1. What is NLP?

NLP = NLU + NLG

NLU: 语音音/文文本 -> 意思(meaning)

NLG: 意思-> 文文本/语音音

2. The Challenge

Multiple Ways to Express (多种表达方方式)

Ambiguity(一词多义)

→ how to solve ambiguity:

从单词到语境(context) interest --- a financial interest in IBM

更新认知

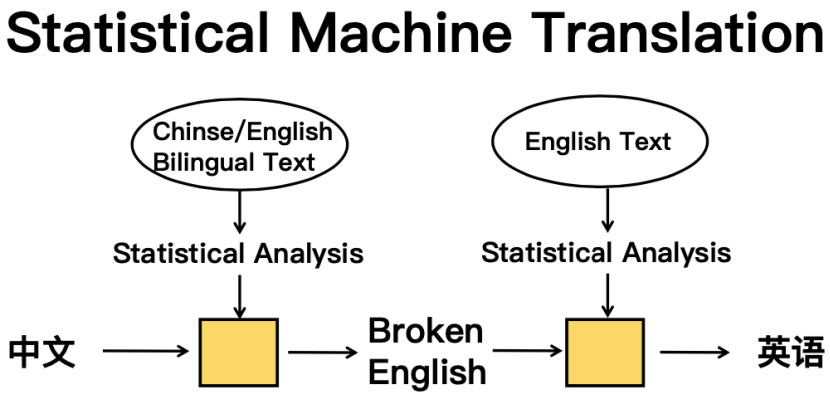
3. Case Study: Machine Translation

请翻译这句话: farok crrrok hihok yorok clok kantok ok-yurp

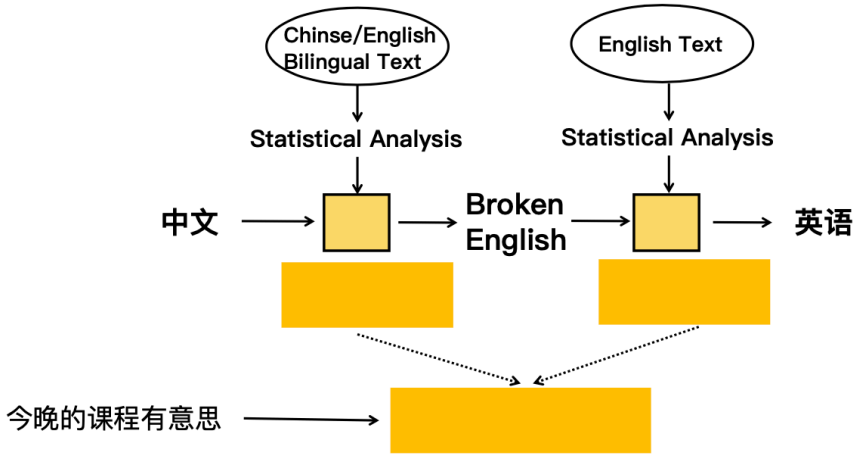
1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

在语料库中进行统计，找单词配对。

缺点：慢、语义语境、上下文考虑不足、语法错误



今晚的课程有意思---tonight of course interesting (broken English)
排列组合一共有 4！中句子
Language Model：求解每一种排列组合的概率---最高的来寻找合适的句子



Translation model → language model→decoding algorithm

4. Statistical MT: Three Problems

- **语言模型（Language Model）**
  - 给定一句英文e, 计算概率（e)
  - 如果是符合英文语法的，p(e)会高
  - 如果是随机语句，p(e)会低
- **翻译模型**
  - 给定一对<c,e>，计算p(f|e)
  - 语义相似度高, 则p(f|e)高
  - 语义相似度低，则p(e|f)低
- **Decoding Algorithm**
  - 给定语言模型，翻译模型和f, 找出最优的使得p(e)p(f|e)最大

语言模型是需要提前训练出来的。
翻译模型，可以起到一个词典的作用。

Decoding = 语言+翻译

4.1 Language Model (语言模型)

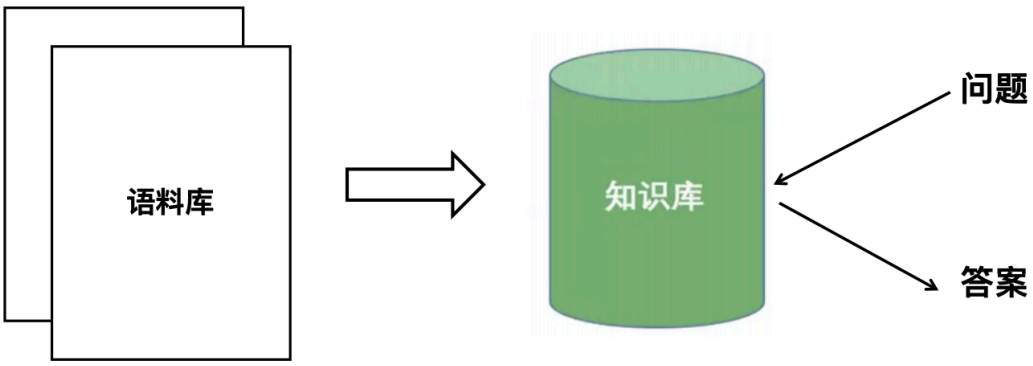
- **对于一个好的语言模型：**
  - p(He is studying AI) > p(He studying AI is)
  - p(nlp is an interesting course) > p(interesting course nlp is an)

- 怎么计算p(.)
  - p(He is studying AI) = p(He)p(is)p(studying)p(AI)
  - P(He is studying AI) = p(He)p(is | He)p(studying | is)p(AI | studying)
  - P(He is studying AI) = p(He)p(is | He)p(studying | he is)
    p(AI | is studying)

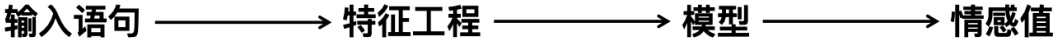
Unigram, bigram, trigram, ---, N-gram 在考虑一个单词的时候，需要参考前面多少个单词
核心是不知道每一个 p

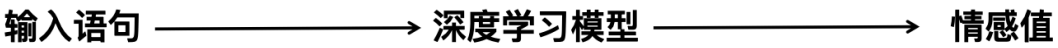
5. NLP 的经典应用场景

Question Answering(问答系统)



Sentiment Analysis(情感分析)





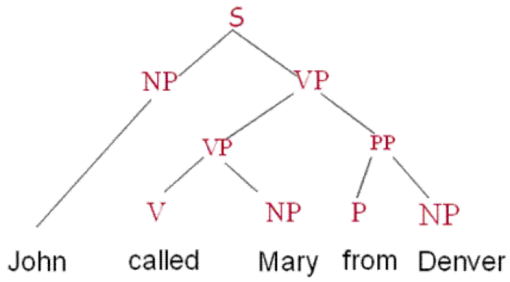
Machine Translation(机器翻译)

Text Summarization(自动摘要)
Chatbot (聊天机器人)
Information Extraction(信息抽取)

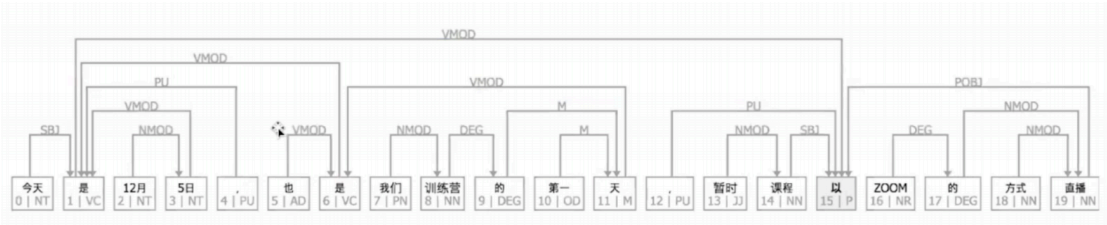
6. NLP 关键技术

四大维度：
Semantic 语义（NLU、机器翻译）
syntax 句子结构（句法分析、依存分析
morphology 单词（分词、pos 词性、NER）
phonetics 声音

word segmentation 分词
part of speech 词性分析，一个单词在不同句子里扮演不同角色
named entity recognition 命名实体识别，抽取关键信息---知识图谱 问答系统
parsing 句法分析



Dependency parsing 依存分析



Relation extraction 关系抽取