词嵌入与神经语言模型

机器学习 · 语言模型专题

学习目标

- 解释词索引、嵌入查表与神经预测怎样联系，说明同一预测损失怎样作用于表示和预测参数
- 辨认CBOW / Skip-gram的预测方向，按学习信号与表示组成比较Word2vec、GloVe和fastText，区分词向量任务与完整语言模型
- 在完整说明支持下跟踪RNN的输入、状态、下一目标与续写，区分静态输入和上下文表示，说明固定窗口与递归状态的限制

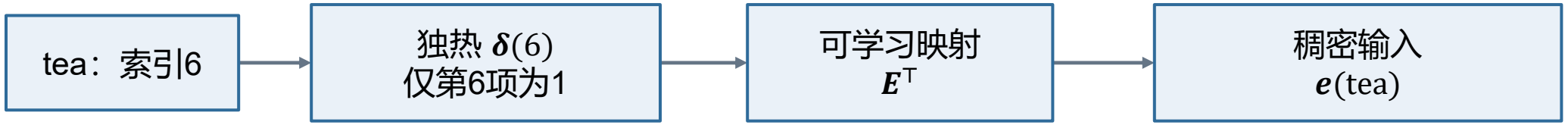
1 词嵌入与相似性

词索引怎样取得可学习向量，向量之间的相似性又怎样衡量？

从独热表示到词嵌入

本讲连接两个问题：词项怎样获得可学习表示？模型怎样利用这些表示和历史预测下一词？向量是否有用，要看它怎样帮助预测学习。

独热表示准确编码词项身份，但不同词的独热向量两两正交，直接余弦都为0，无法表达不同程度的词间关系。



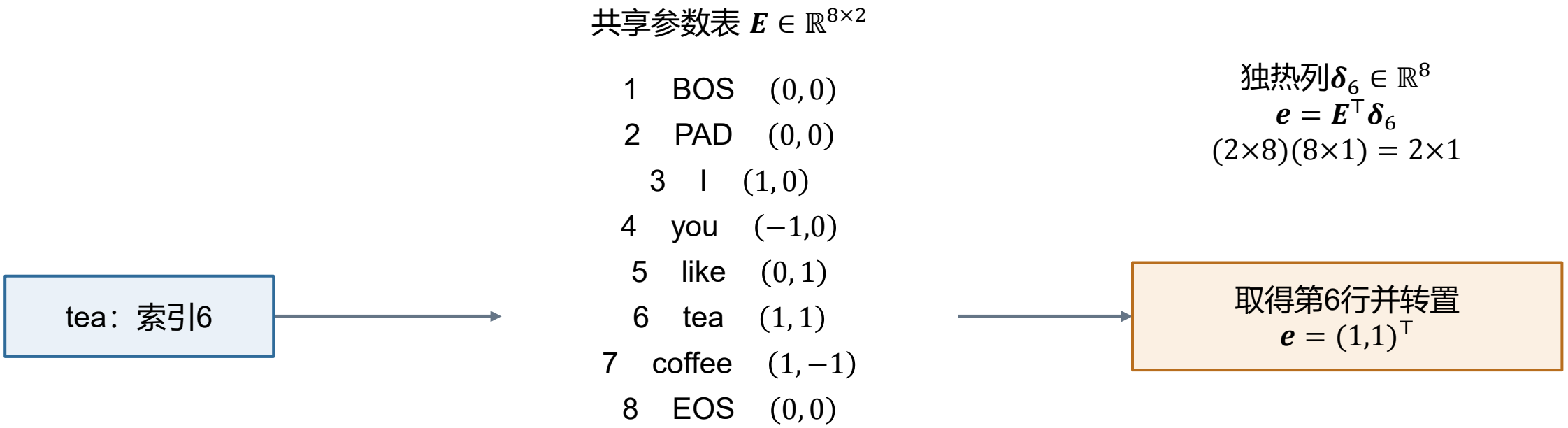
保留词项身份

允许向量关系由数据和目标学习

词嵌入以可训练向量表示词项。预测学习可以调整向量之间的关系；独热向量仍可作为线性选择的输入，索引数值的远近不决定嵌入之间的关系。

词索引与嵌入查表

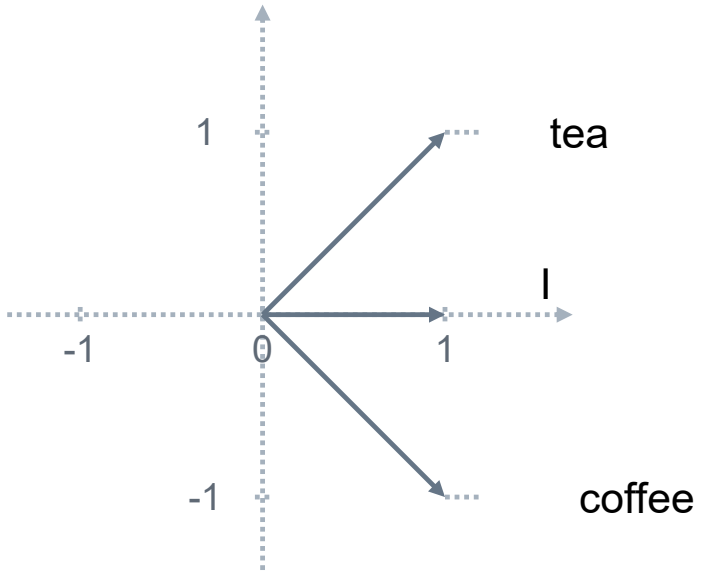
同一词项总是选择共享表中的同一行。两列是表示维度；全部数值与特殊词零向量均为人工约定。



索引查表与独热线性选择等价，实现时无须展开独热向量。实际模型的嵌入参数由训练确定，下一词预测可以提供学习信号。

余弦相似度：人工二维向量示例

余弦相似度比较同维非零向量的方向：$cos(\boldsymbol{a}, \boldsymbol{b}) = \boldsymbol{a}^{\mathsf{T}}\boldsymbol{b}/(\|\boldsymbol{a}\|_2\|\boldsymbol{b}\|_2)$。



图中I与tea的余弦为$1/\sqrt{2}$，说明向量方向可以比较。人工方向不证明真实词义；零向量不参与余弦比较。下一单元说明预测任务怎样训练这些向量。

2 前馈语言模型与共同学习

固定窗口怎样预测下一词，同一预测损失怎样共同学习词表示和模型参数?

统计n-gram与神经语言模型的泛化机制

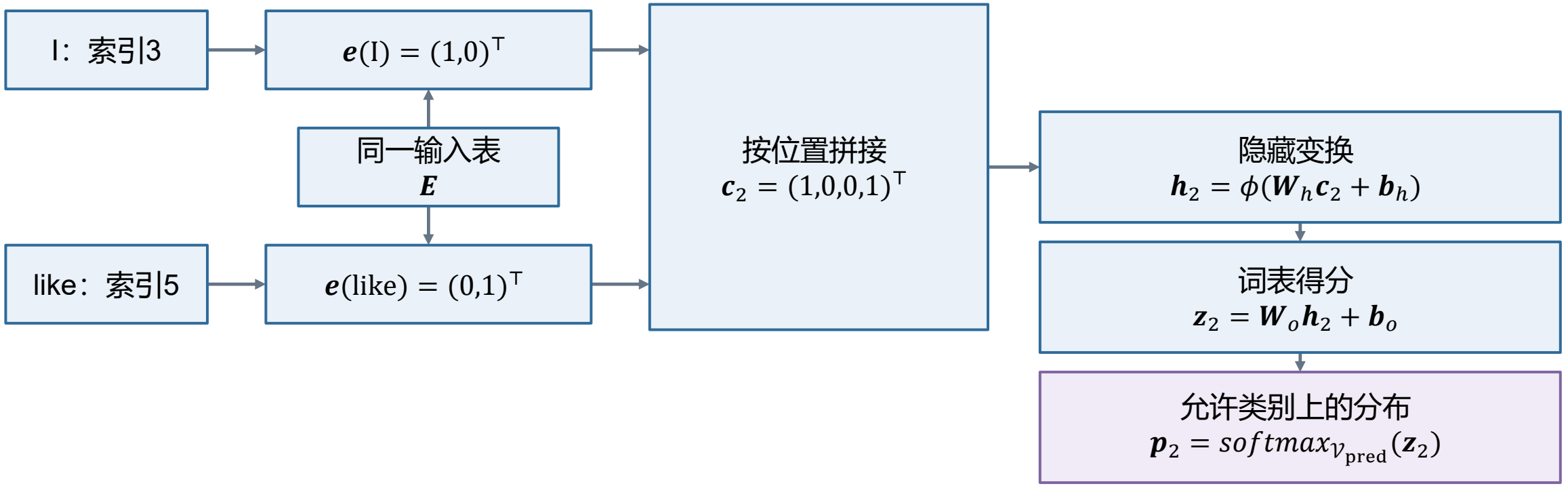
比较维度	统计n-gram模型	前馈神经语言模型
历史输入	离散词组合	按位置组合词嵌入
参数组织	上下文计数与概率估计	共享嵌入与预测网络
泛化机制	平滑、插值与回退	连续表示与共享映射
历史范围	有限窗口	仍是有限窗口

离散历史组合很多，未见组合也需要概率估计。两类模型都保留有限窗口，但共享信息的方式不同。

一个样本更新共享参数，就会影响其他历史的预测。相近表示提供统计联系，但不保证不同历史获得相近预测。

前馈语言模型的下一词预测

前馈神经语言模型按位置组合固定历史窗口。这里读完I like，$k = 2$，目标为tea。



交换为like I时，拼接变成$(0,1,1,0)^{\top}$，顺序不同。$\boldsymbol{c}_2 \in \mathbb{R}^4$、$\boldsymbol{h}_2 \in \mathbb{R}^m$、$\boldsymbol{z}_2, \boldsymbol{p}_2 \in \mathbb{R}^8$；权重形状为$m\times4$与$8\times m$。Softmax仅在允许候选上归一化，BOS / PAD置0，EOS保留。

Bengio神经语言模型的原始结构

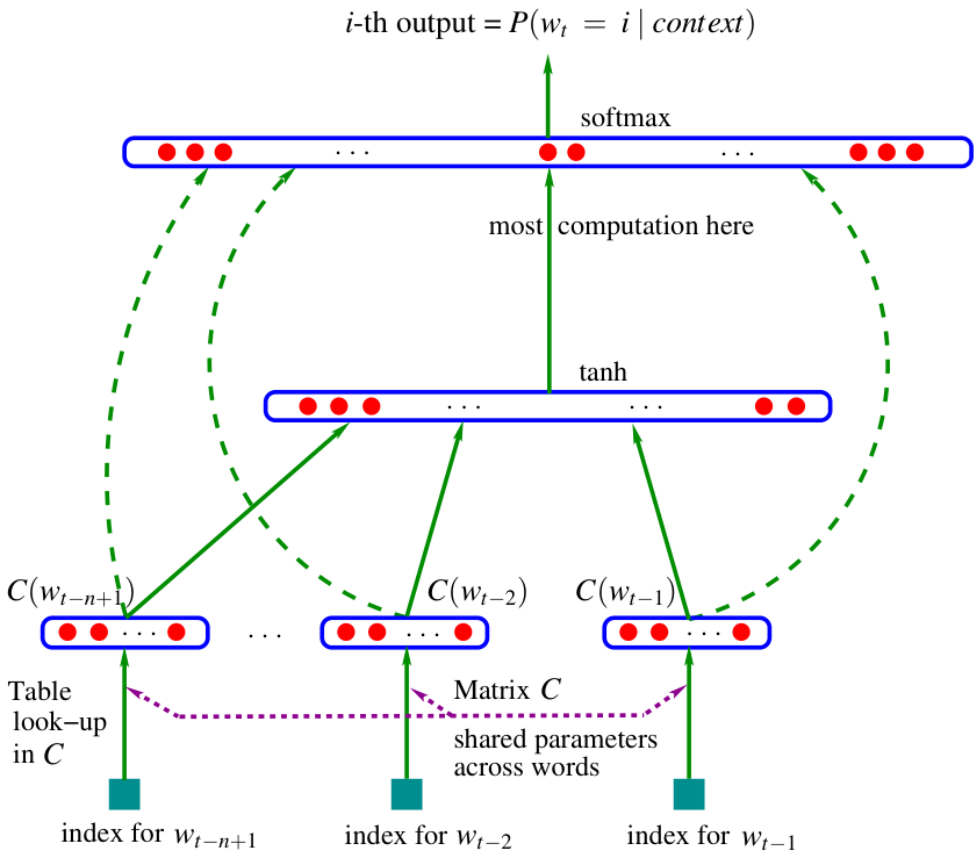


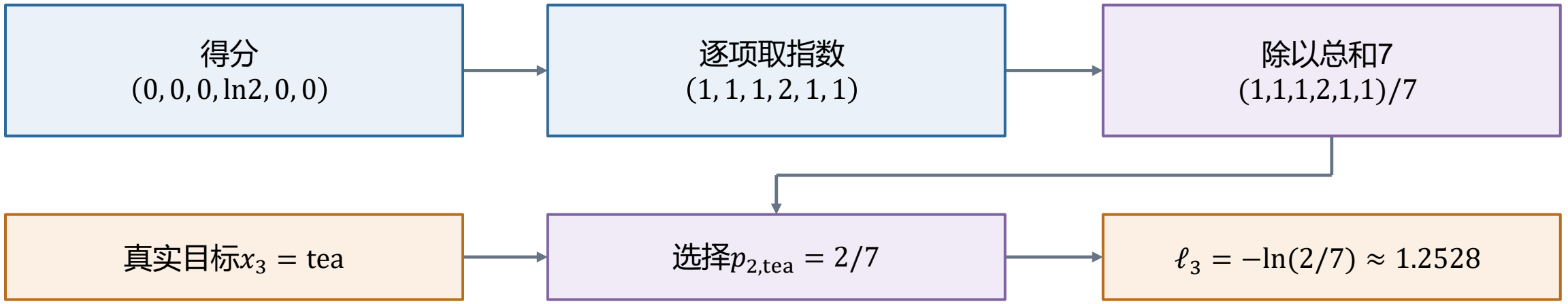
Figure 1: Neural architecture: $f(i, w_{t-1}, \cdots, w_{t-n+1}) = g(i, C(w_{t-1}), \cdots, C(w_{t-n+1}))$ where g is the neural network and $C(i)$ is the i-th word feature vector.

- 共享表C对应本讲E；向量按位置拼接，经tanh与词表输出。
- 绿色虚线为可选直连；本讲采用无直连形式。
- 词特征与预测网络由同一目标共同训练。

Bengio等（2003）Fig.1

下一词概率与负对数似然损失

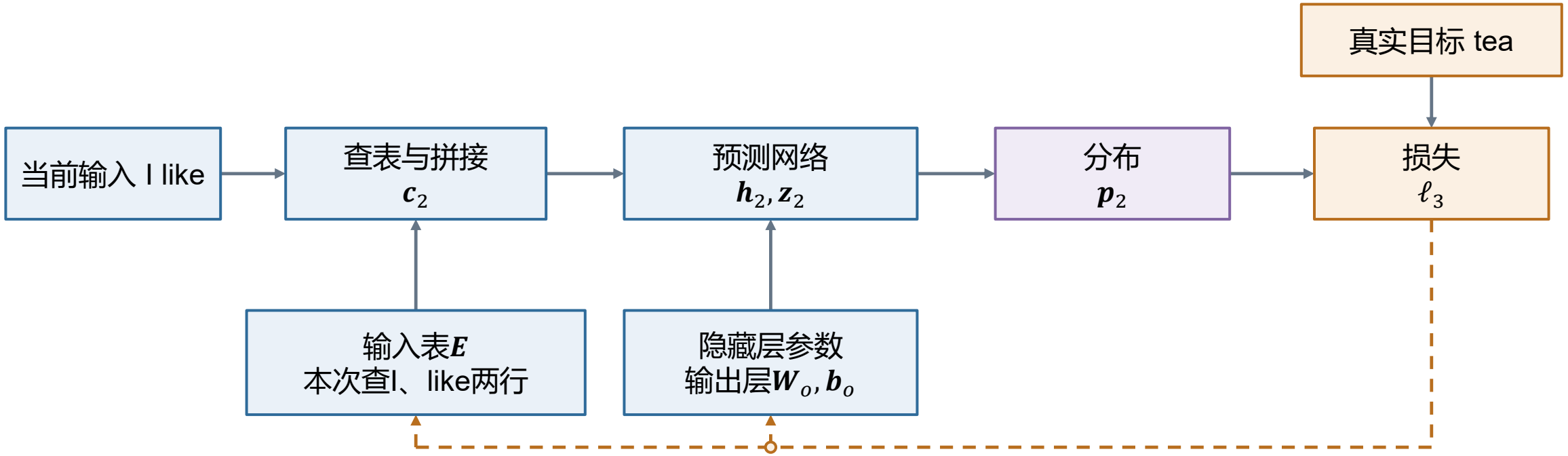
固定历史I like，候选依次为I、you、like、tea、coffee、EOS。以下得分均为人工指定。



$p_{2,i} = \exp(z_{2,i}) / \sum_{r \in \mathcal{V}_{\text{pred}}} \exp(z_{2,r})$。目标来自样本；$\ell_3 = -\ln p_\theta(x_3 \mid x_{0:2})$，是单目标损失。读入$t = 2$后预测$j = 3$；BOS / PAD排除，EOS保留。

词嵌入与预测参数的共同学习

实线表示前向计算依赖，虚线表示这个损失向可训练参数提供的梯度信号。



本例输入表与输出权重不绑定。该样本损失只通过查表路径影响E的I、like行，tea仅作目标，不直接更新E的tea行；输出分类参数也收到梯度。训练目标驱动所用表示与预测参数朝降低预测损失的方向优化，同一输入行跨样本复用。冻结或零梯度时可不变。

3 词向量学习方法

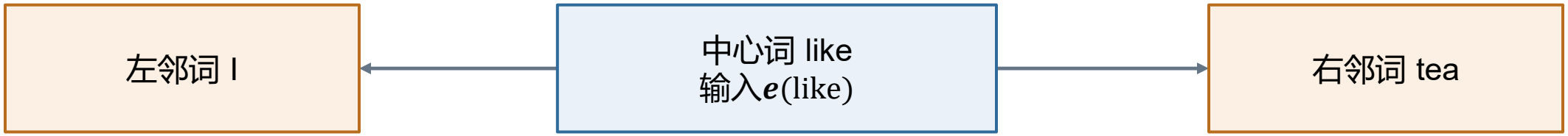
Word2vec、GloVe和fastText分别利用什么学习信号，又怎样组成词表示?

Skip-gram的邻词预测任务

下一词LM以I like预测tea，只使用已读前缀。当前任务是学习可复用词向量：以like预测左右邻词I、tea。

Word2vec用局部预测提供表示学习信号。**分布假设**联系词的意义、功能与使用语境，邻近出现的词不因此同义。

词级序列：I　like　tea

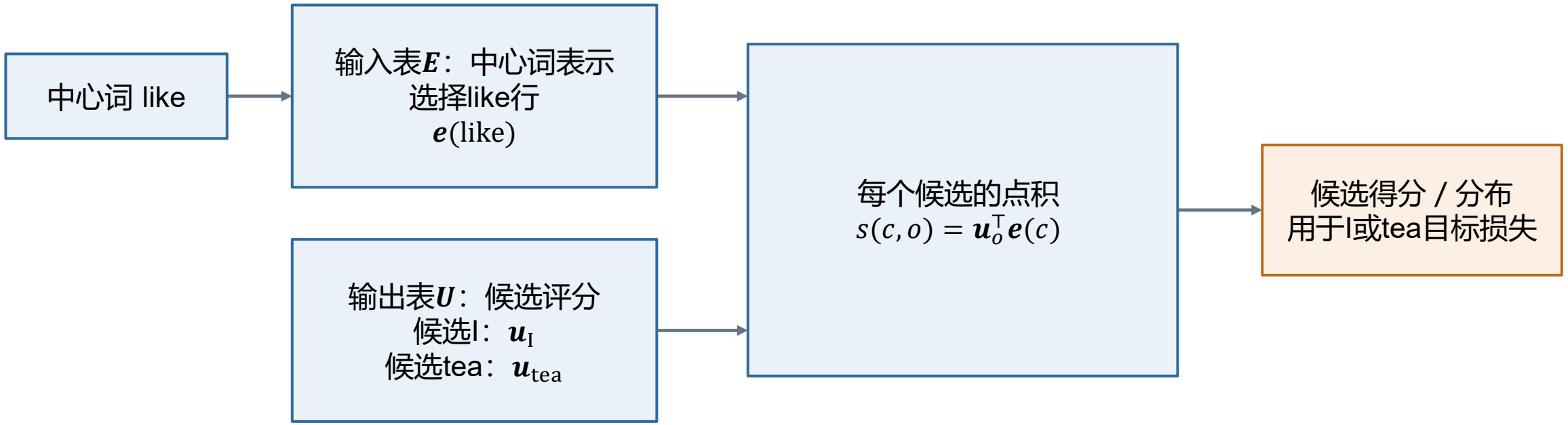


两个样本：(like, I)；(like, tea)

Skip-gram由中心词预测左右窗口内的邻词。本例半径为1，两个样本共同使用like的输入表示；语料中的许多窗口提供学习信号。一个窗口只能说明样本关系，不能证明学得语义效果。

Skip-gram的输入与输出嵌入

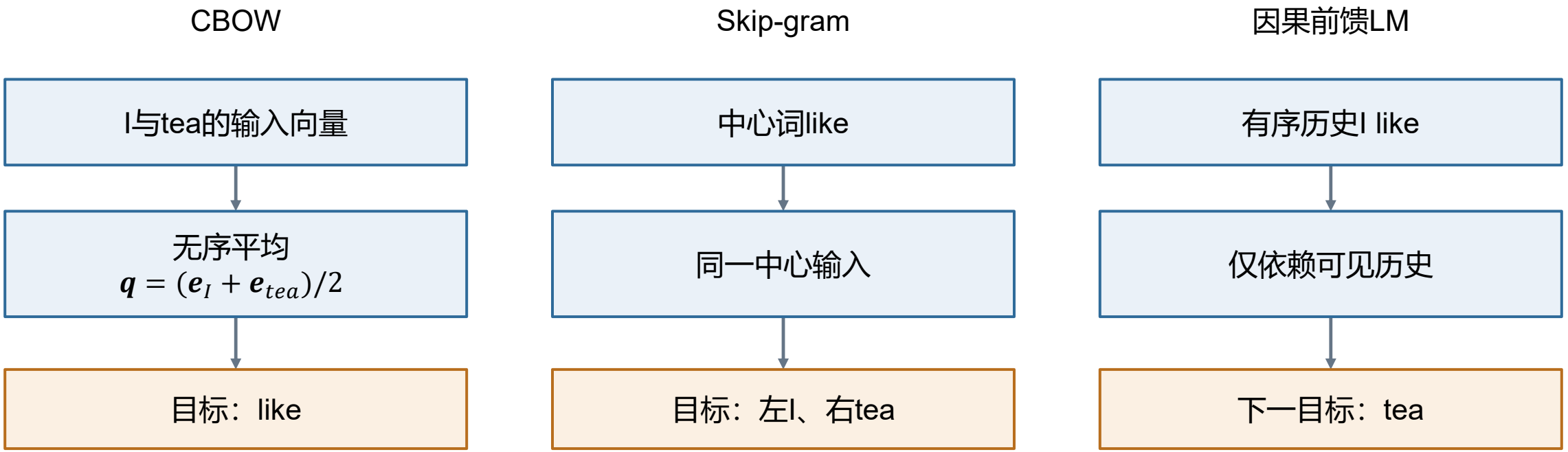
E描述中心词，U为候选邻词评分，两表承担不同职责。图仅画I与tea，其他候选省略；同一词在两表可有不同向量。



两表均有V行、d列，每行转置后得到d维列向量。这里用完整Softmax解释$p_\theta(o \mid c) \propto \exp(s(c,o))$的参数职责；不展开实际训练近似。默认不绑定E与U，提取输入表、输出表或组合时须说明选择；U未填入任何虚构训练数值。

词向量任务与语言模型的预测条件

这些任务都预测词，但使用的条件和预测目标不同。



连续词袋模型（CBOW）聚合语境预测中心词，平均不保留窗口内顺序。CBOW / Skip-gram的左右窗口任务服务表示学习；完整LM还须把历史条件分布与起止事件衔接。词向量可供其他任务复用，局部目标或负采样打分不能直接当整句概率。

Word2vec原始结构：CBOW与Skip-gram

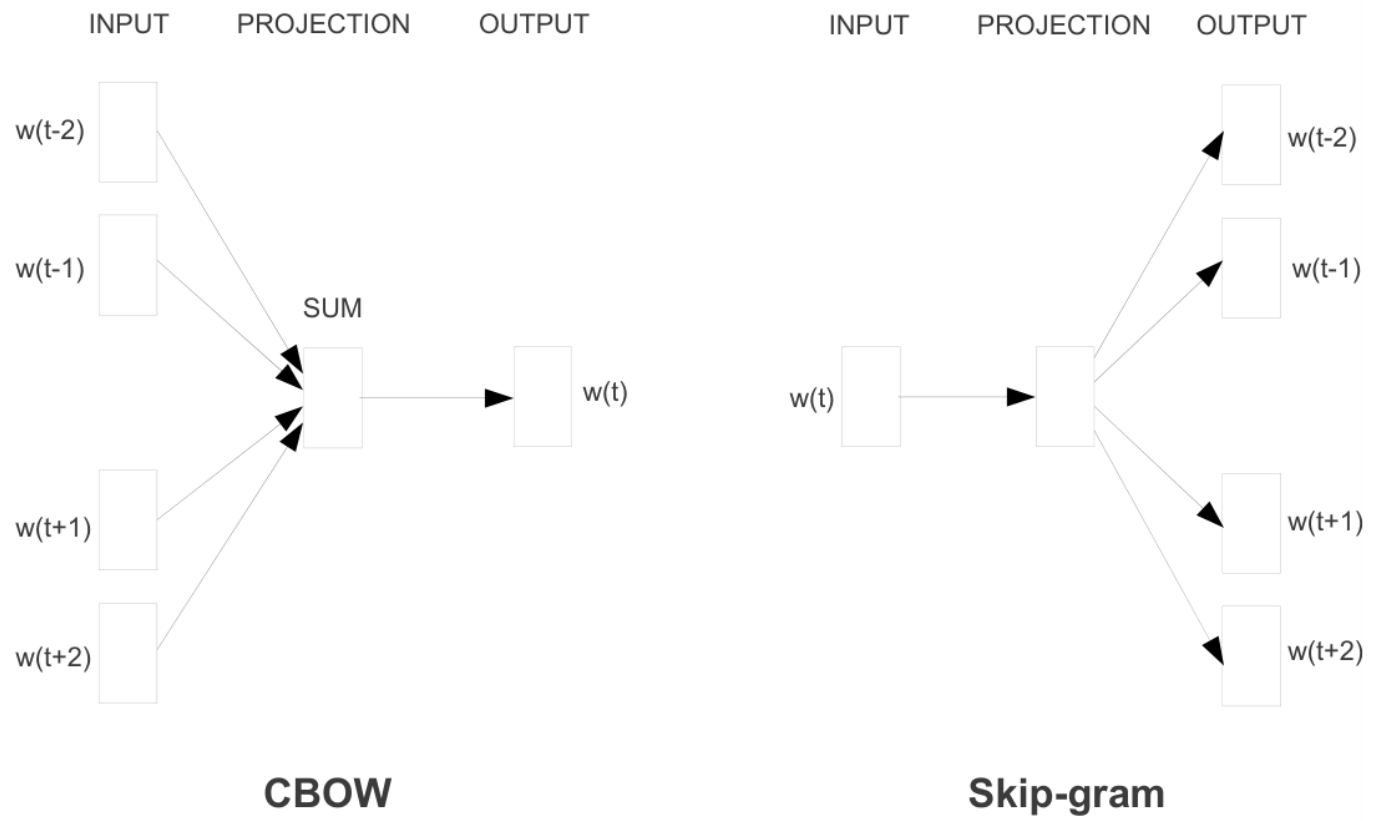


Figure 1: New model architectures. The CBOW architecture predicts the current word based on the context, and the Skip-gram predicts surrounding words given the current word.

- CBOW由左右语境预测中心词；Skip-gram方向相反。
- 原图求和，本讲取平均；二者均不保留顺序，尺度不同。
- 多条输出线是邻词预测任务，不是顺序生成。

Mikolov等（2013）Fig.1

GloVe的设计动机：共现概率比

探查语境词k	ice与steam的关系对照	$P(k \mid \text{ice})/P(k \mid \text{steam})$
solid	更偏向ice	远大于1
gas	更偏向steam	远小于1
water	与两者都有关	接近1

同一语境词对ice与steam的关联强弱，可以通过共现概率比比较。令$P(k \mid i) = C_{ik} / \sum_r C_{ir}$，这里的条件是中心词，不是句子前缀。

对数把概率比转为差值，GloVe用词向量差与语境向量的内积表达这种关系。归一化的行总量项可由偏置吸收，最终用内积与偏置拟合对数计数。

GloVe的共现统计与向量拟合

共现概率比提示应保留词与不同语境的关系。**GloVe**将跨语料汇总的共现计数转为向量拟合约束；下面两句按词切分，左右半径1、等权、不跨句。

C：行是中心词i，列是语境词j

I like tea
I like coffee

左右半径1、不跨句
有向邻接累积到C

	I	like	tea	coffee
I	0	2	0	0
like	2	0	1	1
tea	0	1	0	0
coffee	0	1	0	0

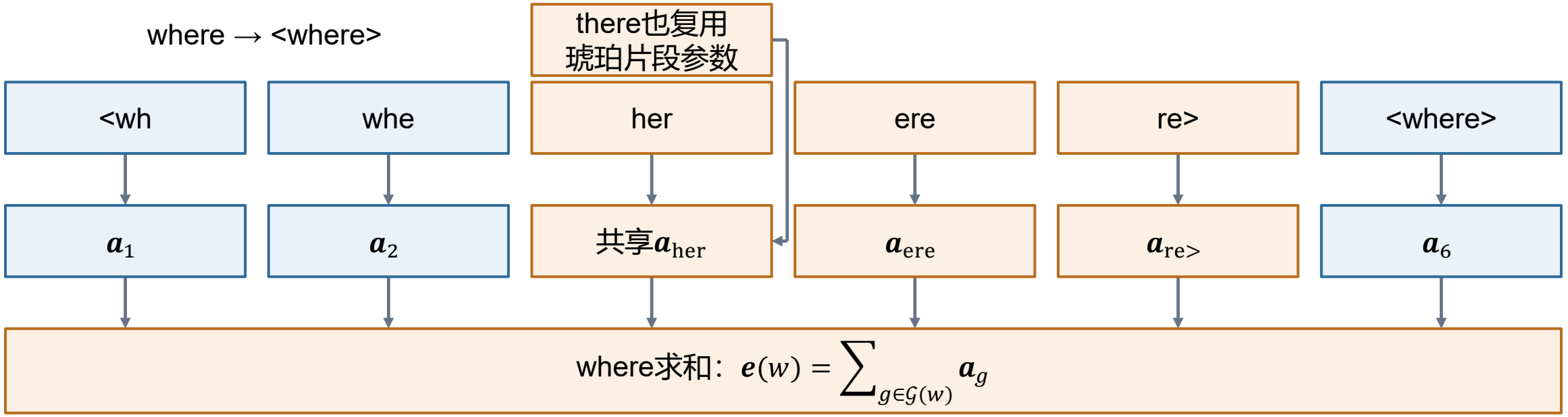
选取$C_{\text{like,I}} = 2$
对数拟合目标：$\ln 2$

$$\boldsymbol{w}_i^{\top}\tilde{\boldsymbol{w}}_j + b_i + \tilde{b}_j \approx \ln C_{ij} (C_{ij} > 0)$$

C_{ij}为共现次数，$\boldsymbol{w}_i, \tilde{\boldsymbol{w}}_j$为词侧、语境侧向量，$b_i, \tilde{b}_j$为偏置。例如like–I计数2，拟合目标为$\ln 2$；like–tea计数1，目标为0。零计数不取对数，拟合并非每项精确相等；本讲不推加权损失。

fastText的字符片段与整词表示

逐词查表为不同拼写设置独立参数，无法直接共享词内部结构。**fastText**加入字符片段共享；where与there加首尾边界符，此处仅显示三元组。



there：<th、the、her、ere、re>＋整词；琥珀标出两词共享的片段参数

where与there共同训练her、ere、re>片段参数，各自整词项独立；各项向量求和。her片段与整词<her>不同。未见词可组合可用片段，质量仍需评价；固定拼写的输入不含句子语境。

词向量方法的学习信号与表示组成

方法	核心设计	主要作用	主要边界
Word2vec	局部预测；逐词表示	从语境预测学习词向量	静态词项表示
GloVe	汇总共现；逐词表示	用聚合统计约束词向量	静态词项表示
fastText	局部预测；字符片段共享	复用词内结构，组合未见词	不直接建模句子历史

三种方法可以从学习信号和表示组成两个维度比较：前者说明训练依据，后者说明参数组成与共享。

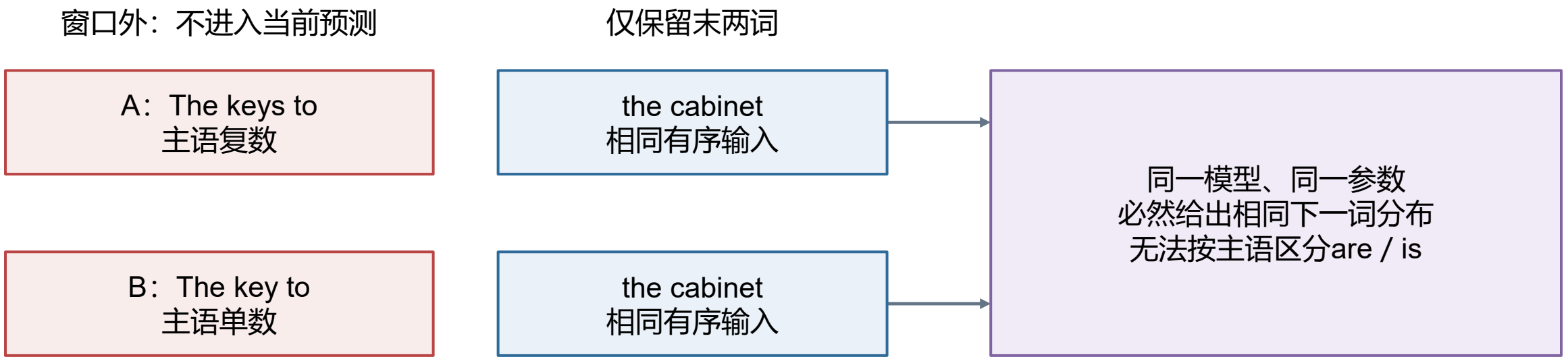
fastText可沿用Word2vec的局部预测路线。共现的“全局”指跨语料汇总；三者不构成性能升级链，也不直接等于完整因果LM。

4 递归语言模型与上下文表示

递归状态怎样利用早期历史，形成随上下文变化的表示？

固定窗口的历史截断

The keys to the cabinet后续接are，The key to the cabinet后续接is。按词切分，两条前缀的末两词相同，主语的单复数却不同。

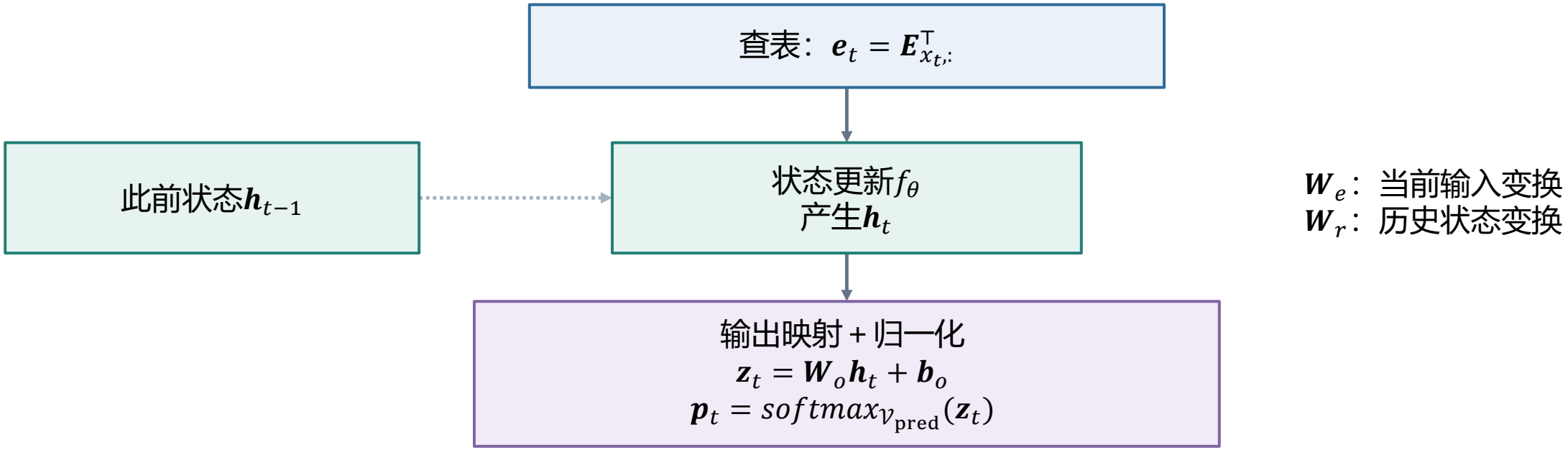


预测需求：A续接are，B续接is；决定区别的keys / key已被窗口舍弃

这是独立于8词人工表的语言依赖示例，只比较此处的语法续接。增大窗口可覆盖本例，但任意固定长度仍会截断更早输入。下一页用递归状态传递历史信息；RNN提供利用路径，不保证学会长期依赖。

RNN语言模型的状态递推

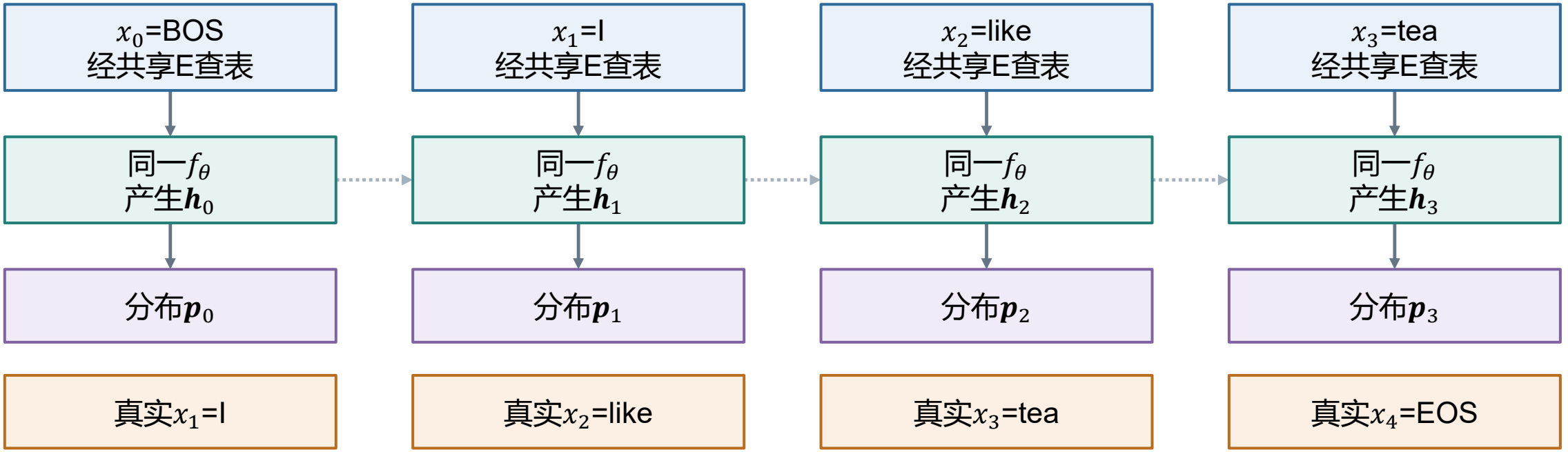
RNN语言模型让当前嵌入与此前状态共同形成新状态：$\boldsymbol{h}_t = \phi(\boldsymbol{W}_e \boldsymbol{e}_t + \boldsymbol{W}_r \boldsymbol{h}_{t-1} + \boldsymbol{b}_h)$。读cabinet时，此前状态可携带keys / key的信息。



$\boldsymbol{W}_e \in \mathbb{R}^{m \times d}$、$\boldsymbol{W}_r \in \mathbb{R}^{m \times m}$、$\boldsymbol{W}_o \in \mathbb{R}^{V \times m}$。参数跨位置共享，状态是依赖前缀的激活；递归提供历史影响路径，但不保证无损保留全部历史。

RNN时间展开与预测目标对齐

I like tea沿时间展开：青绿为状态，紫色为词表分布，琥珀为真实下一词。在当前预测位置，真实下一词用于选取目标概率并计算损失。

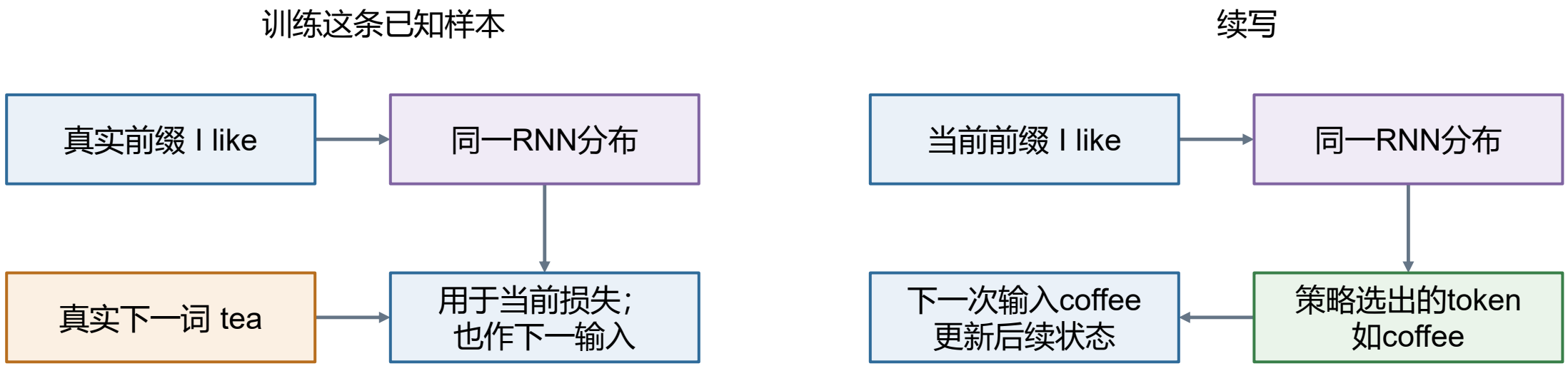


$$L(\theta) = -\frac{1}{4}\sum_{j=1}^{4} \ln p_\theta(x_j \mid x_{0:j-1})$$

$\boldsymbol{h}_2$尚未读tea；真实x_3 = tea从$\boldsymbol{p}_2$选择分量计算损失。四目标含EOS，排除BOS / PAD；$j = t + 1$。

RNN训练与续写的输入来源

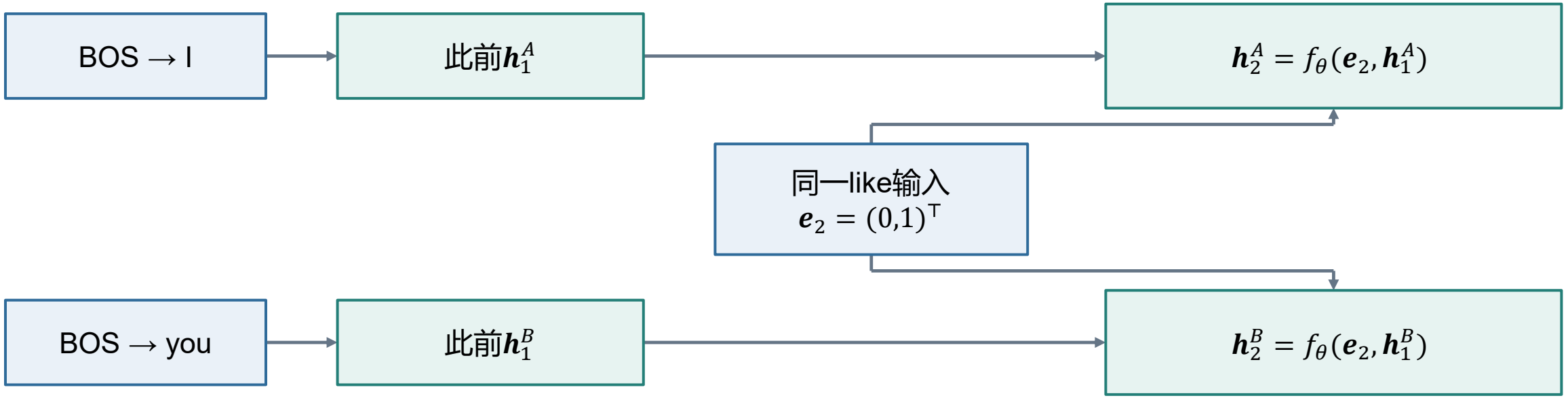
同一个RNN给出紫色分布：训练用琥珀真实词，续写反馈绿色策略选词。分布、真实目标与选词结果分别标明。



训练用真实tea计算损失并提供下一输入；续写若选出coffee，下一次状态更新就使用coffee。分布、真实目标与策略选词是不同对象；选到EOS时停止，具体策略留给第3讲。

同一词的静态嵌入与上下文状态

此处固定读完like的观察点。序列A为I like tea，序列B为you like tea；两者使用相同参数和同一like行。



静态嵌入是词项级共享参数，训练时仍可更新；RNN隐藏状态是依赖已读前缀的动态激活。同一like输入可对应不同状态，但不保证数值必然不同。当前状态未读tea，也不能等同于双向上下文模型的词表示。

输入表示与历史建模的组合

输入表示的组成与历史建模方式可以分别选择并组合。下表列出可采用的结构组合，不表示某次训练的实验结果。

输入表示的组成	历史利用方式	
	固定窗口：显式截取有序历史	递归状态：压缩已读前缀
逐词查表的静态嵌入	可组合	可组合
词内片段组合的静态输入	可组合	可组合

静态词表示可作为上下文网络的输入；fastText改变词内表示的组成，RNN通过状态传递前缀信息。固定窗口舍弃更早输入，递归状态允许其影响预测，但不保证无损记忆。

本讲总结

词表示与历史建模共同服务于下一词预测。

- 通过共享嵌入表取得输入向量，用同一预测损失共同训练嵌入与网络参数
- Word2vec用局部预测，GloVe拟合汇总共现，fastText共享字符片段
- 固定窗口截取后缀，RNN递归状态为更早前缀影响预测提供路径

已有下一词分布后，怎样在给定条件下生成完整序列？第3讲将连接条件输入、逐步反馈与选词策略。