自 然 语 言 处 理

Natural language Processing

强继朋　副教授、博导

扬州大学

强继朋

- **研究方向：**自然语言处理、文本挖掘。
- **个人介绍：**副教授，博士生导师。江苏省青蓝工程优青，IEEE Senior Member, IEEE TETCI 副主编。主持国家级项目2项。一作发表高质量论文20余篇，包括AIJ、TKDE、TACL、TASLP、TKDD、ACL、AAAI、EMNLP等顶级期刊和会议论文。两篇代表性论文入选ESI高被引论文。出版中文和英文专著各一部。已授权发明专利10余项项。

强继朋

主持的科研项目：

(1) 国家语委重点项目“大语言模型语言行为的安全与伦理研究”，2024.01-2026.12

(2) 国家自然科学面上基金“基于无监督学习的文本内容精简方法研究”(62076217)，2021.01-2024.12.

(3) 国家自然科学青年基金“异质文本主题挖掘研究及应用于癌症子类型的发现”(61703362)，2018.01-2020.12.

(4) 江苏省自然科学青年基金“基于吉布斯抽样的异质文本主题模型及应用”(BK20170513)，2017.07-2020.06.

强继朋

出版的专著：

(1) **强继朋**、李云、吴信东. 自动文本简化. 中国科学出版社，2022.

(2) 吴信东、谢飞、**强继朋**、等. 带有通配符和长度约束的模式匹配（英文）. 中国科学出版社，2016.

强继朋

5篇代表性论文：

(1) **Jipeng Qiang**, et al. ParaLS: Lexical Substitution via Pretrained Paraphraser. The 61st Annual Meeting of the Association for Computational Linguistics (ACL), 2023. (CCF A)

(2) **Jipeng Qiang**, et al. Natural language watermarking via paraphraser-based lexical substitution. Artificial Intelligence, 2023, 317:103859. (CCF A, SCI一区)

(3) **Jipeng Qiang**, et al. Transactions of the Association for Computational Linguistics, 2023. (CCF B, SCI一区)

(4) **Jipeng Qiang**, et al. Short Text Topic Modeling Techniques, Applications, and Performance: A Survey. IEEE Transactions on Knowledge and Data Engineering (TKDE), 2022, 34(3): 1427-1445. (CCF A)

(5) **Jipeng Qiang**, Xindong Wu. Unsupervised Statistical Text Simplification. IEEE Transactions on Knowledge and Data Engineering (TKDE), 2021, 33(4):1802-1806. (CCF A)

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/nlp2024/
  - 课程相关的所有信息： 课件、实验、阅读资料、课程设计等

课程安排是暂定的，可能会更改。所有作业均应在周四上午8点前提交。

Week	Date	Topics	Readings	Assignments
第1周	8月29日	NLP介绍	Advances in natural language processing	
第2周	9月5日	语言模型	J & M 3.1-3.4	作业1
第3周	9月12日	文本分类-朴素贝叶斯	J & M 4.1-4.8	
第4周	9月19日	文本分类-逻辑回归	J & M 4.1-4.8	
第5周 第6周	9月26日 10月10日	向量语义: 1. 词的表示 2. 相似度度量 3. TF-IDF 4. Word2Vec	J & M 6.2-6.4, 6.6 Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors Efficient Estimation of Word Representations in Vector Space	作业2

课程信息

- 本课程评分标准：作业（30%）+实验（20%）+期末考试（50%）
  - 作业： 4次
  - 实验： 4次

- 本课程评分标准：系统实现（40%）+ 设计验收（30%）+ 设计报告（30%）

教材

Speech and Language Processing (3rd ed. draft)

Dan Jurafsky and James H. Martin

Here's our August 20, 2024 release!

Individual chapters and updated slides are below;

Here is a single pdf of Aug 20, 2024 book!

要了解最先进的自然语言处理（NLP）技术的一个好方法是通过研究论文和博客文章。

1. https://aclanthology.org/
2. https://arxiv.org/list/cs.CL/recent

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）？

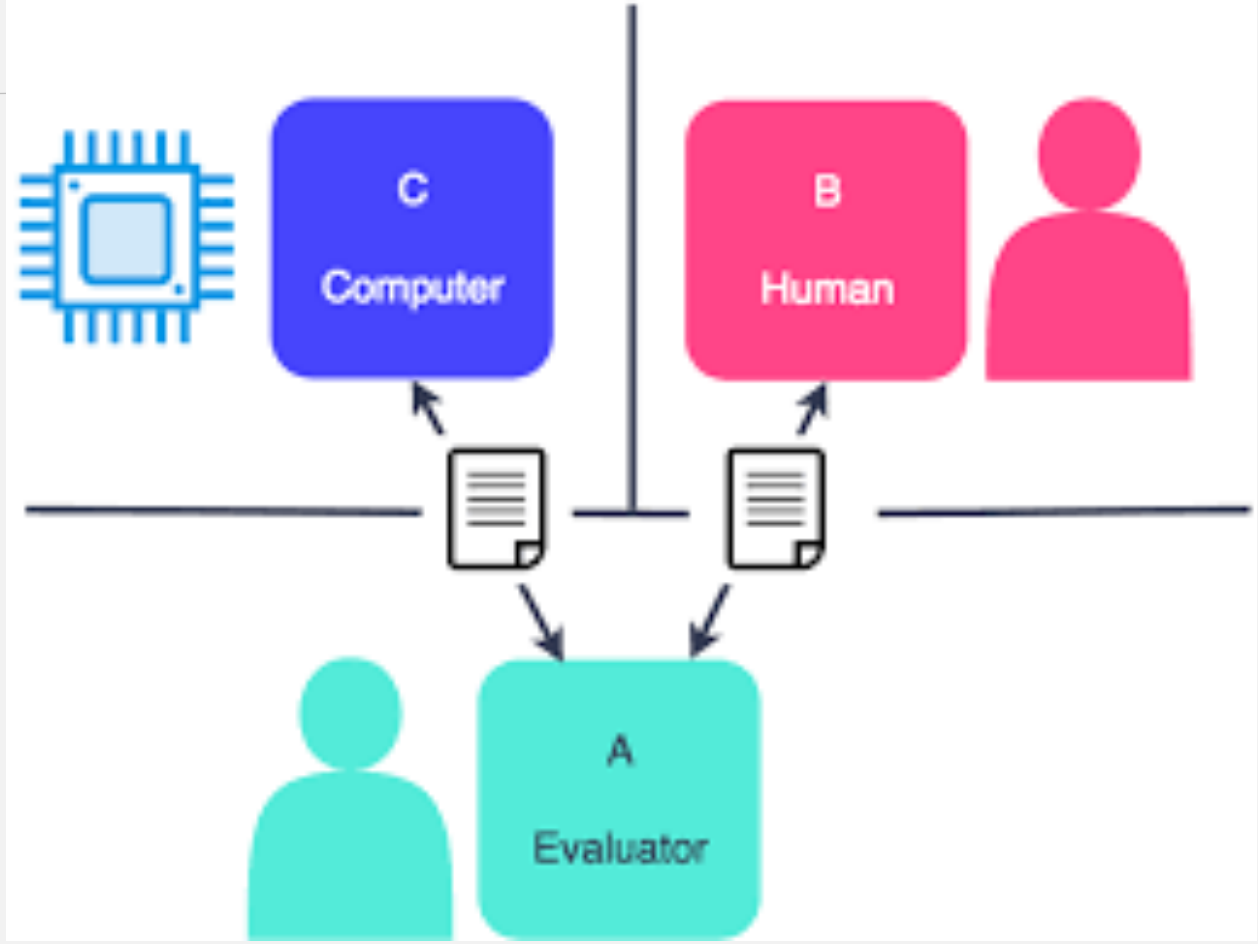
赋予计算机处理人类语言的能力。

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。

人机对话（Conversational agent / Dialogue system）



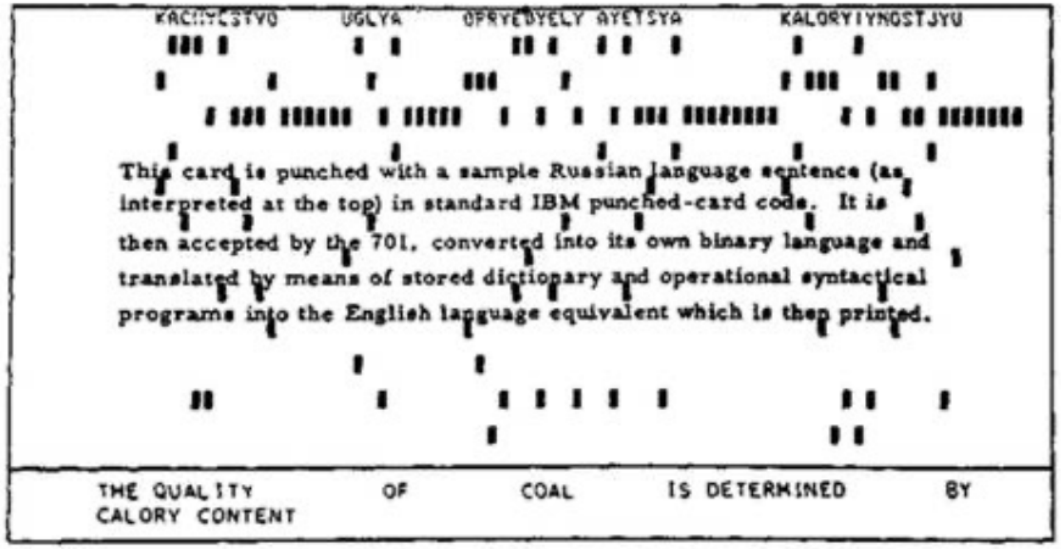
三个参与者：两个人和一台电脑

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

How it started

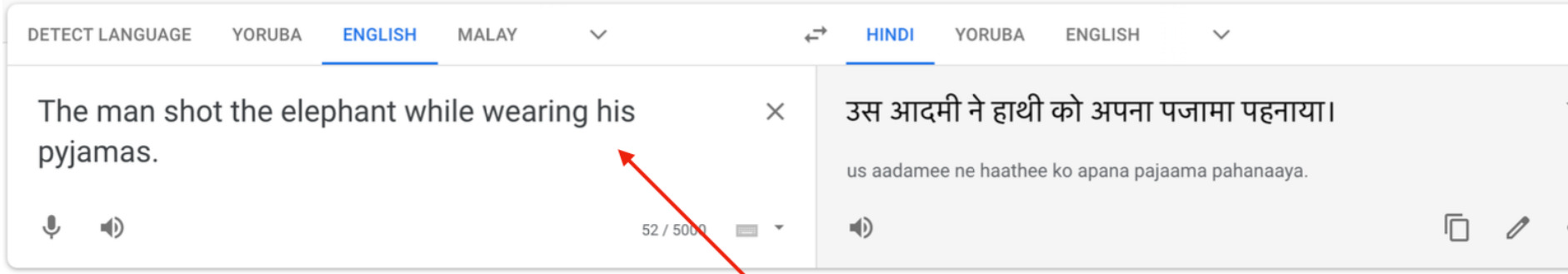


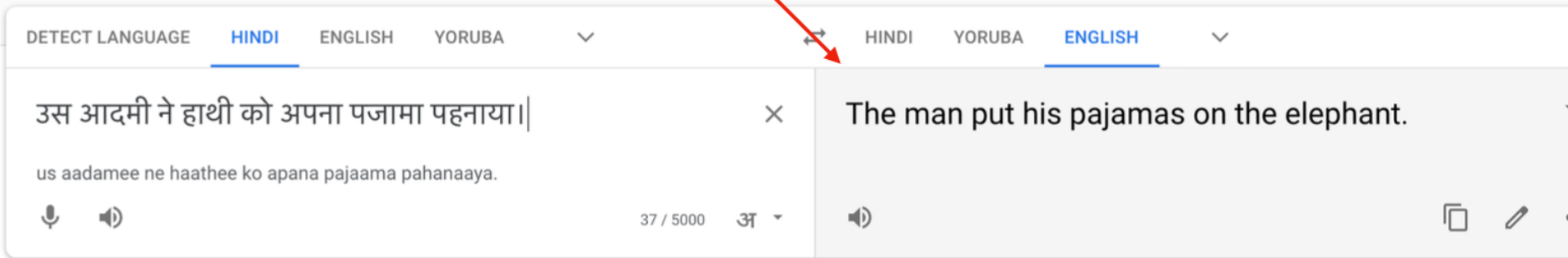
Specimen punched card and below a strip with translation, printed within a few seconds

Georgetown experiment 1954

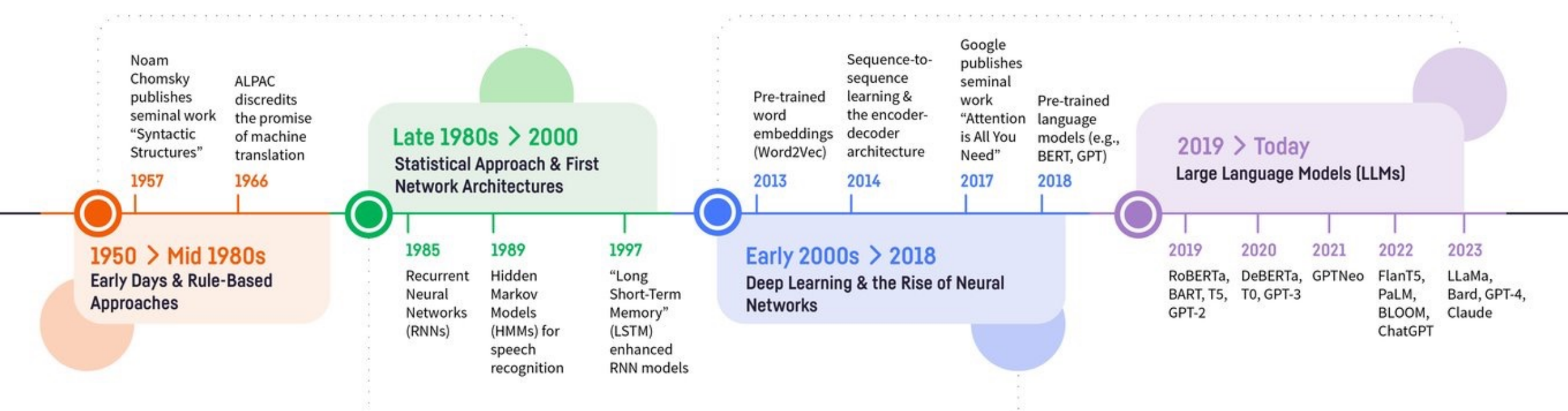
"Within three or five years, machine translation will be a solved problem"

How it's going





NLP发展历程


Noam Chomsky publishes seminal work "Syntactic Structures"
1957
ALPAC discredits the promise of machine translation
1966
1950 > Mid 1980s
Early Days & Rule-Based Approaches
Late 1980s > 2000
Statistical Approach & First Network Architectures
1985
Recurrent Neural Networks (RNNs)
1989
Hidden Markov Models (HMMs) for speech recognition
1997
"Long Short-Term Memory" (LSTM) enhanced RNN models
Pre-trained word embeddings (Word2Vec)
2013
Sequence-to-sequence learning & the encoder-decoder architecture
2014
Google publishes seminal work "Attention is All You Need"
2017
Pre-trained language models (e.g., BERT, GPT)
2018
Early 2000s > 2018
Deep Learning & the Rise of Neural Networks
2019 > Today
Large Language Models (LLMs)
2019
RoBERTa, BART, T5, GPT-2
2020
DeBERTa, T0, GPT-3
2021
GPTNeo
2022
FlanT5, PaLM, BLOOM, ChatGPT
2023
LLaMa, Bard, GPT-4, Claude

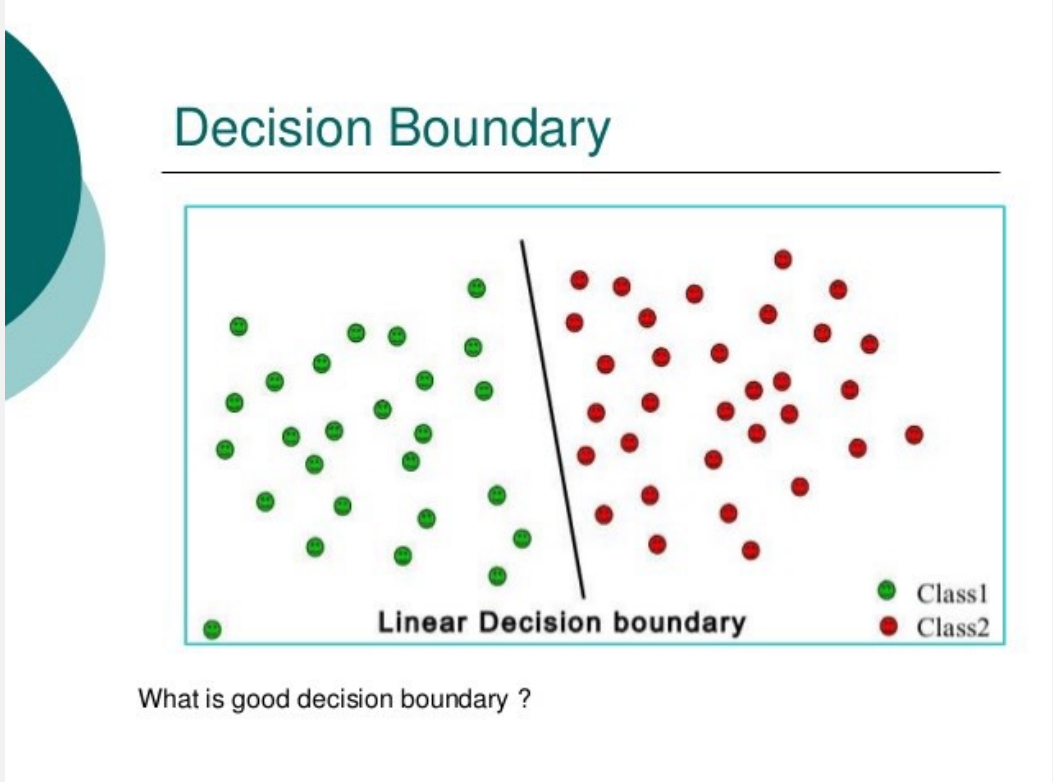
NLP发展历程

- **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”

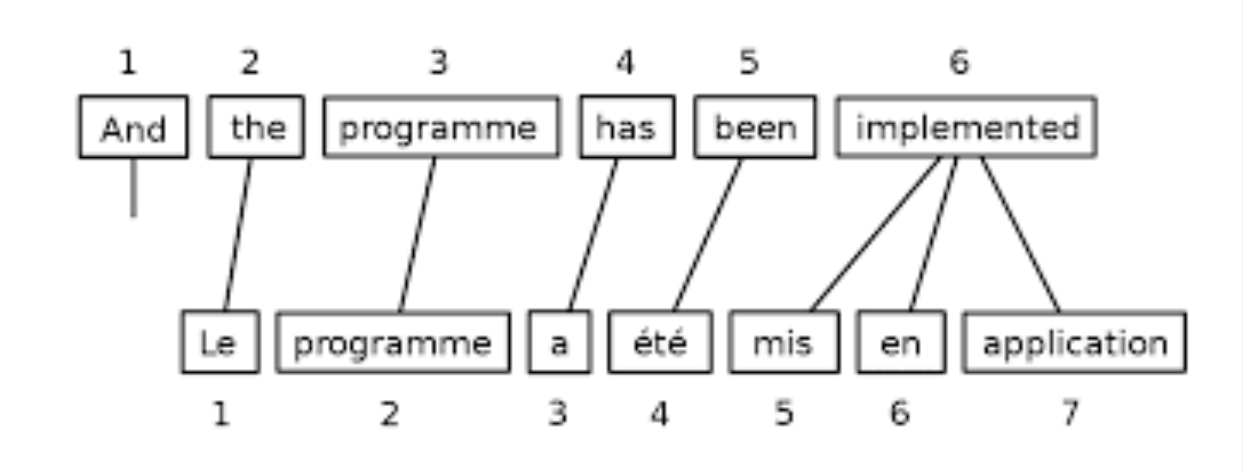
NLP发展历程

- **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

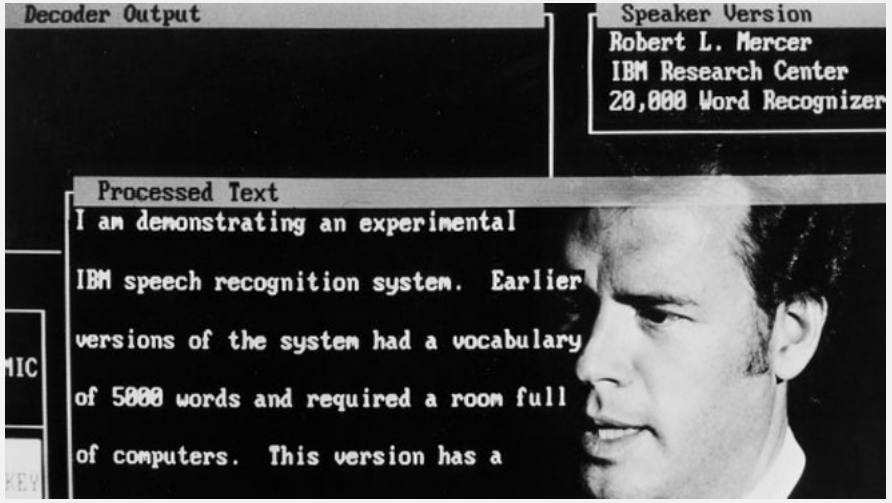


基于统计的方法

IBM的机器翻译模型



语音识别

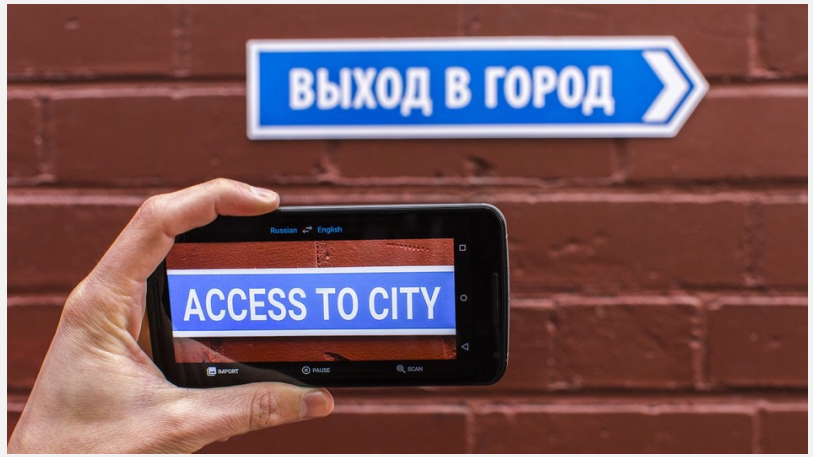


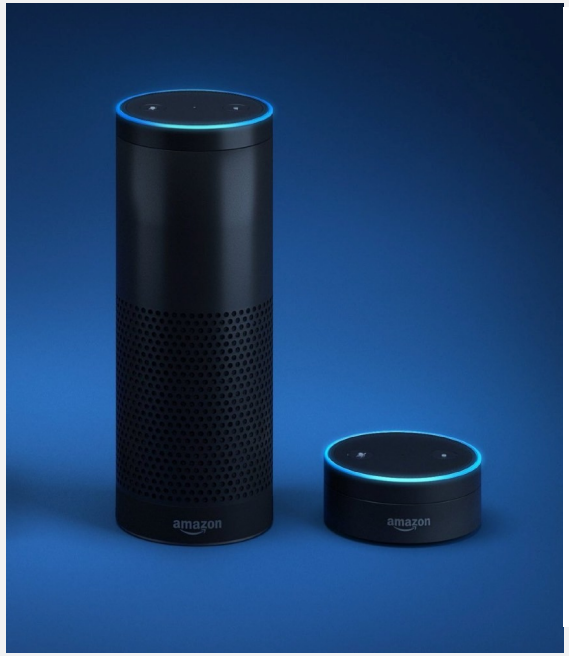
Anytime a linguist leaves the group the (speech) recognition rate goes up.

- 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





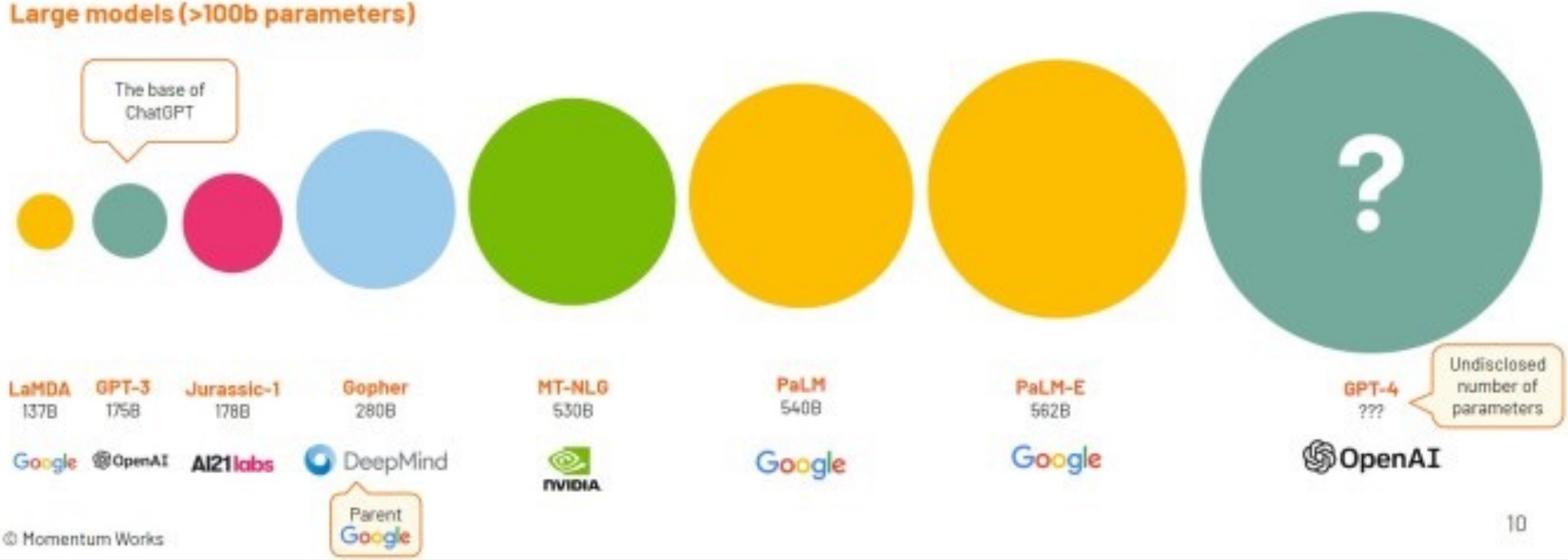
NLP发展历程

- **大模型时代**
  - 单个模型能够解决许多任务
  - 早期的通用人工智能

Large Language Models are becoming very large indeed

Small models (<= 100b parameters)

Large models (>100b parameters)



© Momentum Works

NLP发展历程

what nlp topics should i teach in 2024? give a concise answer.

In 2024, focus on these key NLP topics:

1. **Foundations:** Text preprocessing, word embeddings, syntactic/semantic analysis.
2. **Deep Learning:** Transformer architectures, fine-tuning pretrained models (BERT, GPT).
3. **Advanced Techniques:** Large language models (LLMs), multimodal NLP, explainability.
4. **Applications:** Text simplification, sentiment analysis, machine translation, text generation.
5. **Domain-Specific NLP:** Low-resource languages, biomedical NLP, legal/financial NLP.
6. **Ethics:** Bias in models, ethical AI considerations.
7. **Emerging Trends:** LLM research, conversational AI, NLP for knowledge graphs.

NLP发展历程

what do you know about Jipeng Qiang

Jipeng Qiang is a researcher known for his work in natural language processing (NLP), particularly in areas such as text simplification, sentiment analysis, and deep learning-based NLP methods. He has contributed to research in automatic text simplification, including exploring techniques like lexical simplification and sequence-to-sequence models for simplifying text. His work often focuses on developing and improving algorithms that can understand and generate human language more effectively.

Would you like more specific information about his research or contributions?

CCTV 13
新 闻
人工智能大模型
将如何改变世界？

Part.03

为什么语言很难理解?

为什么语言很难理解?

- 歧义
- 方言
- 口音
- 听者必须推断 - 语用学
- 幽默、讽刺、反讽
- 上下文、依赖关系

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.

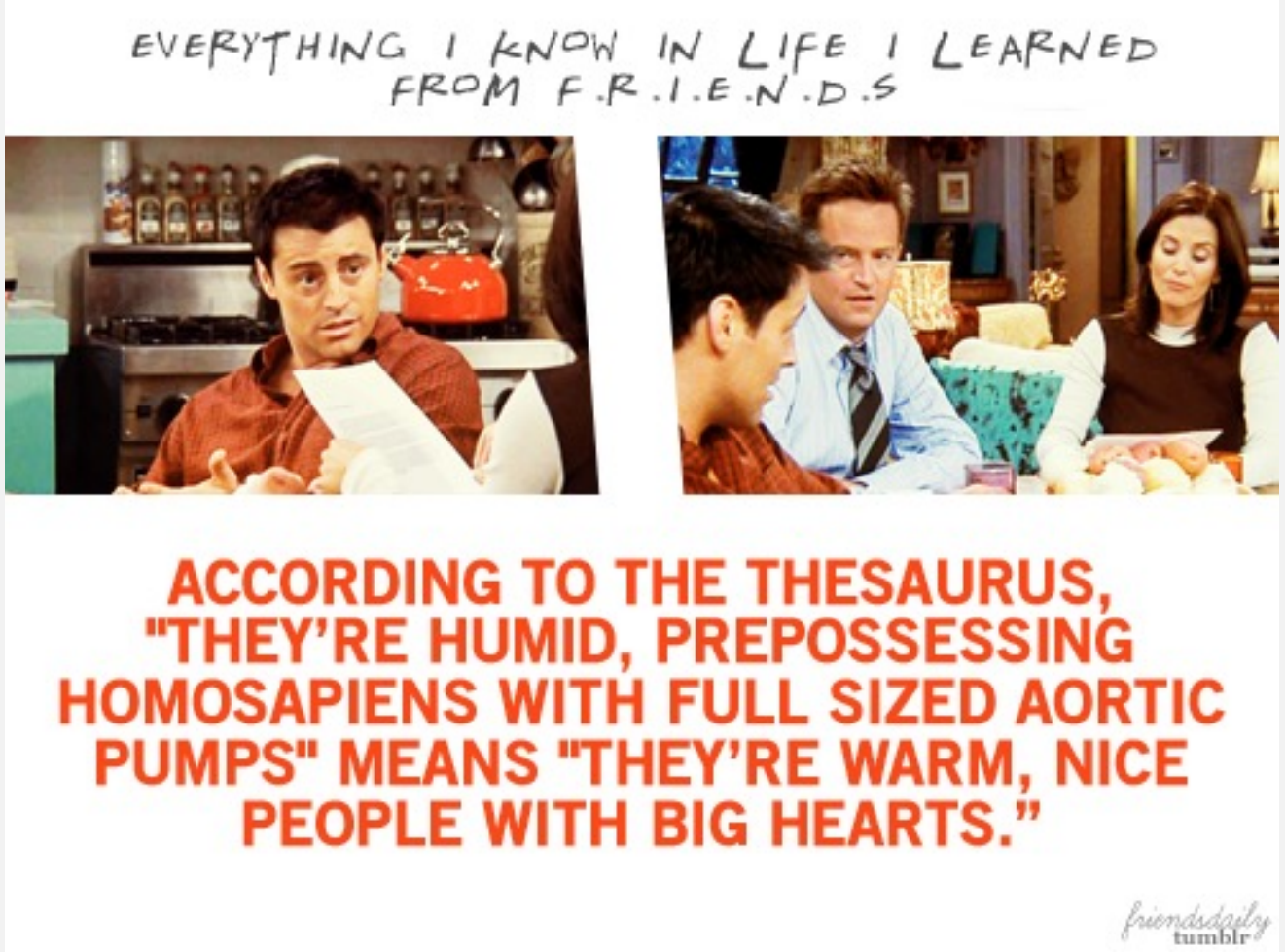
词义消岐（Word sense disambiguation ）

词语的多样性

多个词语具有相同的意思

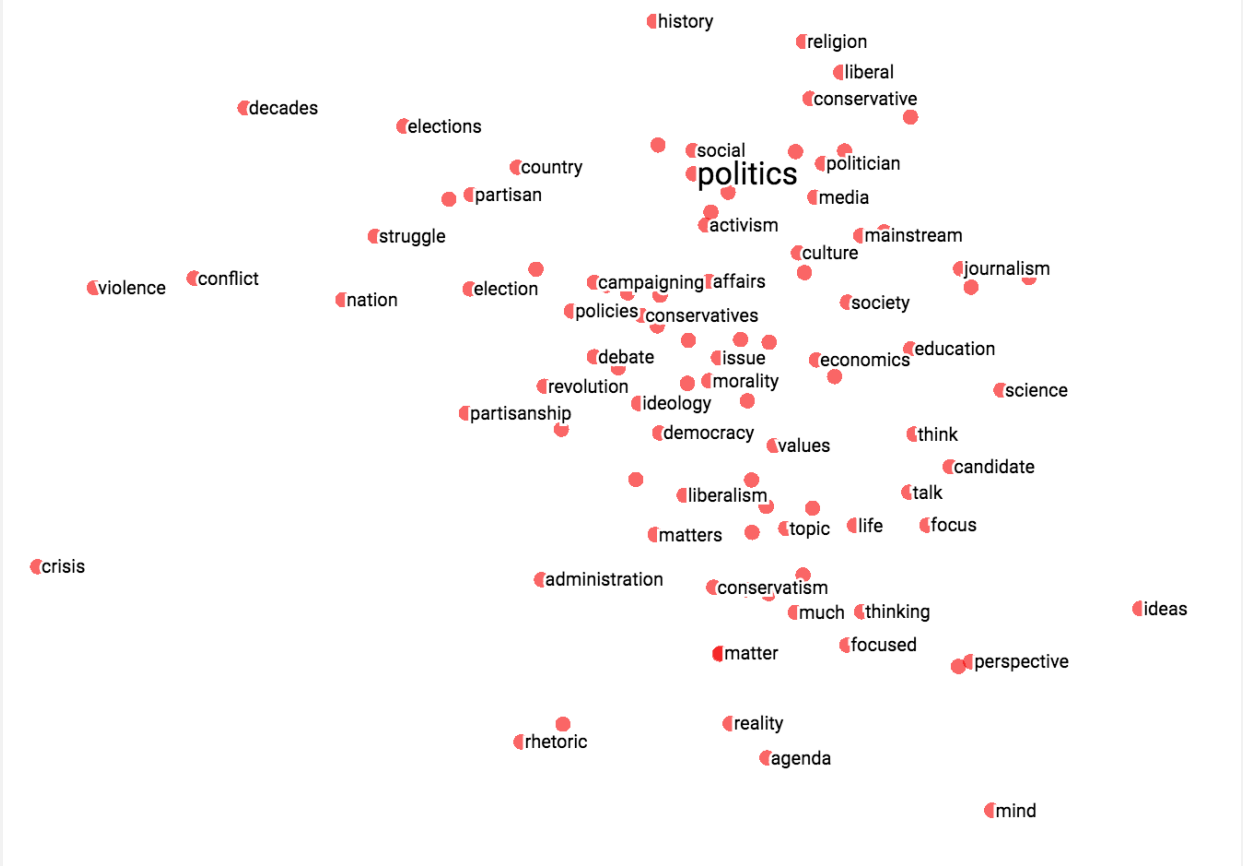
they’re humid, prepossessing homosapiens with full-sized aortic pumps’

'they’re warm, nice people with big hearts.'"



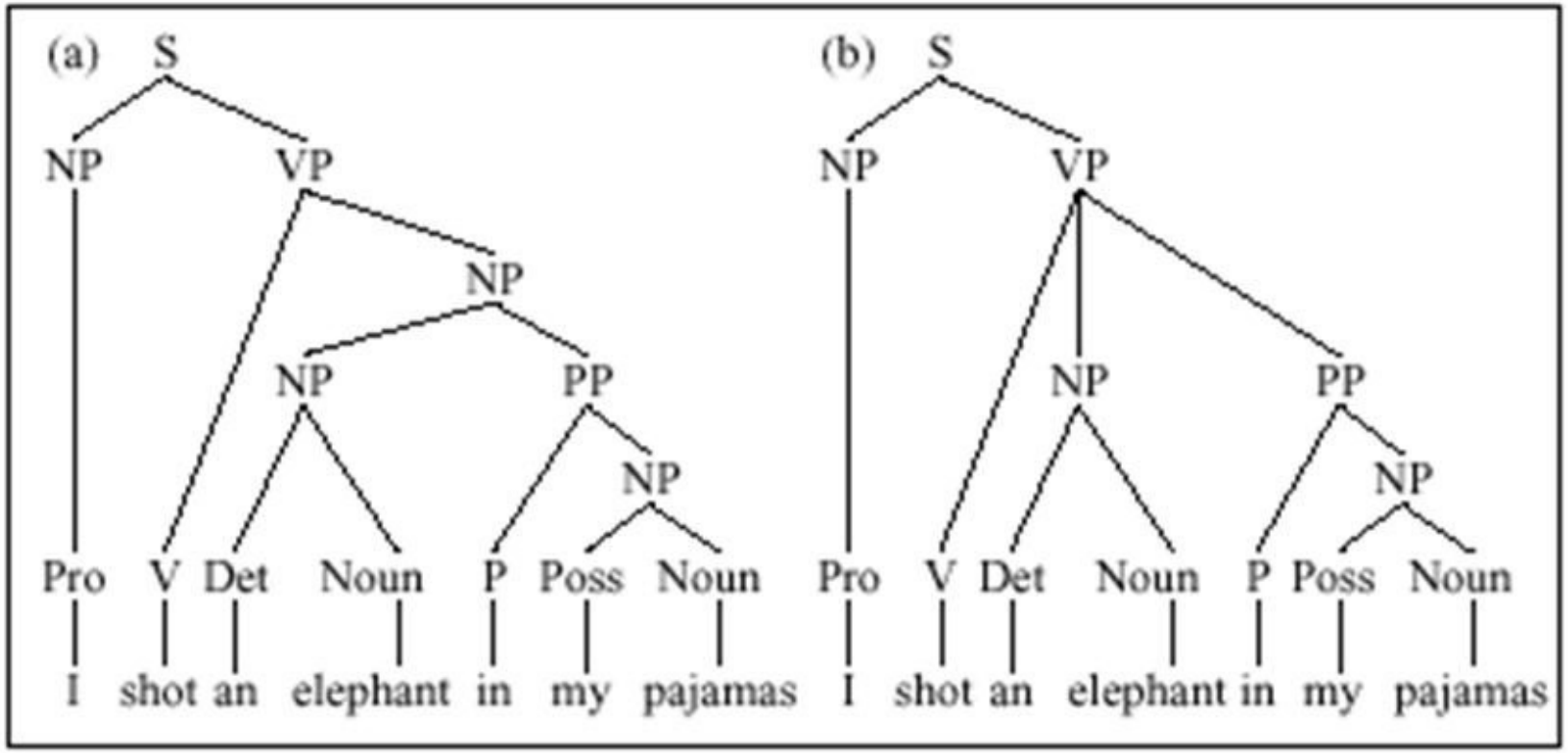
词语的分布式表示

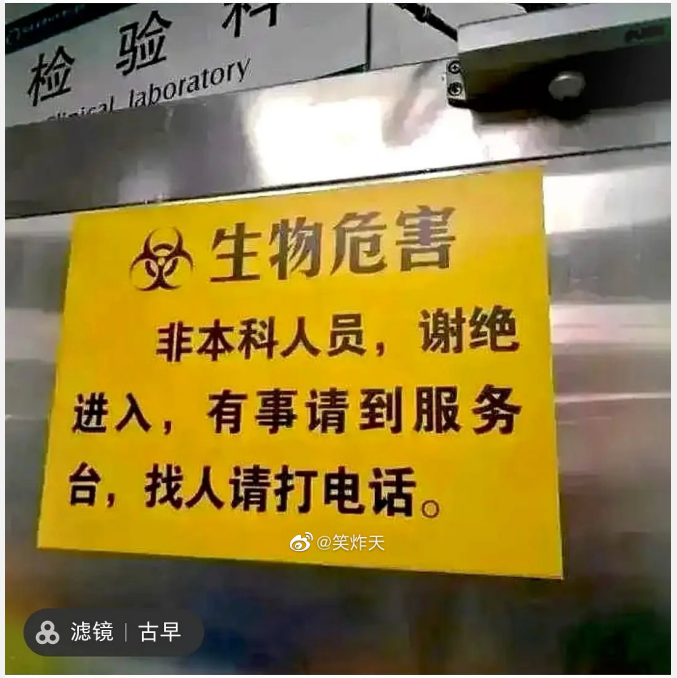
将单词投影到连续向量空间



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！！！

说点什么... 77 4 891

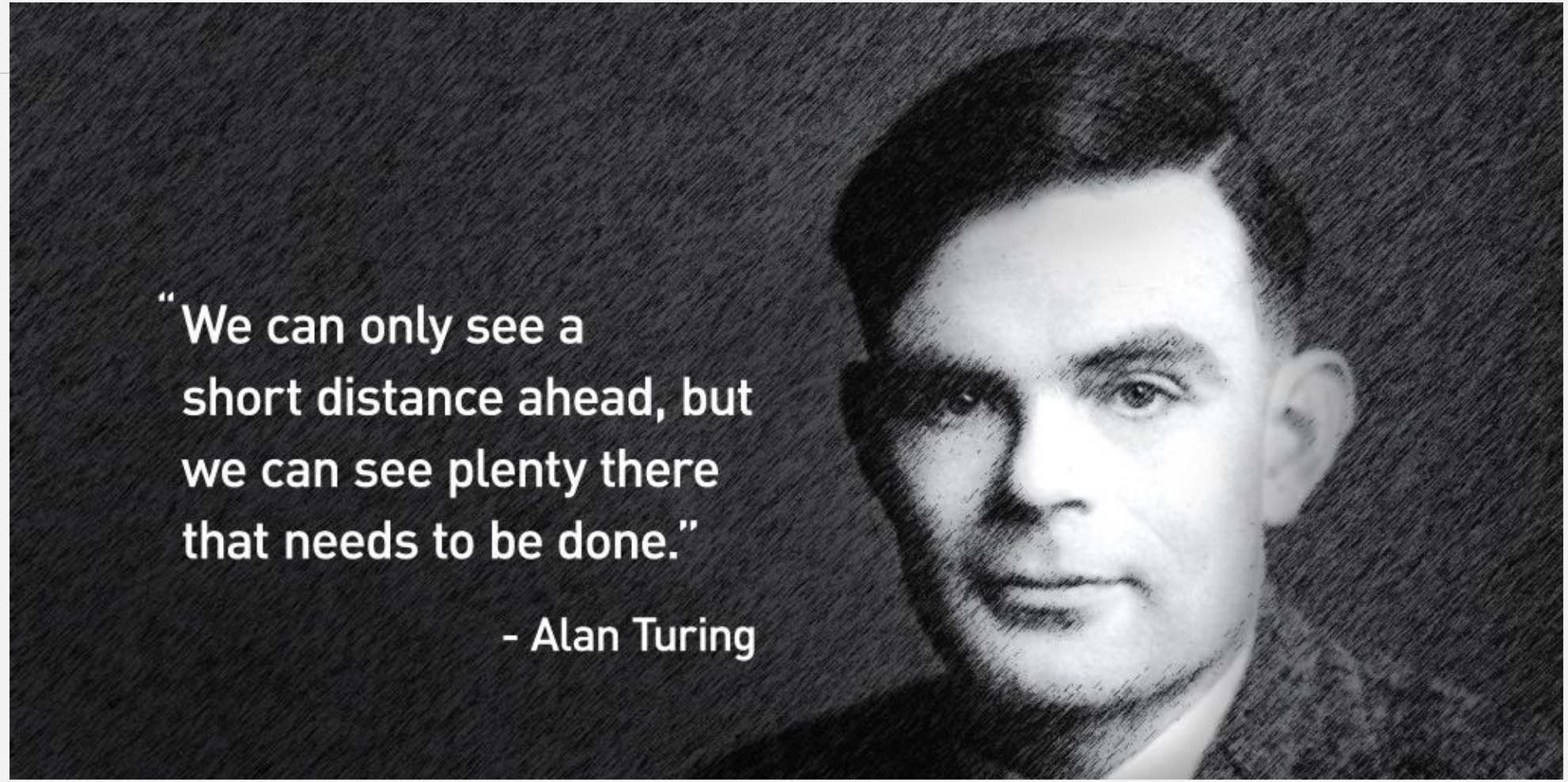
@半糖小哥

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

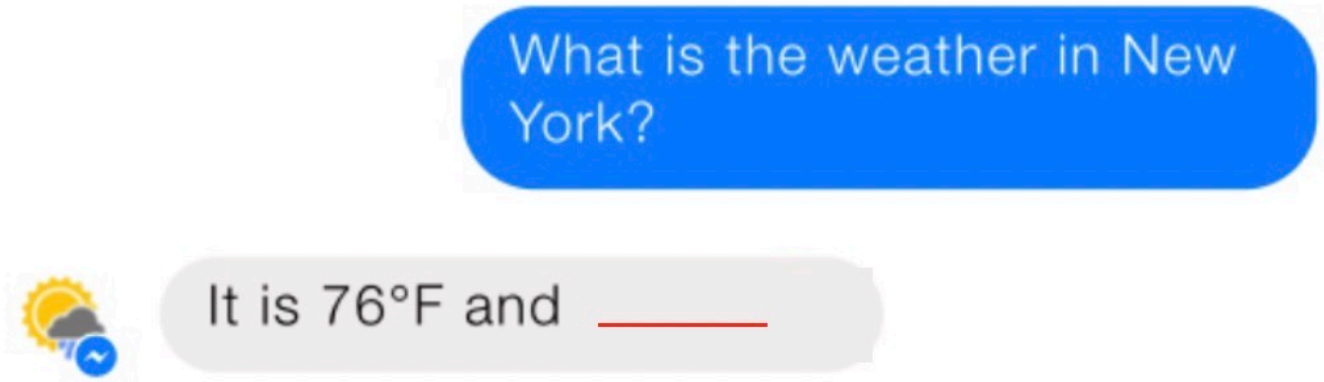
Part.05

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)
6. Transformer
7. 大语言模型

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

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

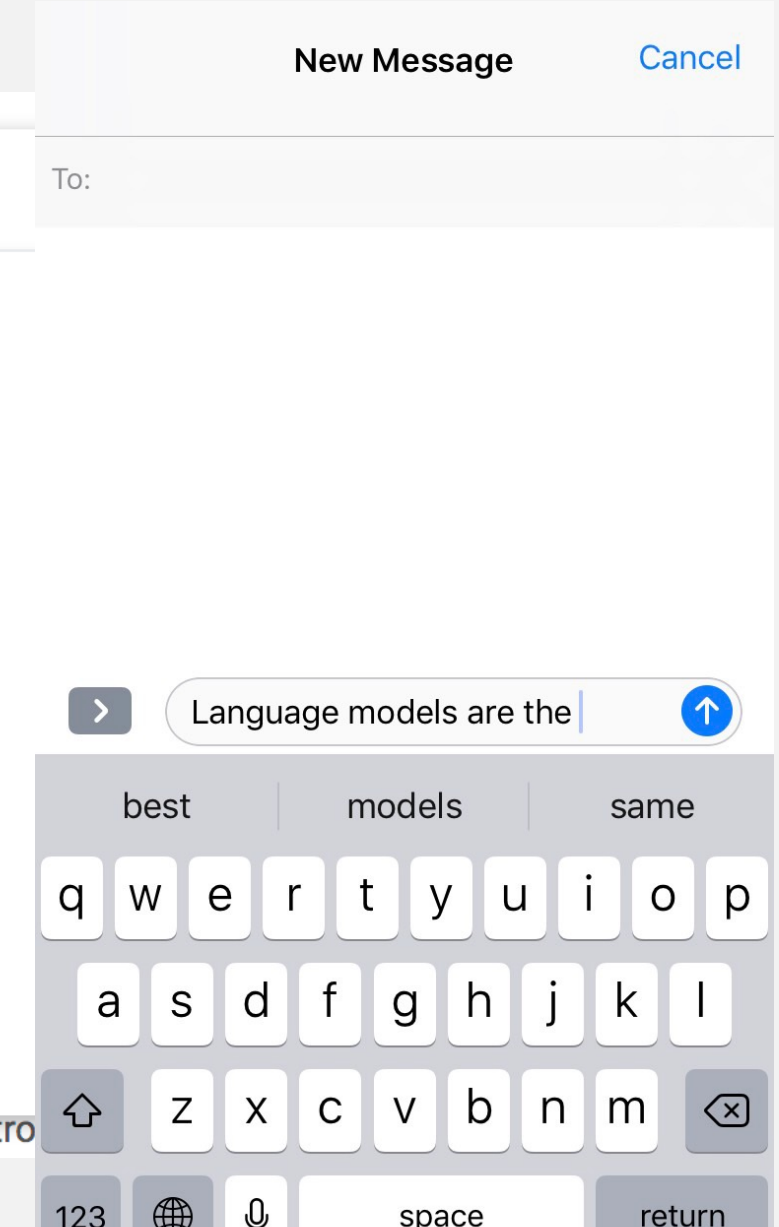
相关的任务：计算下一个单词的概率

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

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