01 语言建模与统计基础

从文本预测到计数模型与独立评价

学习目标

- 解释文本、token、词表与整数编码的关系，说明分词怎样改变预测单位。
- 联系下一 token 的条件概率与序列概率，区分概率分解和 N-gram 假设，并跟踪给定语料的计数与评分。
- 联系似然、负对数似然、交叉熵与困惑度，说明有效位置和平均范围。
- 理解未见符号、未见组合、稀疏计数和有限历史的区别，以及子词和平滑的作用。
- 解释训练、验证、测试的职责，以及困惑度比较的成立条件。

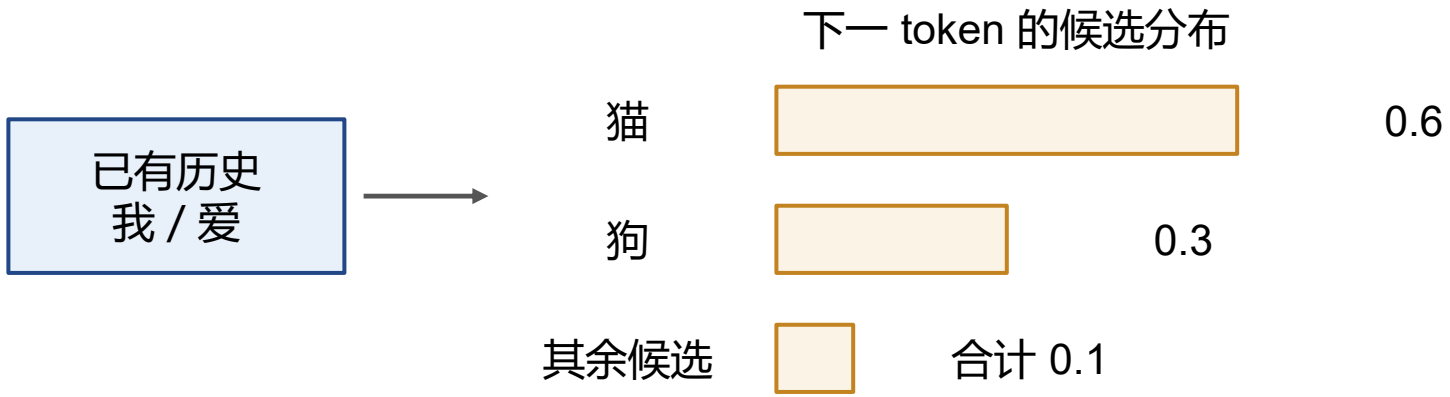
本讲以解释关系和跟踪完整示范为主，不要求独立实现算法或从零训练模型。

1 文本表示与序列概率

从文本表示建立下一token概率与序列评分。

语言模型对下一 token 分配概率

给定“我 / 爱”，**语言模型**为每个候选 **token（词元）**分配概率，输出的是一个分布。



预测读取候选分布；序列评分组合各位置目标的概率；生成按选择规则取出一个 token，再将它加入历史。

图中数值仅为示意，概率合计为 1；输出选择规则留待后续讲次。

分词决定 token 单位

文本	单位选择	序列
机器学习	字符	机 / 器 / 学 / 习
机器学习	词	机器 / 学习
lower	子词	low / e / r，具体划分取决于已有规则

分词（tokenization）把文本转换为有序的离散单位。按字符或词处理“机器学习”，得到的预测位置与单位不同。

子词可为词内片段，也可小至单个字符；具体划分取决于已有规则。改变分词会改变模型预测的事件。

词表与 ID 确定离散类别

词表项	我	你	爱	猫	狗	鱼	EOS
ID	0	1	2	3	4	5	6

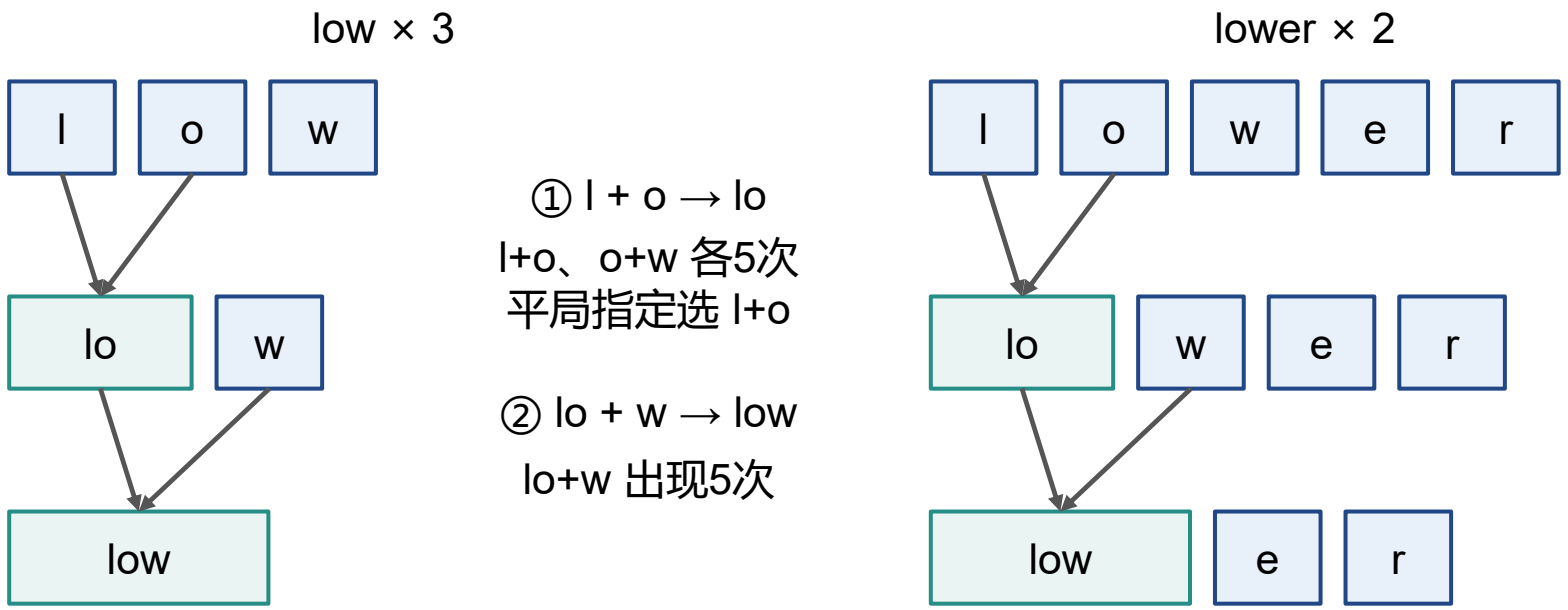
词表 $\mathcal{V}$ 是允许使用的 token 类型集合，$V = |\mathcal{V}|$。表中是本讲计数模型的输出词表及示意ID；一次出现是token实例，ID是词表项的类别编码。

本例词表预先给定；词表也可以从指定训练语料中学习。

“我 / 爱 / 猫”编码为 $(0, 2, 3)$，ID大小和差值不表达语义距离。记有序序列为 $x_{1:T} = (x_1, \ldots, x_T)$，长度为 T，位置 t 的token为 x_t；它不是稠密特征向量。

BPE 从相邻片段学习合并规则

整词粒度可能需要较大词表，字符粒度通常产生较长序列。**字节对编码（BPE）**合并频繁相邻片段，在两者之间作取舍。



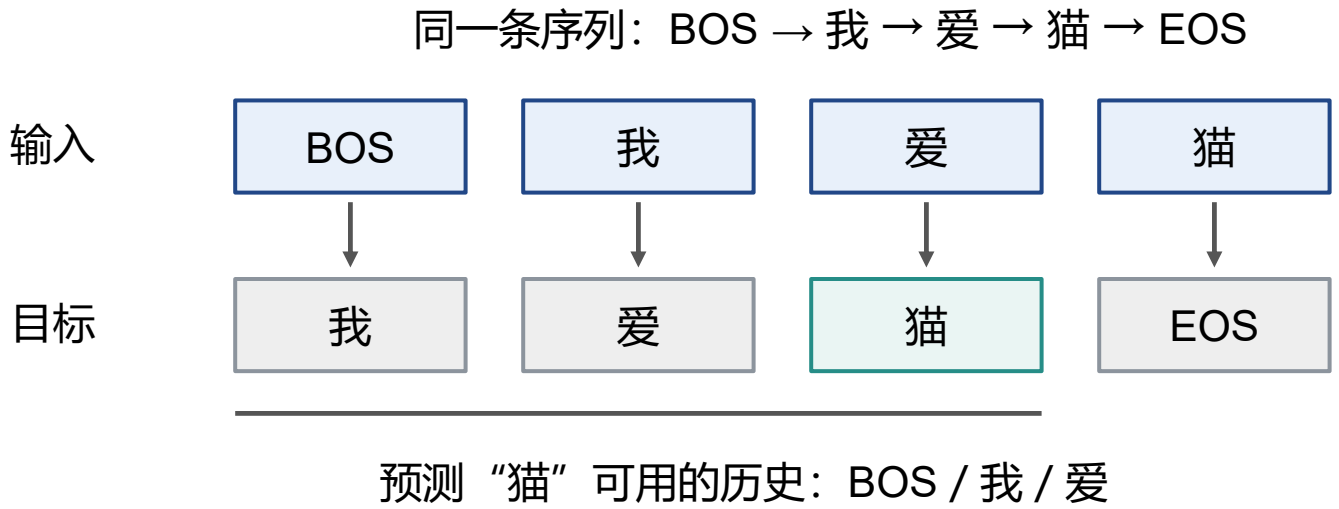
未见词lowe依次应用这两条规则，得到low / e，不重新学习。合并不跨词边界；本例基础字符仅l、o、w、e、r，其他输入仍须覆盖或回退规则。

起始、终止与补齐承担不同职责

标记	职责	本讲处理
BOS	表示序列开始时的输入条件	用来预测第一个内容token，不作为目标
EOS	表示序列结束	作为需要预测的终止目标
PAD	对齐批量中的存储长度	不属于有效目标，不参与损失平均
UNK	在配置该机制时接收未知输入	当前构造不配置；词表外输入须另定规则

特殊标记的拼写、存在与否及职责由模型设定；本讲采用表中输入、目标和补齐规则。

目标移位确定四个有效预测位置



输入与目标错开一个 token，每列学习预测下一个 token。预测“猫”时可用完整前缀 BOS、我、爱，而非仅看上一行的“爱”。

共4个有效目标，EOS是第四个；BOS仅作条件，追加PAD不增加有效目标数。

给定历史限定下一 token 的概率范围

本页固定历史“BOS / 我 / 爱”，问下一目标为“猫”的概率；此处仍可使用完整前缀。

条件概率描述给定事件发生后，另一个事件在这个范围内发生的概率。设 B 是“历史满足指定条件”，A 是“下一 token 为指定候选”，当 $p_*(B) > 0$ 时：

$$p_*(A \mid B) = \frac{p_*(A \cap B)}{p_*(B)}.$$

分子同时满足历史和目标条件，分母限定给定历史的范围。交换 A 与 B 会改变问题；分母为零时这个比值没有定义。

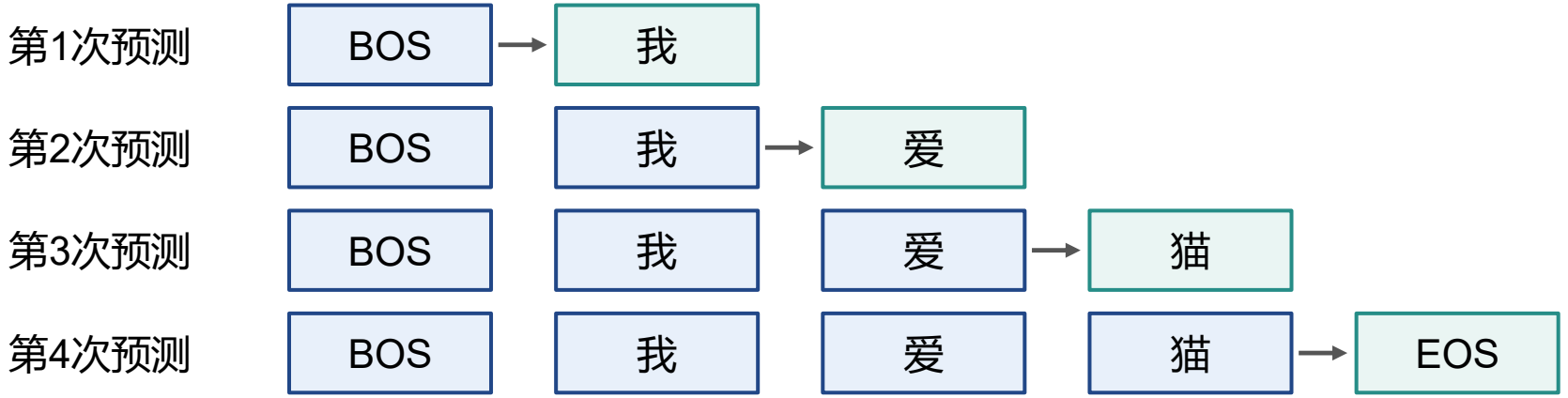
模型用 $p_\theta(v \mid h)$ 表示历史 h 下候选 v 的概率，满足

$$p_\theta(v \mid h) \geq 0, \sum_{v \in \mathcal{V}} p_\theta(v \mid h) = 1.$$

求和对象是同一历史下的全部输出候选。p_* 表示未知目标分布，p_θ 是给定参数元组 θ 的模型，二者身份不同。

链式法则保留完整历史

链式法则逐位置使用完整历史：当前目标向右移动时，前面的token进入条件。



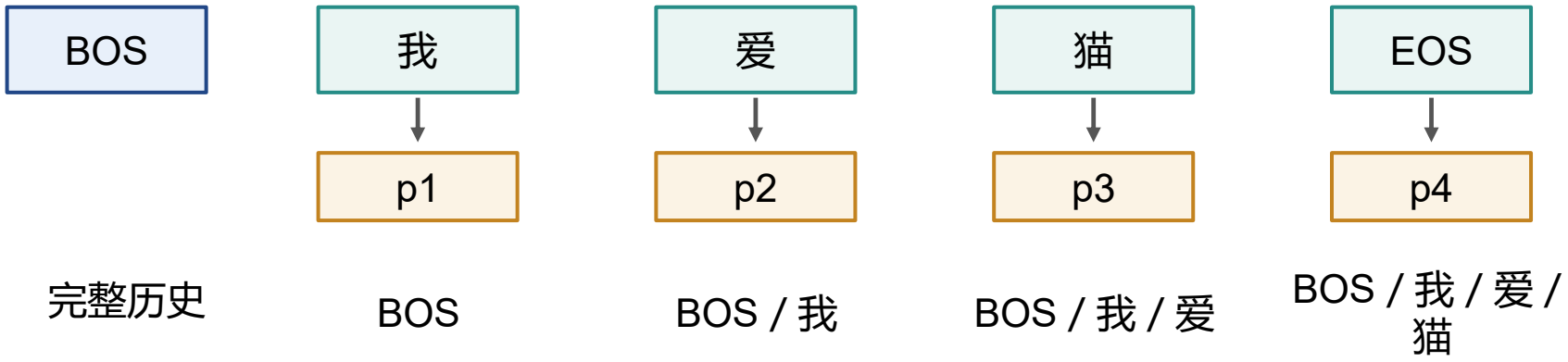
由条件概率的乘法关系，三个位置可逐步拆开：

$$p_*(x_1, x_2, x_3) = p_*(x_1, x_2)p_*(x_3 \mid x_1, x_2)$$
$$= p_*(x_1)p_*(x_2 \mid x_1)p_*(x_3 \mid x_1, x_2).$$

一般地，$p_*(x_{1:T}) = \prod_{t=1}^{T} p_*(x_t \mid x_{<t})$。所需条件概率须有定义；共同给定BOS时，各项保留该条件。**连乘不要求token独立，也没有截短历史。**

条件分布的分解仍需要数值来源

给定 $b = \mathrm{BOS}$，四个目标 $x_{1:4}$ 依次为“我、爱、猫、EOS”；下方四个因子与目标逐列对应。



$$p_\theta(x_{1:4} \mid b) = p_\theta(x_1 \mid b)p_\theta(x_2 \mid b, x_1) \cdot p_\theta(x_3 \mid b, x_1, x_2)p_\theta(x_4 \mid b, x_1, x_2, x_3).$$

加入EOS才包含本次终止。**分解规定需要哪些概率，模型和数据决定其数值。**完整长历史很少重复，下一段用有限历史模型简化估计。

2 计数模型与统计限制

用有限历史与语料计数估计概率，认识模型限制。

N-gram 用有限历史近似完整历史

长历史的重复事件可能很少。**N-gram 模型**用最近至多 $n-1$ 个 token 作为预测条件，以减少需要分别估计的历史类型。

- 一元模型：$n=1$，空历史，使用 $p_\theta(x_t)$。
- 二元模型：$n=2$，一个历史token，使用 $p_\theta(x_t \mid x_{t-1})$。
- 三元模型：$n=3$，两个历史token，使用 $p_\theta(x_t \mid x_{t-2}, x_{t-1})$。



序列开始处按已声明的边界条件处理。以二元模型为例：

$$p_\theta(x_t \mid x_{<t}) = p_\theta(x_t \mid x_{t-1}).$$

同一预测“猫”的问题中，二元模型只保留“爱”。模型内等号成立；对真实语言而言，较早信息可能因此丢失。

训练语料提供十六个有效转移

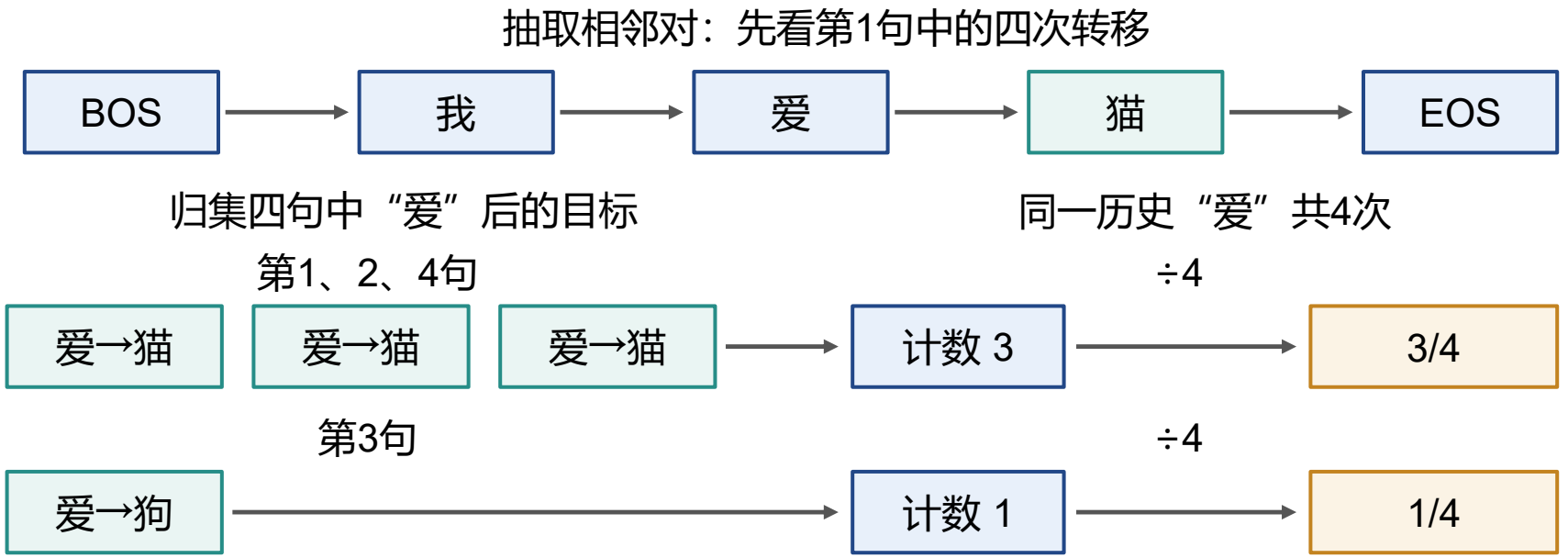
序列编号	输入起点	内容token	终止目标
1	BOS	我 / 爱 / 猫	EOS
2	BOS	我 / 爱 / 猫	EOS
3	BOS	我 / 爱 / 狗	EOS
4	BOS	你 / 爱 / 猫	EOS

逐句取相邻转移，重复行保留。首句BOS→我→爱→猫→EOS含4次；四句各4次，共16次。

BOS只作条件，EOS作为目标计数，序列之间不连接。

相同历史内的相对频率估计条件概率

从四句训练语料抽取相邻转移，再固定历史“爱”，归集对应目标。



计数估计在同一历史内归一化：

$$C(h)=\sum_{v\in\mathcal{V}} C(h,v), \hat{p}(v\mid h)=C(h,v)/C(h), C(h)>0.$$

$C(h,v)$数历史后接目标的次数，$C(h)$汇总该历史的有效转移。EOS之后不跨句计数；$\hat{p}$是语料估计，不是已知真实概率。

条件分布由同一历史的计数归一化

历史 h	非零的 $C(h,v)$	$C(h)$	非零条件概率
BOS	我：3；你：1	4	我：$3/4$；你：$1/4$
我	爱：3	3	爱：1
你	爱：1	1	爱：1
爱	**猫：3；狗：1**	**4**	猫：$3/4$；狗：$1/4$
猫	EOS：3	3	EOS：1
狗	EOS：1	1	EOS：1

表列全部非零计数，未列组合为零。每一行在输出词表“我、你、爱、猫、狗、鱼、EOS”上求和为1；“爱”后4次转移中3次接“猫”、1次接“狗”。

“鱼”虽在词表中，但在训练转移中从未出现；将它取作历史 h 时，$C(h) = 0$。原比例公式没有定义该历史下的分布，不能在表中补写一整行零概率。

局部条件概率组成整段评分

固定上一页的未平滑二元模型，为整句未出现在训练语料中的“你 / 爱 / 狗 / EOS”评分。



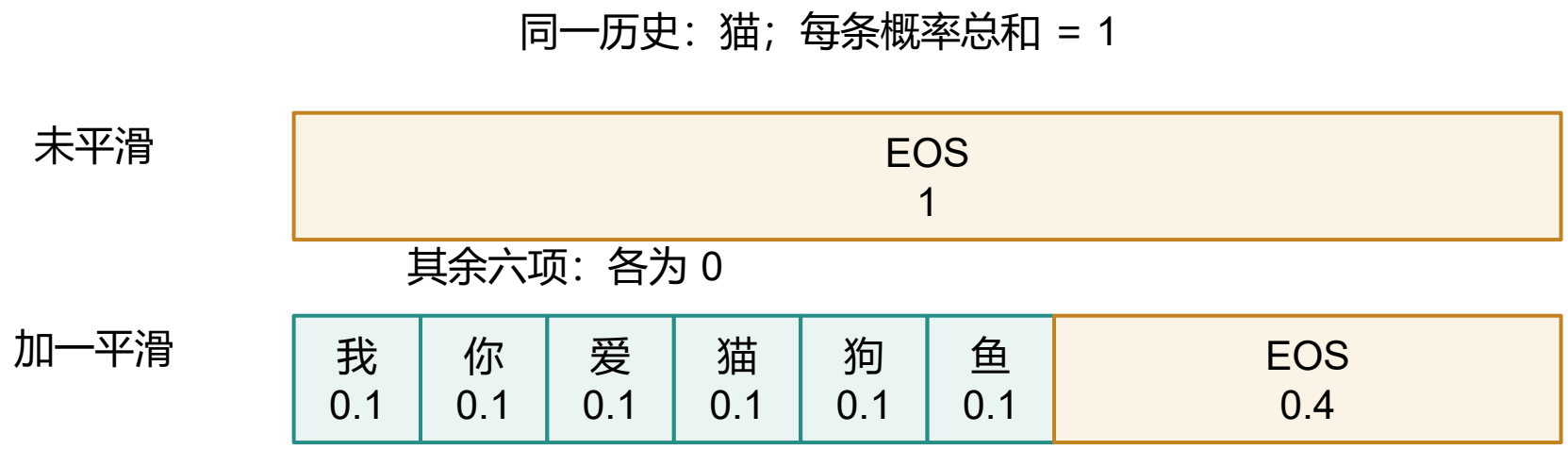
整句未见；四个局部转移均有计数

四个局部转移均有计数，因此序列概率为 $(1/4) \cdot 1 \cdot (1/4) \cdot 1 = 1/16 > 0$；模型不要求整句在训练中出现。

同表给“我 / 爱 / 猫 / EOS”的评分仍为 $(3/4) \cdot 1 \cdot (3/4) \cdot 1 = 9/16$。两段均含4目标，未重新估计参数；这是构造评分，不是句义理解或独立泛化成绩。

平滑改变概率分配，有限历史仍然受限

同一词表、同一历史“猫”，仅示范加一平滑：$V = 7$，历史“猫”的计数为3，三次后面都接EOS。



令 h 为“猫”，则 $p_{+1}(v \mid h) = [C(h, v) + 1]/(3 + 7)$。

平滑重分配概率；“鸟”仍在词表外，“鱼”作历史时原比例仍无定义，二元模型仍只看末词。零历史须由所选规则处理。

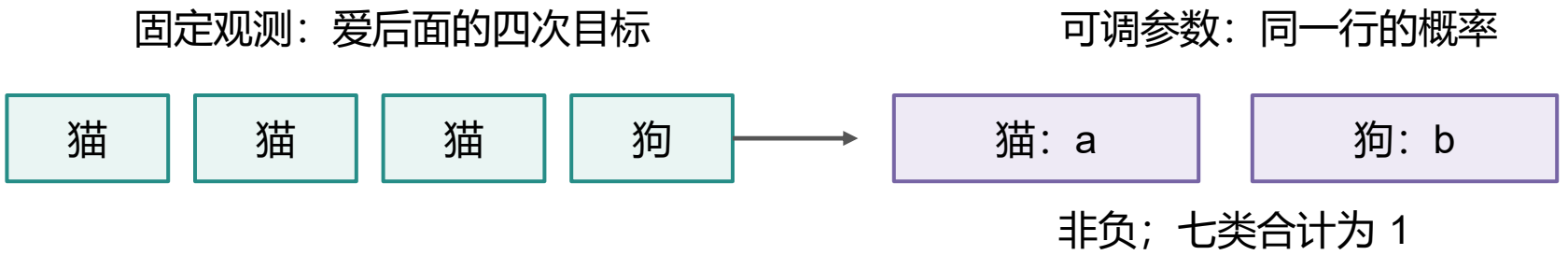
本页仅解释机制，不作为最佳平滑方案；P18、P25继续使用未平滑模型。

3 学习目标与独立评价

联系学习准则与评价指标，明确分数的解释条件。

固定观测比较参数，概率成为似然

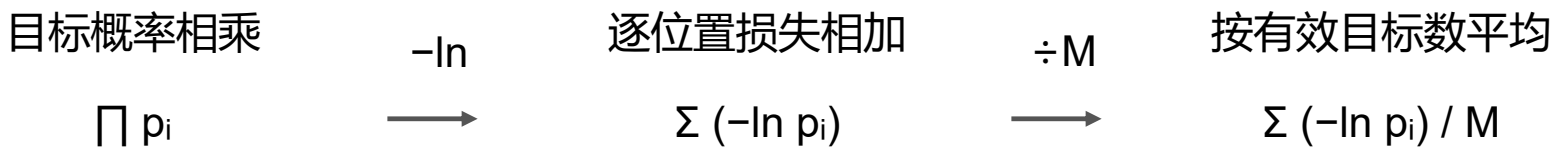
本讲二元模型的参数 θ 就是各历史对应的条件概率表。固定“爱”后3次接猫、1次接狗的观测，调整这一行的概率。



记猫、狗的条件概率分别为 a、b；这一行对观测的**似然**为 a^3b。最大似然在 $a = 3/4, b = 1/4$、其余候选为0时取得，正好得到前面的计数比例。

前面固定概率表给文本评分；现在固定训练文本选择概率表。似然是参数的函数，不是参数的概率分布。

负对数把概率乘积变为有效位置上的平均损失



负对数似然（NLL）把目标概率乘积改写为各位置相加的损失。设 $\mathcal{I}$ 为有效目标位置集合，$M = |\mathcal{I}| > 0$，x_i 是观测目标，h_i 是对应历史：

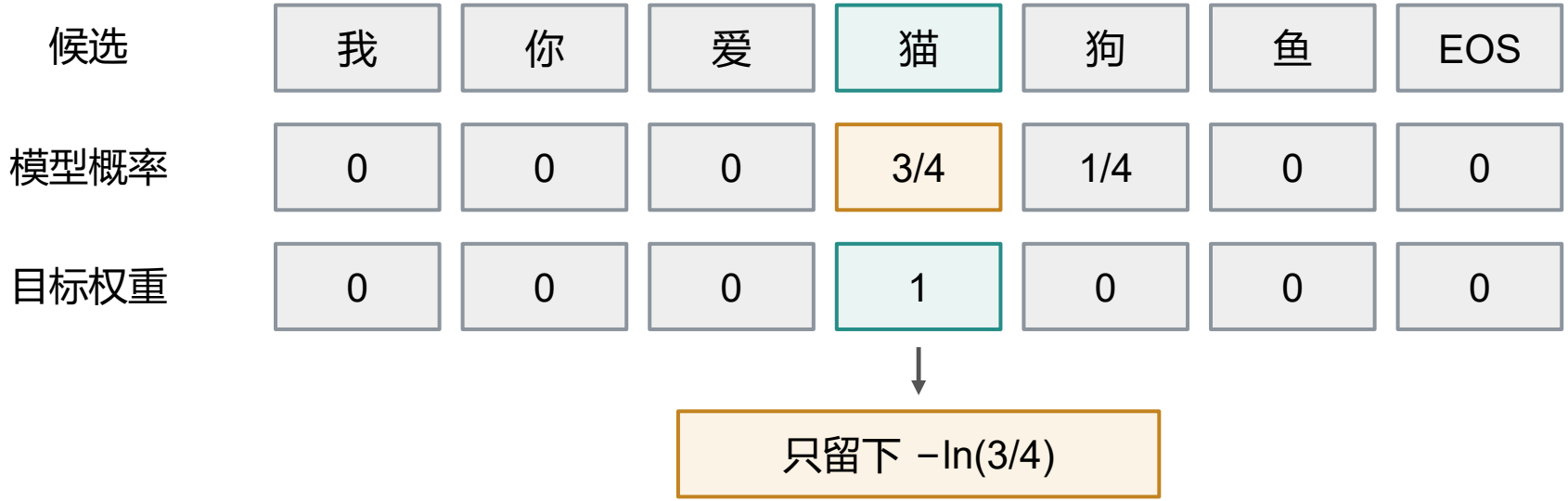
$$NLL_{\mathcal{I}}(\theta) = -\ln \prod_{i\in\mathcal{I}} p_\theta\,(x_i \mid h_i)$$
$$= -\sum_{i\in\mathcal{I}} \ln p_\theta(x_i \mid h_i),$$
$$L_{\mathrm{NLL}}(\theta) = NLL_{\mathcal{I}}(\theta)/M.$$

固定数据、历史和有效位置，最大化似然等价于最小化NLL：对数保持大小次序，负号反转方向；除以同一正数M不改变最优参数。

本讲用自然对数；我 / 爱 / 猫 / EOS共4个目标，BOS不计，PAD排除。目标概率为0时NLL为正无穷；分布无定义时须先补模型规则。

单标签交叉熵与同一位置的 NLL 相等

历史固定为“爱”，模型输出全部七个候选的分布；这次真实目标是“猫”。



交叉熵（CE）用目标分布 q_i 对负对数概率加权；单标签时，真实目标权重为1，其余为0：

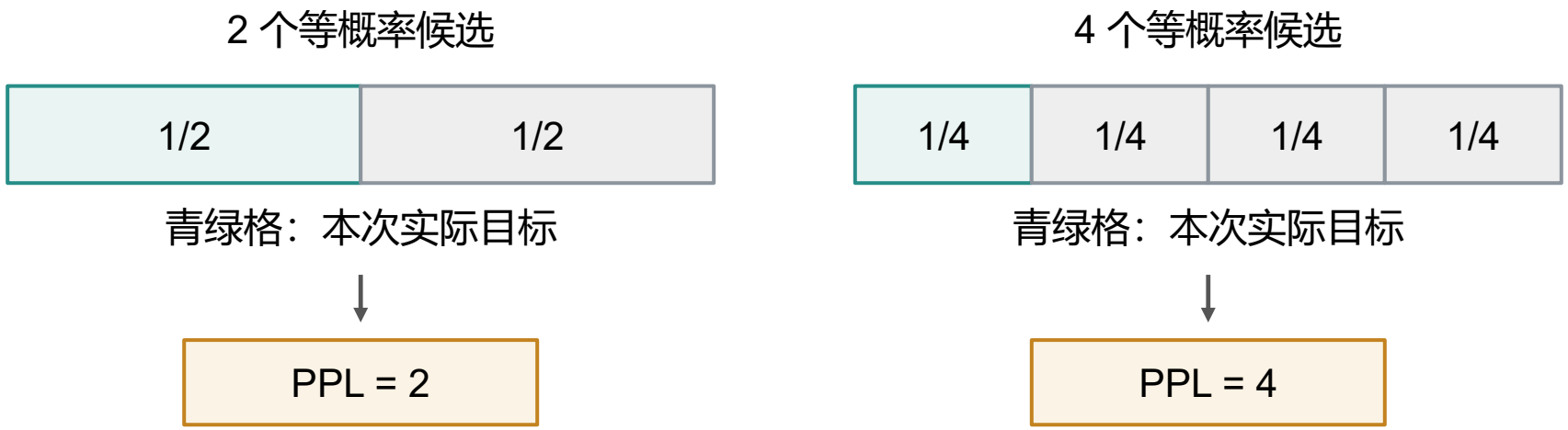
$H(q_i, p_\theta) = -\sum_{v \in \mathcal{V}} q_i\,(v) \ln p_\theta(v \mid h_i) = -\ln(3/4).$

零权重项约定为0。使用同一预测分布、有效位置和平均方式，单标签平均CE等于平均NLL。

困惑度是平均 NLL 的指数

困惑度（PPL）是平均NLL的指数，也是观测目标概率几何平均的倒数：

$PPL = \exp(L_{\mathrm{NLL}}) = \left(\prod_{i \in \mathcal{I}} p_\theta\left(x_i \mid h_i\right)\right)^{-1/M}.$



若每次在K个候选上均匀分配概率，实际目标总得到1/K，则PPL=K。一般PPL是由目标概率形成的等效量，不表示每步真的剩下K个候选。

在同一评价设定下，PPL越低，目标的几何平均预测概率越高。

同一评分确定 NLL、CE 与 PPL

评分序列	我 / 爱 / 猫 / EOS	你 / 爱 / 狗 / EOS
目标概率	$3/4,1,3/4,1$	$1/4,1,1/4,1$
总NLL	$-\ln(9/16) \approx 0.575$	$-\ln(1/16) \approx 2.773$
平均NLL = 平均CE	$-\ln(9/16)/4 \approx 0.144$	$-\ln(1/16)/4 \approx 0.693$
PPL	$(9/16)^{-1/4} \approx 1.155$	$(1/16)^{-1/4} = 2$

复用同一未平滑计数模型，两段都含EOS在内的4个目标。用精确序列概率计算负对数和指数，表中小数只作近似展示。

同一预测概率产生总NLL、平均CE和PPL；有效位置或历史改变时，表中这些数值不能原样沿用。

拟合、选择与独立评价使用不同数据职责

数据	用途	语言模型中的对应
训练集	拟合模型	统计N-gram计数、估计参数；学习数据驱动的分词规则和词表
验证集	选择模型与设置	比较历史长度、平滑设置或其他已声明模型选择
测试集	选择结束后的独立评价	在固定模型及表示规则下，评价未参与拟合和选择的数据

对测试文本应用已经固定的tokenizer，是使用既有表示规则；根据测试文本重新学习词表或合并频率，则让测试信息参与拟合。反复根据测试成绩选择模型，也会改变测试的独立评价职责。

数据职责决定分数能支持的结论。重复或同源文本跨集合可能影响独立评价；随机切token不自动保证独立。结果须结合文本来源与使用目标解释，不能直接推广到所有新领域。

PPL 比较必须保持预测条件可比

条件	不一致时改变了什么
评价文本与来源	观察目标及任务分布可能不同
tokenizer、词表与单位	一个目标token的含义、序列长度和分母可能不同
上下文的使用方式	模型阶数可以不同；须说明窗口、重置与分块策略
BOS/EOS、PAD及有效位置	被评分的事件与平均范围不同

固定评价文本、预测单位和目标统计规则，明确各模型怎样使用上下文；不同阶数的N-gram可以比较PPL。

上述PPL用于说明同一构造模型对两段文本的评分，不构成泛化实验。分词改变时，预测单位与分母也会改变，裸PPL不可直接比较。

PPL描述预测概率，事实与任务质量须另查。自回归PPL不能直接与BERT等掩码目标的分数比较。

本讲总结

- 文本通过分词形成离散token序列；词表确定候选身份，ID承担编码。BPE规则的学习与固定规则的应用具有不同职责，覆盖仍有条件。
- 链式法则把序列概率联系到完整历史下的条件概率；N-gram另加有限历史假设，计数为这些分布提供经验估计。
- 符号能表示不保证组合见过。零组合计数、零历史计数和历史截断需要分别解释；平滑改变概率分配，保留原模型的历史范围。
- 相同有效目标上，似然的负对数形成总NLL；单标签目标的平均CE等于平均NLL，PPL是平均NLL的指数。BOS作输入、EOS终止目标和PAD排除须保持一致。
- 训练用于拟合，验证用于选择，测试用于独立评价。PPL比较须保持文本、单位及目标统计规则可比，并说明上下文使用方式，结论限于相应预测设定。

本讲用N-gram计数模型说明概率语言建模的对象、学习与评价；后续神经语言模型将用不同方式表示历史和计算条件分布。