DMWIS

吉林大学 JILIN UNIVERSITY

自然语言处理

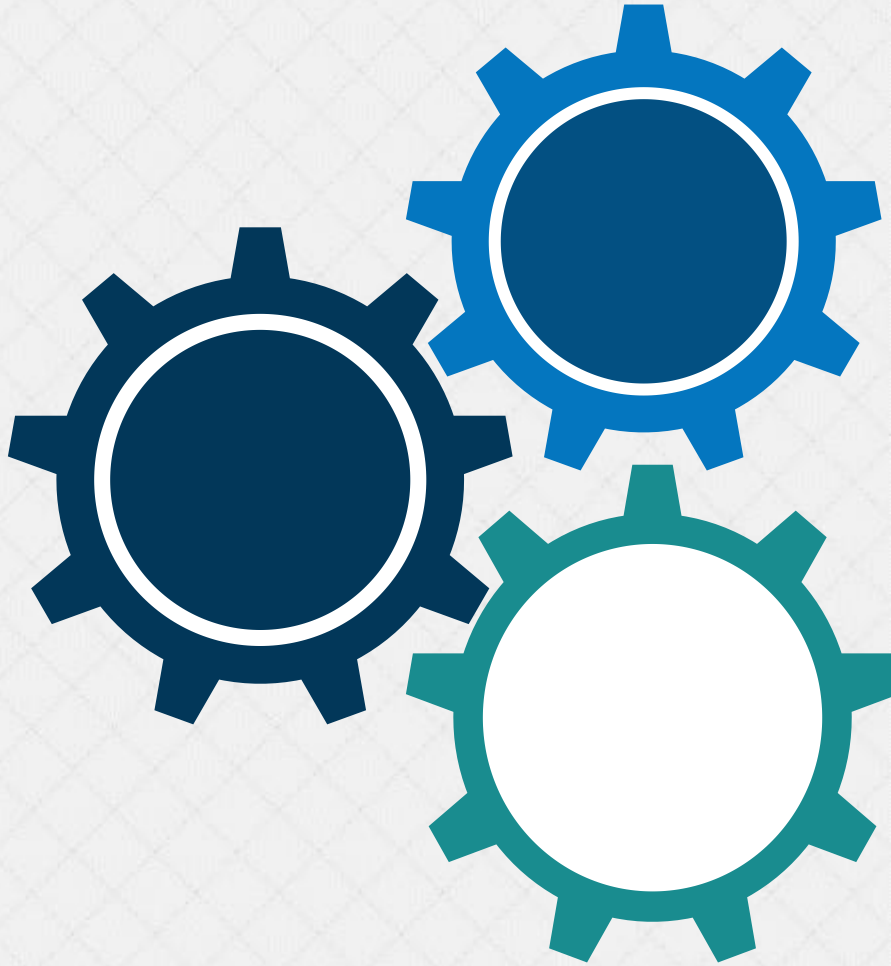
包铁

2023年9月16日

baotie@jlu.edu.cn
Data Mining and Web Information System Group （DMWIS），
College of Computer Science and Technology, Jilin University

主要内容

1. 基础
2. 规则分词
3. 统计分词
4. 语言模型
5. 分词工具



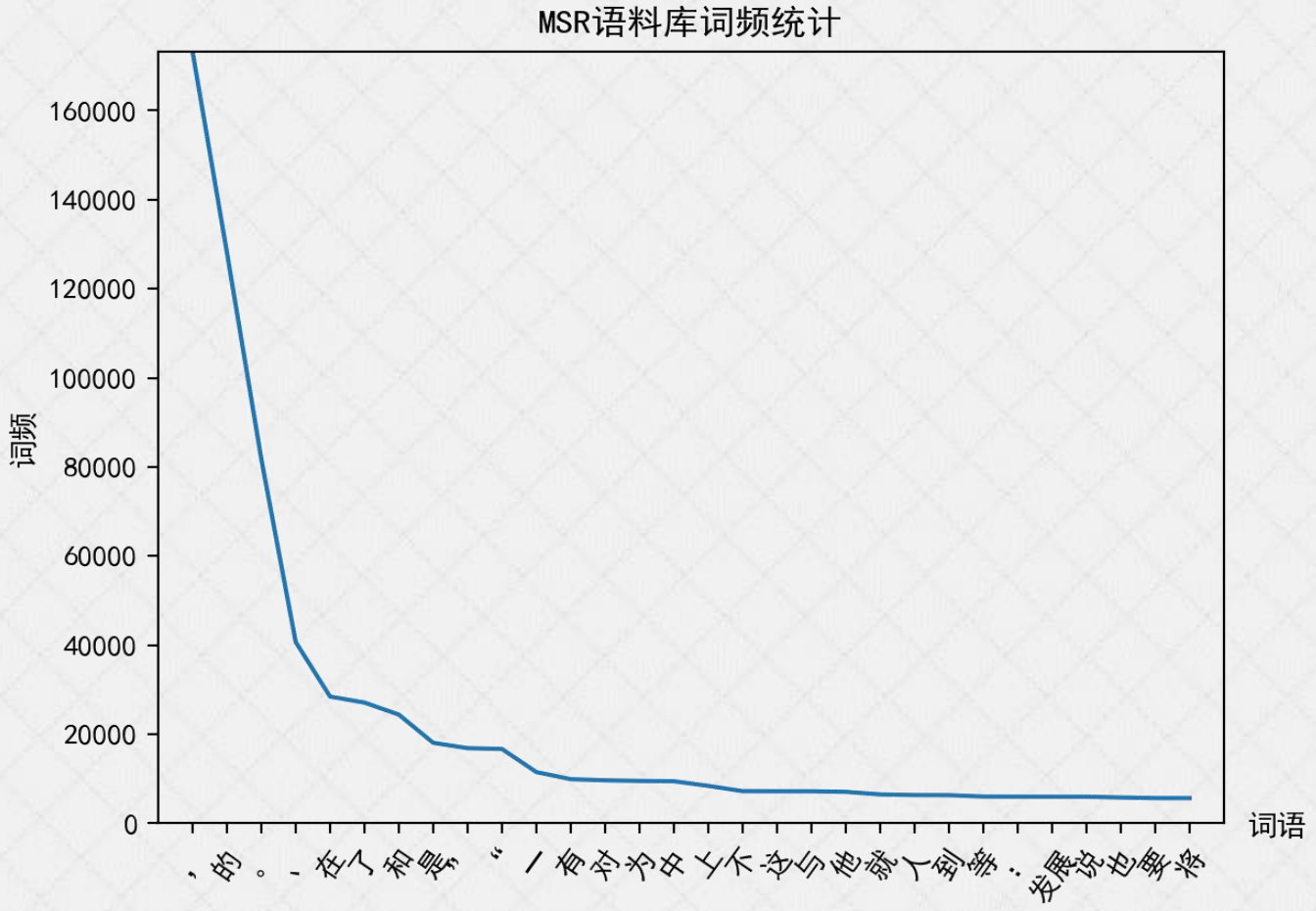
词的定义

- 语言学上，词是具有确定的语义或语法功能的基本单位
  - 定义比较模糊，难以作为计算依据
  - 标准难以统一：“吃饭”、“北京机场”
- 基于词典的中文分词中，词典中的字符串就是词
  - 词典外的字符串不是词，词典可以包含的词数量有限

词的性质-齐夫定律

- 一个单词的词频与它的词频排名成反比

1949年哈佛大学语言学家齐夫



满足幂律分布-长尾效应、二八原则、马太效应等

词典

- 基于搜狗实验室发布的互联网词库（SogouW，15万词条）
- 清华大学开放中文词库（THUOCL）
- HanLP汉语词库（千万级词条）-空格（可选）分隔、词性、词频

```
希望    v       386     n       96
希罕    a       1
希翼    v       1
希腊    ns      19
希腊共和国      ns      1
```

中文分词

- 定义：将连续的字序列，按照一定的规范重新组合成词序列的过程
- 词的抽象定义、词的具体界定难以明确统一的表达
- 中文有高度语境化、隐喻化特点，词的构成边界很难界定，而且需要消除切分中的歧义。

中文分词-切分

- 切分粒度-粗粒度、细粒度
  - 浙江/大学/坐落/在/西湖/旁边。
- 可能产生的歧义切分
  - 交集型-结合/成；结/合成
  - 组合型-站/起/身/来、起身/去/北京
- 早期设计的分词系统-基于词典进行分词
  - 未考虑词汇上下文相关性

中文分词-未登录词

- 词汇的特性
  - 稳固性、常用性
  - 能产性-表示新事物的新词汇不断出现
- 主要构成
  - 大部分为专有名词
  - 也包括通用新词、专业术语
- 基于构词学的方法对于专门领域效果较好-人名，统计学

中文分词常用方法

- 基于词典
  - 将文档中的字符串与词典中的词条逐一匹配，匹配成功则切分，否则不予切分
- 基于语法和规则
  - 分词的同时进行句法、语义分析，利用句法信息和语义进行词性标注，解决分词歧义问题
- 基于统计方法-基于字符串在语料库中出现的统计频率进行判断

具体方法-正向最长匹配

- 从左向右取待切分汉语句的m个字符作为匹配字段，m为词典中最长的词字符串长度。
- 查找词典并进行匹配，如果匹配成功：则将这个匹配字段作为一个词切分，如果匹配不成功：则将这个匹配字段的最后一个字去掉，剩下的作为匹配字段，进行再次匹配，直到切分出所有词。

规则分词

具体方法-正向最长匹配

词典

长江大桥

南京市长

南京市

大桥

长江

江

语句：南京市长江大桥

分词后：南京市长 江 大桥

具体方法-逆向最长匹配

- 从右向左取待切分汉语句的m个字符作为匹配字段，m为词典中最长的词字符串长度。
- 查找词典并进行匹配，如果匹配成功：则将这个匹配字段作为一个词切分，如果匹配 不成功：则将这个匹配字段的最前面一个字去掉，剩下的作为匹配字段，进行再次匹 配。直到切分出所有词。
- 汉语中偏正结构较多，逆向匹配可以适当提高精度。

规则分词

具体方法-逆向最长匹配

词典

长江大桥

南京市长

南京市

大桥

长江

江

语句：南京市长江大桥

分词后：南京市 长江大桥

具体方法-双向最长匹配

- 将正向最大匹配法与逆向最大匹配法组合。先根据标点对文档进行粗切分，把文档分解成若干个句子，然后再对这些句子用正向最大匹配法和逆向最大匹配法进行扫描切分。如果两种分词方法得到的匹配结果相同，则认为分词正确，否则，按最小集处理。

规则分词

具体方法-双向最长匹配

词典

长江大桥

南京市长

南京市

大桥

长江

江

语句：南京市长江大桥

正向分词后：南京市长 江 大桥

逆向分词后：南京市 长江大桥

选择最小集：南京市 长江大桥

中文分词评测

- 准确率（accuracy）：用来衡量一个系统的准确程度的值，可以理解为一系列评测指标（不同任务应采用不同指标）
  - 正确判断数占总测试数的比率
  - 不同场景可能不全面-某项疾病的检测的指标计算
  - 精确率、召回率、F1值
- 混淆矩阵与TP/FN/FP/TN

中文分词评测

- 精确率（Precision）指的是预测结果中正类数量占全部结果的比率
  - $P = \frac{\text{TP}}{\text{TP+FP}}$
- 召回率：所有正类样本中，能回想起的比例
  - $R = \frac{\text{TP}}{\text{TP+FN}}$

预测 \ 答案	P	N
P	TP	FP
N	FN	TN

中文分词评测

- 混淆矩阵
  - TP（True Positive，真阳）：预测是P，答案是P
  - FP（False Positive，假阳）：预测是P，答案是N
  - TN（True Negative，真阴）：预测是N，答案是N
  - FN（False Negative，假阴）：预测是N，答案是P
  - 样本全集 =TP ∪ FP ∪ FN ∪ TN，相互无交集

预测＼答案	P	N
P	TP	FP
N	FN	TN

中文分词评测

- 中文分词P、R、F1计算
  - 标准答案的所有区间构成一个集合A，作为正类
  - 此集合之外的所有区间构成另一个集合（A的补集），作为负类
  - 记分词结果所有单词区间构成集合B

$$TP \cup FN = A$$
$$TP \cup FP = B$$
$$TP = A \cap B$$

$$P = \frac{|A \cap B|}{|B|}$$
$$R = \frac{|A \cap B|}{|A|}$$

中文分词评测

- **中文分词P、R、F1计算**

	单词序列	集合	集合中的元素
标准答案	结婚 的 和 尚未 结婚 的	A	$[1,2],[3,3],[4,4],[5,6],[7,8],[9,9]$
分词结果	结婚 的 和尚 未结婚 的	B	$[1,2],[3,3],[4,5],[6,7,8],[9,9]$
重合部分	结婚 的 ~~和尚未结婚~~ 的	$A \cap B$	$[1,2],[3,3],[9,9]$

$$TP \cup FN = A$$
$$TP \cup FP = B$$
$$TP = A \cap B$$

$$P = \frac{|A \cap B|}{|B|}$$
$$R = \frac{|A \cap B|}{|A|}$$

分词“准确率”为：

$$P = \frac{3}{5} = 60\%$$
$$R = \frac{3}{6} = 50\%$$
$$F_1 = \frac{2 \times 60\% \times 50\%}{60\% + 50\%} = 55\%$$

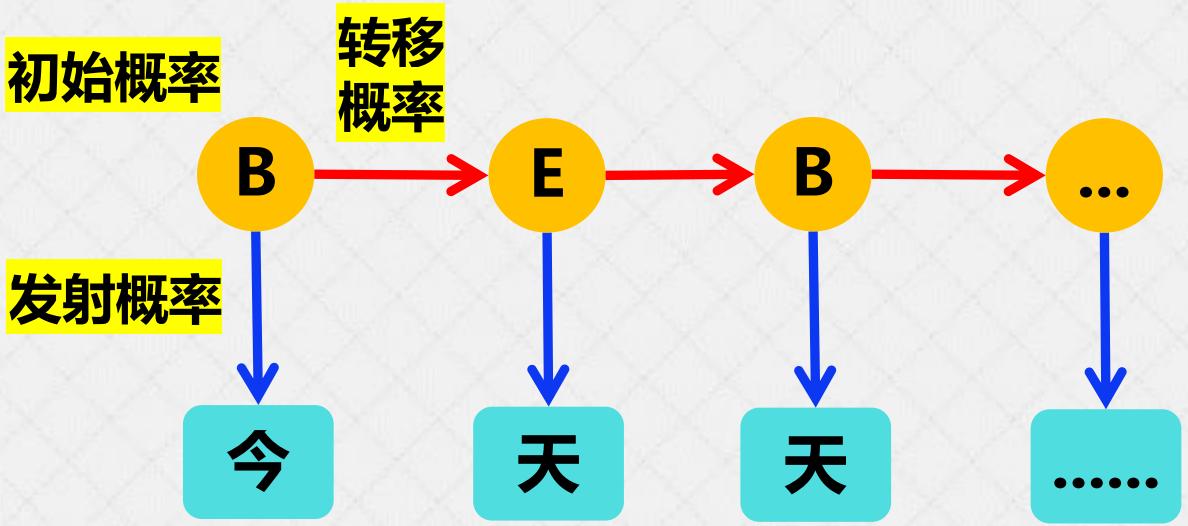
统计分词的含义

- 统计分词
  - 基于人们对中文词语的经验（基于语料库）。在中文文章的上下文中，相邻的字搭配出现的频率越多就越有可能形成一个固定的词。
- 统计分词的主要步骤
  - 建立统计语言模型-学习语料库中的语言知识
  - 对句子进行单词划分，然后对划分结果进行概率计算，获得概率最大的分词方式。 使用统计学习算法，如隐马尔可夫模型、条件随机场等。

统计分词一般步骤

- 构建和使用语言统计模型
  - 标注需分词的语句-B：词开始，M：词中间，E：词结束，S：单字的词

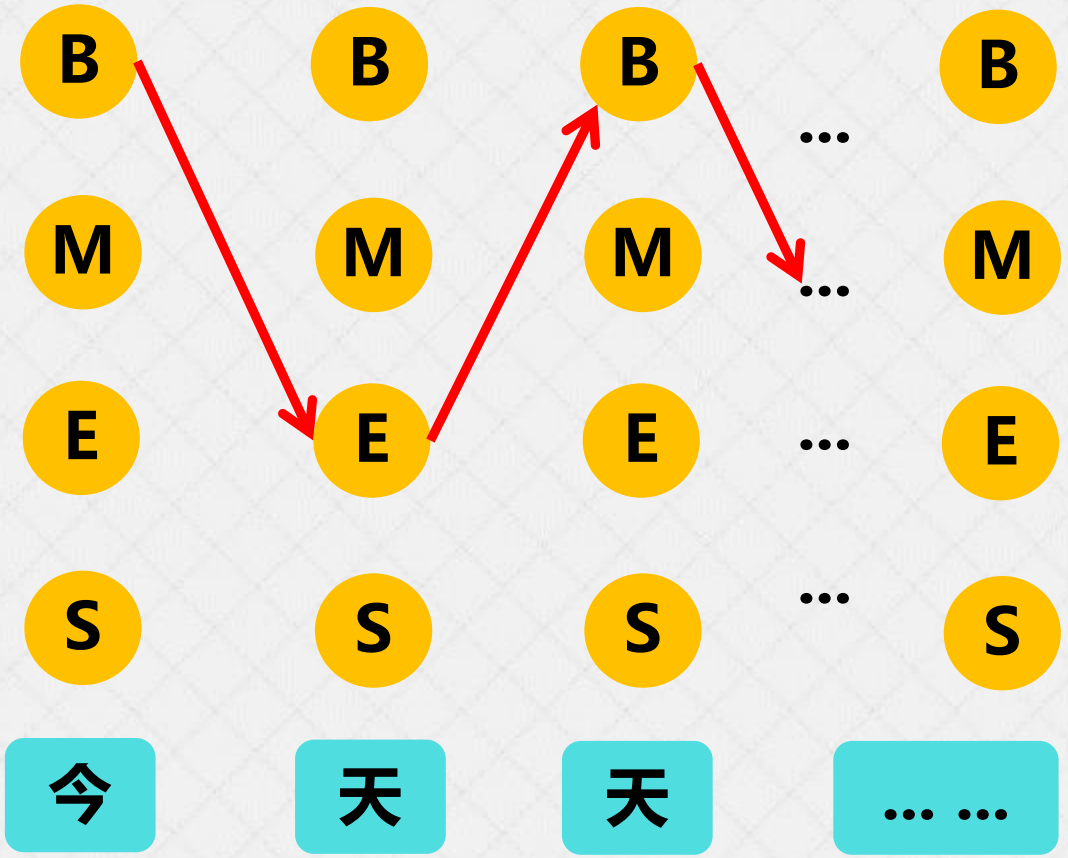
今	天	天	气	非	常	好
B	E	B	E	B	E	S



统计分词一般步骤

- 构建和使用语言统计模型
  - 维特比算法搜索状态序列

今	天	天	气	非	常	好
B	E	B	E	B	E	S



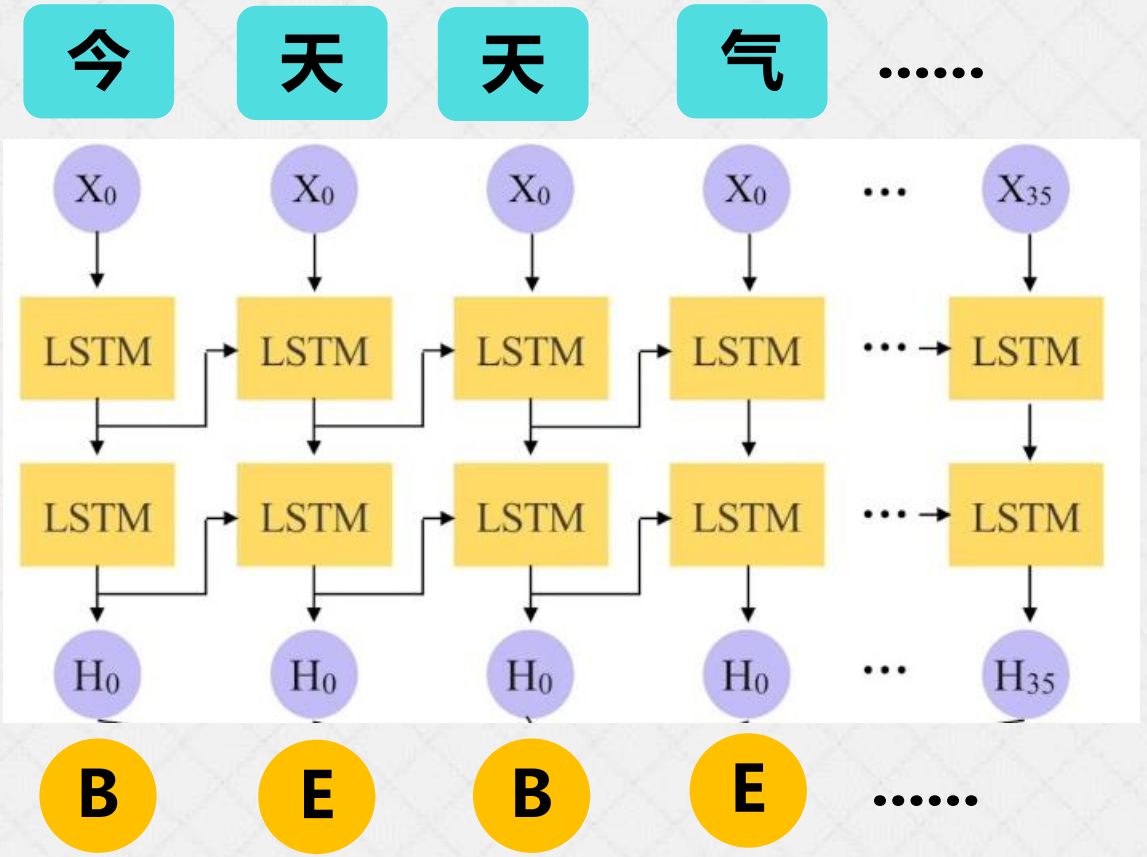
深度神经网络分词

- 基于深度学习的方法，将深度神经网络引入到分词任务中
- 深度学习技术不断发展，越来越多的深度学习模型用于解决NLP中的各种任务，尤其一些应用的高级任务
- 基于深度学习的NLP基础技术也会在后面课程中介绍

统计分词

深度神经网络分词

- 需要一个极大的标注语料库
- 选择一个合适的深度学习模型，训练模型
- 模型训练结束后，就可以输入需分词语句，返回预测结果



语言模型的定义

- **模型**指的是对事物的数学抽象
- **语言模型**（Language Model，LM）指的就是对语言现象的数学抽象
  - 给定一个句子w，语言模型就是计算句子的出现概率$p(w)$的模型
  - 判断一个语言序列是否是正常语句-一般需分词，不直接针对对整个句子

$$P(\text{我是中国人}) > P(\text{中国是我人})$$

语言模型的计算

- 链式法则

$$P(w_1, w_2, \ldots, w_n) = P(w_1)P(w_2|w_1)\cdots P(w_n|w_1, \ldots, w_{n-1})$$

$$P(\text{我是中国人}) = P(\text{我}) * P(\text{是}\mid\text{我}) * P(\text{中}\mid\text{我是}) * P(\text{国}\mid\text{我是中}) * P(\text{人}\mid\text{我是中国})$$

- 马尔科夫假设-针对计算代价大、数据稀疏问题
  - 某个词出现的概率只依赖于前面的有限个词（如n个词）

$$P(w_i|w_1,\ w_2,\ \ldots,\ w_{i-1}) \approx P(w_i|w_{i-k},\ \cdots,\ w_{i-1})$$

语言模型的用途

- 自然语言的生成、模型预训练
- NLP中各种任务均可使用
  - 分词
  - 序列标注
  - 文本分类
  - .....

N-gram语言模型

- 基于马尔可夫假设，即假设当前词出现的概率只依赖于前n-1个词，可以得到

$$P(w_i|w_1, w_2, \ldots, w_{i-1}) = P(w_i|w_{i-n+1}, \ldots, w_{i-1})$$

- NLP中各种任务均可使用
  - n=1 unigram $P(w_i|w_1,\ w_2,\ \ldots,\ w_{i-1}) \approx P(w_i)$
  - n=2 bigram $P(w_i|w_1, w_2, \ldots, w_{i-1}) \approx P(w_i|w_{i-1})$
  - n=3 trigram $P(w_i|w_1, w_2, \cdots, w_{i-1}) \approx P(w_i|w_{i-2}, w_{i-1})$
  - ……

N-gram语言模型计算

- N-gram语言模型的计算-基于语料库统计
  - 如何计算

$$P(w_i|w_{i-k},\ldots,w_{i-1}) = \frac{P(w_{i-k},\ldots,w_{i-1},w_i)}{P(w_{i-k},\ldots,w_{i-1})} = \frac{\mathrm{Count}(w_{i-k},\ldots,w_{i-1},w_i)}{\mathrm{Count}(w_{i-k},\ldots,w_{i-1})}$$

  - 例如：

$$P(w_i|w_{i-1}) = \frac{\mathrm{Count}(w_{i-1},\ w_i)}{\mathrm{Count}(w_{i-1})}$$

N-gram语言模型训练

- <s> 中 国 加 油 </s>
- <s> 我 是 中 国 人 </s>
- <s> 第 一 届 中 国 国 际 进 口 博 览 会 </s>

$$P(国|中)=\frac{Count(中国)}{Count(中)}=\frac{3}{3}=1$$

$$P(人|国)=\frac{Count(国人)}{Count(国)}=\frac{1}{4}=0.25$$

$$P(人|是)=\frac{Count(是人)}{Count(人)}=\frac{0}{1}=0$$

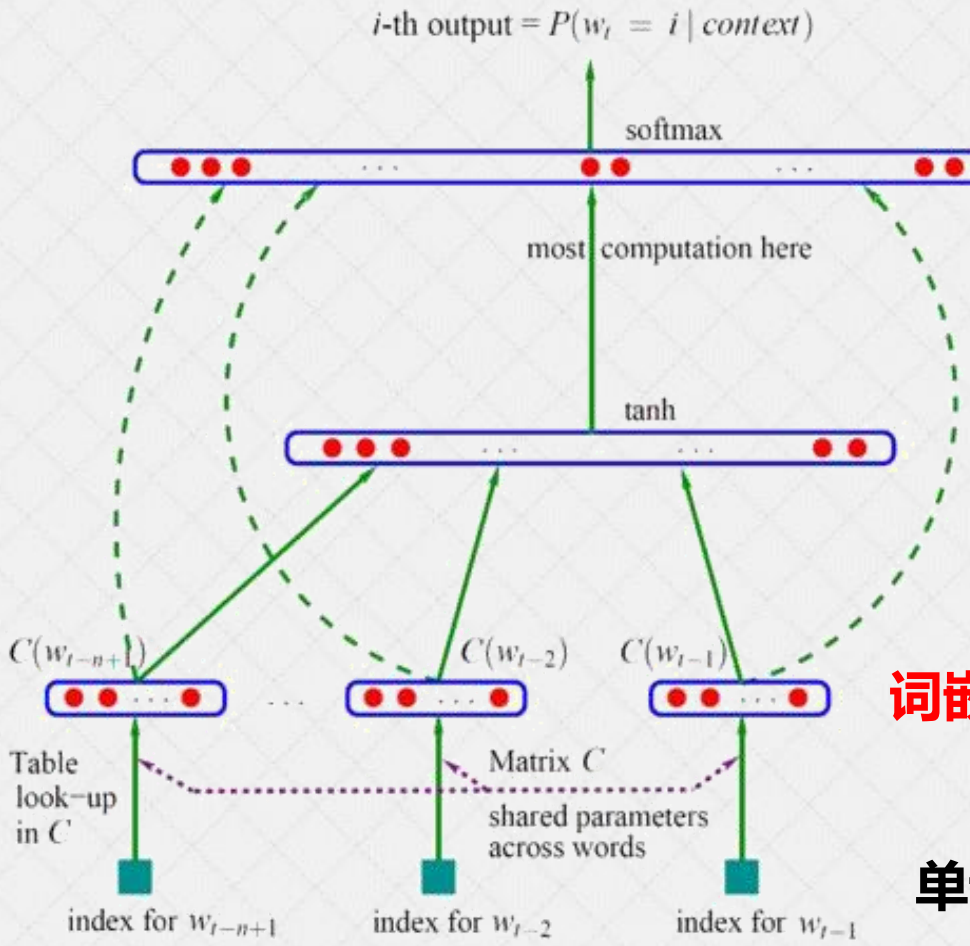
可能出现的问题?

N-gram语言模型总结

- 优点
  - 采用极大似然估计，参数易训练
  - 完全包含了前 n-1 个词的全部信息
  - 可解释性强，直观易理解
- 缺点
  - 缺乏长期依赖，只能建模到前 n-1 个词
  - 数据稀疏，且随着n的增大，参数量过大（平滑法、回退法）
  - 存在OOV的问题（专有词<UNK>、使用subword、哈希法）
  - 基于统计频次，泛化能力差

深度学习语言模型-NNLM

- Bengio 2003, A Neural Probabilistic Language Model



Softmax分类-计算量巨大

DNN深度神经网络

词嵌入-创新

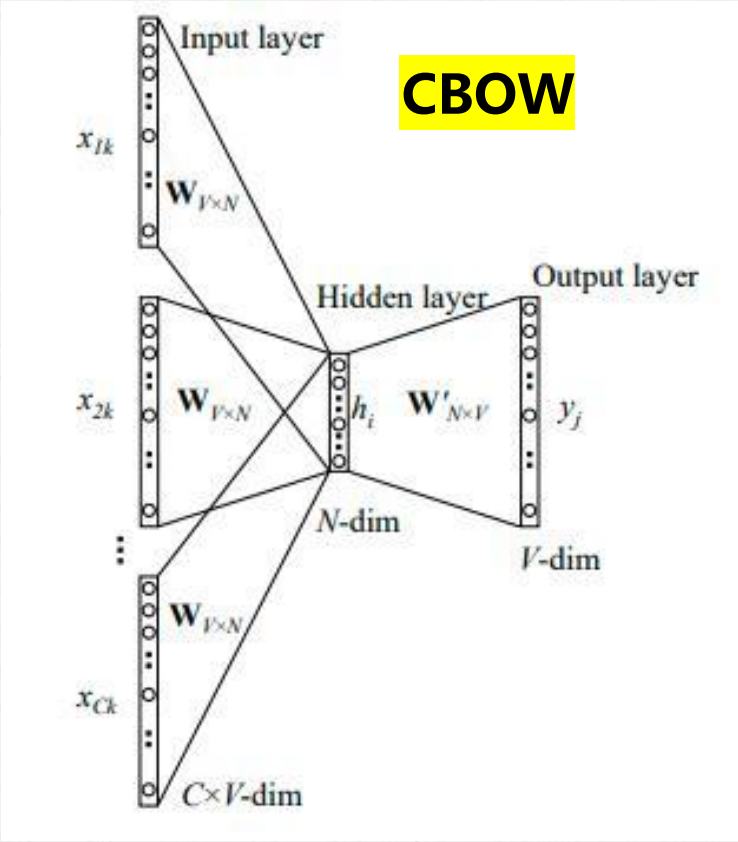
单词的顺序输入

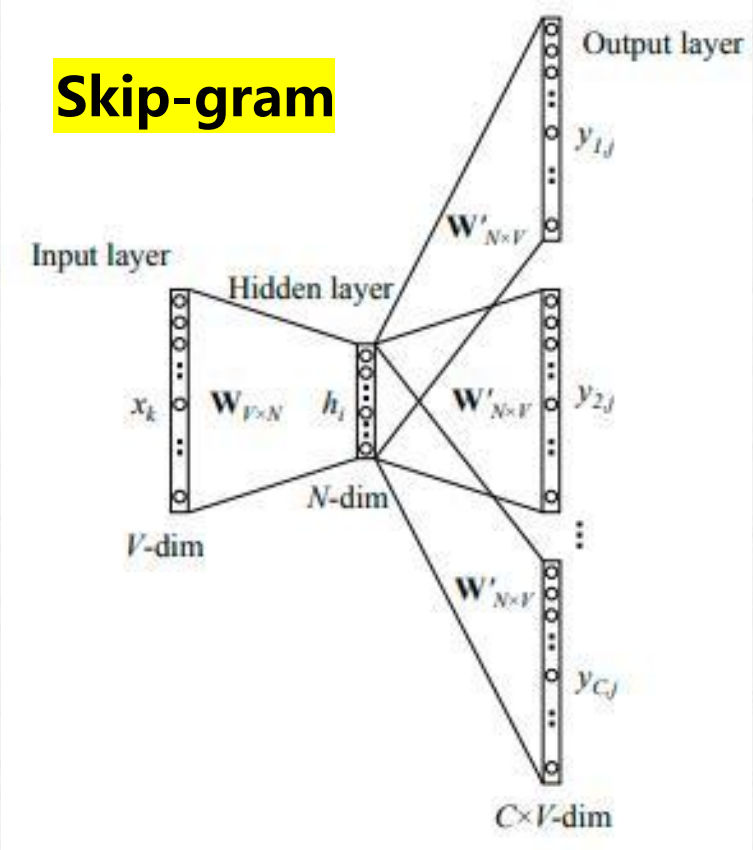
深度学习语言模型-NNLM

- 优点
  - 很好地解决了n-gram的稀疏性问题
  - 参数量比n-gram减少
  - 有一定的泛化性
- 缺点
  - 计算量巨大，在当时难以实现

深度学习语言模型-word2vec

- **Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space**



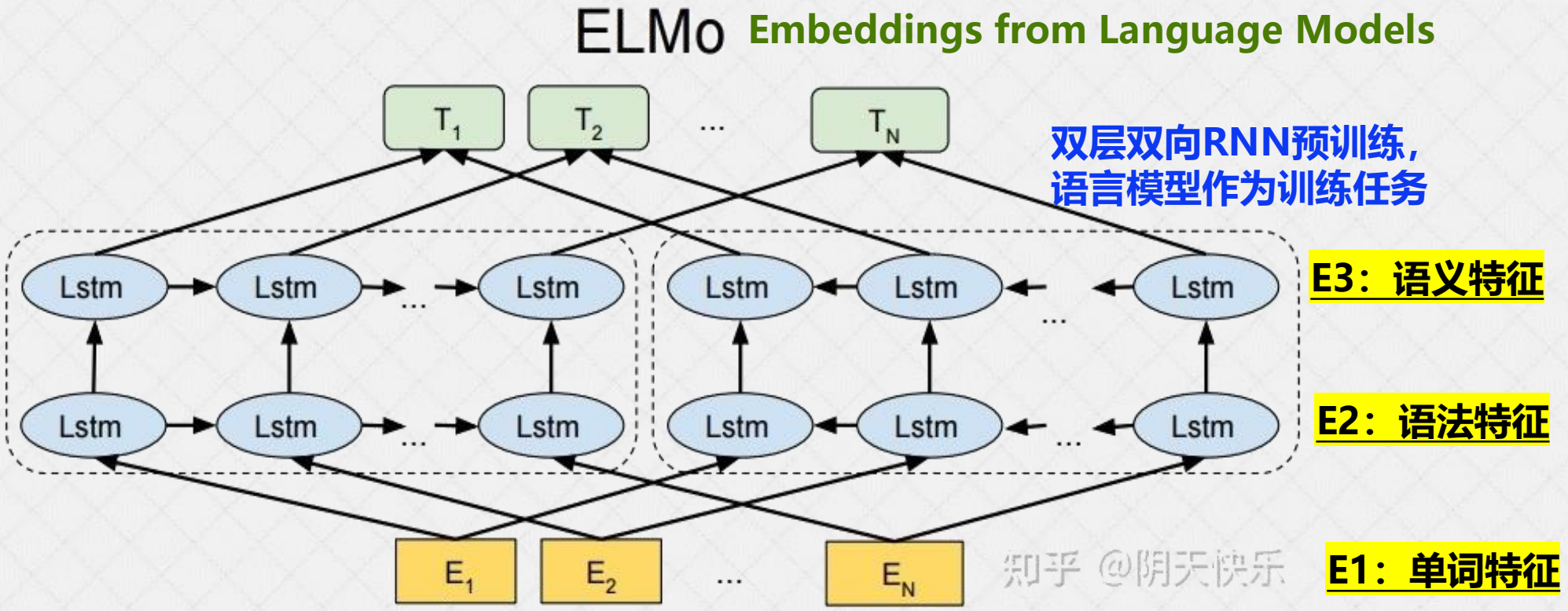


深度学习语言模型-word2vec

- 优点
  - 将语言模型融入到nlp任务
  - 解决了NNLM难以训练的问题
- 缺点
  - 从特征角度来看，无法解决单词的歧义问题，比如：apple，在不同场景有不同意思

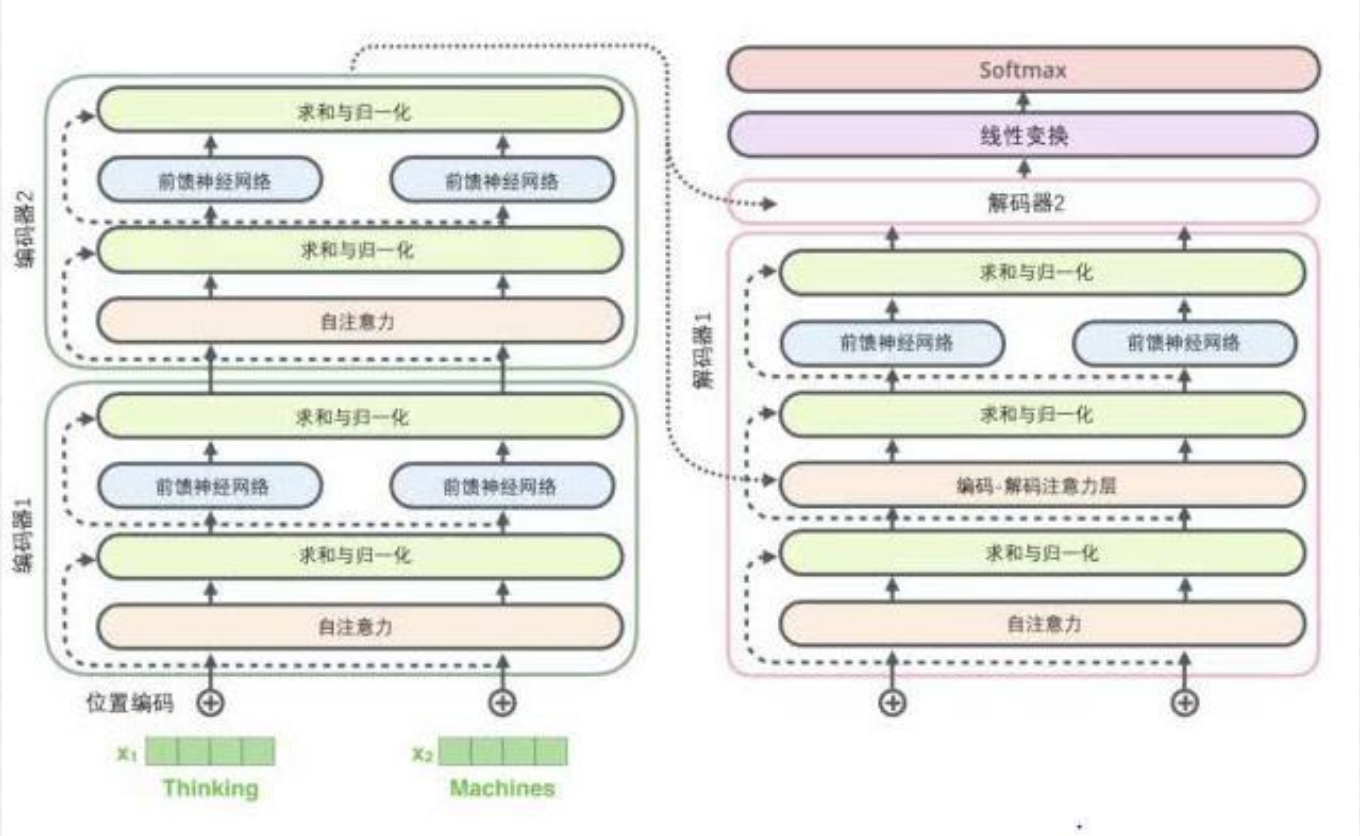
深度学习语言模型-ELMO

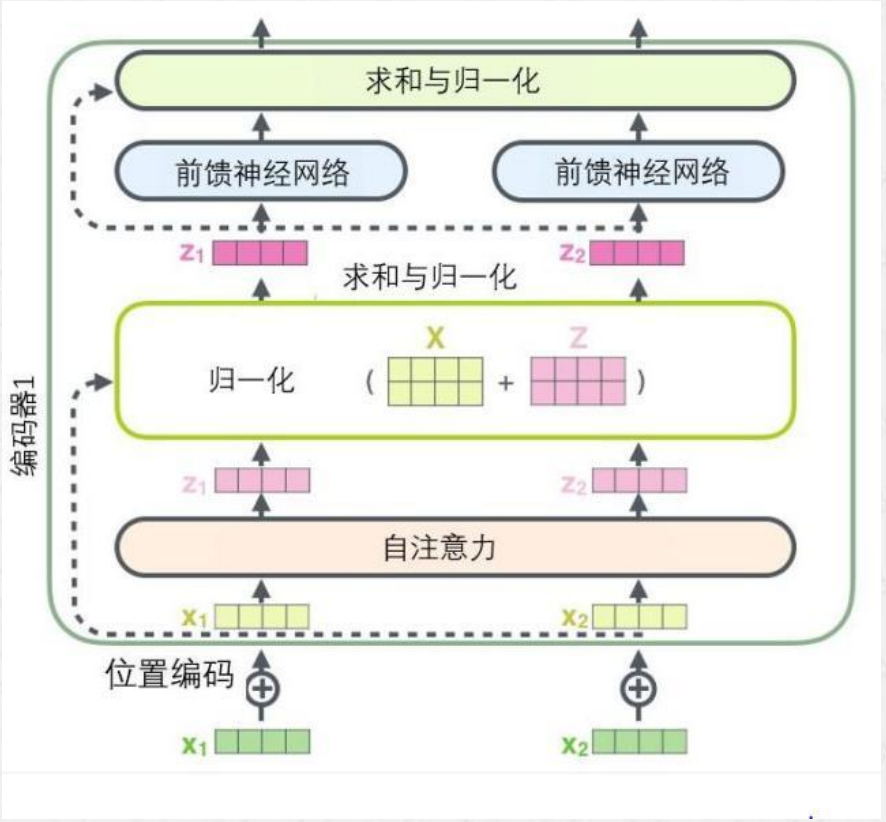
- ELMO Deep contextualized word representations，NAACL2018最佳论文



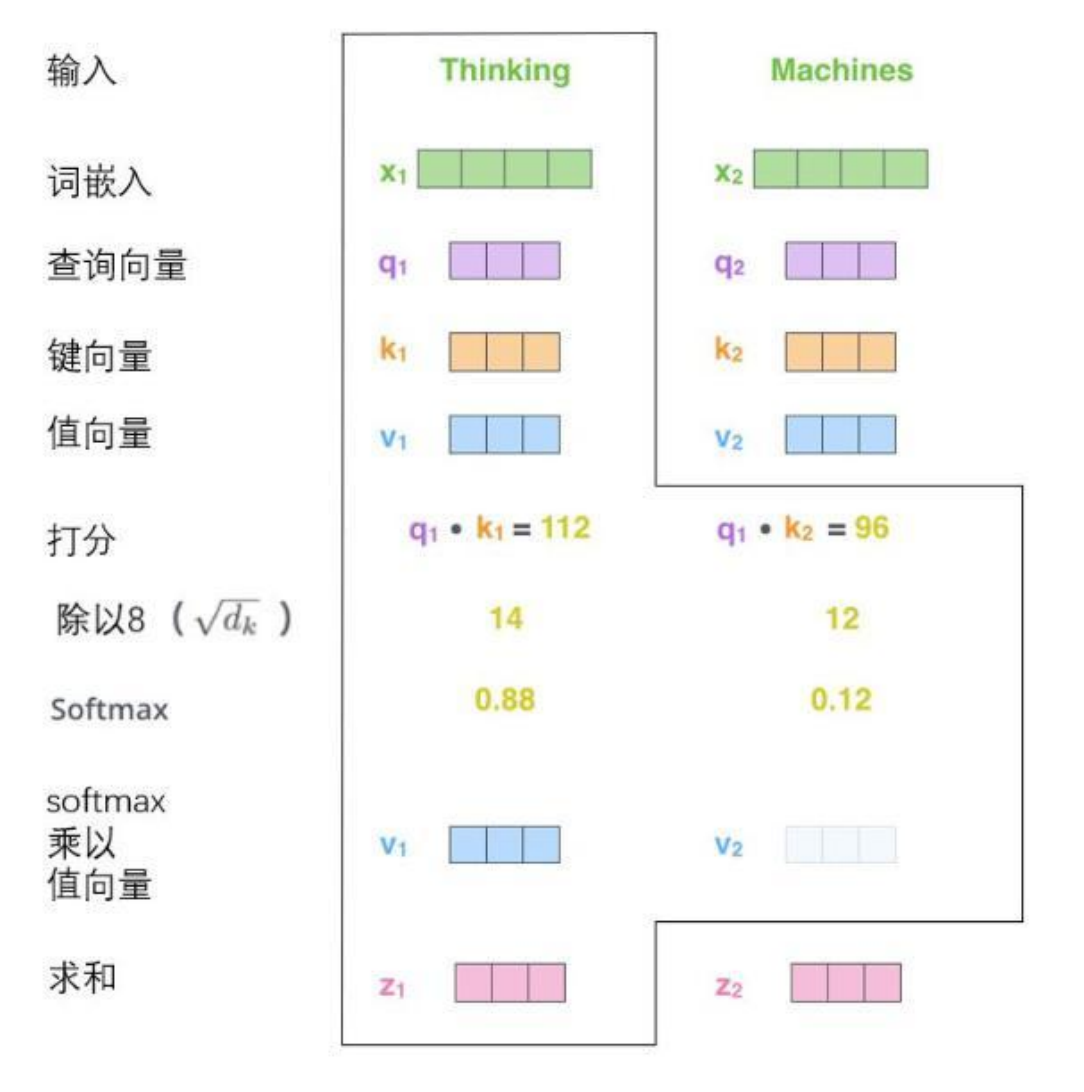
深度学习语言模型-Transformer

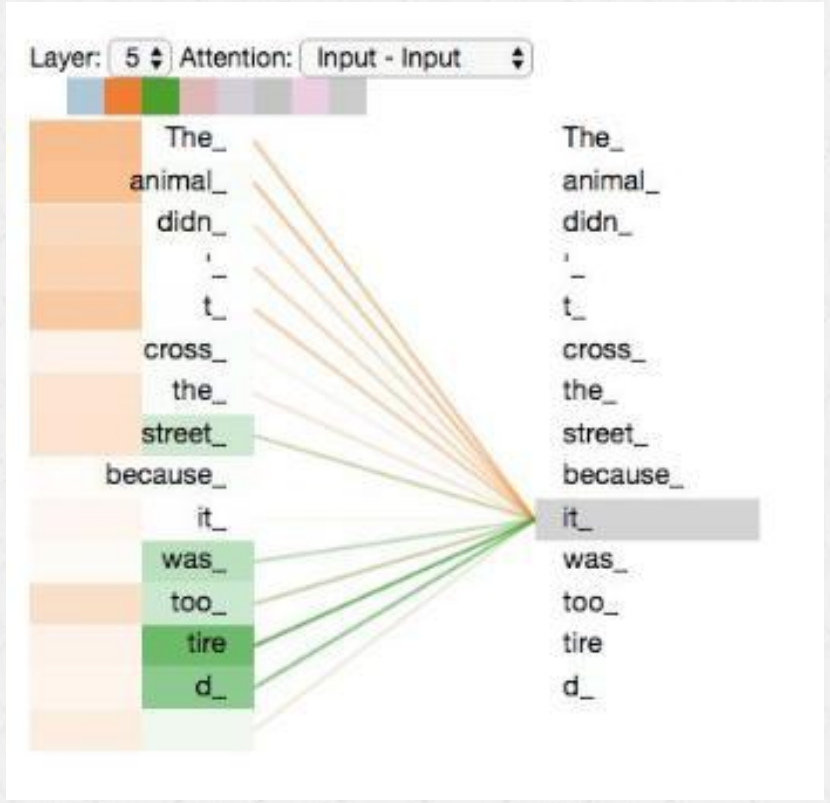
- Transformer Attention is all you need





深度学习语言模型-Transformer



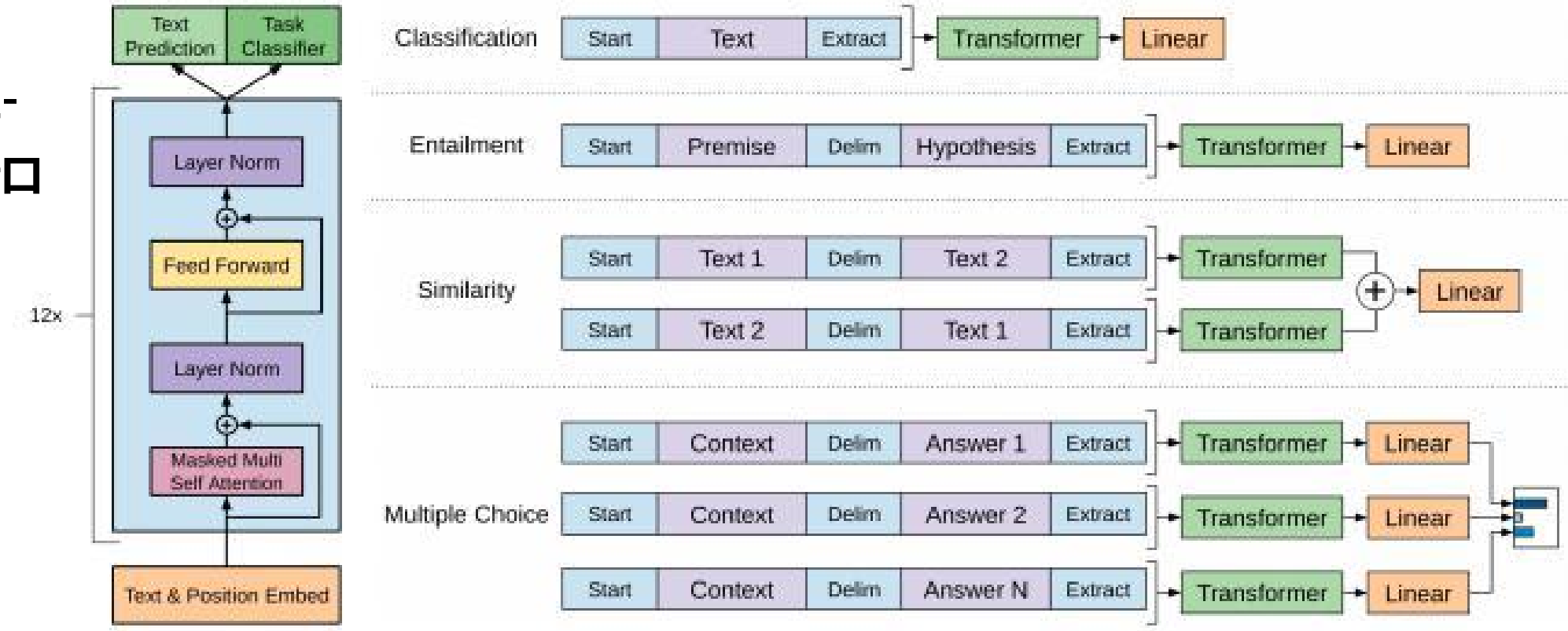


深度学习语言模型-GPT

- GPT Improving Language Understanding by Generative Pre- Training

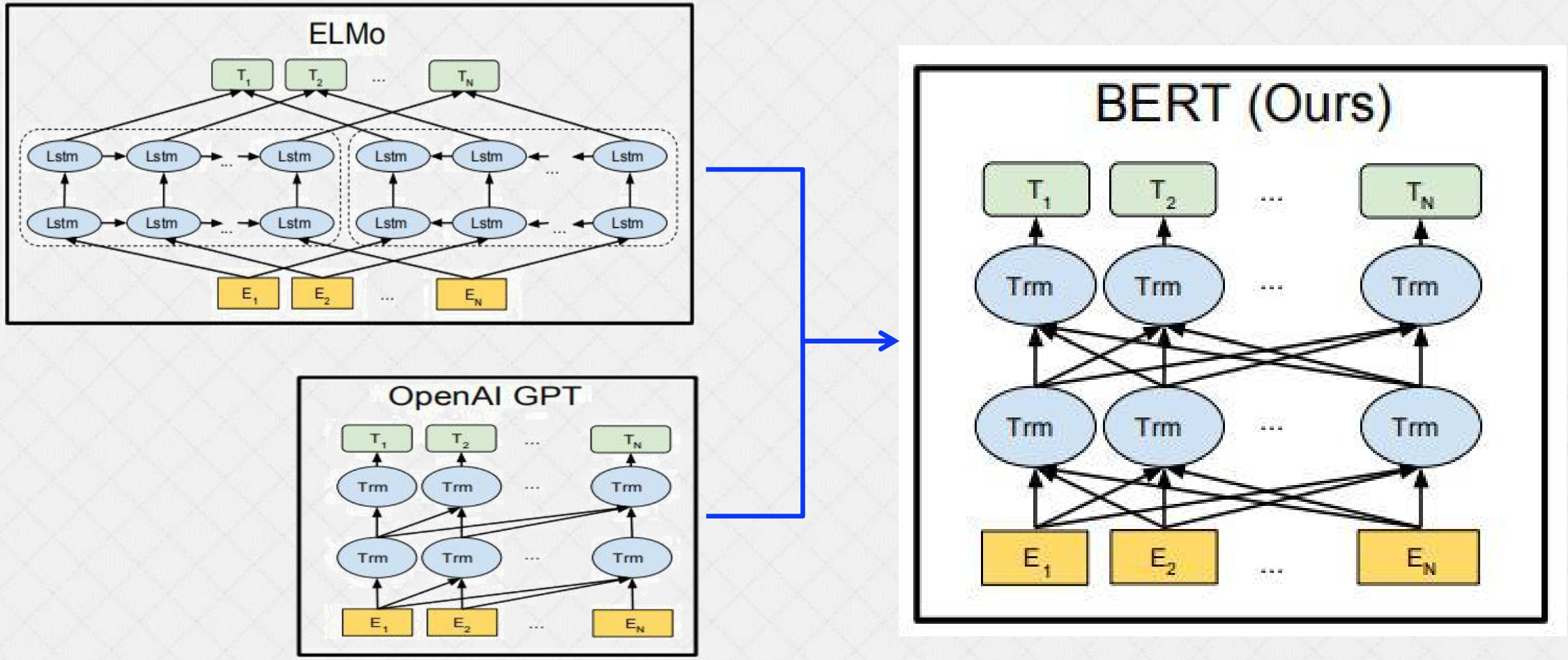
预训练模型-

多种任务接口



深度学习语言模型-Bert

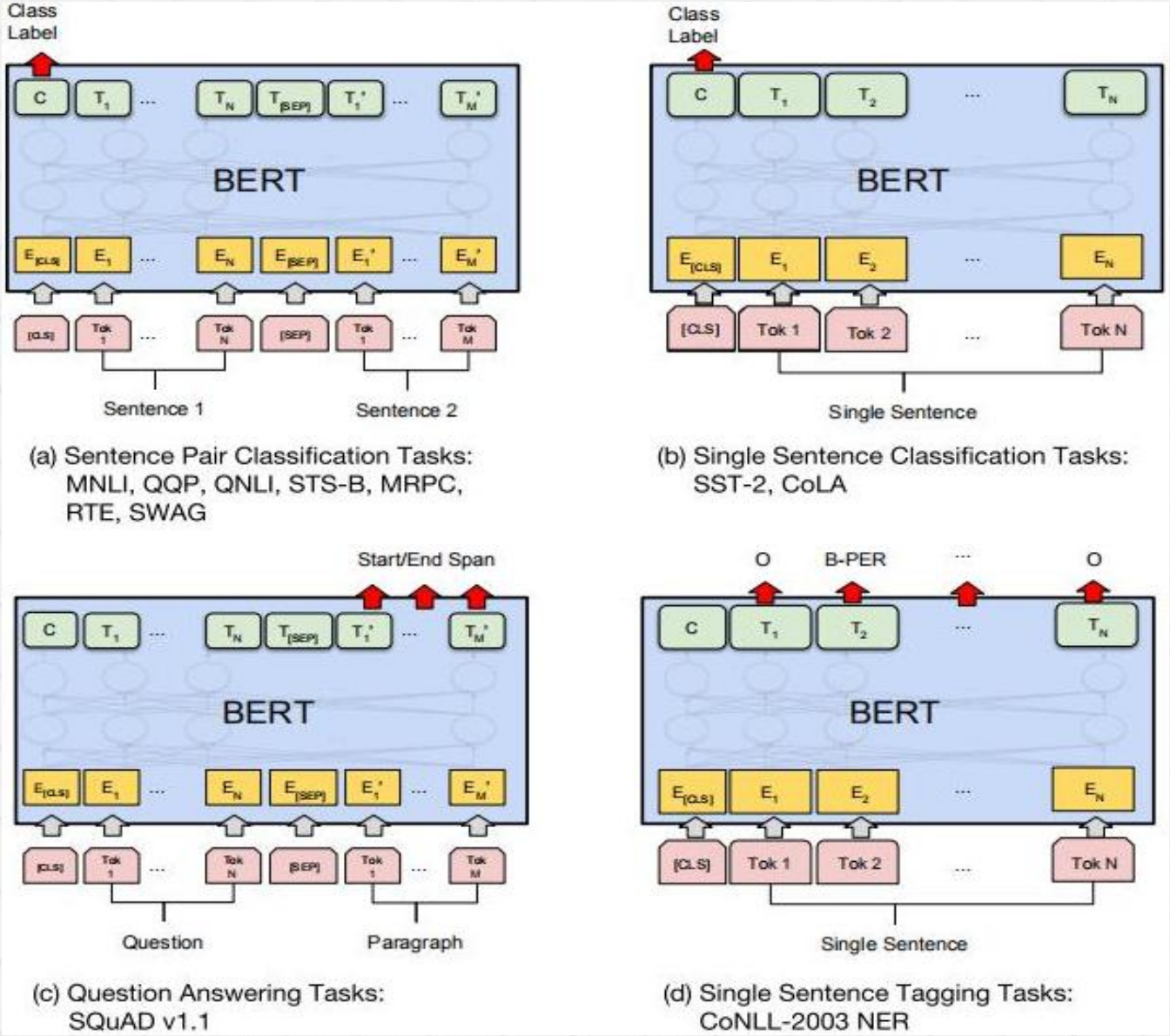
- Bert Pre-training of Deep Bidirectional Transformers for Language Understanding



语言模型

深度学习语言模型-Bert

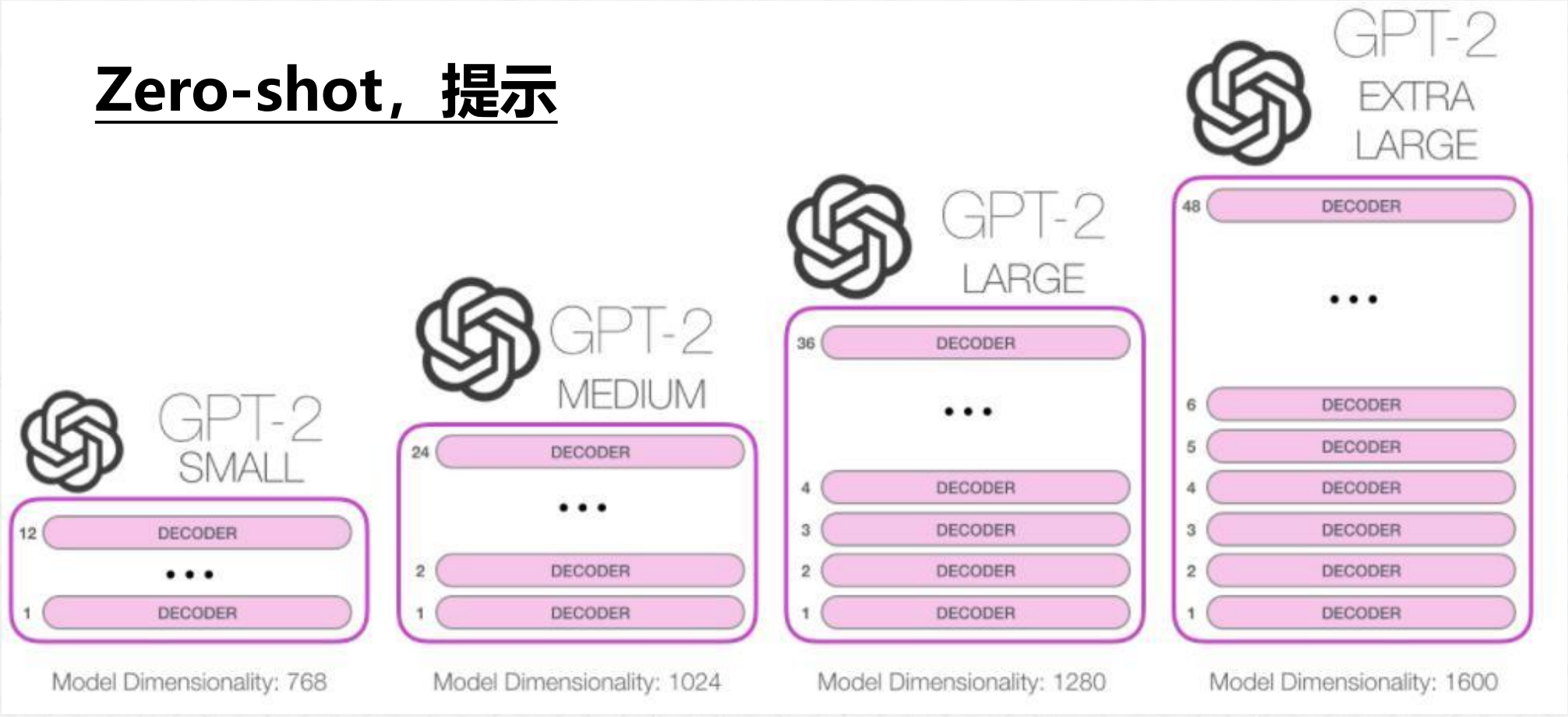
双向网络



深度学习语言模型-GPT2

- GPT2 Language Models are Unsupervised Multitask Learners

Zero-shot，提示



深度学习语言模型-GPT2

深度学习语言模型-GPT-3

- GPT-3 Language Models are Few-Shot Learners
  - 训练数据45T，参数1750亿，模型大小约700G
  - 论文长72页度，训练经费达千万美元
  - Zero-Shot, One-Shot, Few-Shot
  - 特定应用CodeX：GPT-3模型重新训练，使用GitHub数据

深度学习语言模型-GPT-3.5，GPT-4

- GPT-3.5，GPT-4 GPT-4 Technical Report；Training language models to follow instructions with human feedback
  - GPT-3.5引入RLHF(Reinforcement Learning with Human Feedback，基于人类反馈的强化学习)，解决了生成模型的一个核心问题：如何让人工智能模型的产出和人类的常识、认知、需求、价值观保持一致。
  - GPT-4是一个大型的多模态模型（可接收图像和文本输入，输出文本），虽然在许多现实世界的场景中能力不如人类，但在各种专业和学术基准上匹敌人类水平的表现。

深度学习语言模型-GPT-3.5

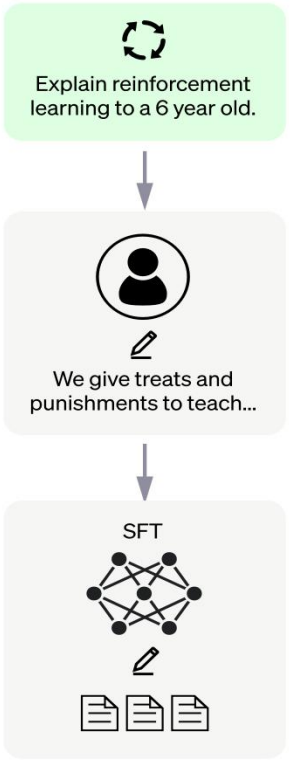
Step 1

Collect demonstration data and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3.5 with supervised learning.



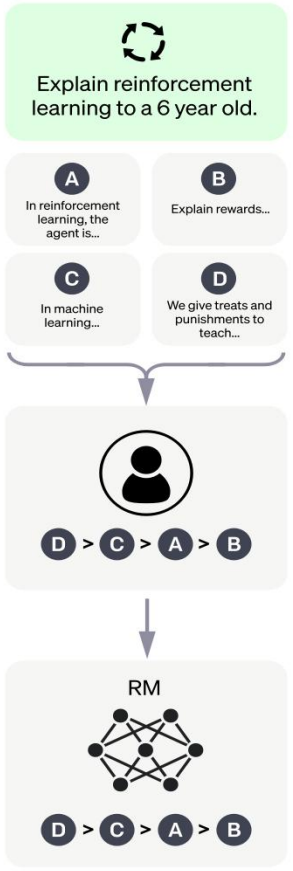
Step 2

Collect comparison data and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.



Step 3

Optimize a policy against the reward model using the PPO reinforcement learning algorithm.

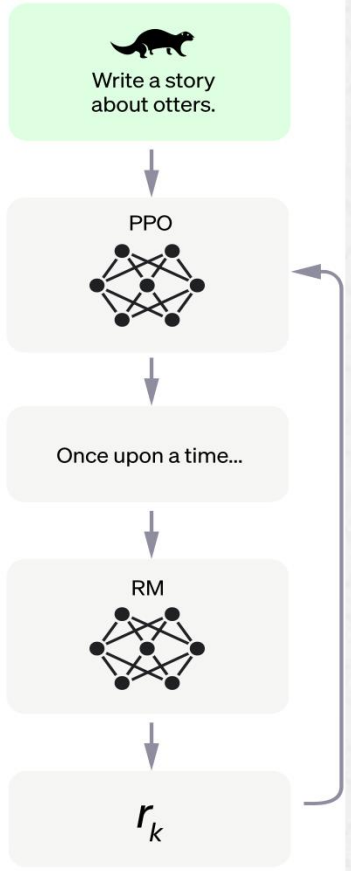
A new prompt is sampled from the dataset.

The PPO model is initialized from the supervised policy.

The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.



分词工具

中文分词

- 工具
  - jieba-结巴
  - 北京大学PKUseg
  - 清华大学THULAC
  - HanLP
  - 哈工大LTP
  - NLTK
  - 斯坦福分词器CoreNLP
- 语料库
  - 1998年《人民日报》语料库 PKU
  - 微软亚洲研究院语料库MSR
  - 香港城市大学提供的CITYU-繁体中文语料库

jieba中文分词

- 支持四种分词模式
  - 精确模式，试图将句子最精确地切开，适合文本分析；
  - 全模式，把句子中所有的可以成词的词语都扫描出来，速度非常快，但是不能解决歧义；
  - 搜索引擎模式，在精确模式的基础上，对长词再次切分，提高召回率，适合用于搜索引擎分词；
  - paddle模式，利用PaddlePaddle深度学习框架，训练序列标注（双向GRU）网络模型实现分词。同时支持词性标注。

PKUseg中文分词

- 一个多领域中文分词工具包
  - 多领域分词。不同于以往的通用中文分词工具，此工具包同时致力于为不同领域的数据提供个性化的预训练模型；
  - 更高的分词准确率；
  - 支持用户自训练模型；
  - 支持词性标注。

HanLP中文分词

- 面向生产环境的多语种自然语言处理工具包，目标是普及落地最前沿的NLP技术
  - 使用深度神经网络；
  - HanLP具备功能完善、性能高效、架构清晰、语料时新、可自定义的特点。

分词实现

- 规则分词实现-代码实例
- jieba分词工具使用-代码实例

课后实践

- 使用PKUseg、HanLP、jieba等工具进行分词，对比三者分词效果
- 研究开源工具源码-jieba、HanLP