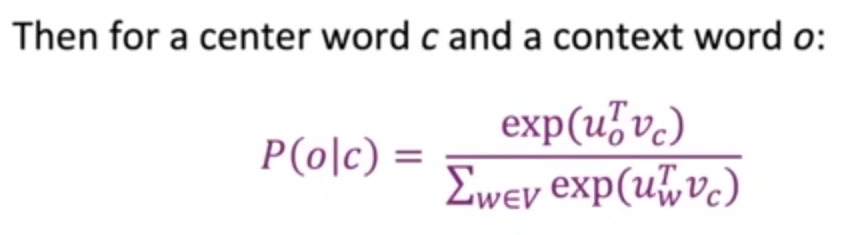
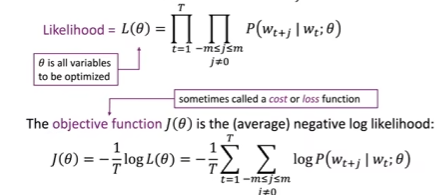
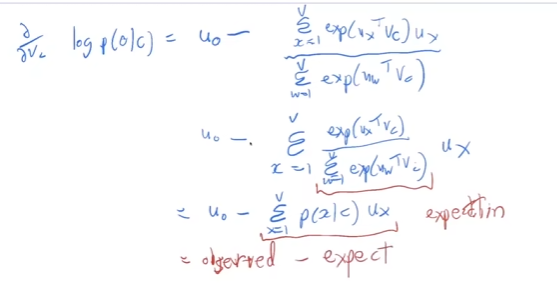
# 一.自然语言处理导引及词向量

## 1.词向量

一个词有两个词向量，一个是用作中心词另一个是用作上下文词







Positive 加上特征 negative减去特征

**2.各种模型**

对于所有的模型，训练复杂度为

其中E是训练epoch的个数，T是训练集中单词的个数，Q是针对每个模型架构进一步定义的。

* 对于NNLM模型：

有四层：输入层、投影层、隐藏层和输出层

N为单词个数、N\*D为投影层的维数、H为隐藏层大小、V为输出层维数

大部分的复杂性是由N×D×H引起的。

* 对于RNNLM模型：

没有投影层，只有输入层、隐藏层和输出层

其中单词表示D与隐藏层H具有相同的维数，同样，通过使用分层softmax，术语H ×V可以有效地简化为H×log2(V)。然后，大部分复杂性来自H ×H。

* 连续词袋模型(Continuous Bag-of-Words)——CBOW

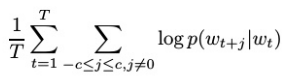
此架构类似于前馈NNLM，其中非线性隐藏层被删除。

CBOW架构根据上下文预测当前单词。复杂度：

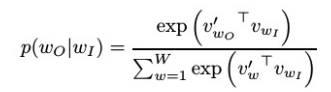
* 连续Skip-gram模型

Skip-gram根据当前单词预测周围单词。

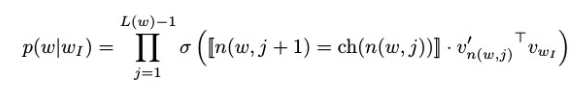
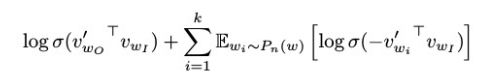
复杂度：

其中C是单词的最大距离。

Skip-gram模型的目标是最大化平均对数概率：较大的c会产生更多的训练样例，因此可以获得更高的精度，但代价是训练时间。基本Skip-gram公式使用softmax函数定义p(wt+j|wt):

w为单词数，计算∇logp(wO|wI)的成本与W成正比，W很大不切实际

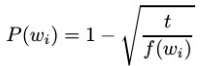
优化：

1. Hierarchical Softmax（L(w)为这条路径的长度、σ(x) = 1/(1 + exp(−x))）
2. 噪声对比估计(Noise contrast Estimation, NCE)：通过逻辑回归将数据与噪声区分开来
3. 负采样(NEG)

用于替换logP(wO|wI)

1. 频繁词的子采样

使用了一种简单的子抽样方法:训练集中的每个词wi被丢弃概率由公式计算



1. **依赖分析**

* 歧义

Prepositional phrase attachment ambiguity

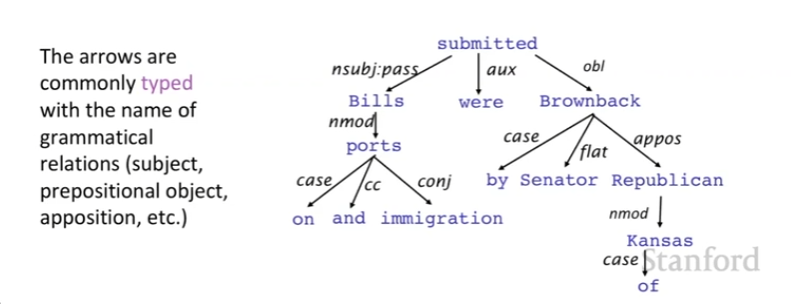
PP attachment ambiguities multiply

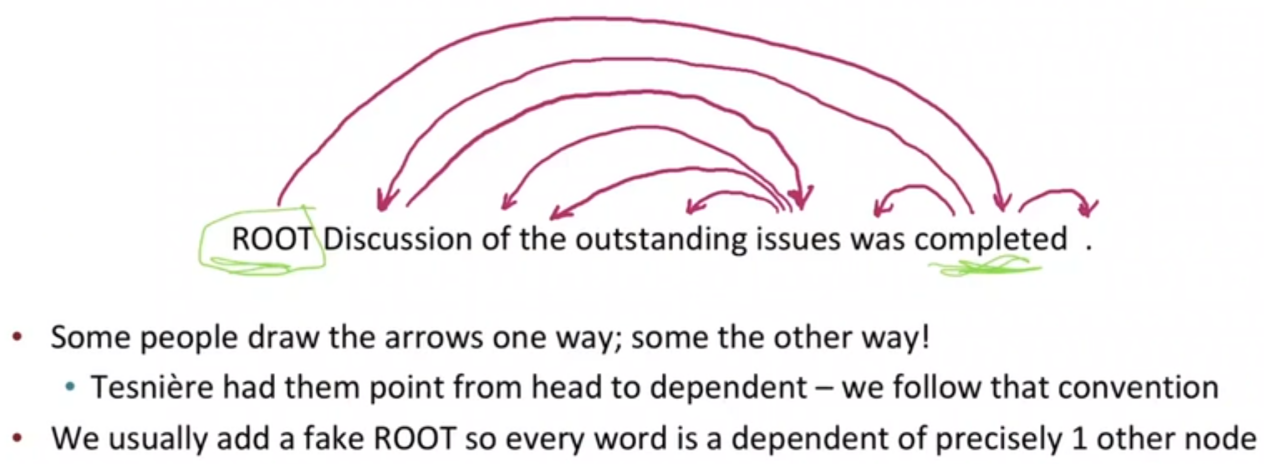
Coordination scope ambiguity

Adjectival/Adverbial Modifier Ambiguity

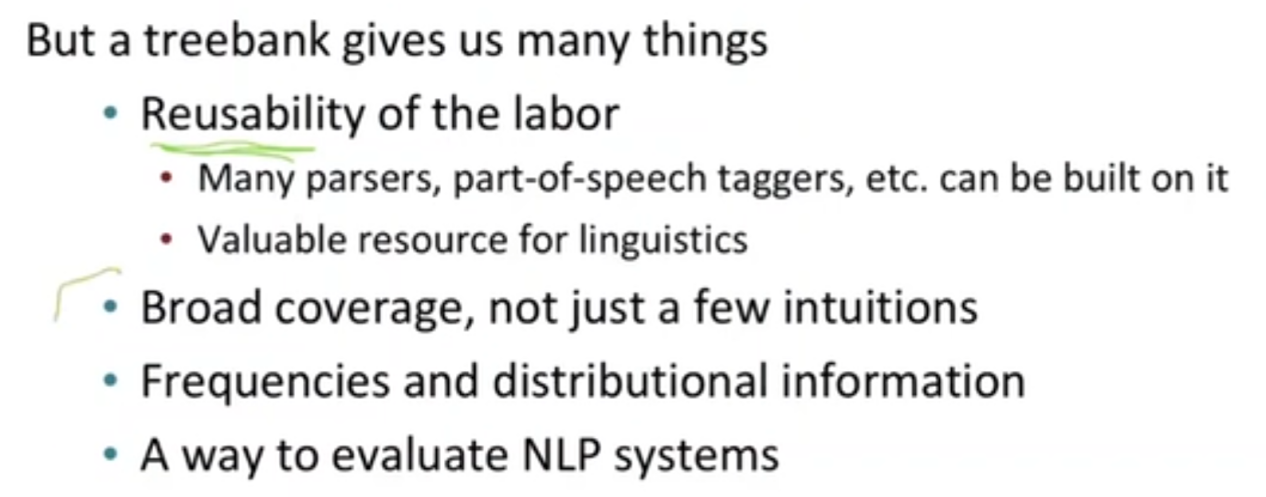
Verb Phrase（VP）attachment ambiguity

* Dependency Grammmar and Dependency Structure

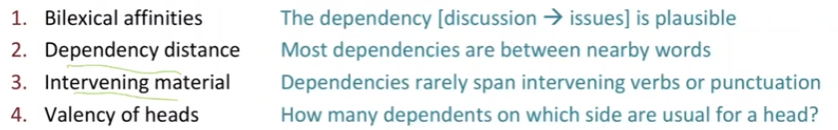




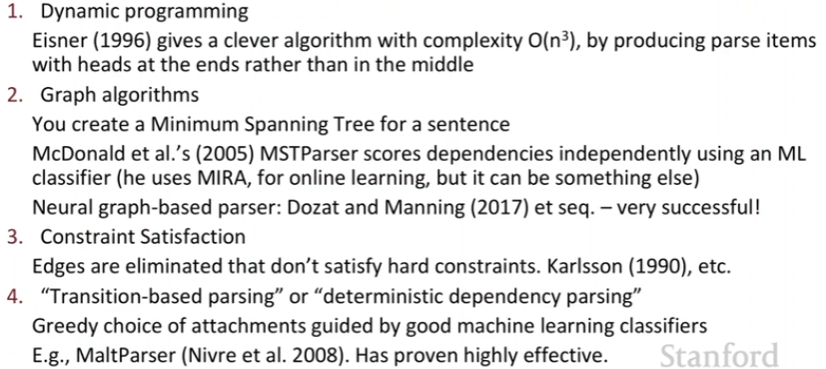
* 为什么使用treebank



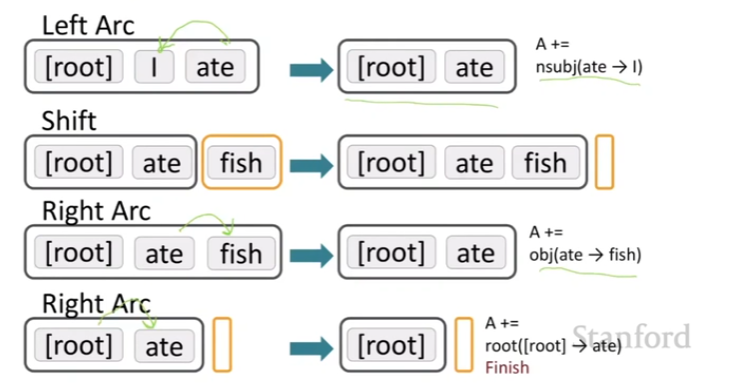
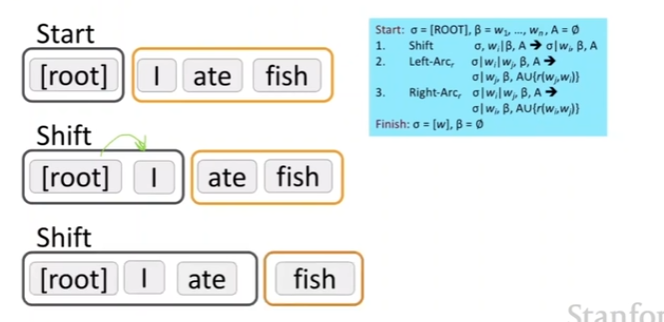
* 依赖关系的信息来源

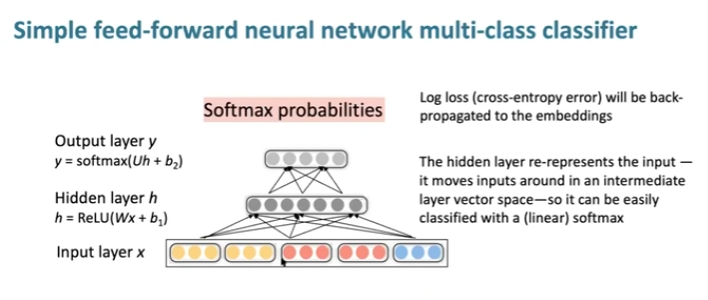


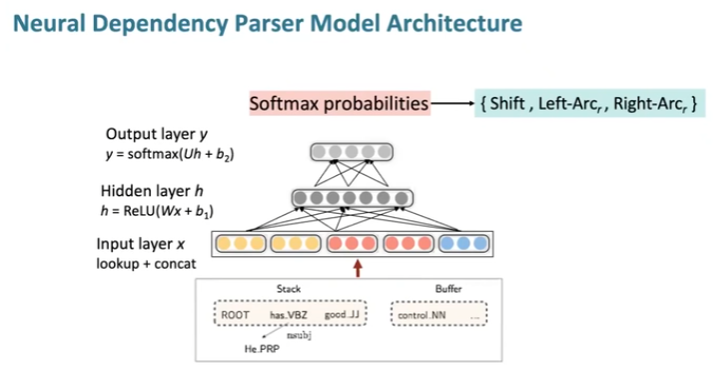
* 构建依赖关系解析器的方法

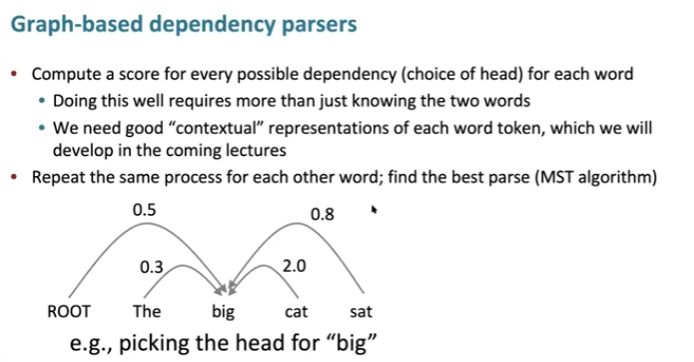


* Basic Transition-based dependency parser



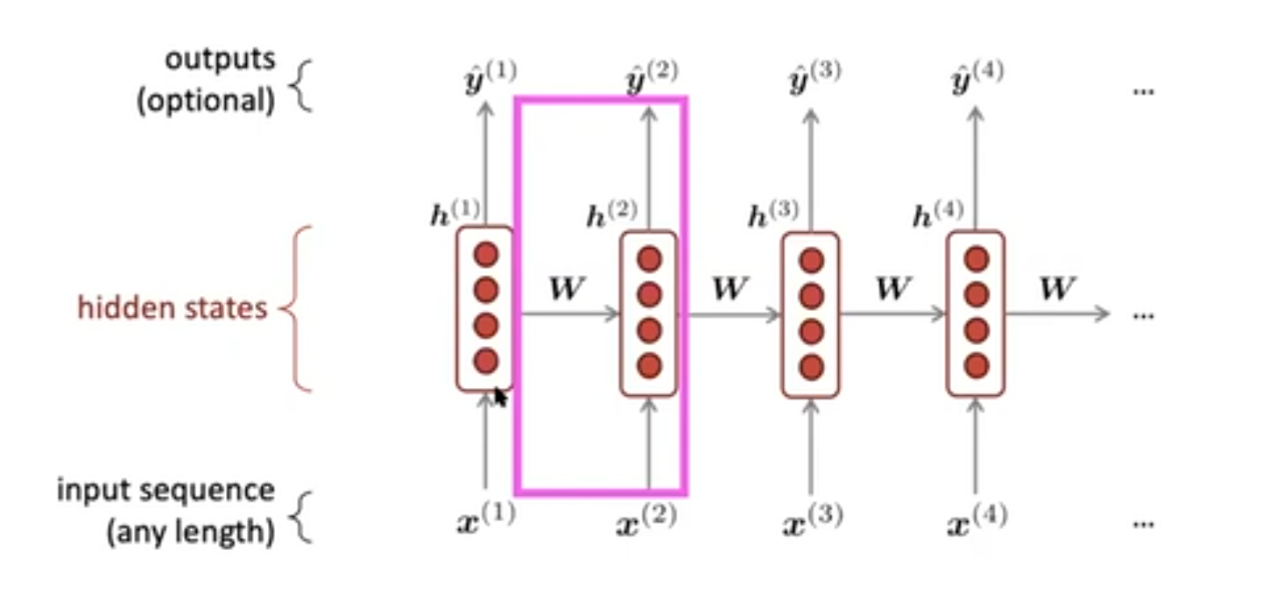




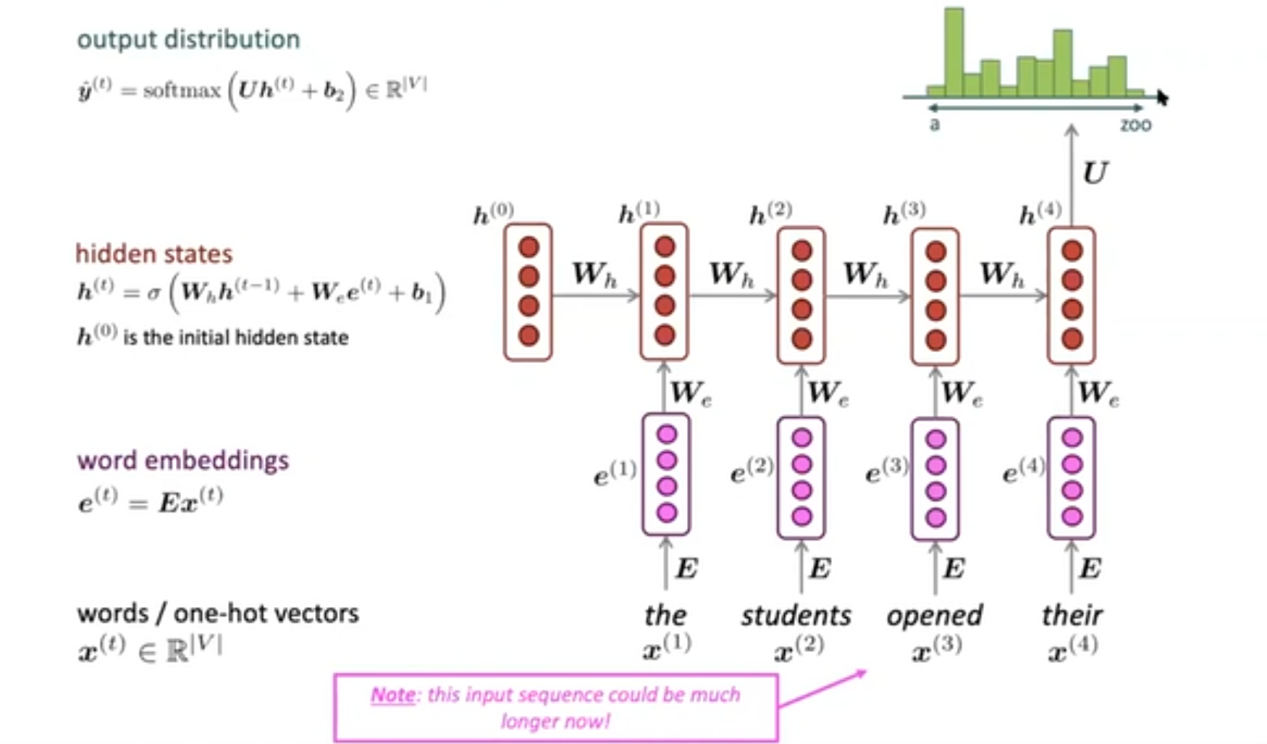


# 语言模型

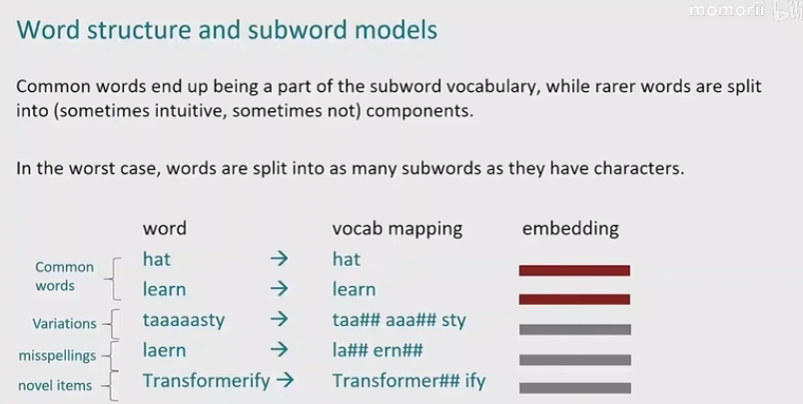
RNN：

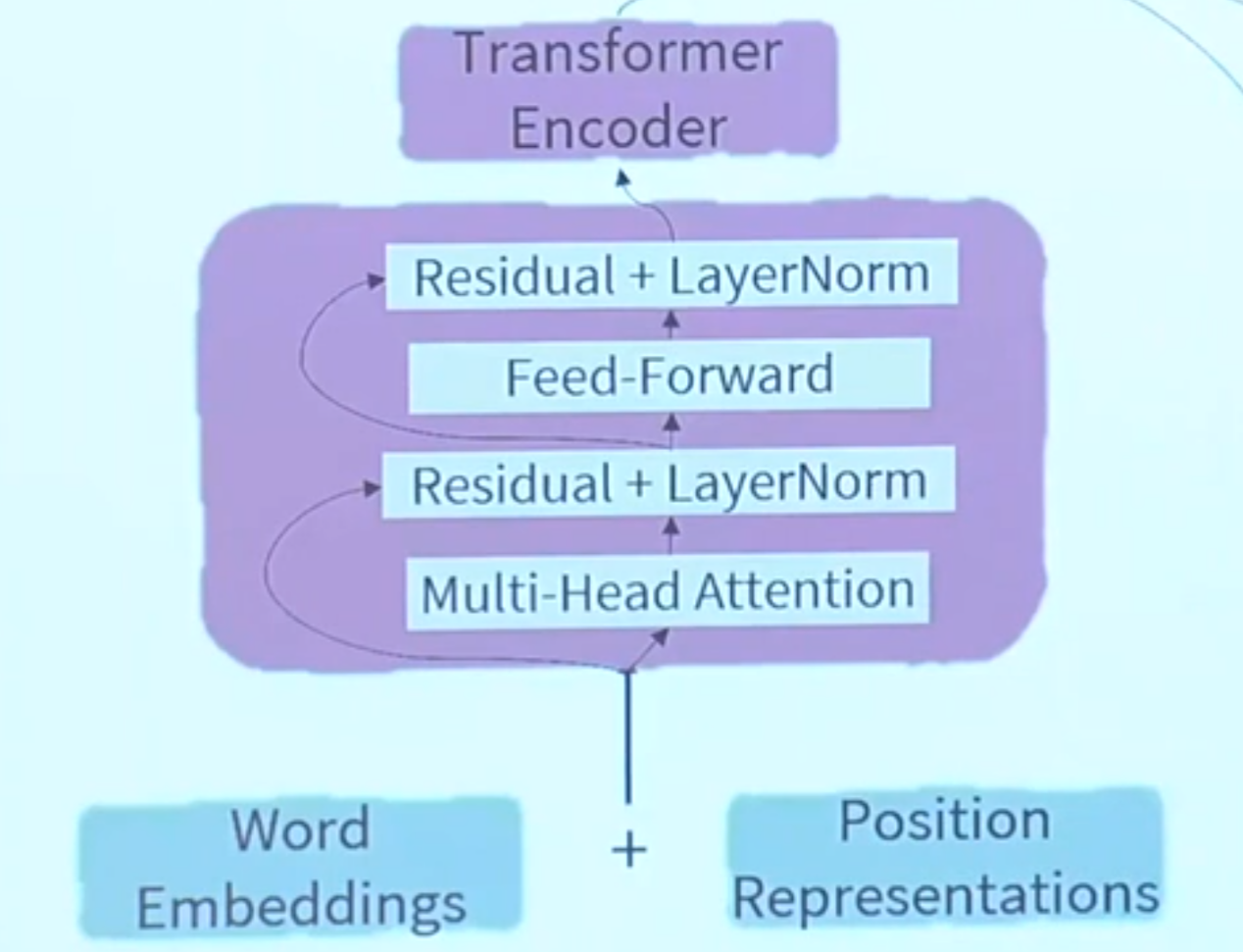


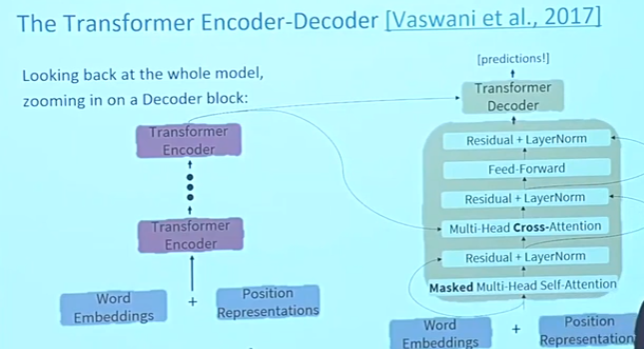
Eg：

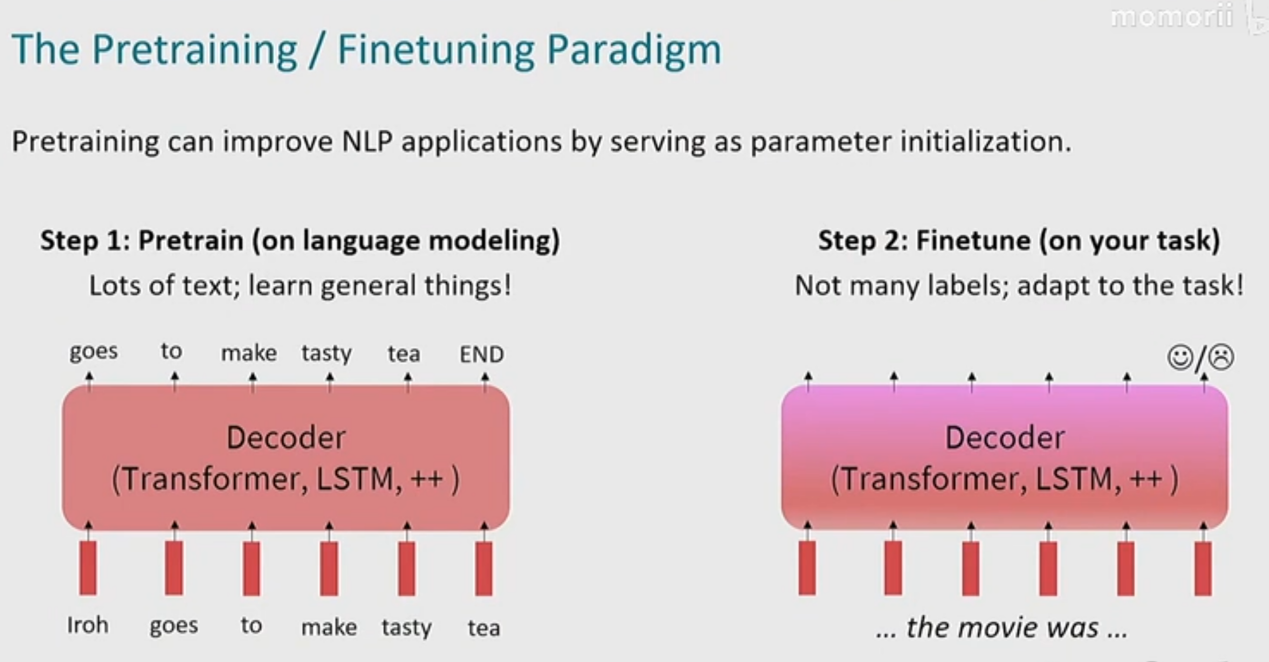


# **Transformer模型、预训练模型**









预训练的三种结构：encoder（编码器）decoder（解码器）

