

我们是怎么做的？



怎么将看到的图片表达出来？

我们需要学习和使用一门语言

认字识词

现实实体到抽象实体的映射，死记硬背

语法识别

将一个个独立的字词串联起来的规则

语义识别

结合语句上下文理解句子表达的含义

在一个阳光明媚的夏天，一个叫melonkid的小男孩，长着一头蓬松的头发，戴着一顶清凉帽，背着一个双肩包，怀里抱着一片西瓜，微笑着。

机器要怎么做？

Traditional Language Model

$$P(w_1, w_2, \dots, w_n) = P(w_1)P(w_2|w_1)P(w_3|w_1, w_2)\dots P(w_n|w_1, w_2, \dots, w_{n-1}) = \prod_i P(w_i|w_1, w_2, \dots, w_{i-1})$$

$$P(w_{next} | \text{圆圆, 爱, 吃}) = \frac{\text{count}(w_{next}, \text{圆圆, 爱, 吃})}{\text{count}(\text{圆圆, 爱, 吃})}$$

当需要预测上下文圆圆爱吃 ____ 后面跟的词时，如果词典中所有单词为圆圆，爱，吃，瓜，梨时

$$P(\text{瓜} | \text{圆圆, 爱, 吃}) > P(\text{梨} | \text{圆圆, 爱, 吃})$$

通过计算每个词在句子中出现的概率，判断词与句子的关系

- 优点：比较简单直观简单，容易理解
- 缺点：依赖整个句子作为上下文，处理长文本时比较耗时。并且，由于模型只是基于历史语料进行概率计算。预测精度完全依赖训练模型的语料库大小。并且，如果需要预测的词在词典中没有出现过时，就无法正确预测

为了解决当词典V较大，模型训练效率低的情况时，人们引入马尔科夫链对上诉算法进行优化，也就是所谓的N-Gram算法简单来说，就是在计算第i个词的概率时，不再参考整个语句只参考当前次前k个词即可。