揚州大學
YANGZHOU UNIVERSITY

自 然 语 言 处 理

Natural language Processing

强继朋 副教授

扬州大学

第1章
NLP 介绍

Part.01

课程介绍

课程信息

- 课程网页：https://yzu-nlp.github.io/
  - 课程相关的所有信息： 课件、实验、阅读资料、课程设计等
- 评分标准：作业（30%）+实验（20%）+期末考试（50%）
  - 作业： 4次
  - 实验： 4次

课程目标

- 了解NLP中不同领域的基本原理
- 理解NLP的理论概念和算法
- 为NLP应用建立模型的实践经验
- 独立完成NLP项目的开发

Part.02

NLP背景介绍

什么是自然语言？

语言分为：

- 人类语言 （自然语言）
- 机器语言

自然语言即为自然地随文化演化的语言如：汉语、英语、法语等

自然语言可以采取不同的形式，如语音或手势

有何特性？

什么是自然语言处理（NLP）？

赋予计算机处理人类语言的能力。

人工智能的一个分支

也可被称为：speech and language processing, human language technology, computational linguistics, and speech recognition and synthesis.

目标是让计算机执行涉及人类语言的有用任务，
例如启用人机交流、改善人际交流、或简单地对文本或语音进行有用处理等任务。

NLP发展历程

数据驱动

规则驱动

2012年~

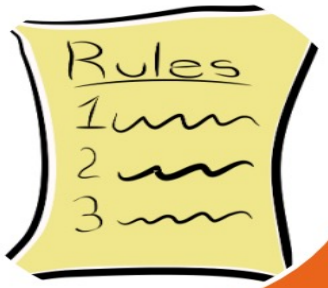
基于深度学习的方法

- 神经网络
- 分布式表示
- 非线性模型

1990年~2012年

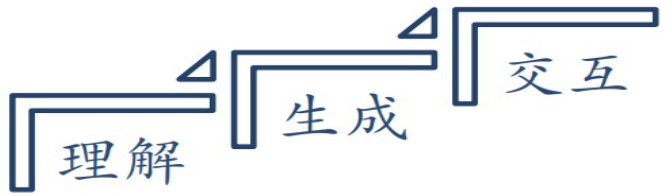
基于统计学习的方法

- 离散表示
- 线性模型



1990年以前

基于规则的方法



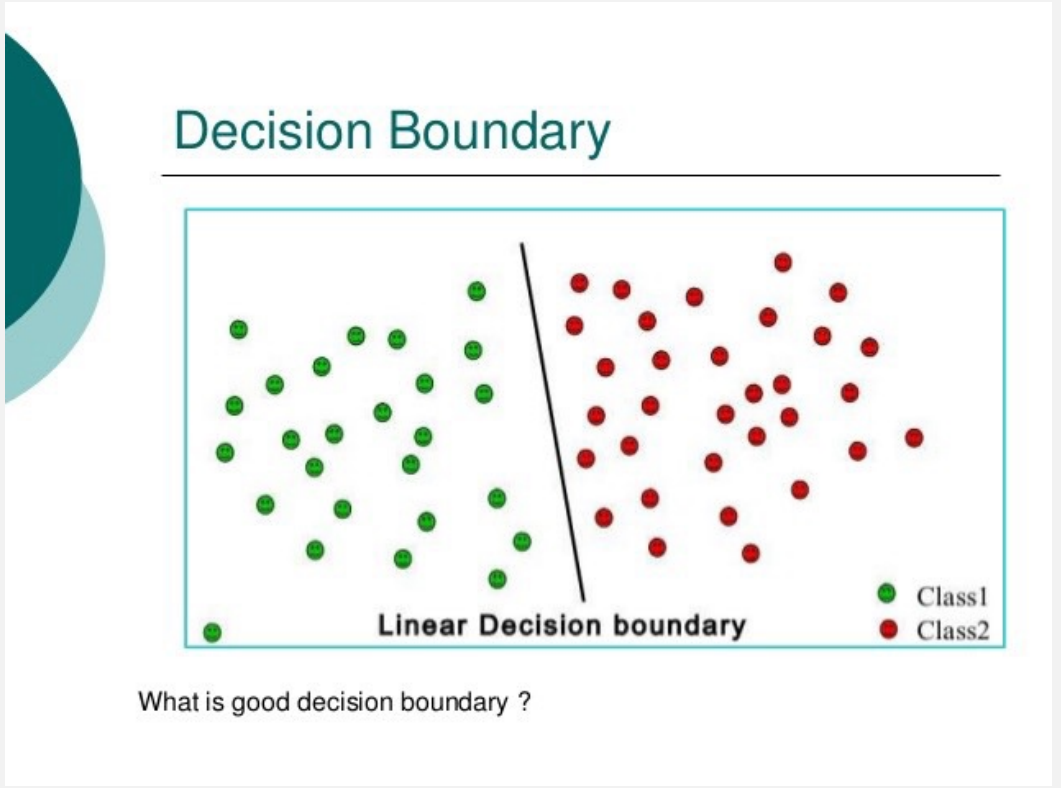
NLP发展历程

- 基于规则的方法（1950-1990）
  - 最古老的NLP方法
  - 利用语言学专家定义的规则
  - 局限于某个领域
  - 解决模糊性具有挑战性

“The spirit is strong, but the flesh is weak”
“The Vodka is good, but the meat is bad”

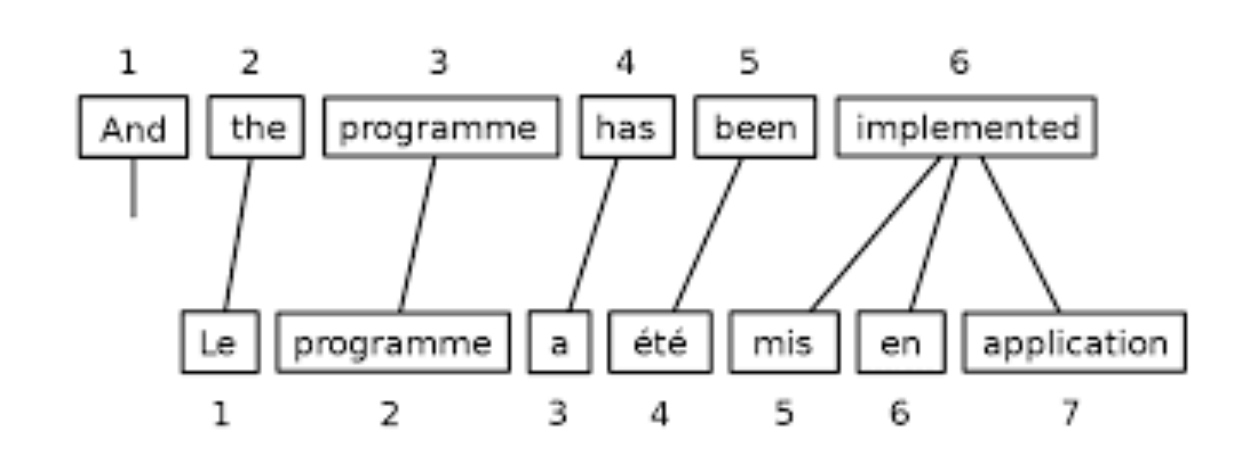
NLP发展历程

- 基于统计的方法（1990s-2010s）
  - 利用传统的机器学习方法
  - 通常采用的是概率模型的方法
    - 人工标注的训练数据
    - 特征工程
    - 训练带参数的模型
    - 应用模型到测试数据

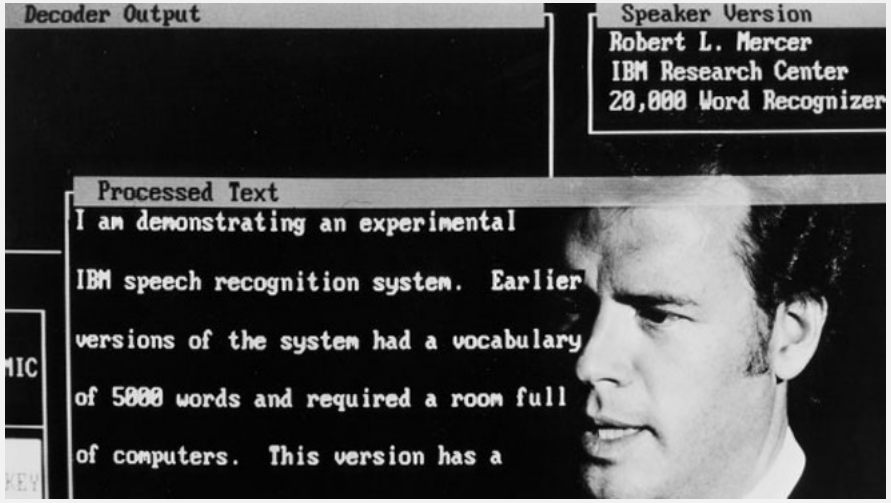


基于统计的方法

IBM的机器翻译模型



语音识别

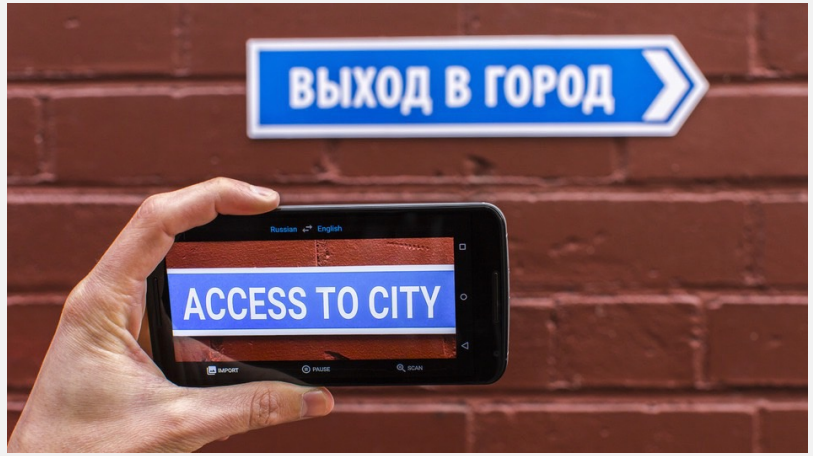


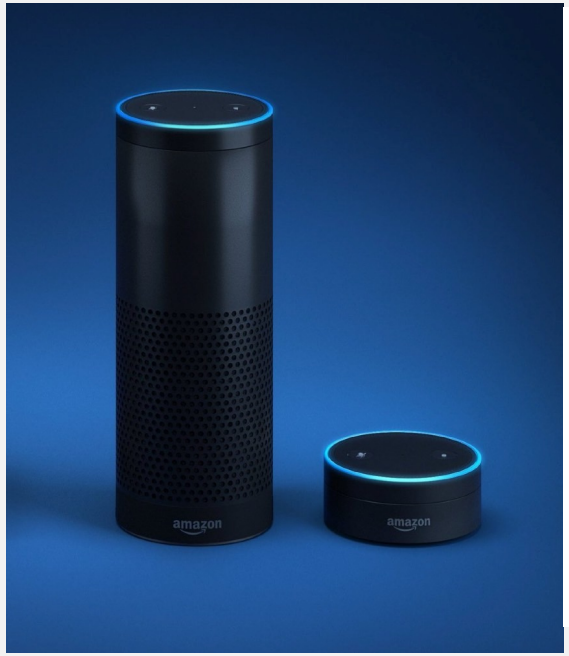
Anytime a linguist leaves the group the (speech) recognition rate goes up.

- Fred Jelinek

NLP发展历程

- 基于深度学习的方法（2010s–）
  - 取得了性能上的大幅度提升
  - 不需要手工定义特征
  - 构造非常大的神经网络
  - 在原始语料上进行预训练

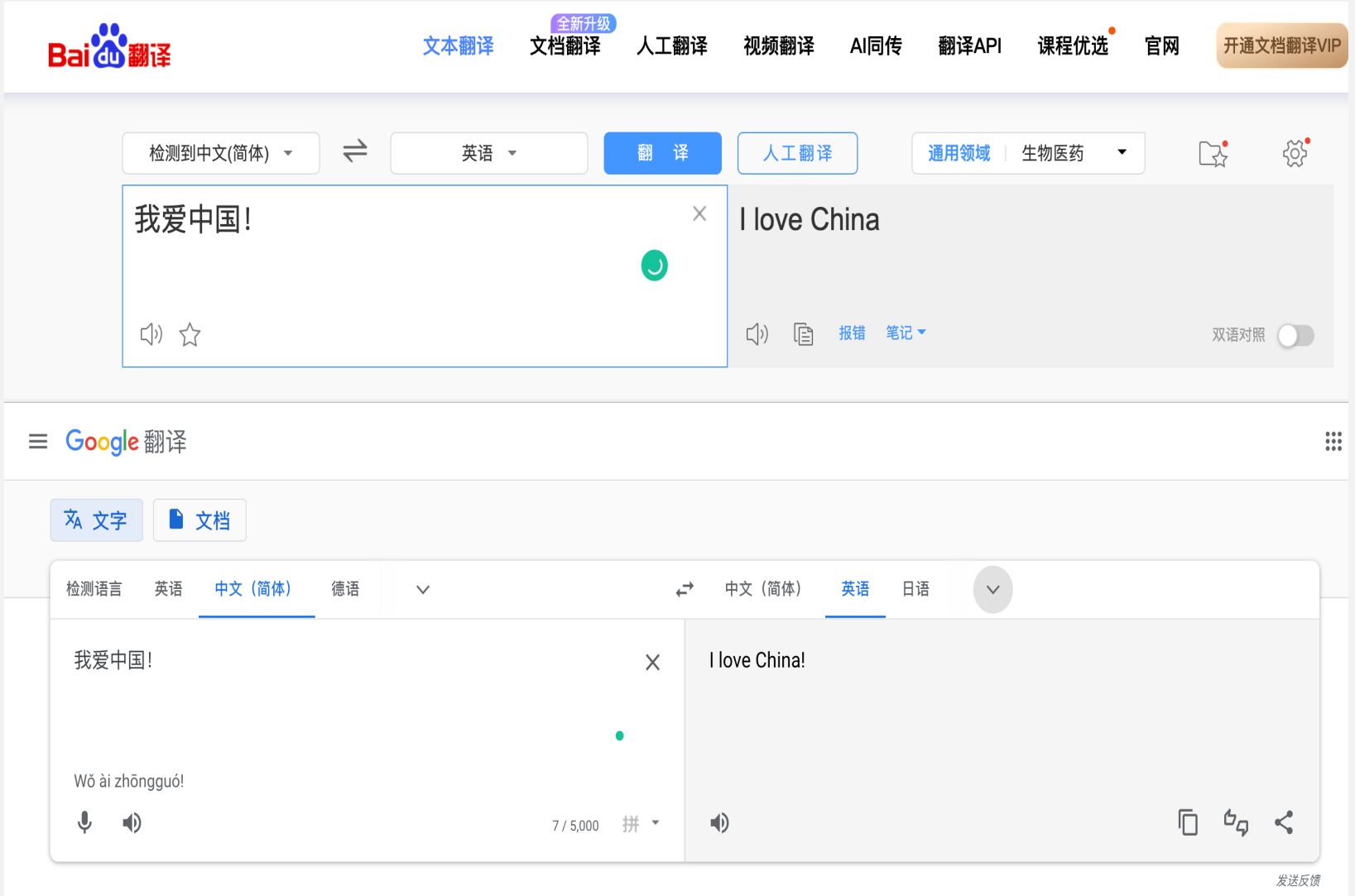


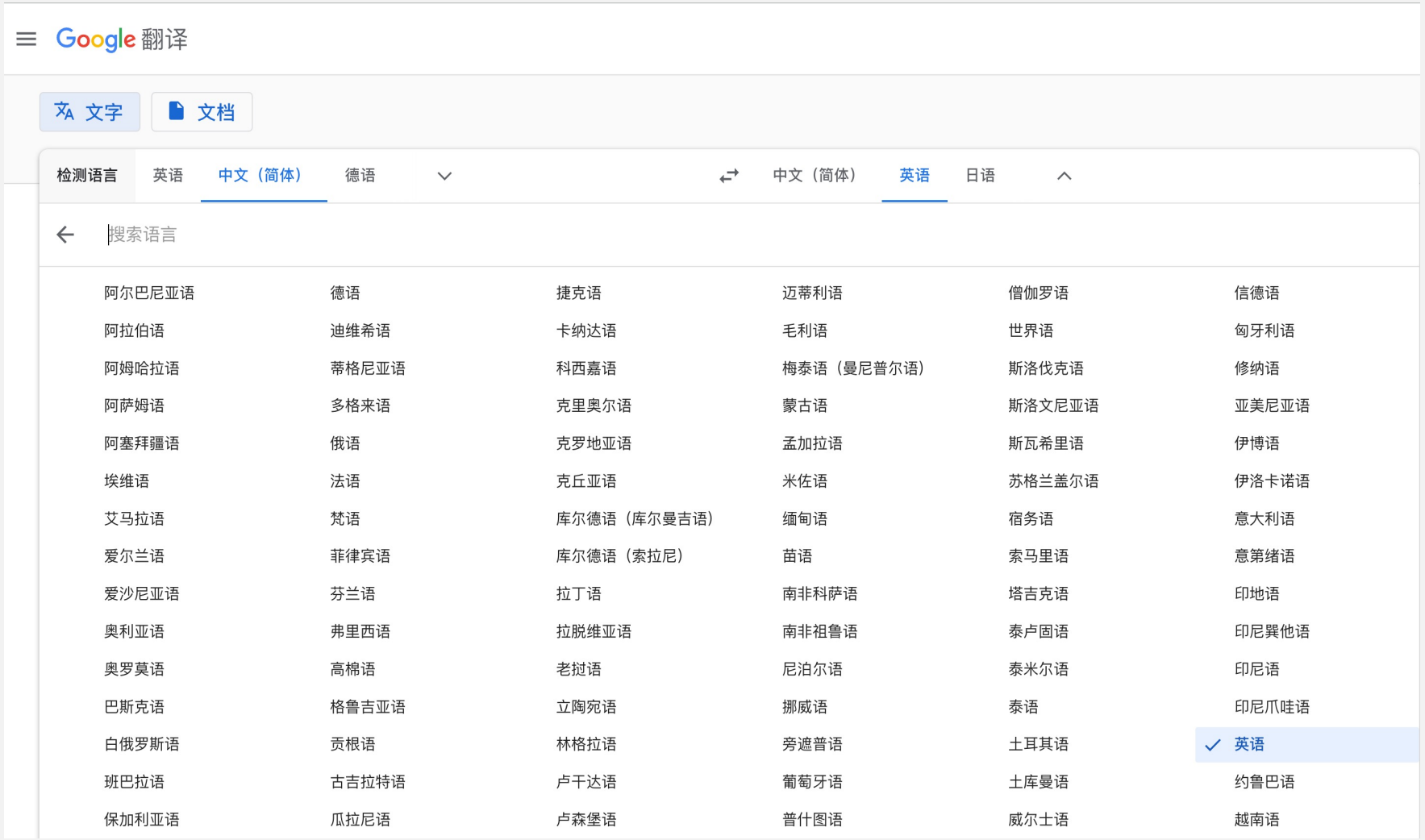


Part.03

NLP应用场景

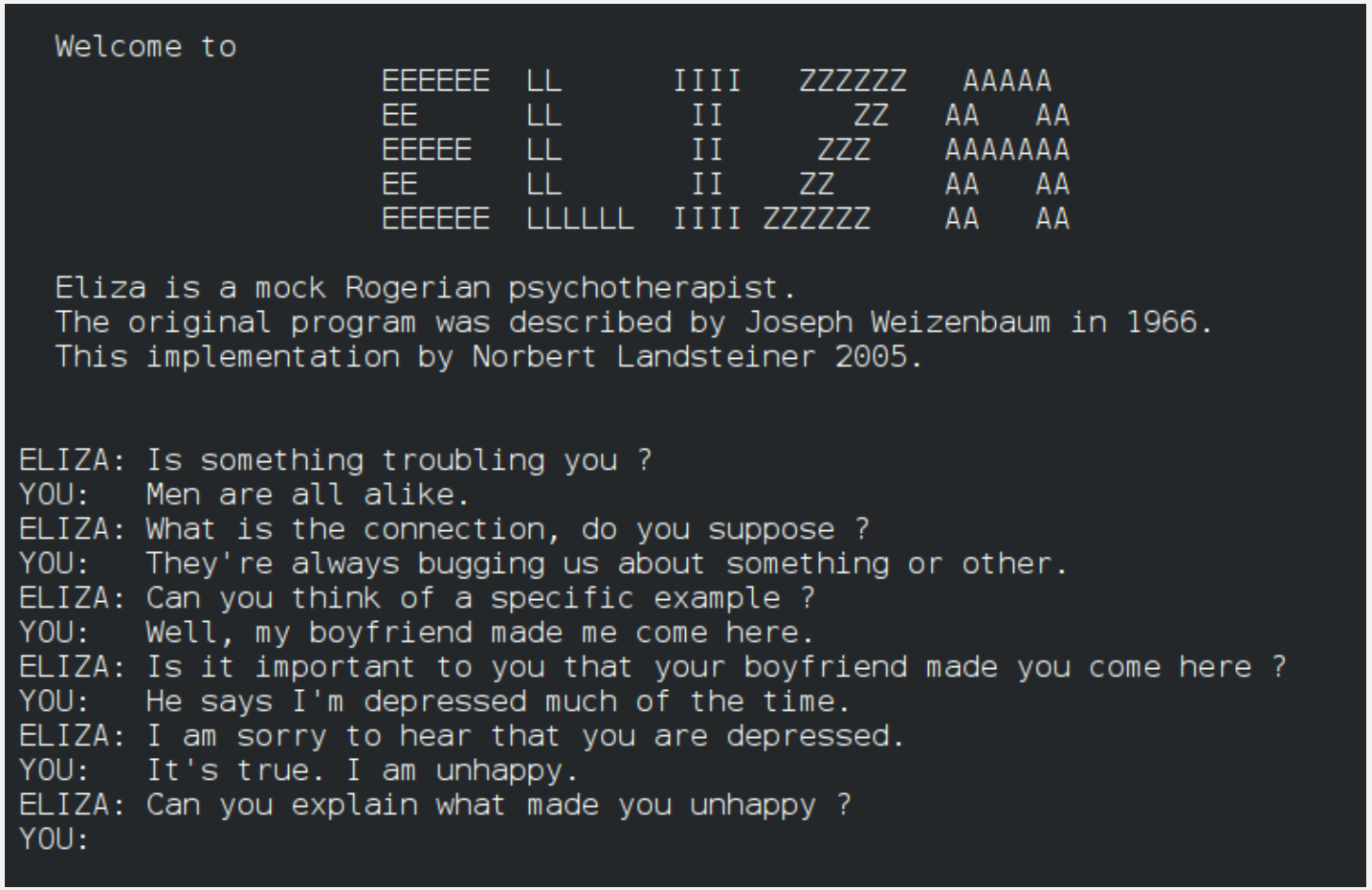
机器翻译（Machine Translation）





机器翻译的任务目标是自动地将一种语言翻译为另一种语言。

人机对话（Conversational agent / Dialogue system）



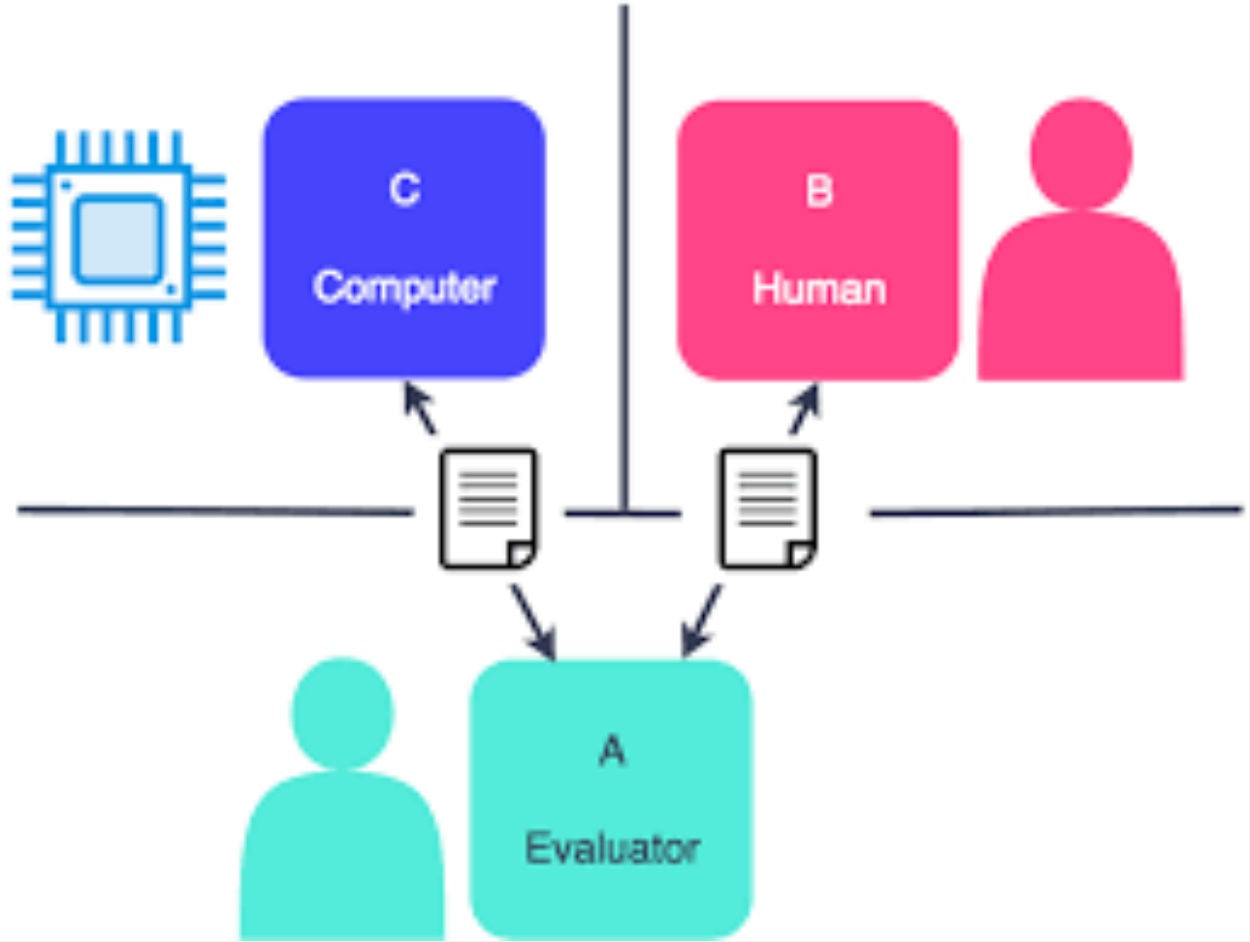
ELIZA 是早期的一个聊天机器人（1966）

使用模式匹配来处理输入
并将其转换为合适的输出

人机对话（Conversational agent / Dialogue system）

汽车智能语音识别系统

人机对话（Conversational agent / Dialogue system）



三个参与者：两个人和一台电脑

中间的人扮演裁判的角色，
并向另外两个参与者提出问题，
裁判通过审视参赛者的回答，
以判断哪个是真人哪个是机器。

两位参赛者(机器和人类)的任务为：
让裁判相信自己是人类，
并判定另一位参赛者是机器。

图灵测试

自动问答（Question Answering）

自动问答（Question Answering）

The first recorded travels by Europeans to China and back date from this time. The most famous traveler of the period was the Venetian Marco Polo, whose account of his trip to "Cambaluc," the capital of the Great Khan, and of life there astounded the people of Europe. The account of his travels, Il milione (or, The Million, known in English as the Travels of Marco Polo), appeared about the year 1299. Some argue over the accuracy of Marco Polo's accounts due to the lack of mentioning the Great Wall of China, tea houses, which would have been a prominent sight since Europeans had yet to adopt a tea culture, as well the practice of foot binding by the women in capital of the Great Khan. Some suggest that Marco Polo acquired much of his knowledge through contact with Persian traders since many of the places he named were in Persian.

How did some suspect that Polo learned about China instead of by actually visiting it?

Answer: through contact with Persian traders

自动问答（Question Answering）

简单 – 定义性问题

- 例如：时间地点任务等事实性问题。
- 解决途径-网络搜索

复杂 – 推理性问题

- 根据已知得出结论
- 综合多源信息进行总结归纳
- 解决途径-信息提取，词义消歧等

例子?

情感分析

Loves the German bakeries in Sydney. Together with my imported honey it feels like home	Positive
@VivaLaLauren Mine is broken too! I miss my sidekick	Negative
Finished fixing my twitter...I had to unfollow and follow everyone again	Negative
@DinahLady I too, liked the movie! I want to buy the DVD when it comes out	Positive
@frugaldougal So sad to hear about @OscarTheCat	Negative
@Mofette briliant! May the fourth be with you #starwarsday #starwars	Positive
Good morning thespians a bright and sunny day in UK, Spring at last	Positive
@DowneyisDOWNEY Me neither! My laptop's new, has dvd burning/ripping software but I just can't copy the files somehow!	Negative

Part.04

NLP的挑战

为什么语言很难理解？

多少话人都看不懂还想
让机器懂!!!

词义的多样性

一个词语具有多个意思

The fisherman went to the *bank*.

bank[1]

/baNGk/

noun

plural noun: **banks**

1. the land alongside or sloping down to a river or lake.
   "willows lined the bank"
   synonyms: edge, side, shore, coast, embankment, bankside, levee, border, verge, boundary, margin, rim, fringe; More

1. a financial establishment that invests money deposited by customers, pays it out when required, makes loans at interest, and exchanges currency.
   "I paid the money straight into my bank"
   synonyms: financial institution, merchant bank, savings bank, finance company, trust company,

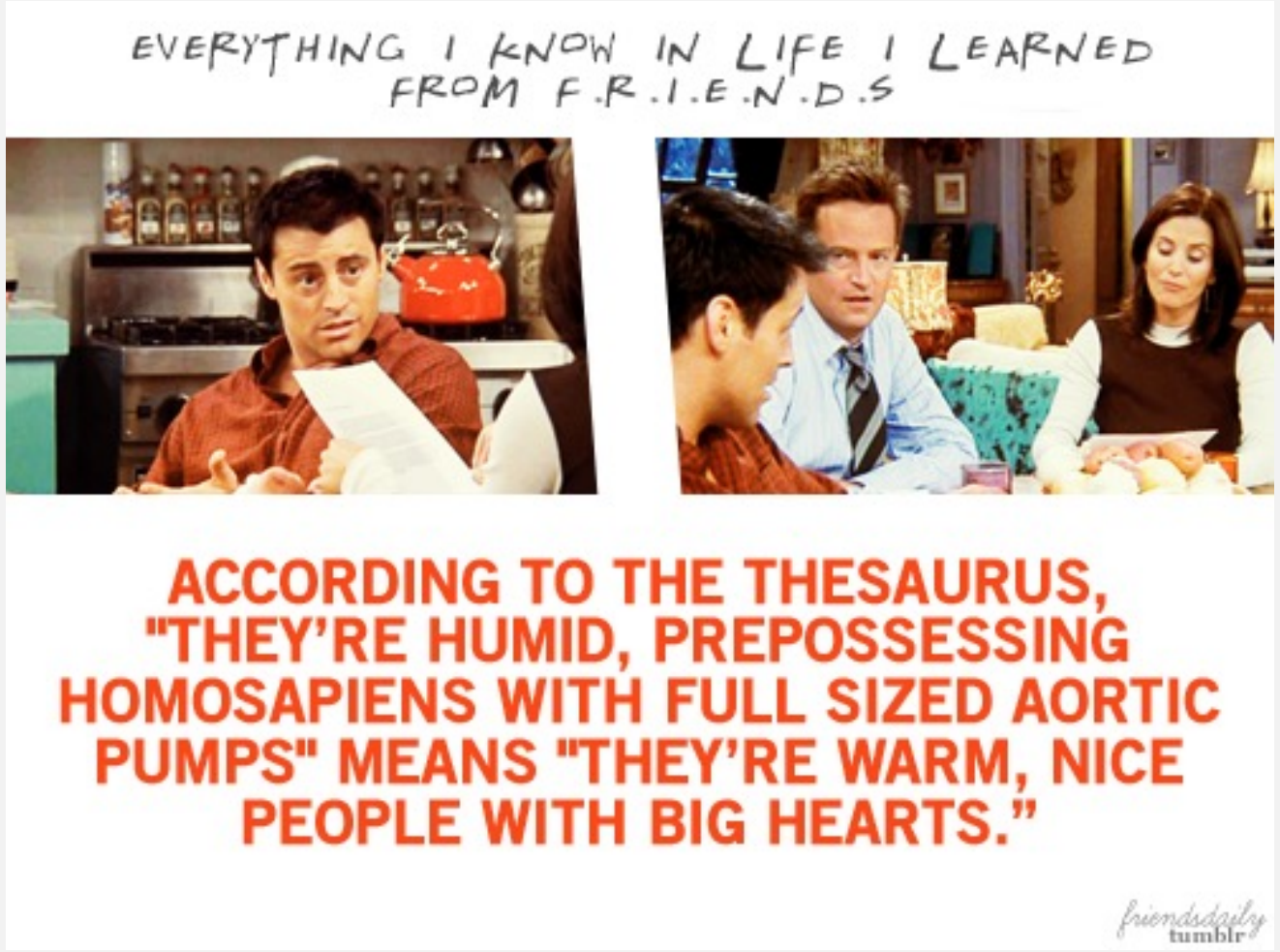
词义的多样性

The fisherman went to the *bank*. He deposited some money.

词义消岐（Word sense disambiguation ）

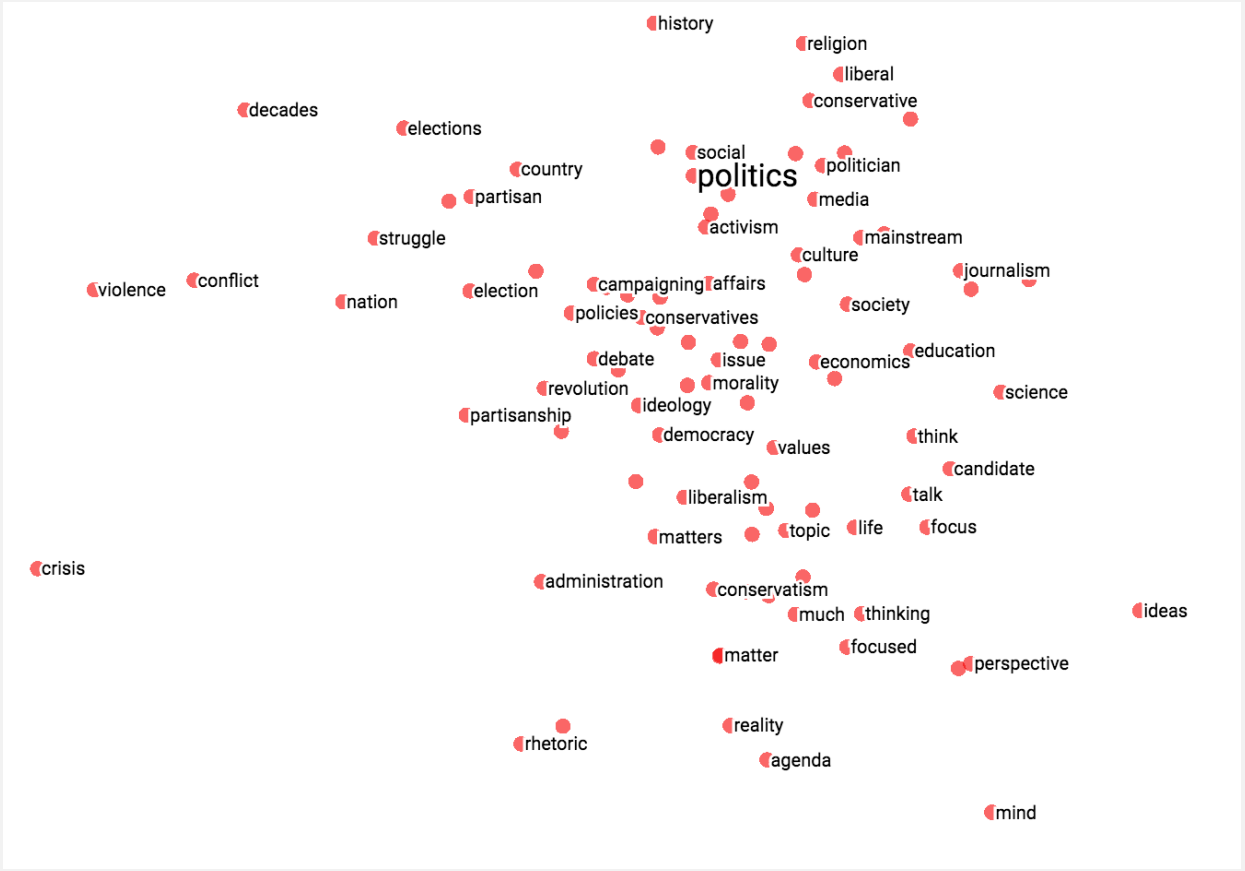
词语的多样性

多个词语具有相同的意思



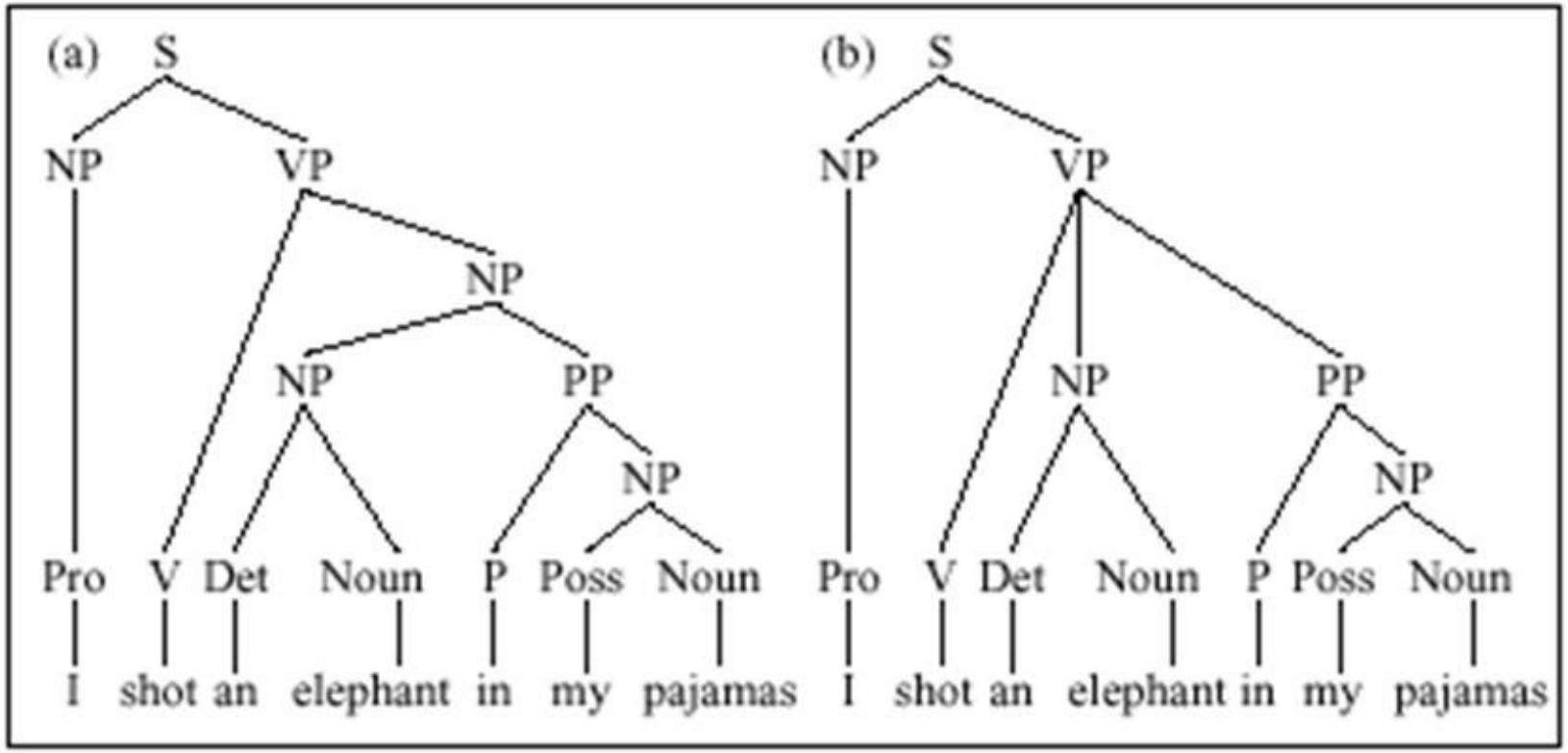
词语的分布式表示

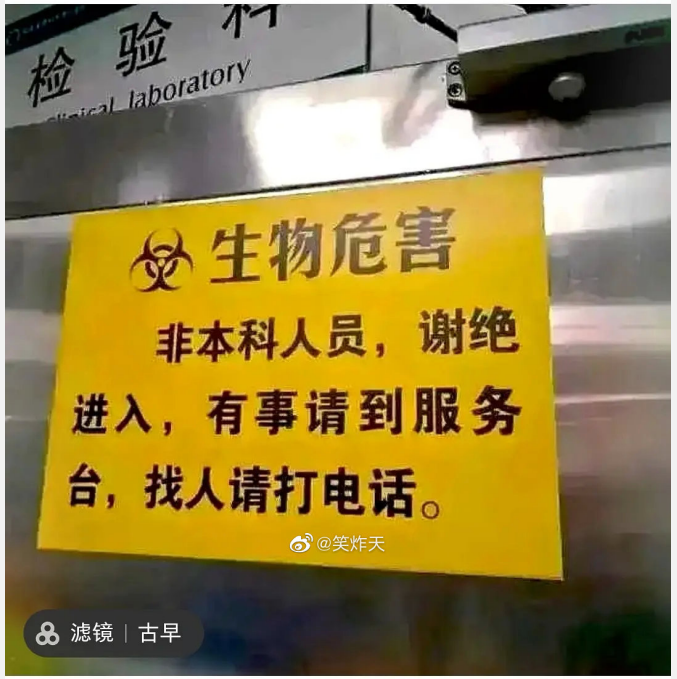
将单词投影到连续向量空间



句法的模糊性

I shot an elephant in my pajamas





没考上本科怎么了？

考上大专怎么了？大专怎么了？？？你们一个个本科不得了了，第一学历大专怎么了？多的是专升本，后来又考研读博出国留学，照样成才，照样过的好！不比你们这些差！

不停说大专没有用，大专第一学历不行，你们只不过就是一些幸运儿罢了，人生比较顺利，考了个本科，天天在这里叫唤大专不行不行！

很多人不是想考大专不是想这样，很多是人生比较坎坷曲折，没有考上本科，不得了了，你本科也不见得拍怎么样！！！

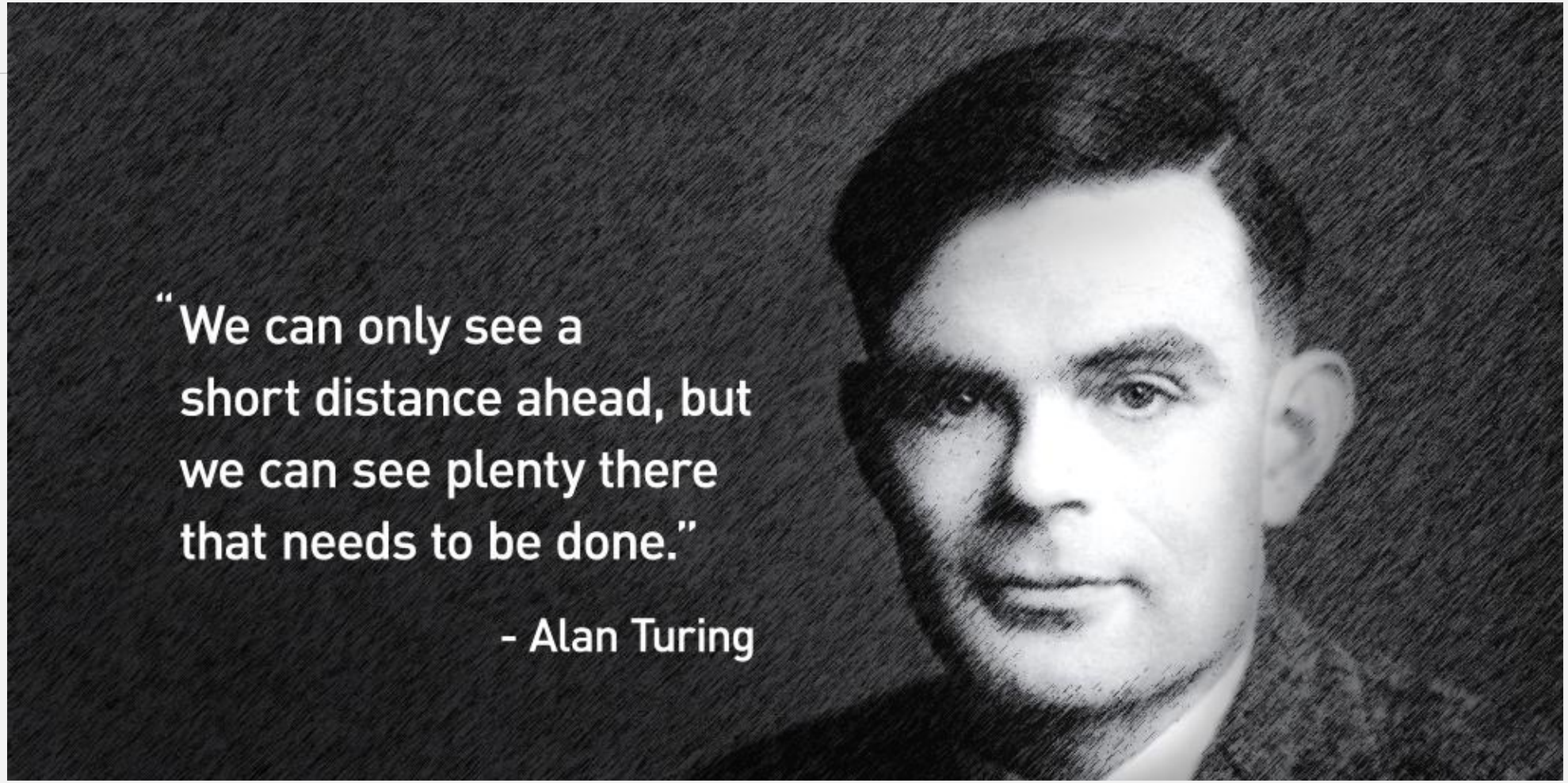
篇章的歧义（Discourse ambiguity）

Alice invited Maya for dinner but she cooked her own food

she = Alice or Maya?

… and brought it with her. Maya

… and ordered a pizza for her guest. Alice


"We can only see a
short distance ahead, but
we can see plenty there
that needs to be done."
- Alan Turing

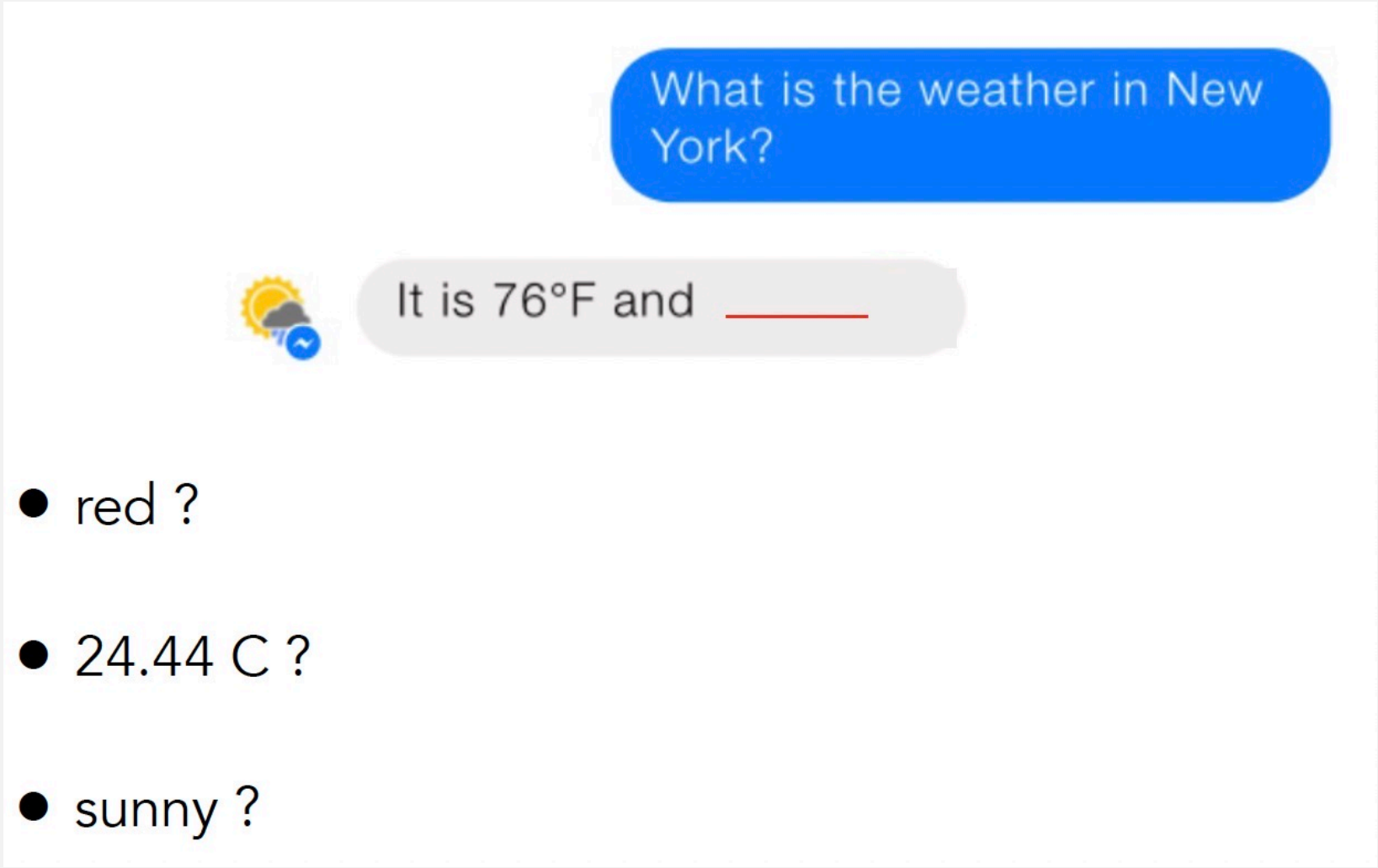
Part.05

课程内容

课程内容

1. 语言模型
2. 文本分类
3. 向量语义和词嵌入
4. 神经网络语言模型
5. 序列处理语言模型 (RNN, LSTM, Transformer)
6. 机器翻译和编码器-解码器模型
7. 预训练语言模型

语言模型（Language models ）



- red ?
- 24.44 C ?
- sunny ?

语言模型（Language models ）

Today, in New York, it is 76 F and red

vs

Today, in New York, it is 76 F and sunny

- Both are grammatical
- But which is more likely?

语言模型（Language models ）

目标：计算一个句子或者一个单词序列的概率

- $P(W) = P(w_1,w_2,w_3,w_4,w_5...w_n)$

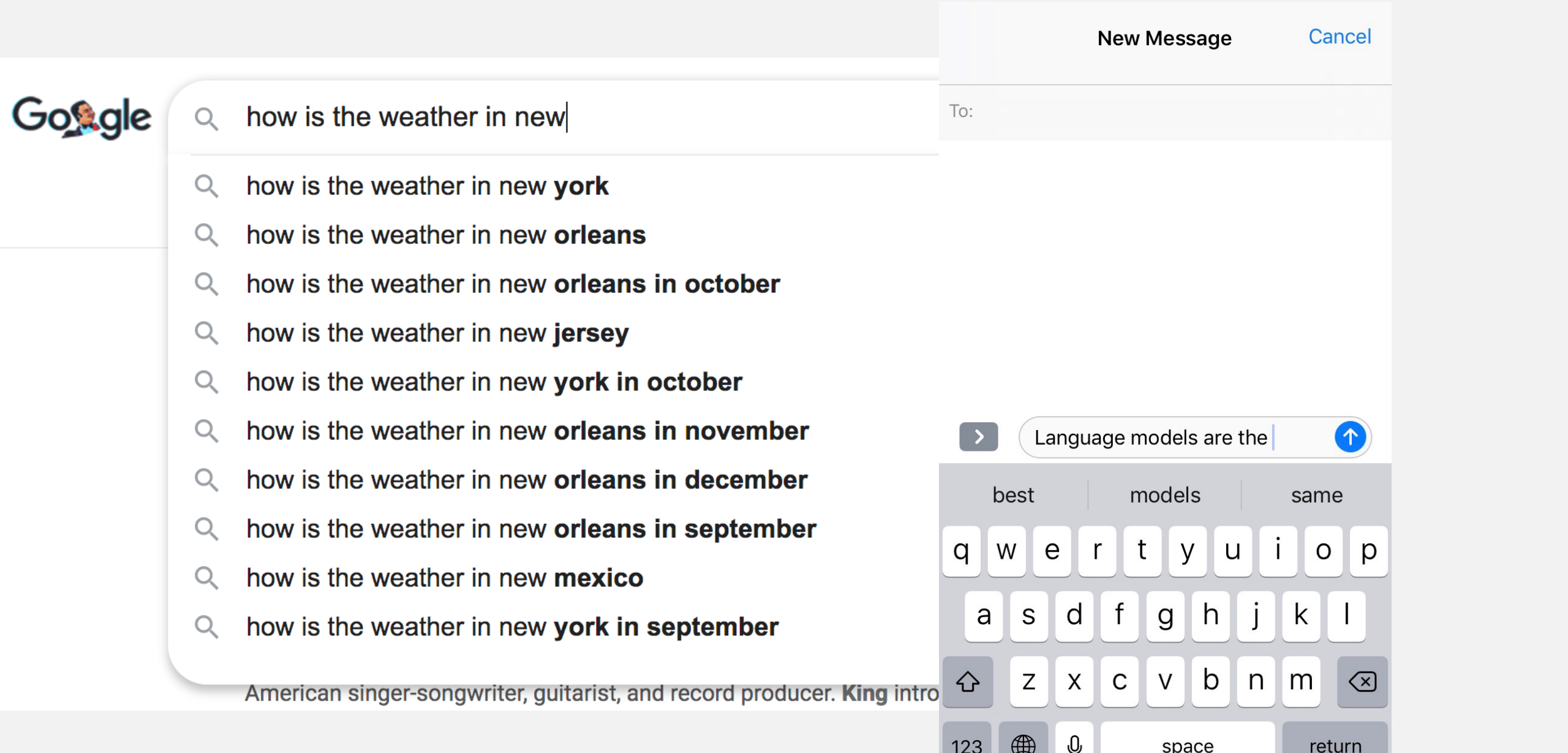
相关的任务：计算下一个单词的概率

- $P(w_5|w_1,w_2,w_3,w_4)$

一个用于计算下列两种概率的模型：

- $P(W)$ or $P(w_n|w_1,w_2...w_{n-1})$ 则被称为语言模型

语言模型（Language models ）



谢 谢