

# NLP Bootcamp (1)

2019年01月22日

# Today's Agenda

- 课程介绍
  - 课程安排
  - 课程团队
  - 评分标准
  - 沟通机制
  - 课程结构
  - Capstone项目
  - 技术文章编写
- 自然语言处理介绍
  - 什么是NLP?
  - NLP的应用场景
  - NLP领域关键技术
- 算法复杂度介绍
  - 简单的几个例子

# Course Info

- 直播+录播
- 周期：18周
- 老师答疑时间：每天固定时间（本周三开始）
- 一周一次Review Session（针对于某一个话题）
- 课程材料/视频利用gitlab维护
- 建议每周学习时间：15个小时以上

# Requirements

- 核心课程项目: 40%
- 聊天机器人项目: 15%
- Capstone项目: 25%
- 技术文章 (4篇) : 每篇1–3%
- 理论作业: 10%
- Bonus (up to 5%): Best Capstone项目, Best技术文章

总分 $\geq$ 75%, 即可以获得学位和招聘推荐服务

# How to Better Communicate?

- 定期群内答疑
- 有任何课程上的疑惑/帮助, 请第一时间联系班主任
- 如有需要, 可定期组织线下活动
- 遇到技术问题时:
  - 首先尝试自己去解决, 解决不了的来询问助教/老师
- 直播课程: 多互动

# Bootcamp Objective

- ✓• 培养AI思维、建模能力
- ✓• 扎实的AI基础，不做调参侠
- ✓• 解决问题能力
- ✓• 通过项目深入理解NLP核心技术

# How to succeed in Bootcamp? ↴

## Check List:

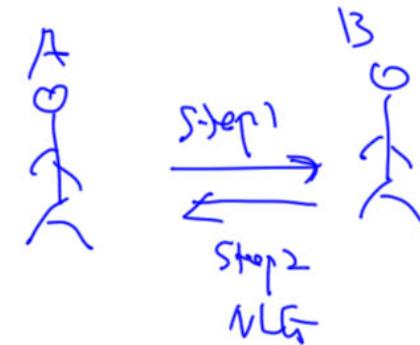
- Learn : 按时参加直播 + 观看录播视频
- Read : 养成读文章（英文）的习惯，接触课程延伸的内容
- Code : 完成项目，一定要自己完成，写过1万行AI程序保证有本质提升
- Write : 养成写文章习惯，梳理思路，自我总结
- Discuss : 群里多交流，不怕提出低级问题
- Collaboration : 鼓励项目合作，这些人很有可能成为AI路上的好朋友  
*开放式*

# AI工程师必备的核心技能



# Introduction : the Start of Journey

# What is NLP?



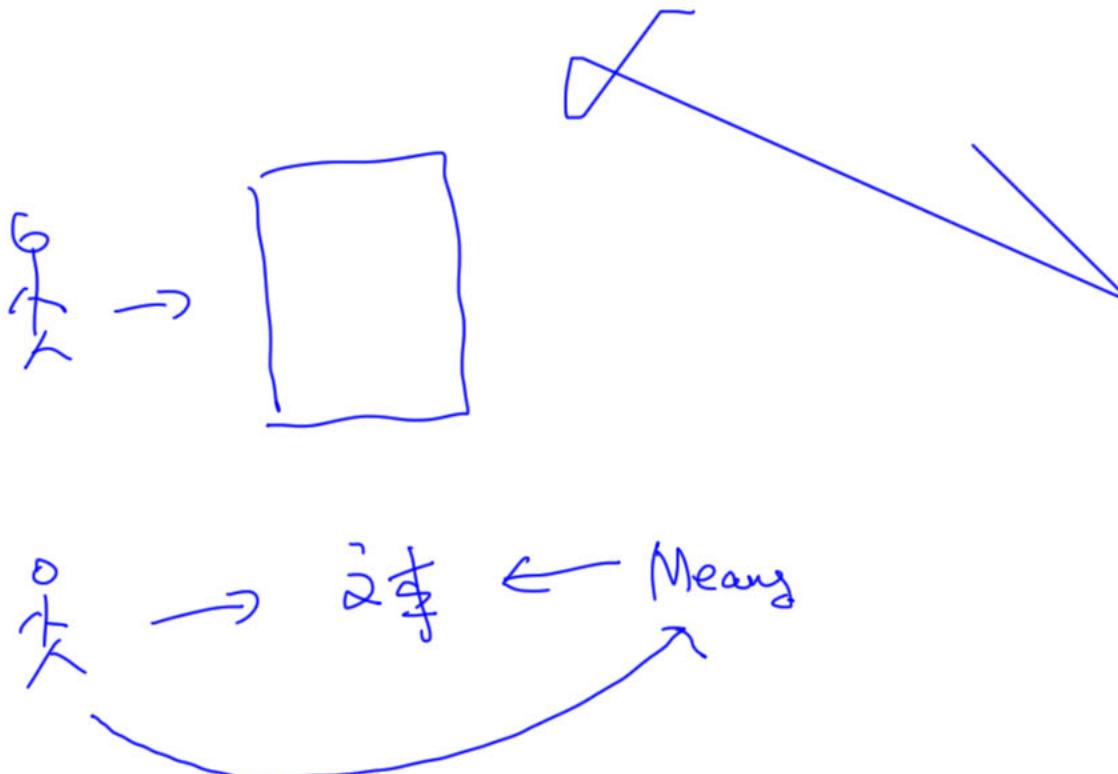
NLP = NLU + NLG

(Natural Language Understanding)

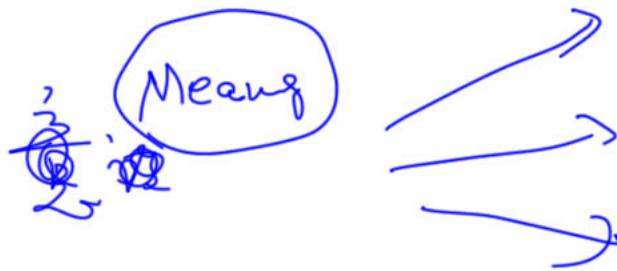
- NLU: 语音/文本 -> 意思 (meaning)
- NLG: 意思-> 文本/语音

(Natural Language Generation)

# Why NLP is Harder (i.e. than Computer Vision)



# The Challenge : Multiple Ways to Express (多种表达方式)

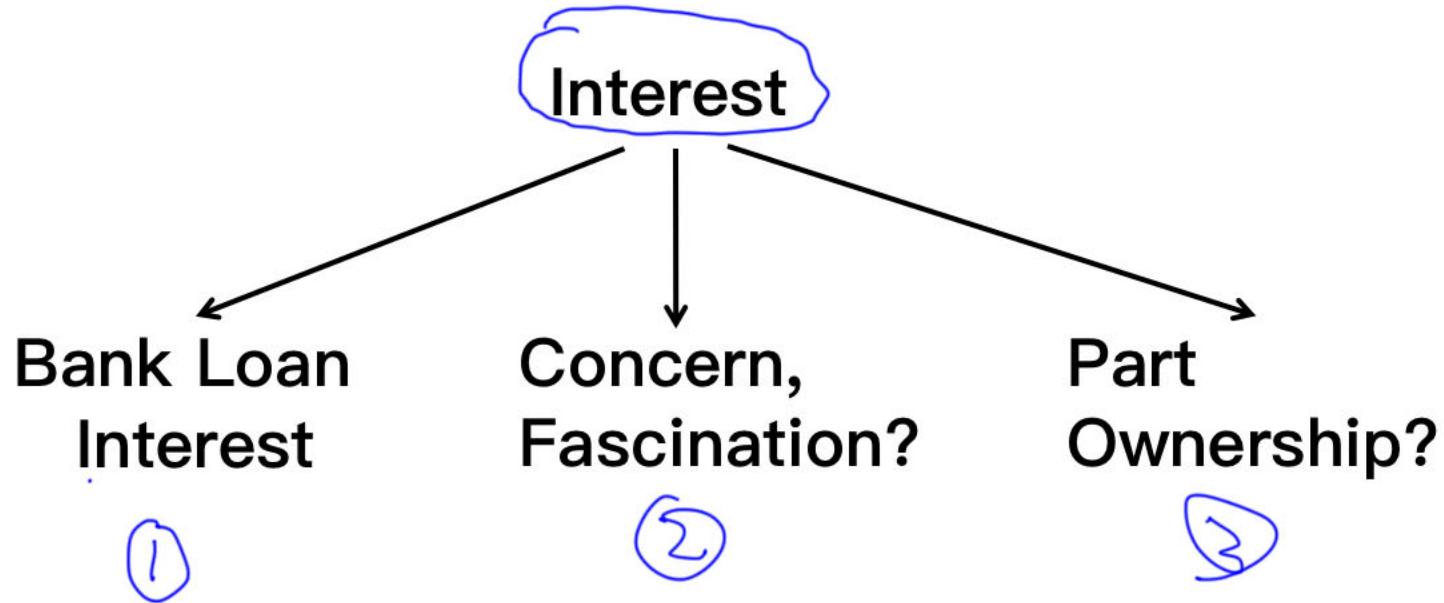


- 贪心科技开设了NLP训练营
- 贪心科技新出了NLP训练营
- 新出的NLP训练营是贪心科技出的
- .....

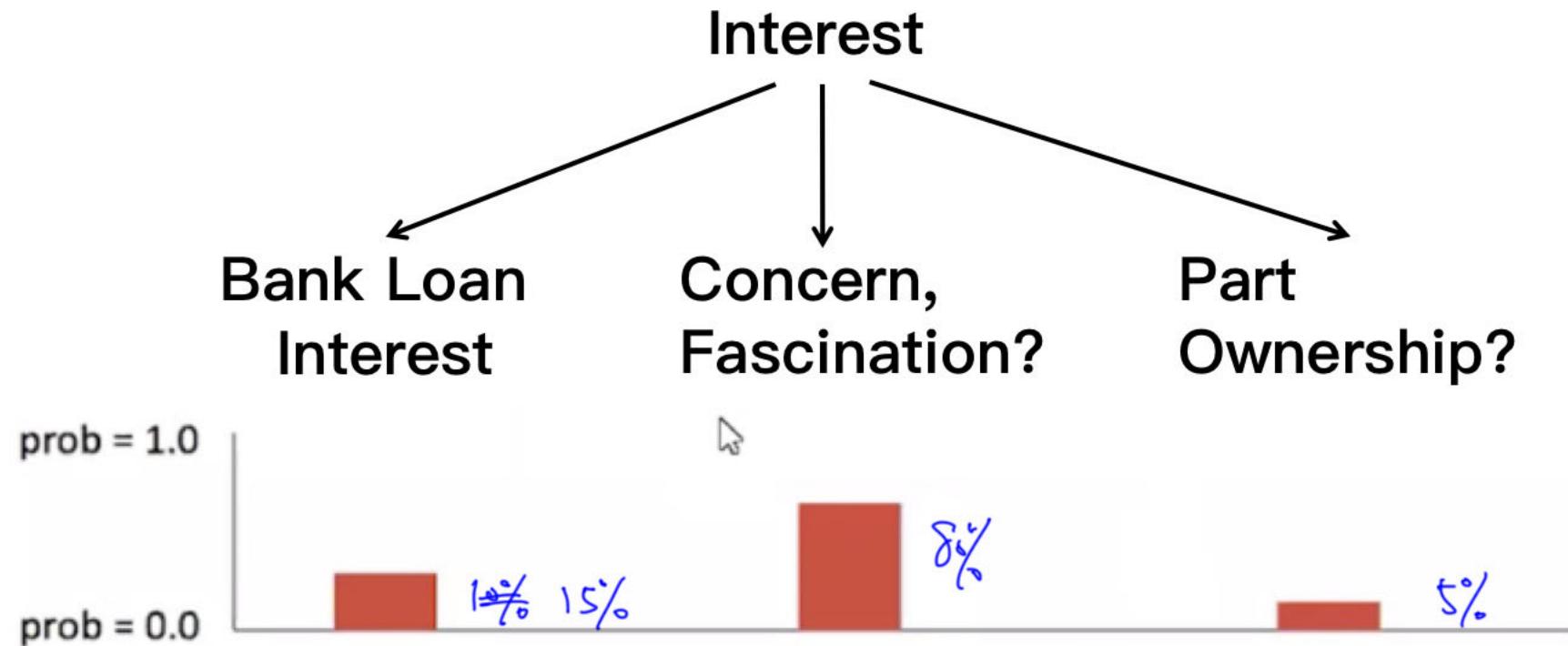
# The Challenge : Ambiguity (一词多义)

- 今天参观了苹果公司
  - 现在正好是苹果季节
- +  
Fruit

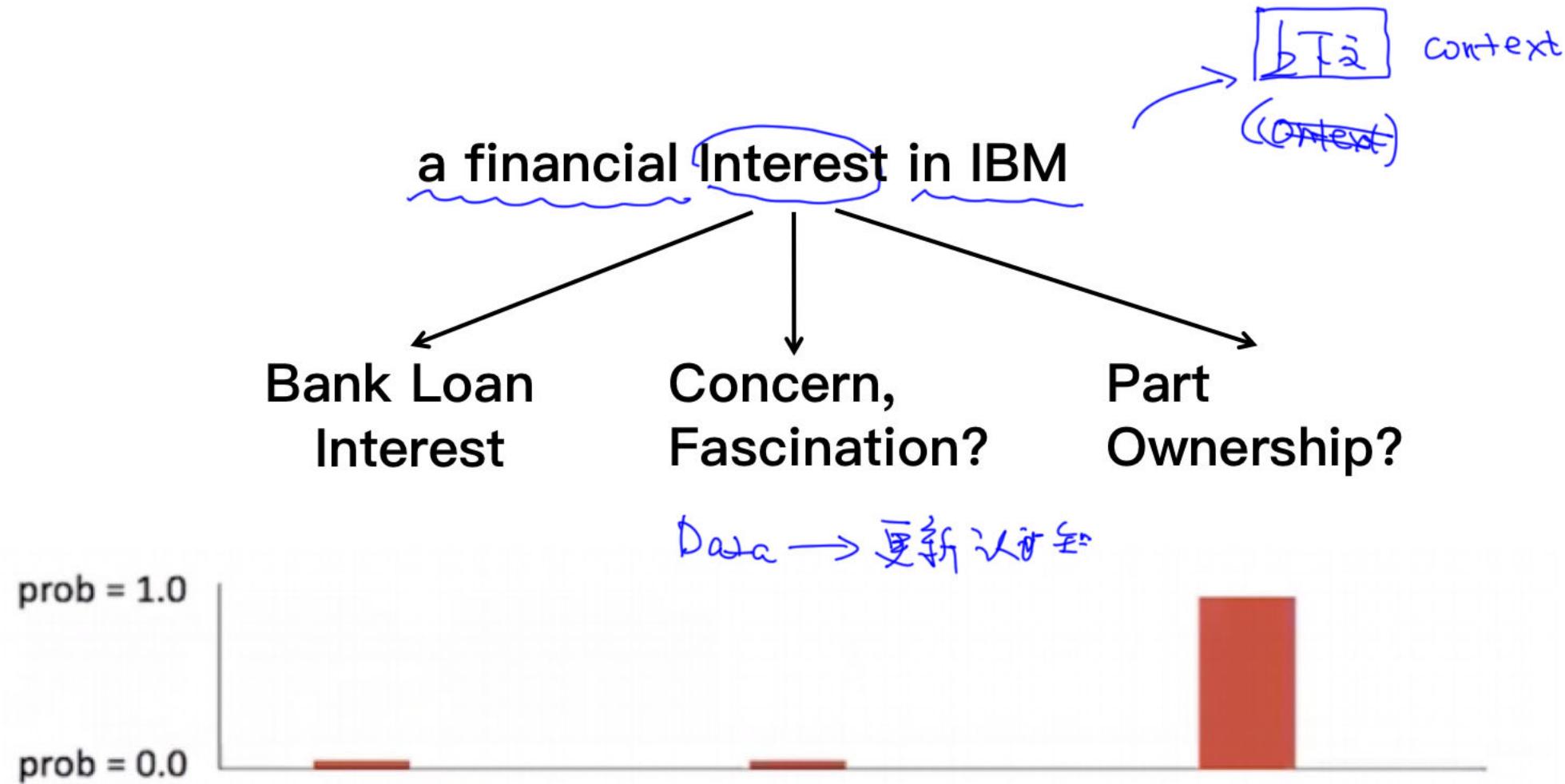
# How to Solve Ambiguity?



# Solving Ambiguity: Learning from Data



# Solving Ambiguity: Learning from Data



# Today's Case Study: Machine Translation



百度wifi翻译机 热 人工翻译 下载翻译插件 下载翻译app 登录

检测到中文



英语

翻 译

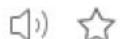
人工翻译



贪心科技旗下的贪心学院通过人工智能技术，打造最懂用户的自适应学习平台，为每一位用户提供个性化的职业教育服务。我们追求最精炼的AI教育内容和个人量身定制的课堂。|

X

Greedy College of Greedy Science and Technology, through artificial intelligence technology, builds the self-adaptive learning platform that knows users best, and provides personalized vocational education services for every user. We pursue the most refined AI education content and individual tailored classroom.



报错

双语对照



如果现在让你写一个机器翻译系统，怎么实现？

请翻译这句话: [farok crrrok hihok yorok clok kantok ok-yurp] → 甲乙

jjat

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

# Today's Case Study: Machine Translation

问题：

- 模 (AI/模型)

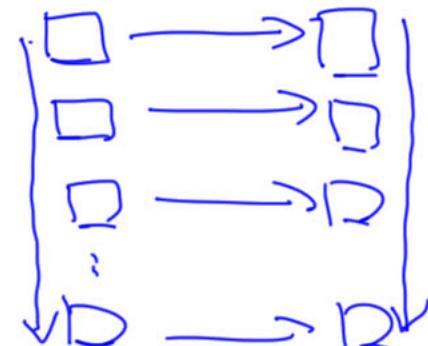
- 语义

- 上下文

- 语法不对

- 规则统计

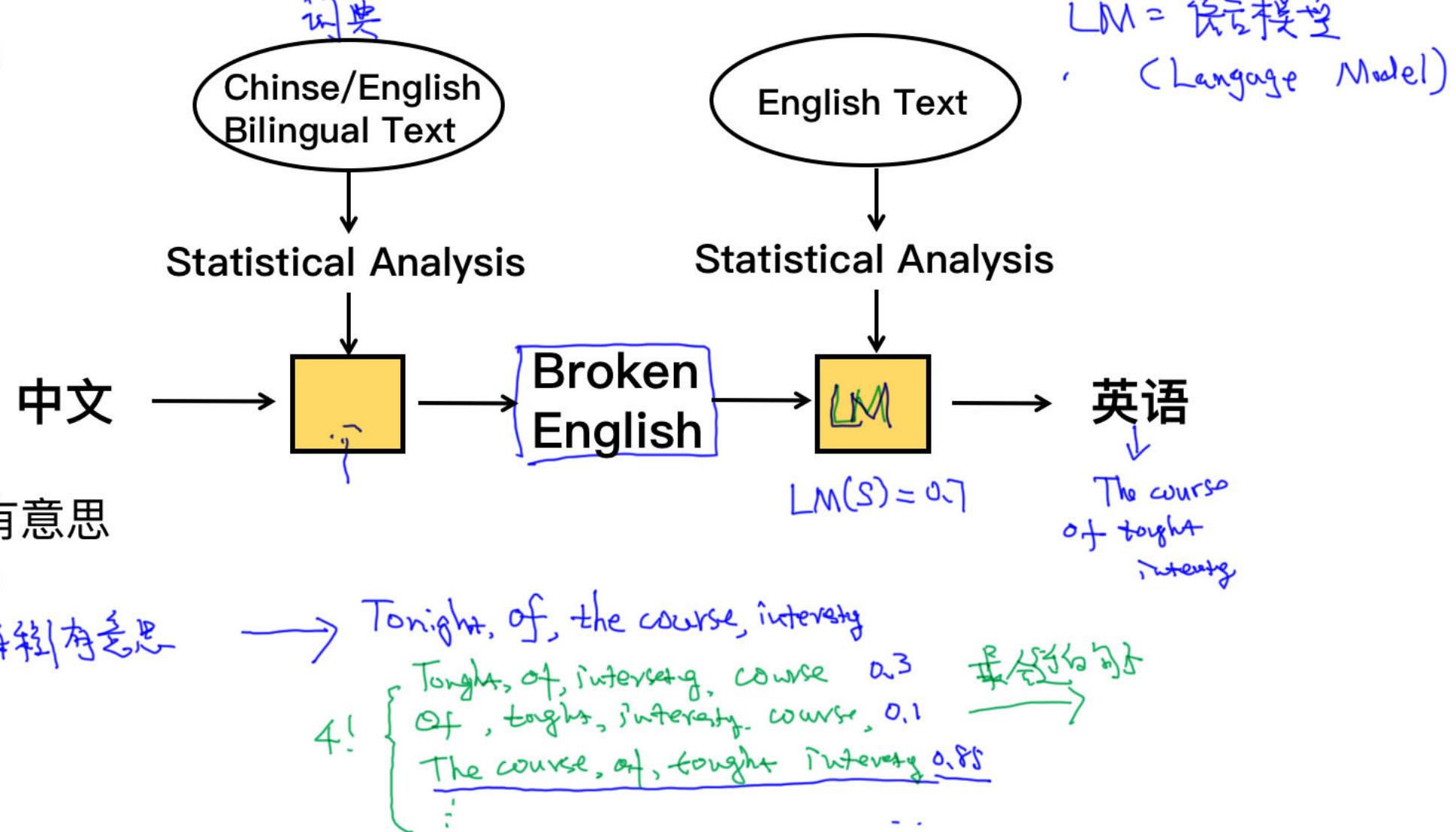
在前一张PPT里所实现的方法有什么缺点？



# Statistical Machine Translation

一语法 (语义)

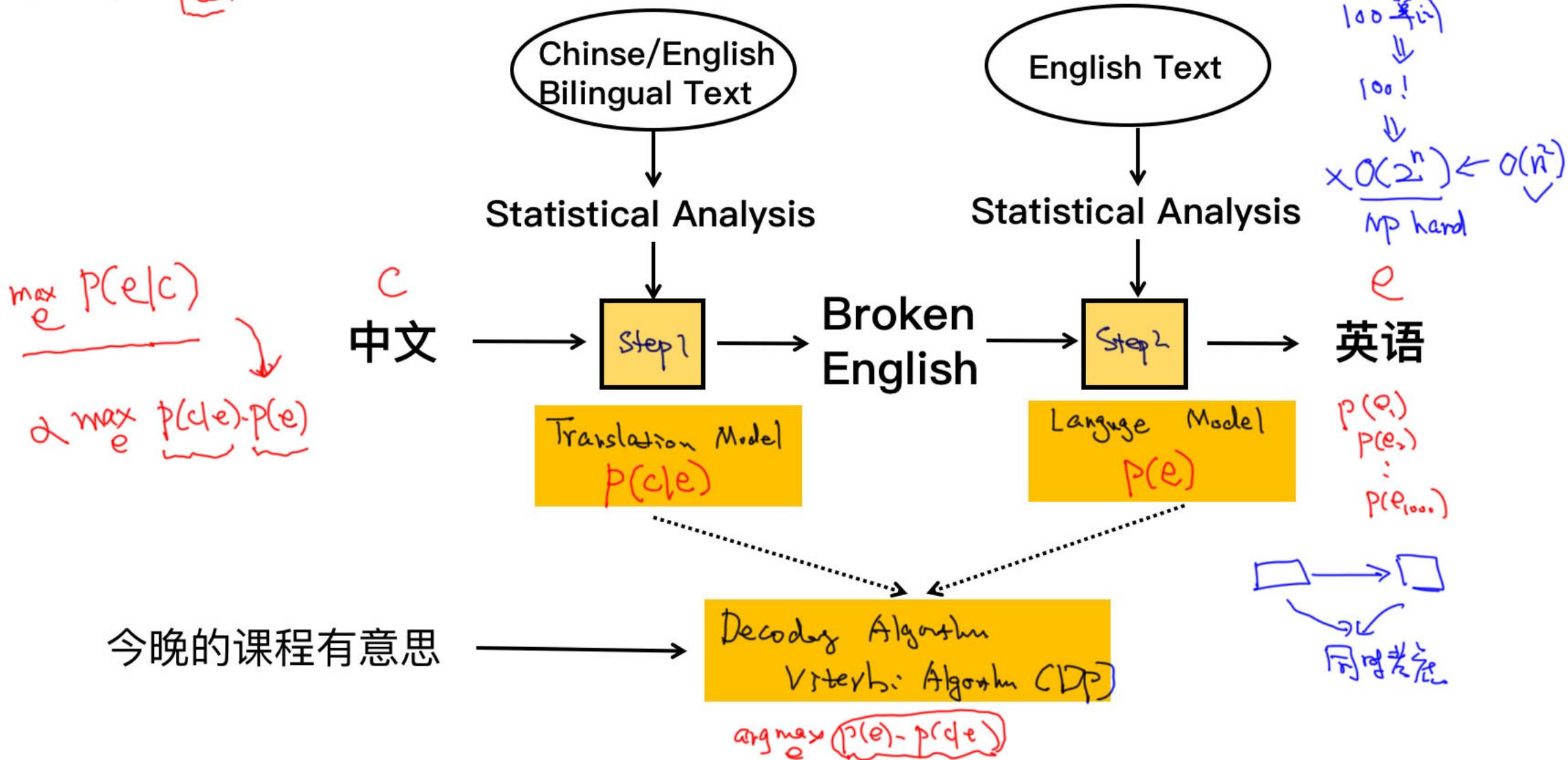
100个单词  
↓  
100句



贝叶斯定理

$$P(e|c) = \frac{P(c|e) \cdot P(e)}{P(c)}$$

# Statistical Machine Translation



# Statistical MT: Three Problems

## • 语言模型 (Language Model) ← 人话?

- 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 ← (ie. 词典)

- 给定一对 $\langle c, e \rangle$ , 计算 $p(f|e)$
- 语义相似度高, 则 $p(f|e)$ 高
- 语义相似度低, 则 $p(f|e)$ 低

$$D(n) \rightarrow O(n^p)$$

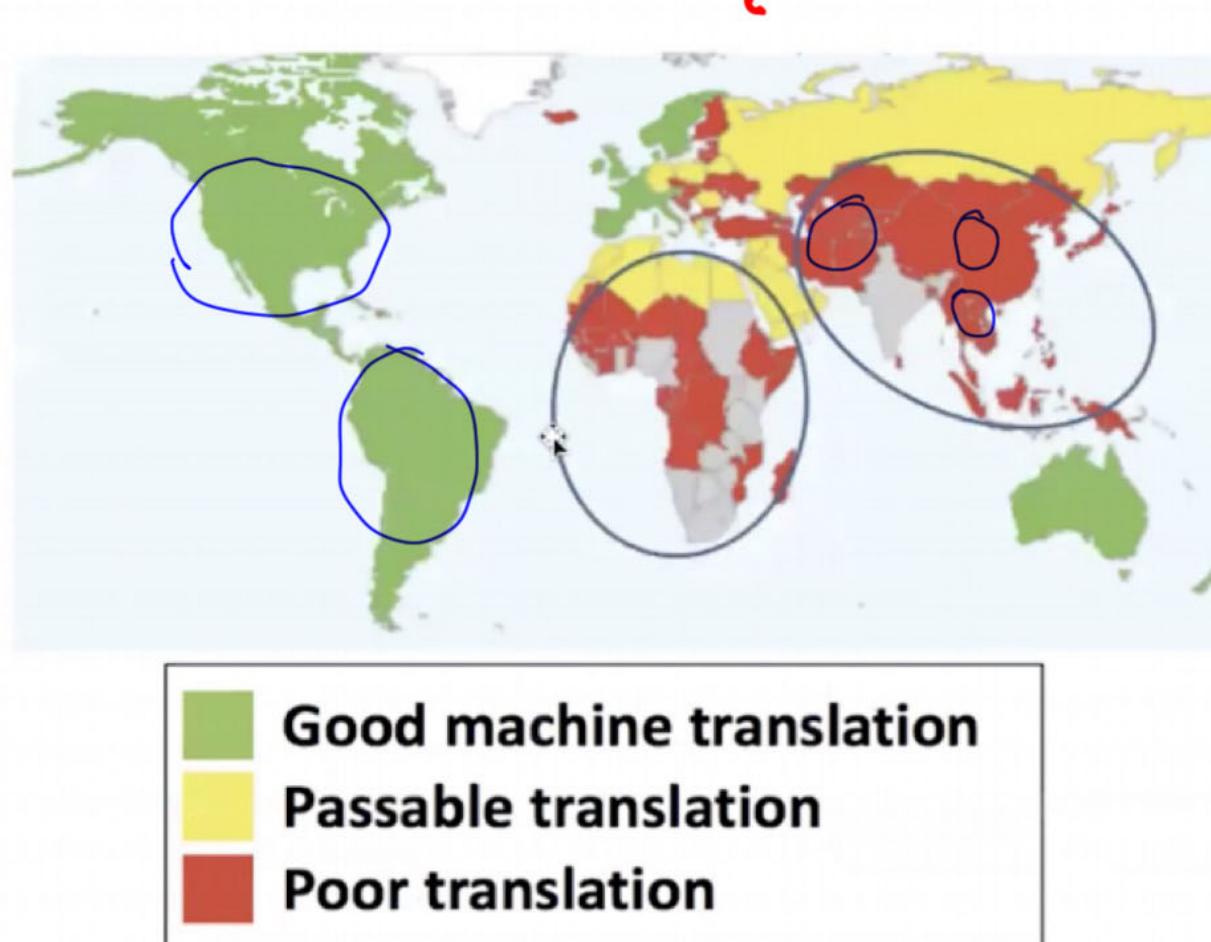
## • Decoding Algorithm . ie. Viterbi: (翻译 + 翻译)

- 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

# Language Model (语言模型)

- 对于一个好的语言模型: (训练好)
    - $p(\text{He is studying AI}) > p(\text{He studying AI is})$
    - $p(\text{nlp is an interesting course}) > p(\text{interesting course nlp is an})$
  - 怎么计算 $p(\cdot)$
- Uni-gram  $\leftarrow$   $p(\underline{x_1} \underline{x_2} \underline{x_3} \underline{x_4}) = p(\underline{\text{He}})p(\underline{\text{is}})p(\underline{\text{studying}})p(\underline{\text{AI}})$
- Bi-gram  $\leftarrow$   $P(\underline{\text{He}} \underline{\text{is}} \underline{\text{studying}} \underline{\text{AI}}) = p(\underline{\text{He}})p(\underline{\text{is}} | \underline{\text{He}})p(\underline{\text{studying}} | \underline{\text{is}})p(\underline{\text{AI}} | \underline{\text{studying}})$
- Tri-gram  $\leftarrow$   $P(\underline{\text{He}} \underline{\text{is}} \underline{\text{studying}} \underline{\text{AI}}) = p(\underline{\text{He}})p(\underline{\text{is}} | \underline{\text{He}})p(\underline{\text{studying}} | \underline{\text{he is}})$   
 $p(\underline{\text{AI}} | \underline{\text{is studying}})$
- $p(x, y) = p(x) \cdot p(y)$  if  $x, y$  相互独立
- $p(x, y) = p(x) \cdot p(y|x)$   
 $= p(y) \cdot p(x|y)$
- :
- N-gram

# Machine Translation: Difficulty Level by Language



# NLP的经典应用场景

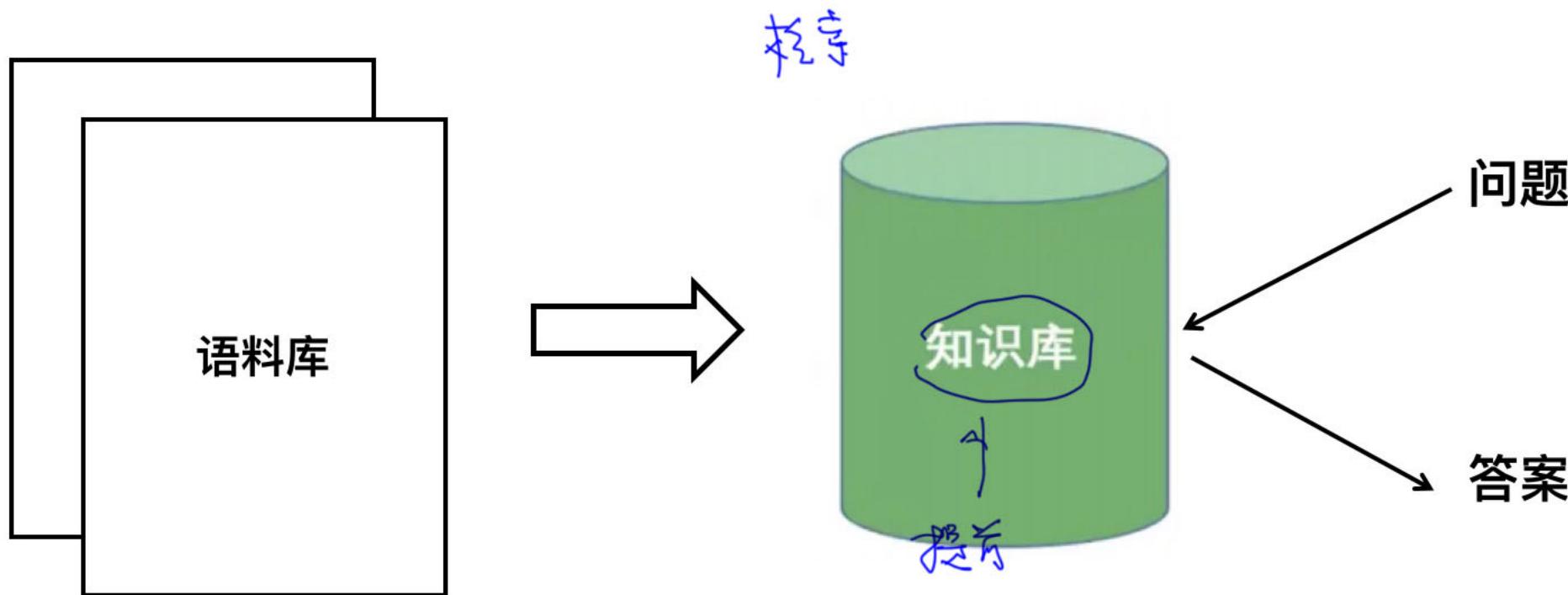
# Question Answering(问答系统)

I<sup>t</sup>BM

Watson.



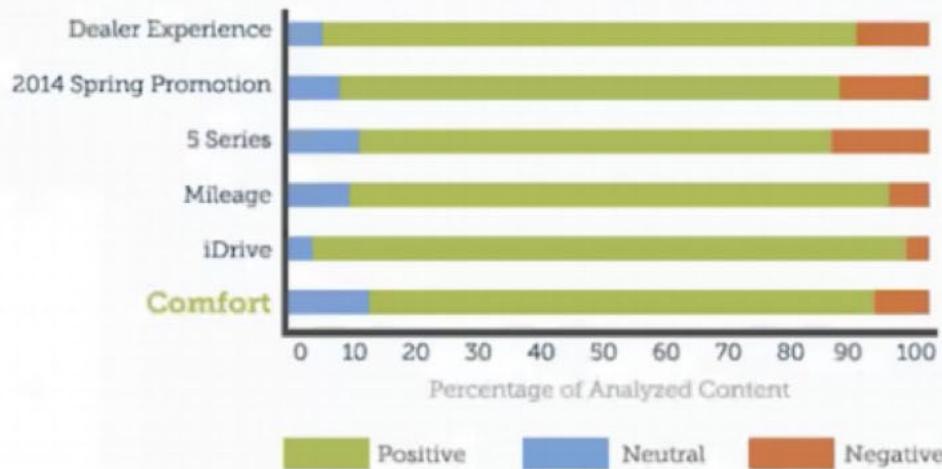
# Question Answering(问答系统)



# Sentiment Analysis(情感分析)

第二个项目

## BMW Sentiment Analysis Example



- 股票价格预测
- 舆情监控
- 产品评论
- 事件监测

Tweet analyzed as positive for **Comfort**



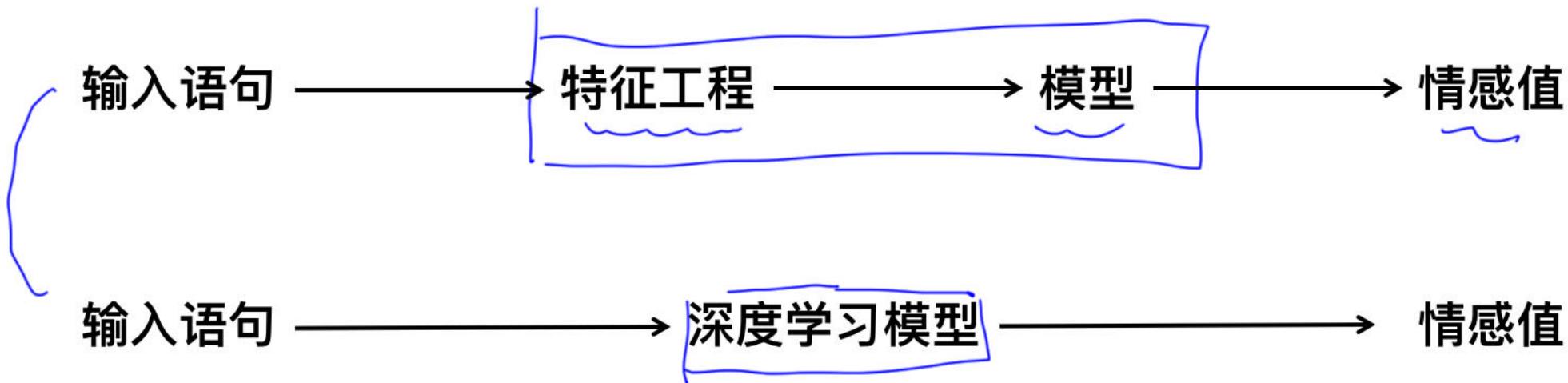
**Sarah Smith**  
@sarahsmith

<3 BMW. So comfy. Def checking out the new 3 series next wk.

2 Jan 14

Reply Retweet Favorite More

# Sentiment Analysis(情感分析)



# Machine Translation(机器翻译)

All

Books

News

Shopping

More

Settings

Toc

About 509,000,000 results (0.39 seconds)

Sep2Sep

Deep Learning

Chinese - detected ▾



English ▾



今天是自然语言处理训练营第一天 Edit

Jīntiān shì zìrán yǔyán chǔlǐ xùnliàn  
yíng dì yī tiān

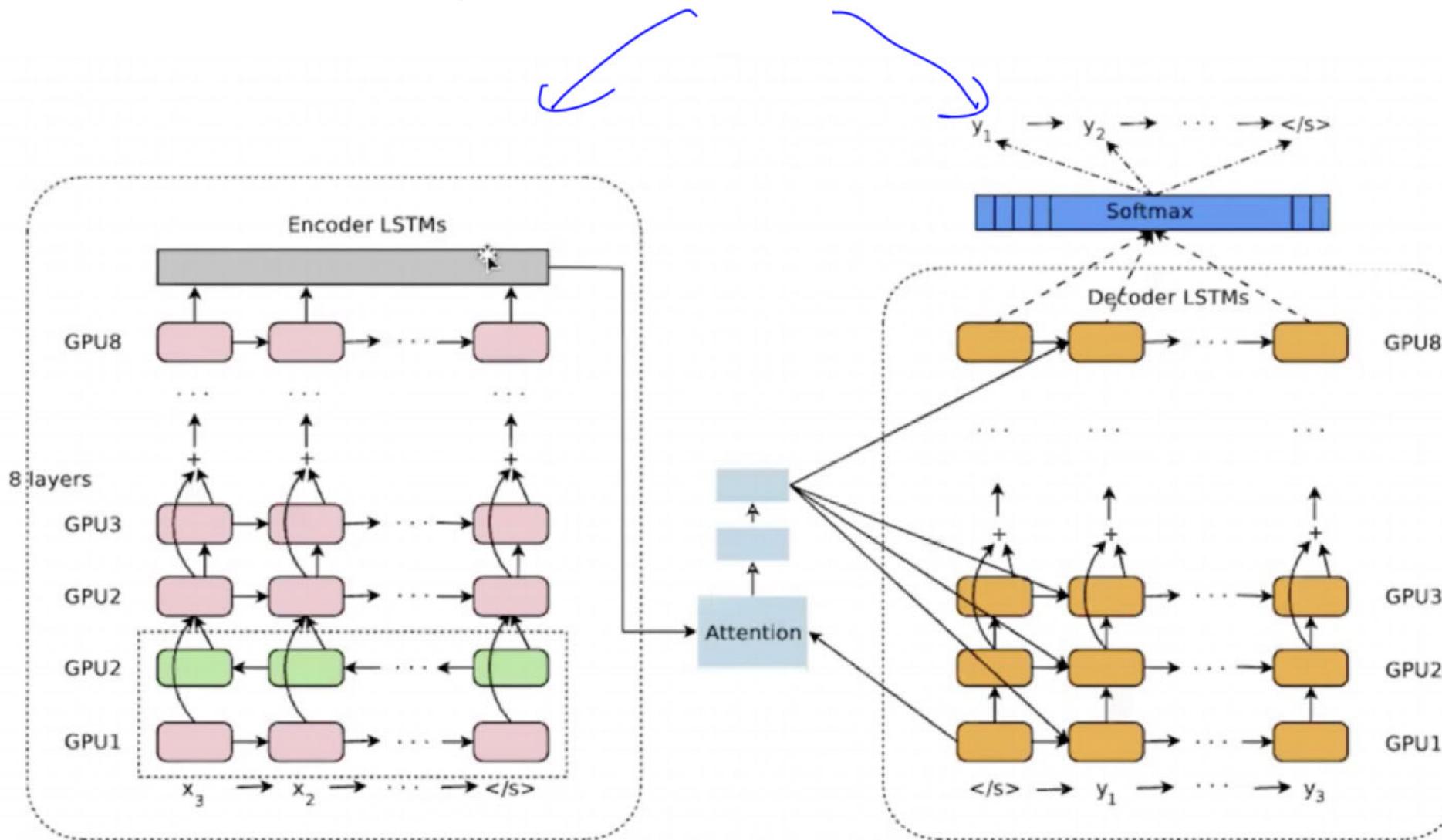
Today is the first day of the Natural  
Language Processing Training Camp.

[Open in Google Translate](#)

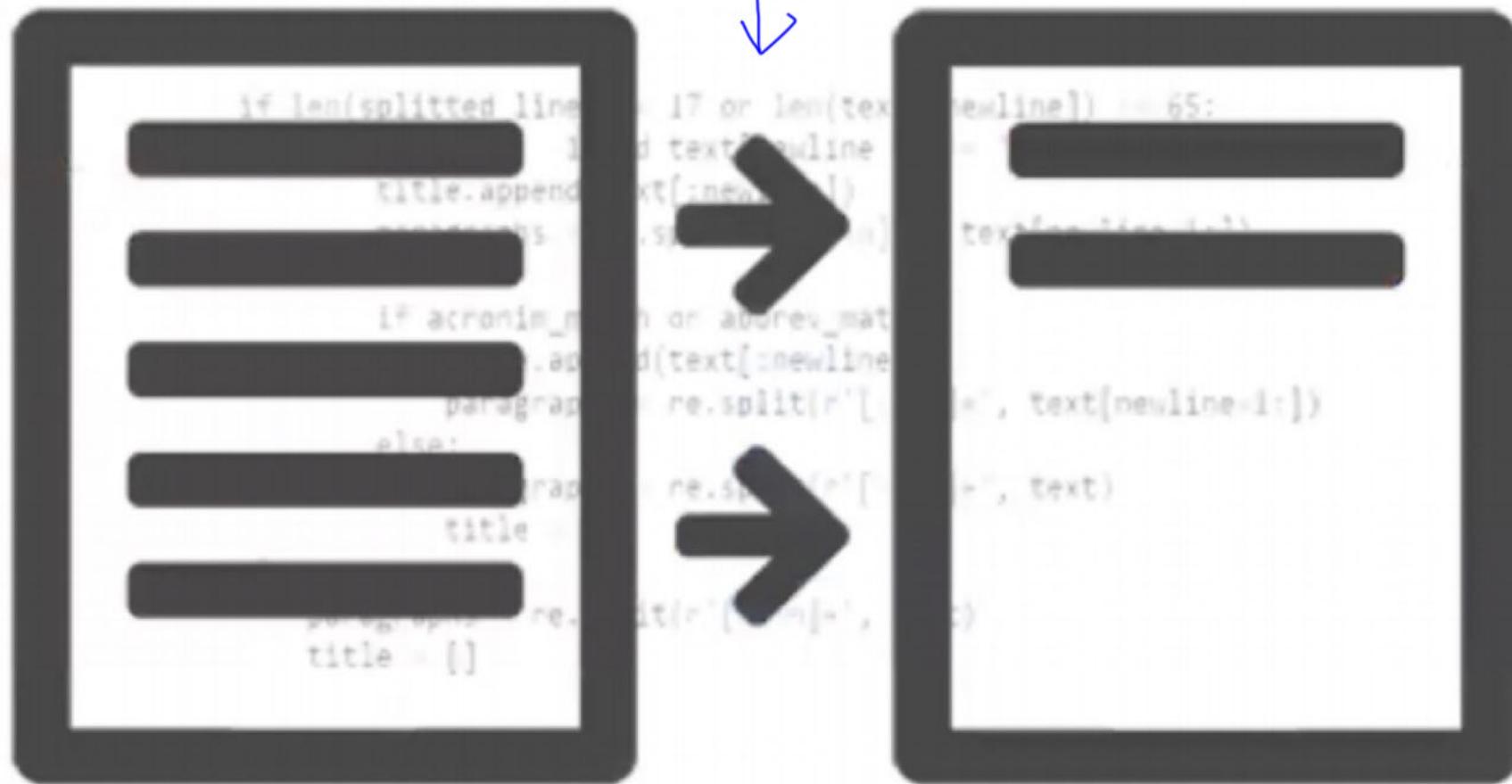
[Feedback](#)

# Google 机器翻译

GPU



# Text Summarization(自动摘要)



# Chatbot (聊天机器人)



无聊了，随便聊一聊



想定一个机票



不知道自己要做啥

# Information Extraction(信息抽取)

Dan Jurafsky



## Information Extraction

Subject: **curriculum meeting**

Date: January 15, 2012

To: Dan Jura

NER ←  
Relation Extraction ←

NLP

Event: Curriculum  
Date: Jan-16-2012  
Start: 10:00am  
End: 11:30am  
Where: Gates 159

Hi Dan, we've now scheduled the curriculum meeting.

It will be in Gates 159 tomorrow from 10:00-11:30.

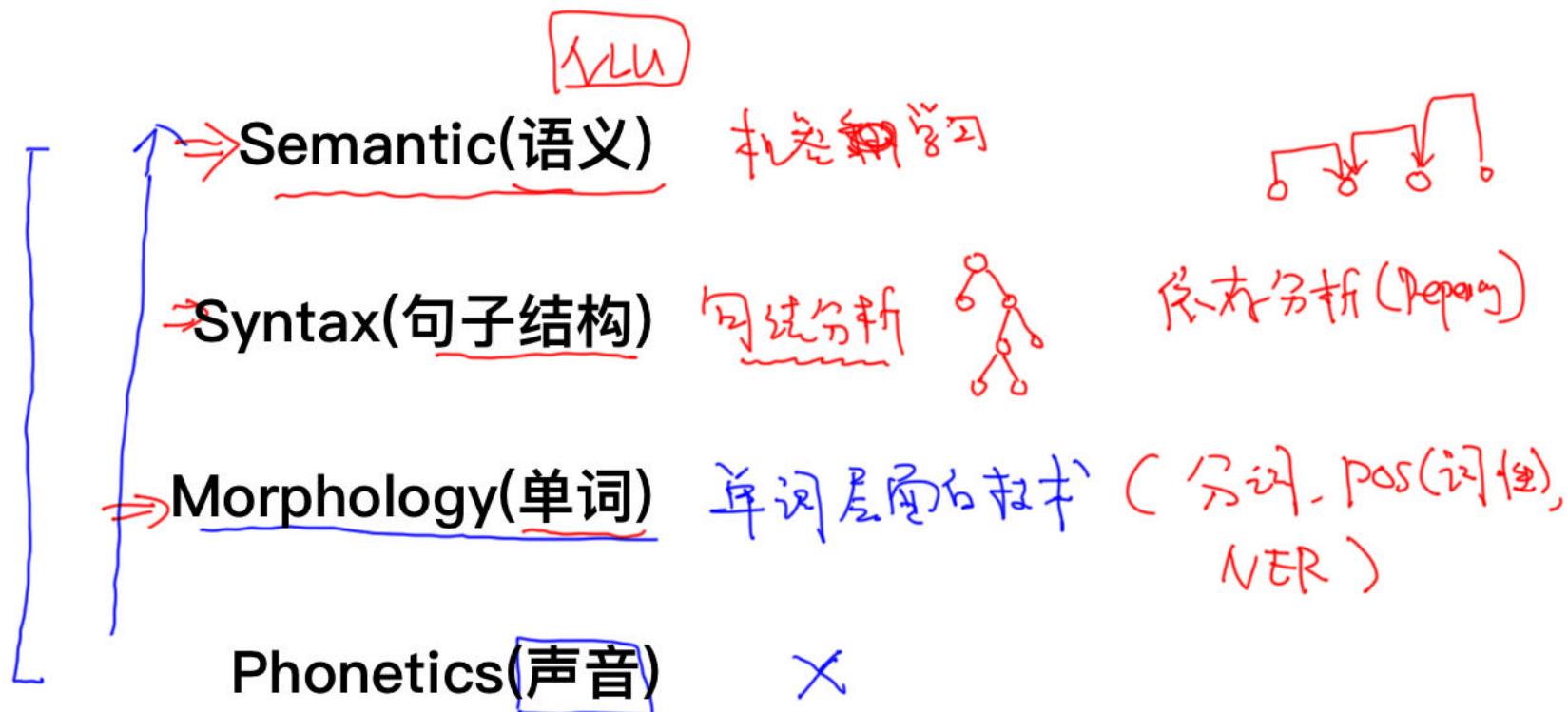
-Chris

Create new Calendar entry

# NLP关键技术



# 自然语言处理技术四个维度



# Word Segmentation(分词)

今天是自然语言处理训练营第一次课



今天 是 自然语言处理 训练营 第一次 课

# Part-of-Speech (词性) ← 特征

今天是1月22日，也是我们训练营的第一天，暂时课程以ZOOM的方式直播

# Named Entity Recognition

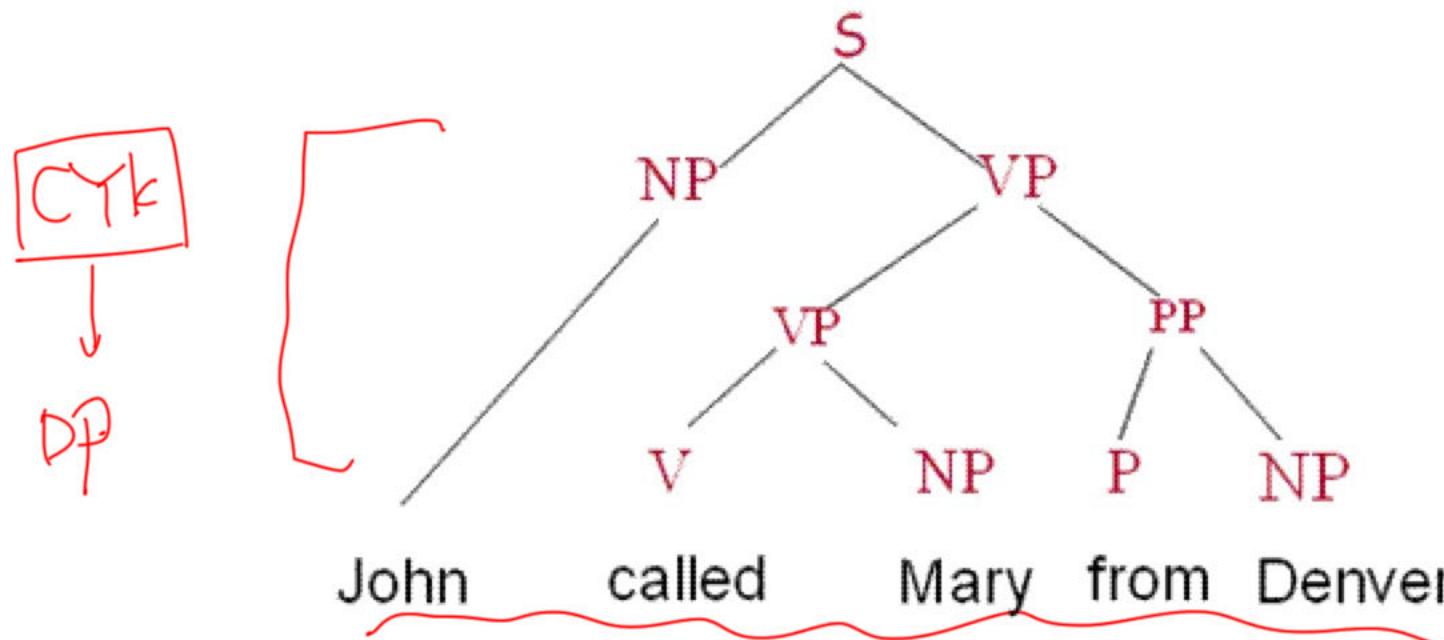
## (命名实体识别)

今天是1月22日，也是我们训练营的第一天，暂时课程  
以ZOOM的方式直播

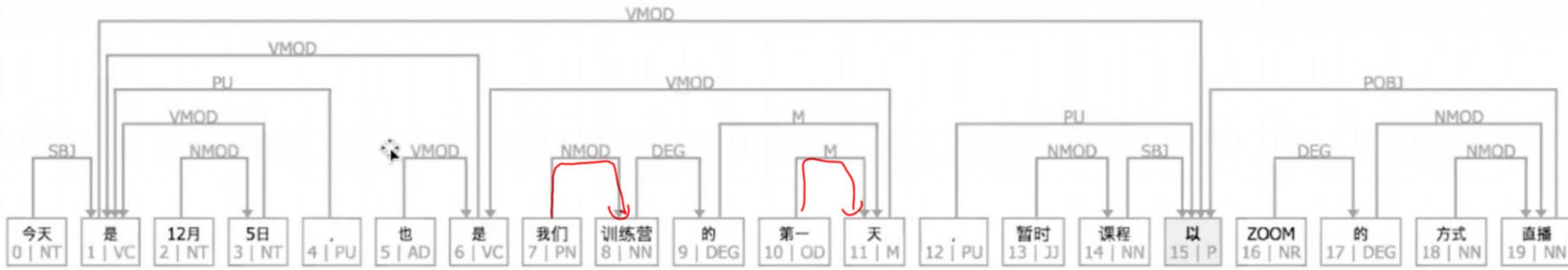
1月22日  
ZOOM  
Medical

实体 / 词类

# Parsing (句法分析)



# Dependency Parsing (依存分析)



# Relation Extraction(关系抽取)

0→6

关系抽取

