ 平淡/ 了/ 。

这/ 篇/ 文章/ 写/ 得/ 太平/ 淡/ 了/ 。

4.5 语言模型应用举例

采用基于语言模型的分词方法

➢ **方法描述**：

设对于待切分的句子$S = z_1 z_2 \ldots z_m$，$W = w_1 w_2 \ldots w_k$ $(1 \le k \le n)$ 是一种可能的切分。那么，

$$
\begin{aligned}
\hat{W} &= \arg\max_{W} p(W \mid S) \\
&= \arg\max_{W} p(W) \times p(S \mid W) \\
&\cong \arg\max_{W} p(W)
\end{aligned}
$$

最基本的做法是以词为独立的统计基元，但效果不佳。

4.5 语言模型应用举例

具体实现时，可把汉语词汇分成如下几类：

(1) 分词词典中规定的词；

(2) 由词法规则派生出来的词或短语，如：干干净净、非党员、副部长、全面性、检查员、看不出、克服了、走出来 、洗个澡 …

(3) 与数字相关的实体，如：日期、时间、货币、百分数、温度、长度、面积、重量、电话号码、邮件地址等；

(4) 专用名词，如 ：人名、地名、组织机构名。

占未登录词的95%！

4.5 语言模型应用举例

进一步做如下约定，把一个可能的词序列 ***W*** 转换成词类序列 $C = c_1 c_2 \Lambda\ c_N$ ，即：

- 专有名词：人名**PN**、地名**LN**、机构名**ON**分别作为一类；
- 实体名词中的日期 **dat**、时间**tim**、百分数**per**、货币**mon** 等作为一类；
- 对词法派生词**MW**和词表词**LW**，每个词单独作为一类。

4.5 语言模型应用举例

<u>例如</u>：

3月14日下午3点比尔盖茨在北京发表讲话，决定从今年起微软亚洲研究院将大规模招收研发人员，其中，80％将从中国科学院大学所培养的应届博士或硕士毕业生中选拔，年薪不低于3万美元。

Just a joke ☺

4.5 语言模型应用举例

日期dat　时间tim　人名PN　地名LN

3月14日下午3点比尔盖茨在北京发表讲话，决定从今年起微软亚洲研究院将大规模招收研发人员，其中，80%将从中国科学院大学所培养的应届博士或硕士毕业生中选拔，年薪不低于3万美元。

机构名ON

百分数per　货币数mon

4.5 语言模型应用举例

词序列变为类序列：

dat/ tim/ PN/ 在/ LN/ 发表/ 讲话/ ，/ 决定/ 从/ tim/ 起/ ON/ 将/ 大/ 规模/ 招收/ 研发/ 人员/ ，/ 其中/ ，/ per/ 将/ 从/ 中国/ 科学院/ 大学/ 所/ 培养/ 的/ 应届/ 博士/ 或/ 硕士/ 毕业生/ 中/ 选拔/ ，/ 年薪/ 不/ 低于/ mon/ 。

4.5 语言模型应用举例

那么， $$\hat{C} = \arg\max_{C} p(C \mid S)$$

$$= \arg\max_{C} p(C) \times p(S \mid C) \qquad \ldots (5\text{-}22)$$

语言模型 生成模型

$p(C)$ 可采用三元语法：

$$p(C) = p(c_1) \times p(c_2 \mid c_1) \prod_{i=3}^{N} p(c_i \mid c_{i-2}c_{i-1}) \qquad \ldots (5\text{-}23)$$

$$p(c_i \mid c_{i-2}c_{i-1}) = \frac{count(c_{i-2}c_{i-1}c_i)}{count(c_{i-2}c_{i-1})} \qquad \ldots (5\text{-}24)$$

4.5 语言模型应用举例

生成模型在满足独立性假设的条件下，可近似为：

$$p(S \mid C) \approx \prod_{i=1}^{N} p(s_i \mid c_i) \qquad \dots (5\text{-}25)$$

该公式的含意是，任意一个词类 c_i 生成汉字串 s_i 的概率只与自身有关，而与其上下文无关。例如，如果“教授”是词表里的词，那么

$p(s_i\text{=教授}|c_i\text{=LW})=1$，否则，$p(s_i \mid c_i) = 0$。

4.5 语言模型应用举例

词　类	生成模型 $p(S\|C)$	语言知识来源
词表词 (LW)	若*S*是词表词，$p(S\|\mathrm{LW})=1$，否则为0；	分词词表
词法派生词 (MW)	若*S*是派生词，$p(S\|\mathrm{MW})=1$，否则为0；	派生词词表
人名 (PN)	基于字的2元模型	姓氏表，中文人名模板
地名 (LN)	基于字的2元模型	地名表、地名关键词表、地名简称表
机构名 (ON)	基于词类的2元模型	机关名关键词表，机构名简称表
其他实体名 (FT)	若*S*可用实体名词规则集*G*识别，$p(S\|G)=1$，否则为0。	实体名词规则集

本章小结

- n 元语法的基本概念
  - uni-gram，bi-gram，tri-gram

- 数据平滑方法

- 语言模型的应用

习题

4-1. 请阅读有关文献，了解除了本讲义介绍的数据平滑方法以外的其它平滑方法；请对Good-Turing 平滑方法进行简要的评价，阐述你个人的观点。

4-2. 利用汉语切分和标注语料（注意版权的合法性），尝试用 bi-gram 实现一个简单的汉语自动分词程序。

Thanks

谢谢！

4.3 数据平滑

(2) 减值法/折扣法(Discounting)

<u>**基本思想**</u>：修改训练样本中事件的实际计数，使样本中(实际出现的)不同事件的概率之和小于1，剩余的概率量分配给未见概率。

4.3 数据平滑

①Good-Turing 估计

I. J. Good 于1953 年引用 Turing 的方法来估计概率分布。

假设 N 是原来训练样本数据的大小， n_r 是在样本中正好出现 r 次的事件的数目(此处事件为 n-gram)，即出现 1 次的n-gram有 n_1个， 出现 2 次的 n-gram 有 n_2个, ……， 出现 r 次的有 n_r 个。

4.3 数据平滑

那么， $N=\sum_{r=1}^{\infty} n_r r$ … (5-6)

由于，$N=\sum_{r=0}^{\infty} n_r r^* = \sum_{r=0}^{\infty} (r+1) n_{r+1}$ 所以，$r^* = (r+1)\dfrac{n_{r+1}}{n_r}$

那么，**Good-Turing** 估计在样本中出现 ***r*** 次的事件的概率为：

$$p_r = \frac{r^*}{N} \quad \dots (5\text{-}7)$$

4.3 数据平滑

实际应用中，一般直接用 $\boldsymbol{n_{r+1}}$ 代替 $\boldsymbol{E(n_{r+1})}$，$\boldsymbol{n_r}$ 代替 $\boldsymbol{E(n_r)}$。这样，原训练样本中所有事件的概率之和为：

$$\sum_{r>0} n_r \times p_r = 1 - \frac{n_1}{N} < 1 \qquad \ldots (5\text{-}8)$$

因此，有 $\frac{n_1}{N}$ 的剩余的概率量就可以均分给所有的未见事件 (**r = 0**)。

Good-Turing 估计适用于大词汇集产生的符合多项式分布的大量的观察数据。

有关证明和推导，请参阅：A. Nadas. on Turing's Formula for Word Probabilities. In *IEEE Trans. on ASSP*-33, Dec. 1985. Pages 1414-1416.

4.3 数据平滑

举例说明：假设有如下英语文本，估计 2-gram概率：

<BOS>John read Moby Dick<EOS>
<BOS>Mary read a different book<EOS>
<BOS>She read a book by Cher<EOS>
……

从文本中统计出不同 2-gram 出现的次数：

<BOS> John	**15**
<BOS> Mary	**10**
……	
read Moby	**5**
……	

4.3 数据平滑

假设要估计以 read 开始的 2-gram 概率，列出以read 开始的所有 2-gram，并转化为频率信息：

r	n_r	r^*
1	2053	0.446
2	458	1.25
3	191	2.24
4	107	3.22
5	69	4.17
6	48	5.25
7	36	保持原来的计数

$$r^* = (r+1)\frac{n_{r+1}}{n_r}$$

$n_{r+1} = 0$

4.3 数据平滑

得到 r^* 后，就可以应用公式(5-7) 计算概率：

$$p_r = \frac{r^*}{N} \qquad \ldots (5\text{-}7)$$

其中，N 为以 read 开始的bigram的总数(样本空间)，即 read出现的次数。那么，以 read开始，没有出现过的 2-gram的概率总和为：

$$p_0 = \frac{n_1}{N}$$

以 read作为开始，没有出现过的 2-gram的个数等于：

$$n_0 = |V_T| - \sum_{r>0} n_r$$

其中，$|V_T|$ 为语料的词汇量。

4.3 数据平滑

那么，没有出现过的那些 以 read 为开始的2-gram的概率平均为：$\frac{p_0}{n_0}$。

注意：$\sum_{r=0}^{7} p_r \neq 1$

因此，需要归一化处理：

$$\hat{p}_r = \frac{p_r}{\sum_r p_r}$$

r	n_r	r^*
1	2053	0.446
2	458	1.25
3	191	2.24
4	107	3.22
5	69	4.17
6	48	5.25
7	36	—

4.3 数据平滑

②Back-off (后备/后退) 方法

S. M. Katz 于 1987 年提出，所以又称 Katz 后退法。

基本思想：当某一事件在样本中出现的频率大于阈值*K* (通常取 *K* 为0 或1)时，运用最大似然估计的减值法来估计其概率，否则，使用低阶的，即 (*n*-1)gram 的概率替代 *n*-gram 概率，而这种替代需受归一化因子α的作用。

4.3 数据平滑

Back-off 方法的另一种理解:

对于每个计数 $r > 0$ 的 n 元文法的出现次数减值，把因减值而节省下来的剩余概率根据低阶的 (n-1)gram 分配给未见事件。

4.3 数据平滑

以2元文法模型为例, 说明**Katz**平滑方法:

对于一个出现次数为 $r = c(w_{i-1}^{i})$的 2元语法 w_{i-1}^{i}，使用如下公式计算修正的概率:

$$p_{katz}(w_i|w_{i-1}) = \begin{cases} d_r \dfrac{C(w_{i-1}w_i)}{C(w_{i-1})} & if C(w_{i-1}w_i) = r > 0 \\ \alpha(w_{i-1})p_{ML}(w_i) & if C(w_{i-1}w_i) = 0 \end{cases}$$

其中，$p_{\mathrm{ML}}(w_i)$表示 w_i 的最大似然(maximum likelihood)估计概率。这个公式的意思是，所有具有非零计数 r 的 2元语法都根据折扣率d_r (0<d_r<1)被减值了，折扣率 d_r 近似地等于 r^*/r，减值由Good-Turing估计方法预测。

4.3 数据平滑

那么，如何确定$\alpha(w_{i-1})$呢？

$$\sum_{w_i} p_{katz}(w_i|w_{i-1}) = 1$$

$$\Downarrow$$

$$\sum_{w_{i:r=0}} p_{katz}(w_i|w_{i-1}) + \sum_{w_{i:r>0}} p_{katz}(w_i|w_{i-1}) = 1$$

$$\Downarrow$$

$$\sum_{w_{i:r=0}} \alpha(w_{i-1}) p_{ML}(w_i) + \sum_{w_{i:r>0}} p_{katz}(w_i|w_{i-1}) = 1$$

4.3 数据平滑

那么，如何确定$\alpha(w_{i\text{-}1})$呢？

$$\sum_{w_{i:r=0}} \alpha(w_{i-1}) p_{ML}(w_i) + \sum_{w_{i:r>0}} p_{katz}(w_i|w_{i-1}) = 1$$

$$\Downarrow$$

$$\alpha(w_{i-1}) = \frac{1 - \sum_{w_{i:r>0}} p_{katz}(w_i|w_{i-1})}{\sum_{w_{i:r=0}} p_{ML}(w_i)}$$

4.3 数据平滑

③绝对减值法 (Absolute discounting)

Hermann Ney 和 U. Essen 1993年提出。

<u>基本思想</u>：从每个计数 *r* 中减去同样的量，剩余的概率量由未见事件均分。

设 *R* 为所有可能事件的数目(当事件为 *n*-gram 时，如果统计基元为词，且词汇集的大小为 *L*，则 $R=L^n$)。

4.3 数据平滑

那么，样本出现了 r 次的事件的概率可以由如下公式估计：

$$p_r = \begin{cases} \dfrac{r-b}{N} & \text{当 } r > 0 \\[2ex] \dfrac{b(R-n_0)}{Nn_0} & \text{当 } r = 0 \end{cases} \qquad \ldots (5\text{-}10)$$

其中，n_0 为样本中未出现的事件的数目。b 为减去的常量，$b \le 1$。

4.3 数据平滑

$b(R - n_0)/N$ 是由于减值而产生的剩余概率量。

b 为自由参数，可以通过留存数据(heldout data)法求得 b 的上限为：

$$b \le \frac{n_1}{n_1 + 2n_2} < 1 \qquad \ldots \text{(5-11)}$$

H. Ney and U. Essen. Estimating Small Probabilities by Leaving-one-Out. In *Proc. Eurospeech* '1993. Pages 2239-2242.

④线性减值法 (Linear discounting)

基本思想：从每个计数 r 中减去与该计数成正比的量(减值函数为线性的)，剩余概率量 α 被n_0个未见事件均分。

$$p_r = \begin{cases} \dfrac{(1-\alpha)r}{N} & \text{当 } r > 0 \\ \dfrac{\alpha}{n_0} & \text{当 } r = 0 \end{cases} \qquad \ldots (5\text{-}12)$$

自由参数 α 的优化值为：$\dfrac{n_1}{N}$

绝对减值法产生的*n*-gram 通常优于线性减值法。

4.3 数据平滑

◆ 四种减值法的比较

- **Good-Turing 法：** 对非0事件按公式削减出现的次数，节留出来的概率均分给0概率事件。
- **Katz 后退法：** 对非0事件按Good-Turing法计算减值，节留出来的概率按低阶分布分给0概率事件。
- **绝对减值法：** 对非0事件无条件削减某一固定的出现次数值，节留出来的概率均分给0概率事件。
- **线性减值法：** 对非0事件根据出现次数按比例削减次数值，节留出来的概率均分给0概率事件。

4.3 数据平滑

(3) **删除插值法** (Deleted interpolation)

基本思想：用低阶语法估计高阶语法，即当 3-gram 的值不能从训练数据中准确估计时，用 2-gram 来替代，同样，当 2-gram 的值不能从训练语料中准确估计时，可以用 1-gram 的值来代替。插值公式：

$$p(w_3 \mid w_1 w_2) = \lambda_3 p'(w_3 \mid w_1 w_2) + \lambda_2 p'(w_3 \mid w_2) + \lambda_1 p'(w_3) \quad \ldots (5\text{-}13)$$

其中，$\lambda_1 + \lambda_2 + \lambda_3 = 1$

4.3 数据平滑

- λ_1，λ_2，λ_3 的确定：

　　将训练语料分为两部分，即从原始语料中删除一部分作为留存数据(heldout data)。

第一部分用于估计 $p'(w_3 | w_1 w_2)$ ，$p'(w_3 | w_2)$ 和 $p'(w_3)$ 。

第二部分用于计算 $\lambda_1, \lambda_2, \lambda_3$：使语言模型对留存数据的困惑度最小。

4.3 数据平滑

- 各种平滑方法的详细介绍和比较请参阅：

  Chen, Stanley F. and Joshua Goodman. 1998. An Empirical Study of Smoothing Techniques for Language Model. Available from the website:

  http://www-2.cs.cmu.edu/~sfc/html/publications.html

- SRI 语言模型工具：

  http://www.speech.sri.com/projects/srilm/

- CMU-Cambridge 语言模型工具：

  http://mi.eng.cam.ac.uk/~prc14/toolkit.html