南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

STA-5007: Advanced Natural Language Processing

Lecture 1: Introduction

陈冠华 CHEN Guanhua

Department of Statistics and Data Science

Biography

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

- Assistant professor in Stat-DS
  - Email: chengh3@sustech.edu.cn
  - Office: Business Building 319
- PhD from HKU, Bachelor & Master from Tsinghua
- Large language models and natural language processing
  - Data synthesis, reasoning LLMs
  - LLM-based agents, multimodal LLMs

Biography

- Why faculty?
- Why SUSTech?
- What is your motivation?

Frontiers of LLMs and NLP

- Better models in terms of efficacy, efficiency, safety, robustness
  - How to define ‘better’
- Extension to multimodalities like image, video, audio, graph, 3D, time series, actions
- AI for science/engineering, AI + X

- “人工智能+”科学技术
- “人工智能+”产业发展
- “人工智能+”消费提质
- “人工智能+”民生福祉
- “人工智能+”治理能力
- “人工智能+”全球合作


中华人民共和国中央人民政府
www.gov.cn

首页 | 简 |

首页 > 信息公开 > 国务院文件 > 科技、教育 > 科技

字号：默认 大 超大 | 打印 收藏 留言

索 引 号： 000014349/2025-00070
主题分类： 科技、教育\科技
发文机关： 国务院
成文日期： 2025年08月21日
标 题： 国务院关于深入实施“人工智能+”行动的意见
发文字号： 国发〔2025〕11号
发布日期： 2025年08月26日

国务院关于深入实施“人工智能+”行动的意见

国发〔2025〕11号

各省、自治区、直辖市人民政府，国务院各部委、各直属机构：

国务院关于深入实施“人工智能+”行动的意见_科技_中国政府网

Course Information

- STA 5007: Advanced natural language processing
  - 3 credits, major elective course
  - Tuesday 3-4 (even week, L1-111) and Thursday 9-10 (every week, L3-206)
- Open office hour
  - Thursday 15:00-17:00, appoint or after class
- Teaching assistant:
  - 阮志文 Email: 12431111@mail.sustech.edu.cn
  - 李乙侠 Email: liyx2023@mail.sustech.edu.cn
  - 赖鹏 Email: 12432270@mail.sustech.edu.cn
  - 郑剑杰 Email：12432284@mail.sustech.edu.cn

Grading

- 5 minutes' sharing, 5%
- Quiz (5%)
- Homework, 40%
  - Four assignments x 10%
- Project, 25%
- Final exam, 25%

5 Minutes' Sharing

Everything related to NLP

- Interesting news, models, software
- Efficiency tools
- Views and insights
- Please add your references

- Start from week 3
- PPT should be sent to me by email before the weekend
- Sampled students will share in the class, while the rest will submit their ppt via BB system.

References

- Weibo '爱可可爱生活', '宝玉xp'
- Wechat Subscription Account 'PaperWeekly' '深度学习自然语言处理', '机器之心', '量子位', '新智元', etc.
- 智源社区
- Hacker News

Reference Books

- Speech and Language Processing (3rd draft) [download]
- 《大语言模型》, 赵鑫等, 高等教育出版社，2024
- 《大规模语言模型-从理论到实践》，张奇等，电子工业出版社，2024

A good way to learn about state-of-the-art NLP concepts is through research papers and blog posts

Is it necessary to learn in the classroom?

What is your core strength in the LLM era?

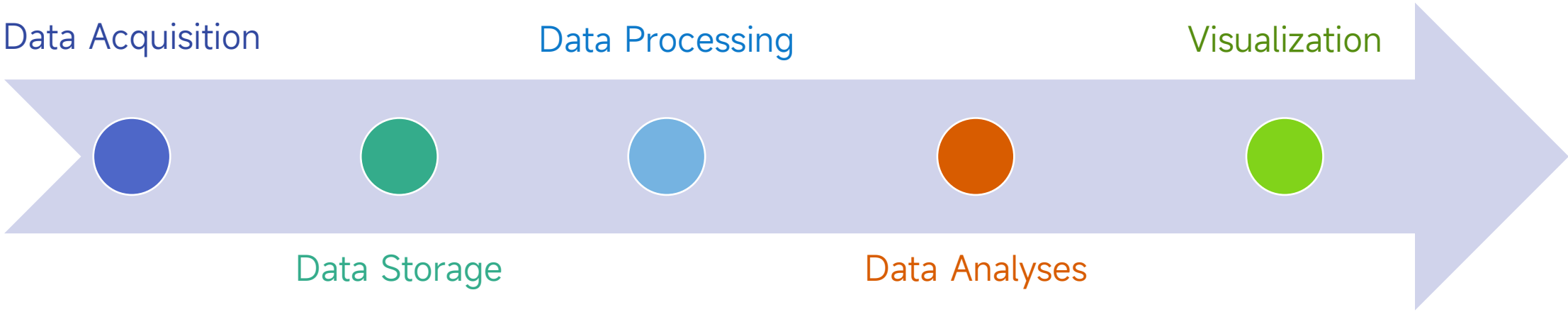
Course Survey

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

- Permanent link
  - https://www.wjx.cn/vm/m9KeRy0.aspx#
  - Anonymous
  - Please finish before 21:00, Sep 15
- Any suggestions, questions or problems

Course Goals

In Data Science, we focus on the following steps of data journey



★ Our focus in this course

Natural Language Processing

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

- Natural language
  - English, Chinese, Spanish, Hindi, etc.
  - Spoken, written, sign language
  - Language-centric or vision-centric world model
- Natural language processing
  - Building computer programs to analyze, understand and generate human language
  - Acquisition of these capabilities from knowledge and data
- An interdisciplinary field
  - Computer Science, Statistics
  - Linguistics, Psychology, Cognitive Science, etc.
  - NLP ≈ computational linguistics
  - NLP = LLM ?

Some concepts

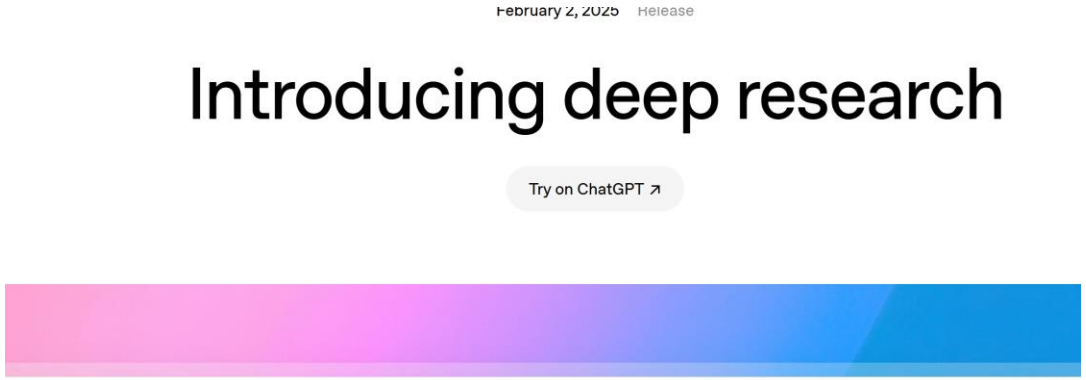
- AI
- Machine learning
- Deep learning
- NLP/CV/graphics/speech/ robotics/time-series, etc.

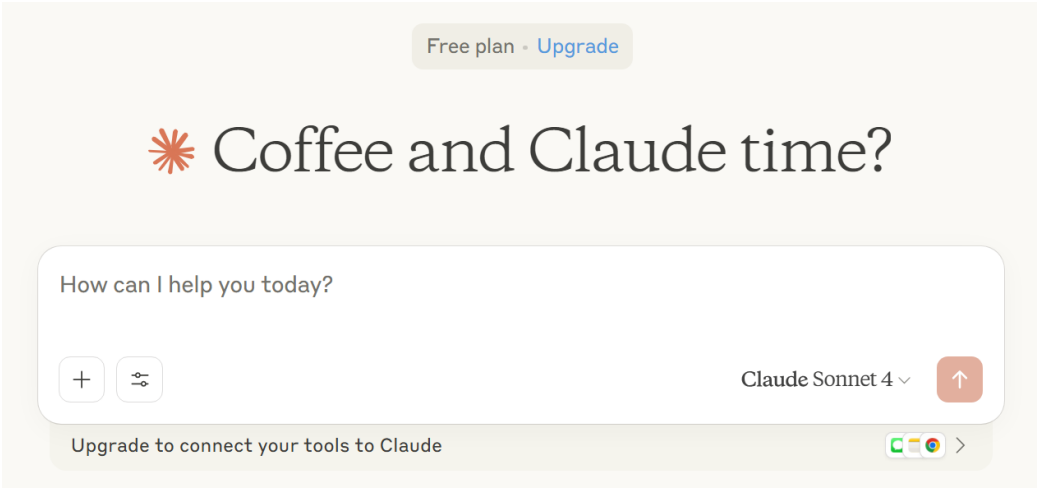
Any exciting NLP applications?

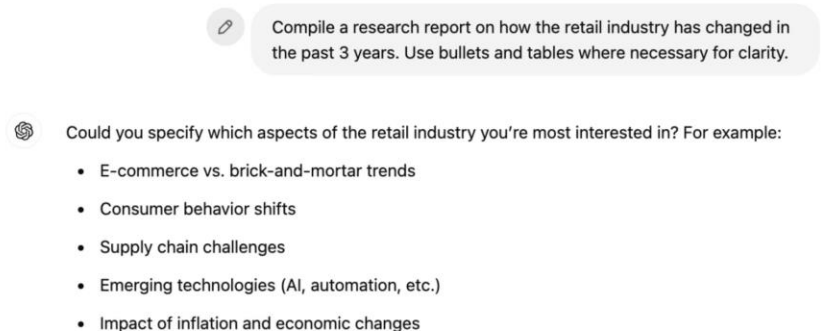
Natural Language Processing

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

- Aid human-machine communication







Natural Language Processing

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

- Aid human-human communication

Machine Translation

Emotional Agents

grammarly

Grammatical Error Correction

Speech-to-text

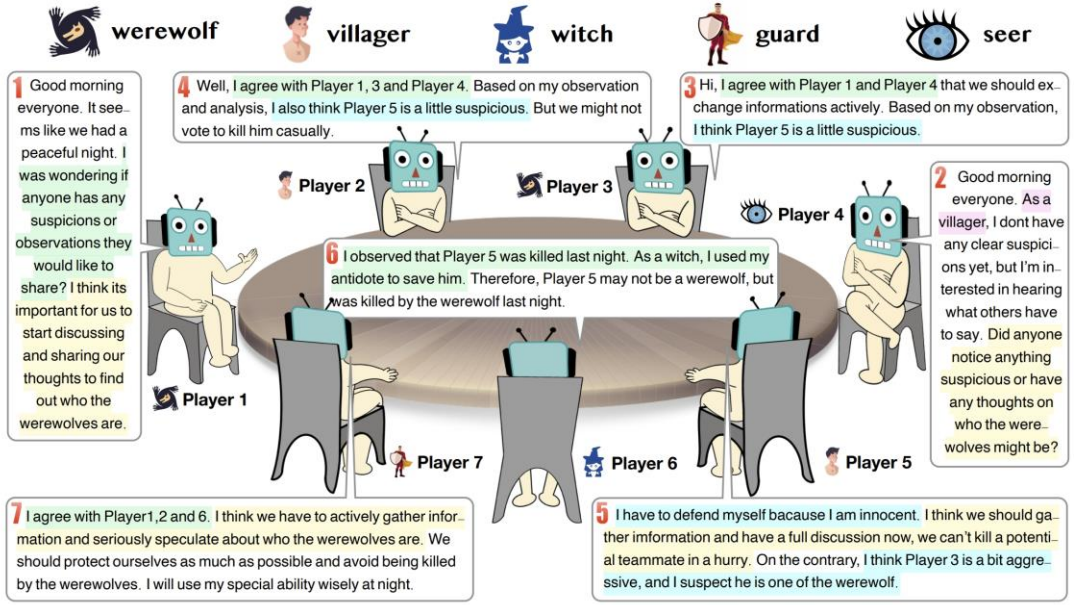
Natural Language Processing

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

- Aid machine-machine communication?

Stanford SmallVille

https://arxiv.org/abs/2304.03442



Natural Language Processing

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Natural language VS. Formal language

- Natural Language evolves naturally through use and repetition by humans for communication.
- Formal Language is designed by humans for specific, precise applications, such as programming language, mathematical notation, musical notations

```
import data.rat.basic
       data.nat.parity
       tactic

lemma even_if_square_even {n : ℕ} (hn2 : 2 ∣ (n*n)) : 2 ∣ n :=
begin
  by_contra hc,
  have hmod2 : n % 2 = 1, from nat.not_even_iff.mp hc,
  set k := n / 2 with hk,
  have hn : n = 1 + 2*k,
  { rw [←nat.mod_add_div n 2, hmod2] },
  have hnn : n*n = 1 + 2*(2*k + 2*k*k),
  { rw hn, ring },
  rw [nat.dvd_iff_mod_eq_zero, hnn] at hn2,
  norm_num at hn2,
end
```

Mathematics in Lean — Mathematics in Lean v4.19.0 documentation

Formal Language

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Seed Prover IMO 2025

Seed Prover solved 4 out of 6 problems in IMO 2025 durint the context, with the following breakdown:

- **Day 1**: Fully solved P2 (geometry) and P3 (number theory), fully solved P1 (combinatorics) after the compe
- **Day 2**: Fully solved P4 (number theory) and P5 (combinatorics / algebra)

Details

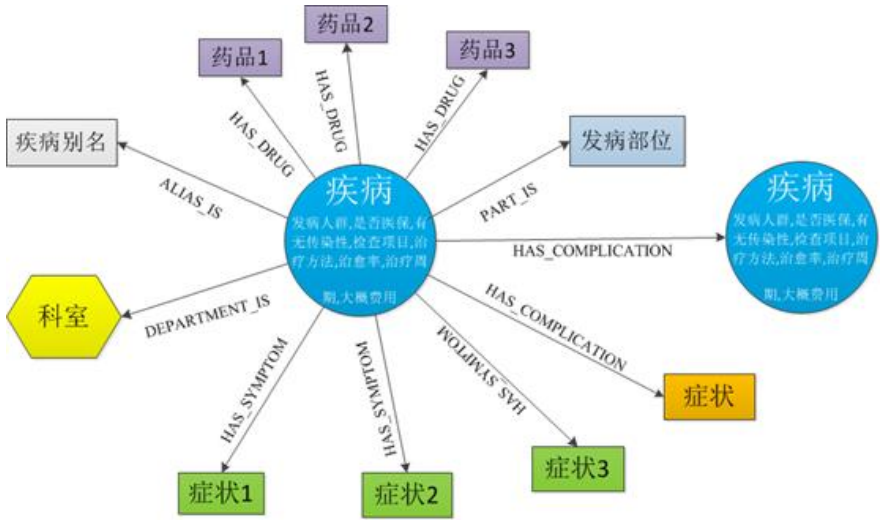
- P1 (combinatorics) Lean : Fully proved after the competition, this is not scored by the IMO.
- P2 (geometry) NL : Generated and verified in 2 seconds using Seed-Geometry system
- P3 (number theory) NL Lean: Solved in 3 days, with a 2000-line formal proof
- P4 (number theory) NL Lean: Solved in 3 days, with a 4000-line formal proof
- P5 (combinatorics / algebra) NL Lean: Solved in 1 day, with a proof slightly different from known human solutions

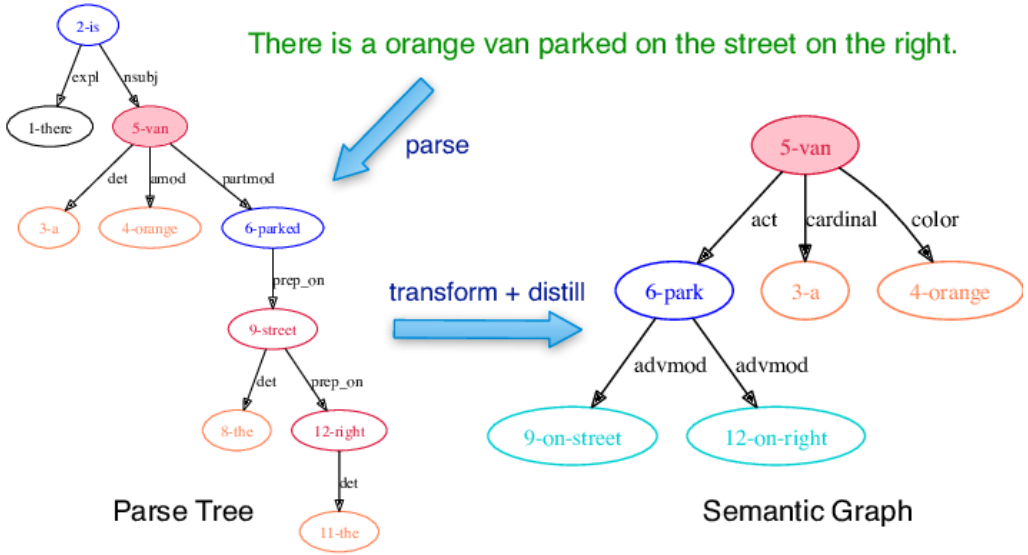
P1,3,4,5 are compiled under Lean v4.14.0.

ByteDance-Seed/Seed-Prover

Natural Language Processing

- Analyze and understand languages
  - Semantic parsing
  - Text classification
  - Named entity recognition
  - Relation recognition





- Semantic parsing converts natural language to
  - Logical form
  - Code
  - Structure data like json/xml

NLP Tasks

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Understanding	Generation	Analyses
Text classification	Machine translation	Part-of-speech tagging
Information extraction	Question answering	Dependency parsing
Named entity recognition	Text summarization	Constituency parsing
Relation recognition	Dialogue	Lexical normalization
Search engine	Data-to-Text Generation	Word sense disambiguation
Recommendation system	Grammatical error correction	.....
.....	.....	

NLP-progress

Application vs. Task ?

Discriminative vs. generative ?

NLP Tasks

Pre-LLM era VS. LLM era

- Knowledge-intensive tasks
- Reasoning-intensive tasks

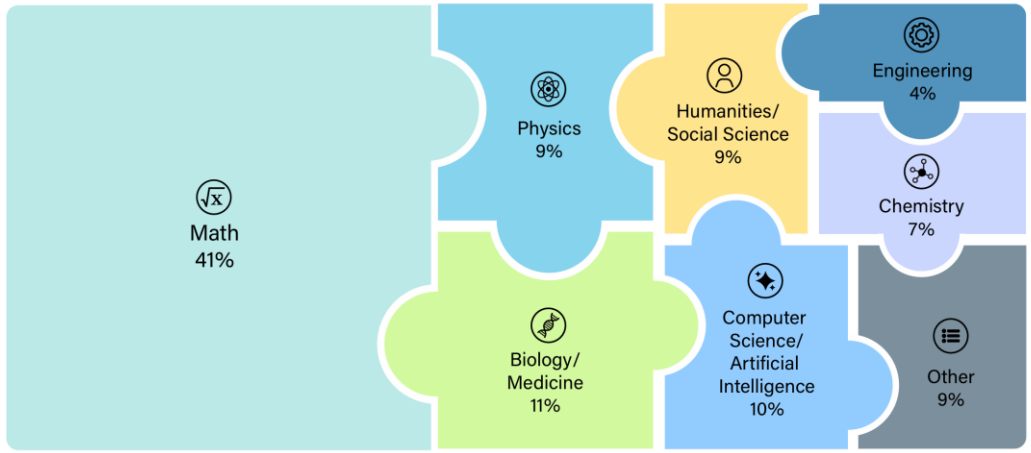
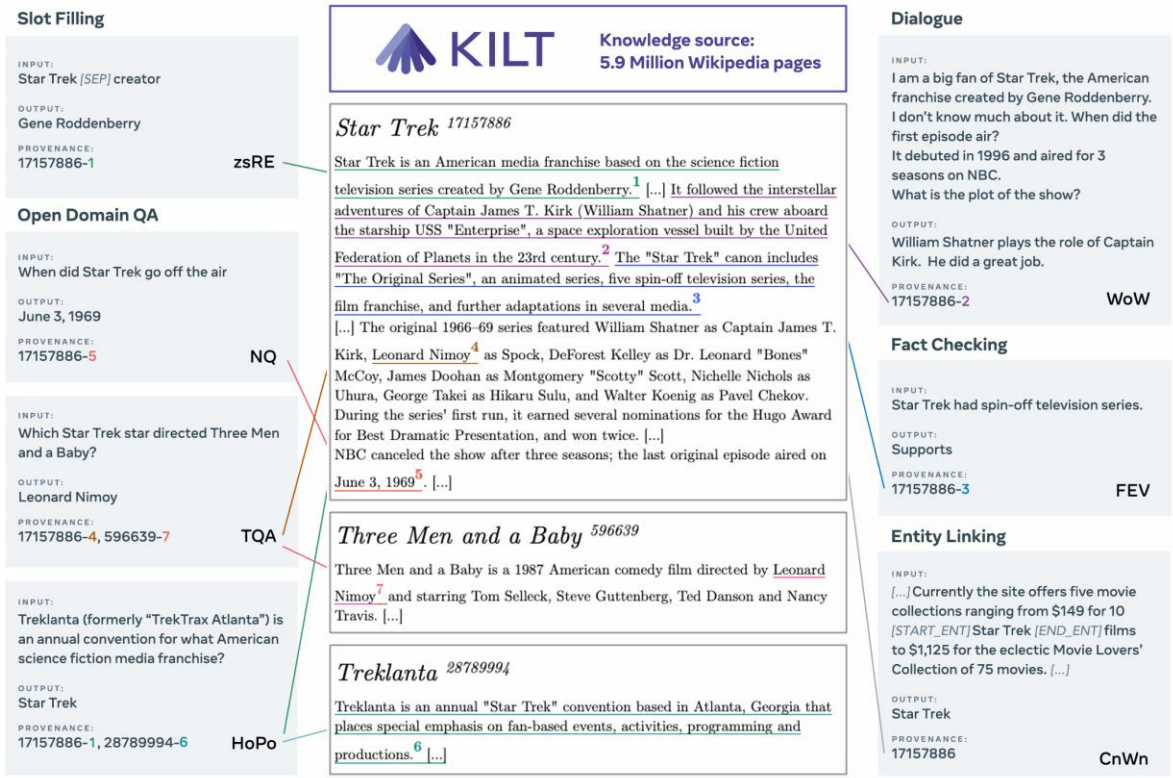


Figure 3: HLE consists of 2,500 exam questions in over a hundred subjects, grouped into high level

[Humanity's Last Exam]



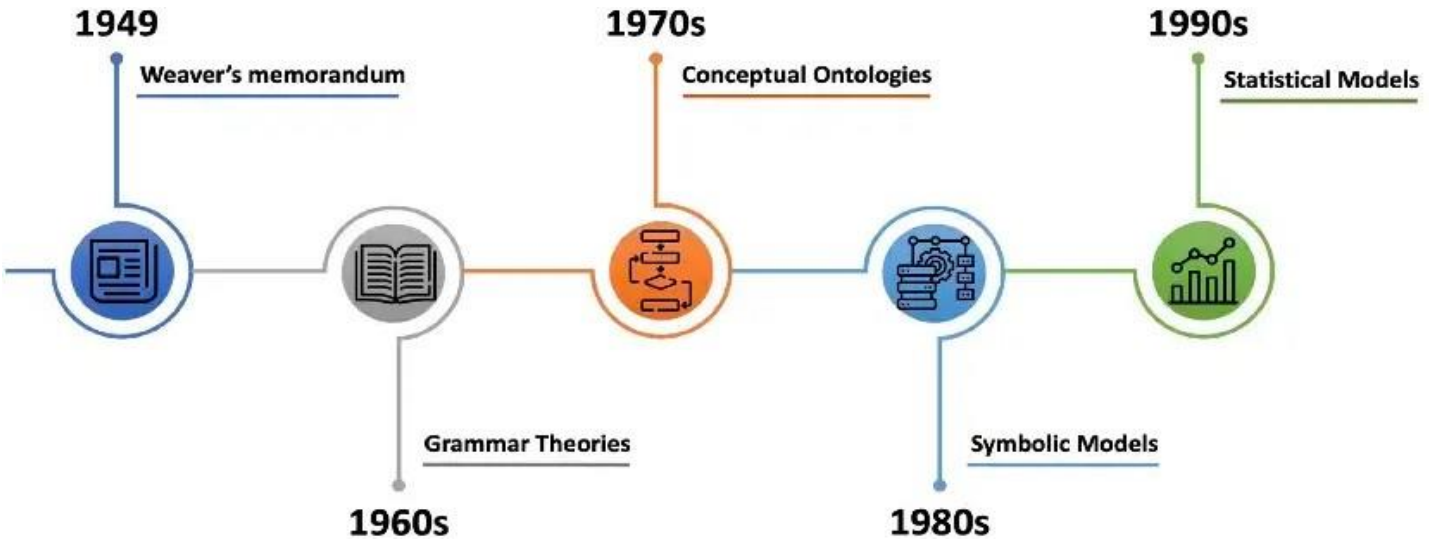
[KILT: a Benchmark for Knowledge Intensive Language Tasks - ACL Anthology]

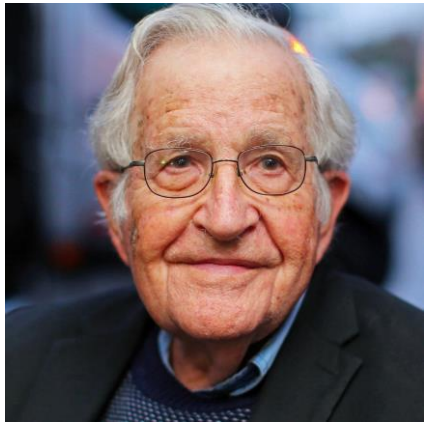
Syllabus

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Week	Date	Name	Lecture	HomeWork	Week	Date	Name	Lecture	HomeWork
1	9.11	Lecture 1	Introduction to NLP		9	11.6	Lecture 13	RAG	Quiz
2	9.16	Lecture 2	Deep learning basics		10	11.11	Lecture 14	LLM Evaluation	
2	9.18	Lecture 3	Language Model		10	11.13	Lecture 15	Efficient LLM Inference	
3	9.25	Lecture 4	Transformer model (1)	HW1 release	11	11.20	Lecture 16	LLM agent (1)	
4	9.30	Lecture 5	Transformer model (2)		12	11.25	Lecture 17	LLM agent (2)	HW4 release
4	10.2	Lecture 6	Skip due to holiday		12	11.27	Lecture 18	Scaling Law	
5	10.9	Lecture 7	Pretrained language model		13	12.4	Lecture 19	Reasoning LLMs	
6	10.14	Lecture 8	Learn to do research	HW2 release	14	12.9	Lecture 20	Multimodal AI	
6	10.16	Lecture 9	Introduction to LLM		14	12.11	Lecture 21	Guest talk (2)	
7	10.23	Lecture 10	LLM coding example	Project release	15	12.18	Lecture 22	NLP Future Trend	
8	10.28	Lecture 11	Guest talk (1)		16	12.23	Lecture 23	Presentation for project	
8	10.30	Lecture 12	LLM Data Synthesis	HW3 release	16	12.25	Lecture 24	Presentation for project	

A Brief History of NLP





Avram Noam Chomsky

https://medium.com/@antoine.louis/a-brief-history-of-natural-language-processing-part-1-ffbcb937ebce

那些年，不得不提的韦弗备忘录(Weaver Memorandum)

知乎：如何系统地学习乔姆斯基的语言学理论？

Context-Free Grammar

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

- A set of formal rules used to describe the structure of strings in a formal language
  - Any non-terminal symbol (a placeholder for a syntactic structure) can be replaced (or "expanded") using a grammar rule regardless of the symbols surrounding it.

- **Non-terminals (N):** `{<S>, <NP>, <VP>, <N>, <V>, <Det>}`
- **Terminals (Σ):** `{the, a, cat, dog, man, loves, feeds, sees}`
- **Start Symbol (S):** `<S>`
- **Production Rules (P):**
  1. `<S> -> <NP> <VP>` // A sentence is a Noun Phrase followed by a Verb Phrase.
  2. `<NP> -> <Det> <N>` // A Noun Phrase is a Determiner followed by a Noun.
  3. `<VP> -> <V> <NP>` // A Verb Phrase is a Verb followed by a Noun Phrase.
  4. `<Det> -> the | a` // A Determiner is "the" or "a".
  5. `<N> -> cat | dog | man` // A Noun is "cat", "dog", or "man".
  6. `<V> -> loves | feeds | sees` // A Verb is "loves", "feeds", or "sees".

1. `<S>`
2. `<NP> <VP>` (Rule 1)
3. `<Det> <N> <VP>` (Rule 2)
4. `the <N> <VP>` (Rule 4: chose `the`)
5. `the cat <VP>` (Rule 5: chose `cat`)
6. `the cat <V> <NP>` (Rule 3)
7. `the cat sees <NP>` (Rule 6: chose `sees`)
8. `the cat sees <Det> <N>` (Rule 2)
9. `the cat sees a <N>` (Rule 4: chose `a`)
10. `the cat sees a man` (Rule 5: chose `man`)

Context-Free Grammar

【西湖大学 张岳老师 | 自然语言处理在线课程 第十章 - 2节】概率上下文无关文法（Probabilistic context free grammar）

2444 1 2022-04-25 13:17:55 未经作者授权，禁止转载

Probabilistic Context Free Grammar

WestlakeNLP

- Context Free Grammars (CFG)

  Formally, a CFG is a 4-tuple: <N,Σ,R,S>.

  - N: the set of non-terminals (i.e. A,B,C,…)
  - Σ: the set of terminals (i.e. $\alpha,\beta,\gamma,\ldots$)
  - R: the set of production rules (i.e. A->BC, A->γ,…)
  - S: the start symbol

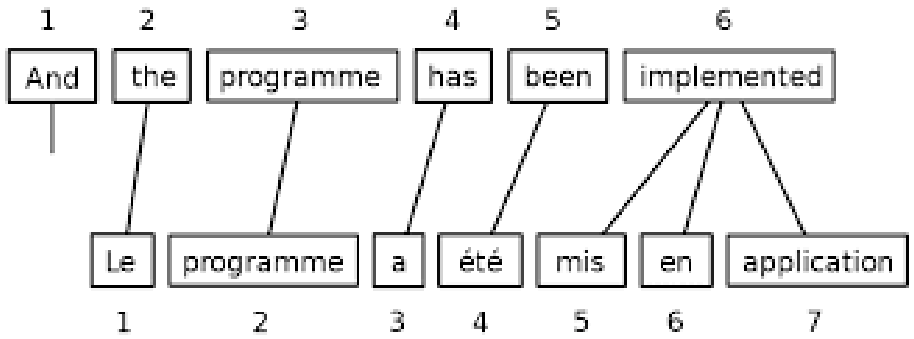
https://www.bilibili.com/video/BV1gv411g735/

Lectures - Natural Language Processing - A Machine Learning Perspective / Spring 2023

Statistical Learning

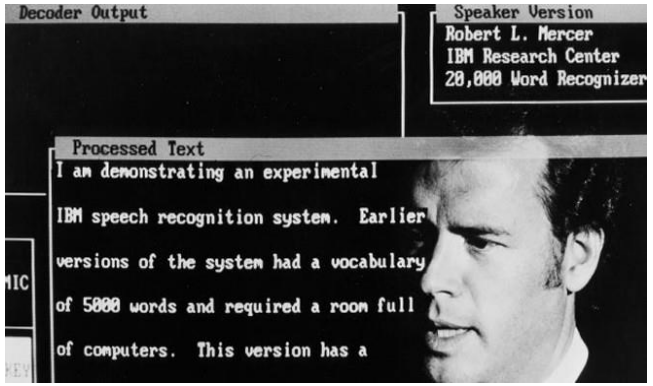
南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

IBM translation model



Expectation-maximization (EM) algorithm

Speech recognition

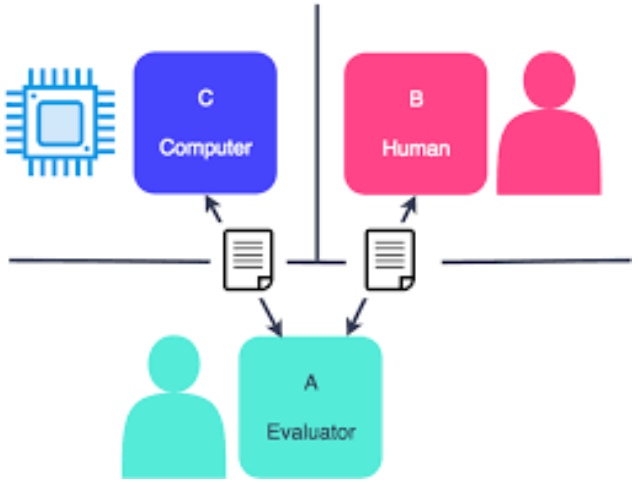


Hidden Markov Model (HMM) algorithm

Anytime a linguist leaves the group the (speech) recognition rate goes up
- Fred Jelinek 1998

Turing Test

- Ability to understand and generate language → intelligence

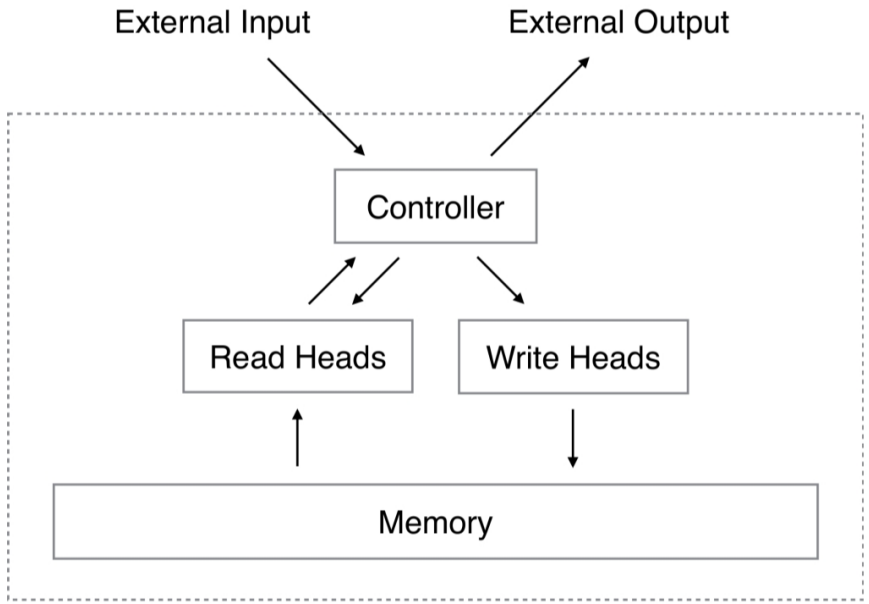


A. M. Turing (1950) Computing Machinery and Intelligence. *Mind 49:* 433-460.

COMPUTING MACHINERY AND INTELLIGENCE

By A. M. Turing

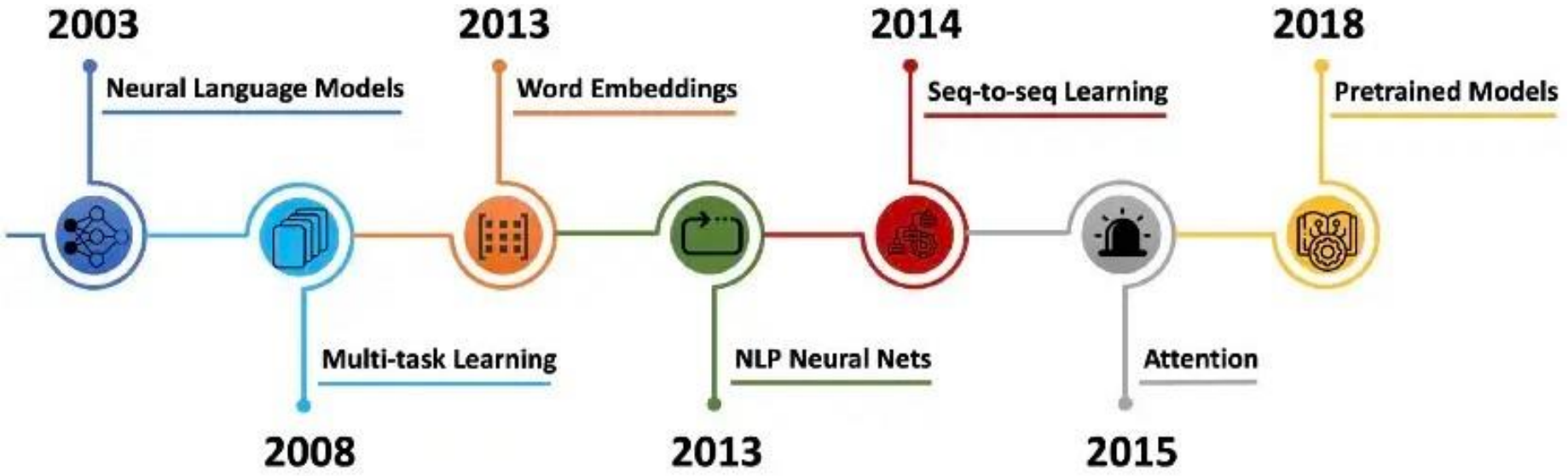
1. The Imitation Game



(Neural) Turing Machine

A Brief History of NLP

Yoshua Bengio



2018: BERT 2019: T5, RoBERTa 2020: GPT-3 2022: ChatGPT

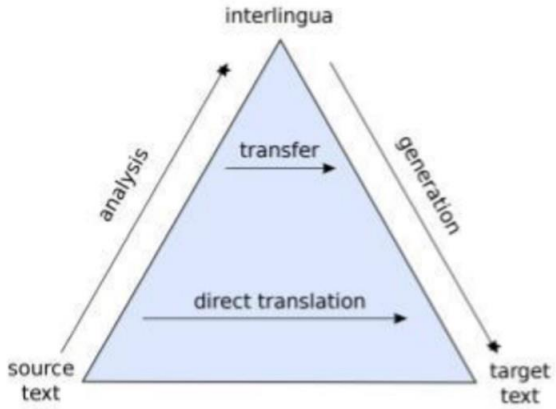
知乎：Yoshua Bengio为什么能跟Hinton、LeCun相提并论

https://medium.com/@antoine.louis/a-brief-history-of-natural-language-processing-part-2-f5e575e8e37

Symbolic and Probabilistic NLP

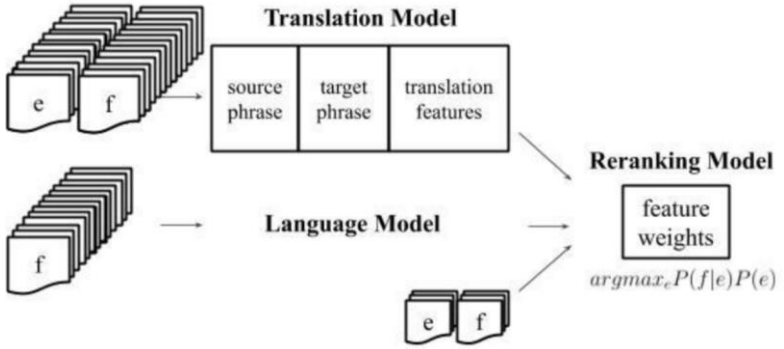
南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Logic-based/Rule-based NLP



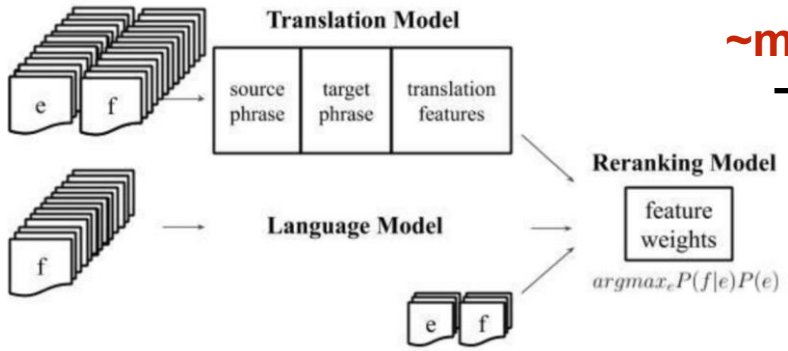
~1990s

Statistical NLP



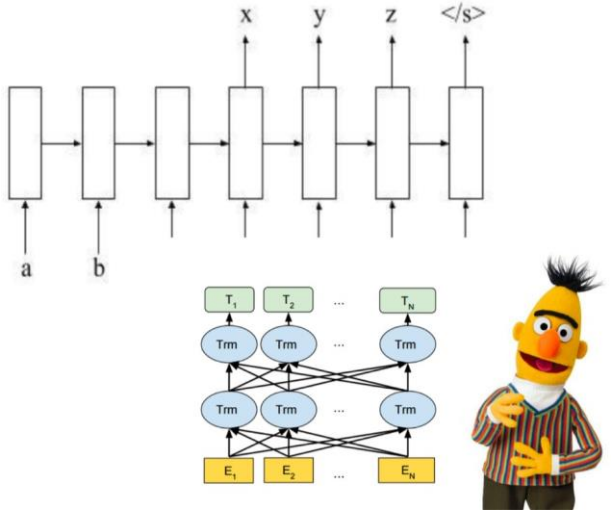
Probabilistic and Connectionist NLP

Engineered Features/Representations



~mid 2010s

Learned Features/Representations



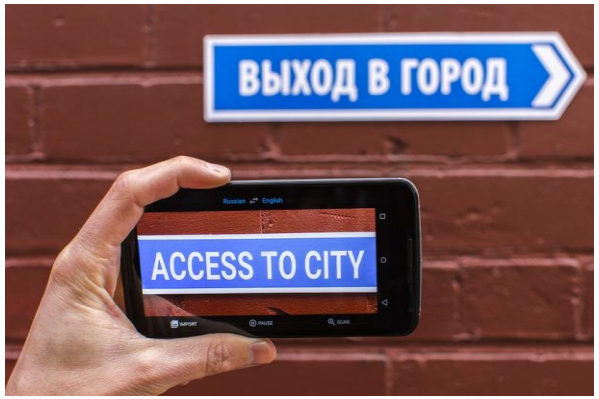
The Era of Deep Learning

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

- Significant advances in core NLP technologies
- Essential ingredient
  - Large-scale supervision
  - Lots of compute
- Reduced manual effort - less/zero feature engineering

GPU

TPU



36M sentence pairs

Russian: Машинный перевод - это круто!

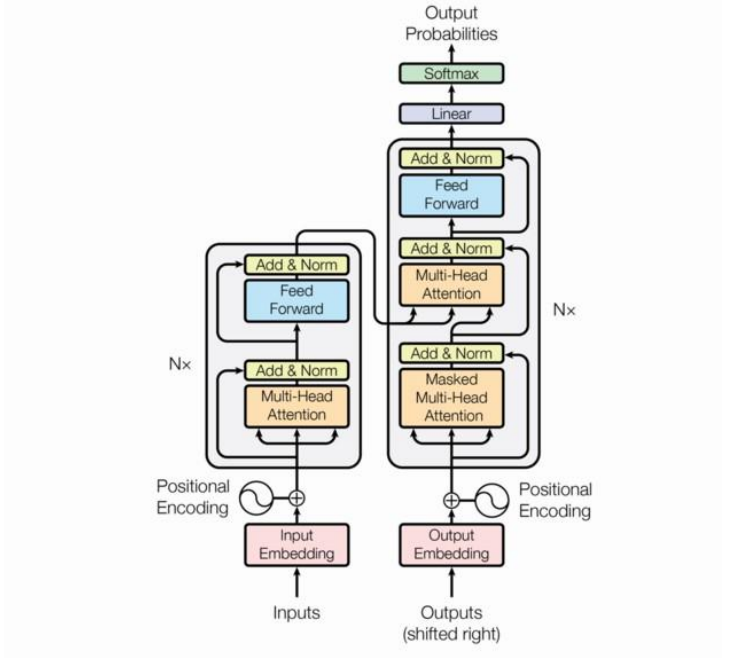
English: Machine translation is cool!

Pipeline VS. End-to-end

The nine steps are

1. Prepare data (45 minutes)
2. Run GIZA++ (16 hours)
3. Align words (2:30 hours)
4. Get lexical translation table (30 minutes)
5. Extract phrases (10 minutes)
6. Score phrases (1:15 hours)
7. Build lexicalized reordering model (1 hour)
8. Build generation models
9. Create configuration file (1 second)

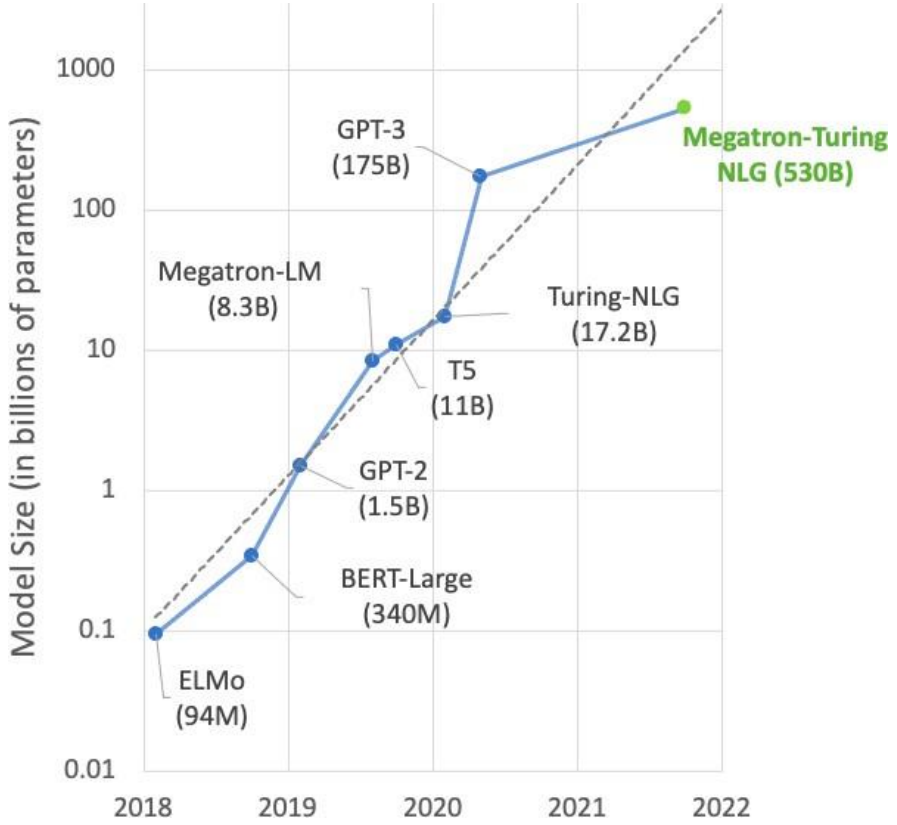
Pipeline method with SMT



End-to-end method with NMT

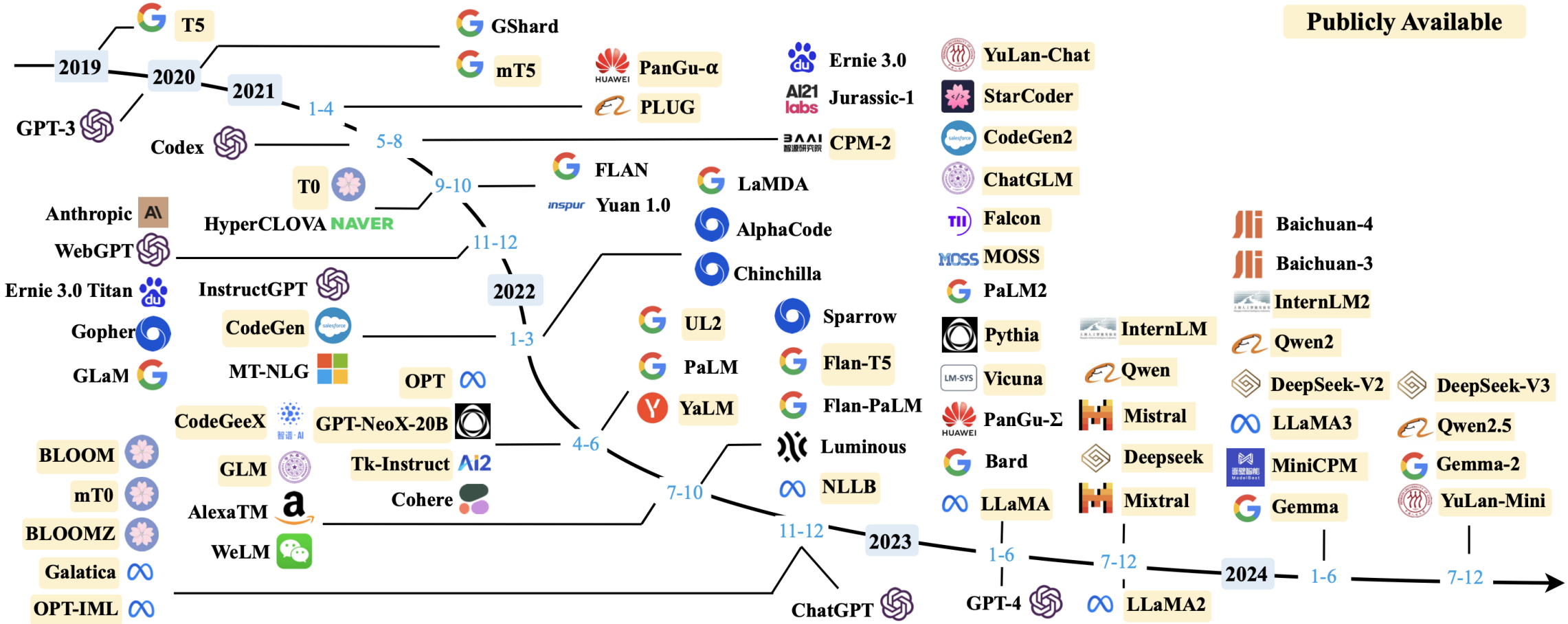
The Era of Pre-training / LLMs

BERT, ELMo, ERNIE…



- Leverages a lot of unlabeled text
- Model size increased by 1000+

The Era of Pre-training / LLMs



A Brief History of NLP

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Discussion:

What are you most excited about NLP?

Why is NLP so Difficult

- ##自然语言处理太难了##

分享一件有趣的事，一位北京大哥点了油渣儿菜，上菜后抱怨这菜怎么是苦的，原来他不知道儿菜是一种味道略苦的蔬菜，误以为是油渣儿，菜！
感叹南北文化、饮食差异甚是有趣 @安宁庄前街的英文是啥

原文：Your state or province
机器翻译：贵州省
原文：PEARL Harbor
翻译：蚌埠
原文：New York
翻译：新乡

Why is NLP so Difficult

- ##自然语言处理太难了##

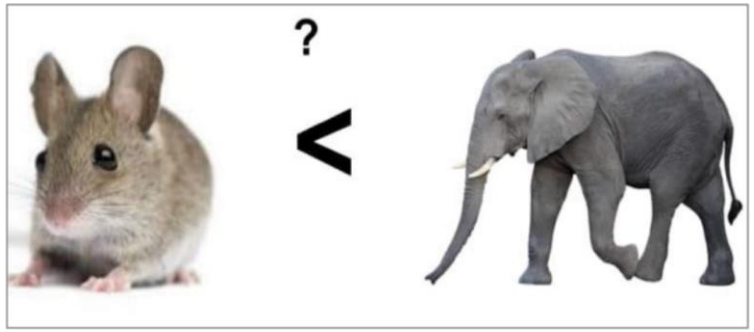
Why is NLP so Difficult

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Ambiguous	Dialects, accents	Abbreviation
Expressivity	Listener has to infer - pragmatics	Humor or irony

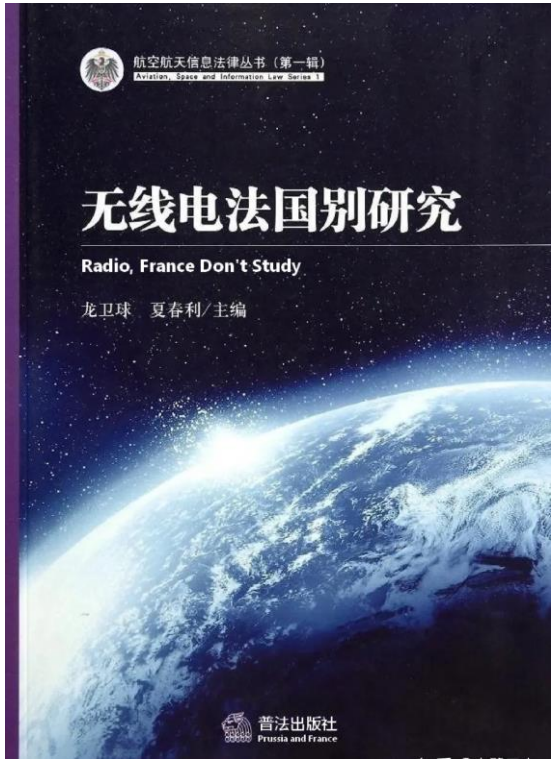
Unmodeled Variables

- World knowledge
  - I dropped the glass on the floor and it broke
  - I dropped the hammer on the glass and it broke



Challenges of Words

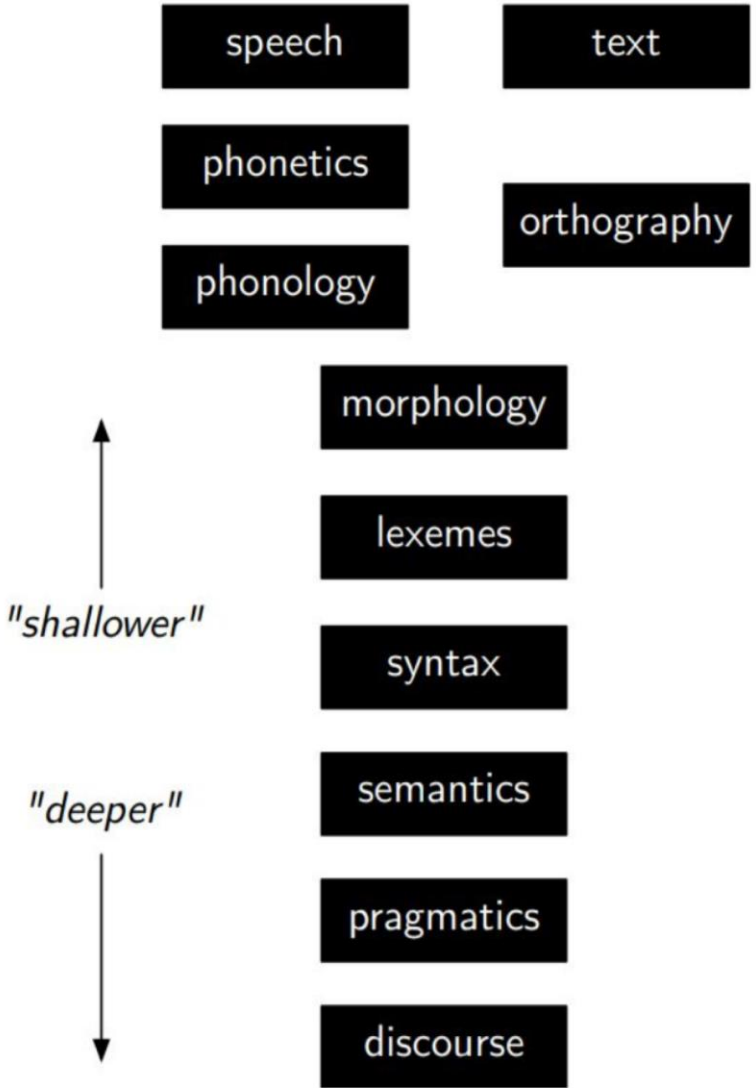
- Segmenting text into words
  - E.g., Chinese example，无线电/法国/别研究 or 无线电法/国别研究
- Morphological variation
  - E.g., un-finish-ed, scient-ist
- Multiword expressions
  - E.g., take out, make up
- New words and changing meanings
  - E.g., covid
  - E.g., Bachelor: a young knight -> an academic degree



Levels of Linguistic Knowledge

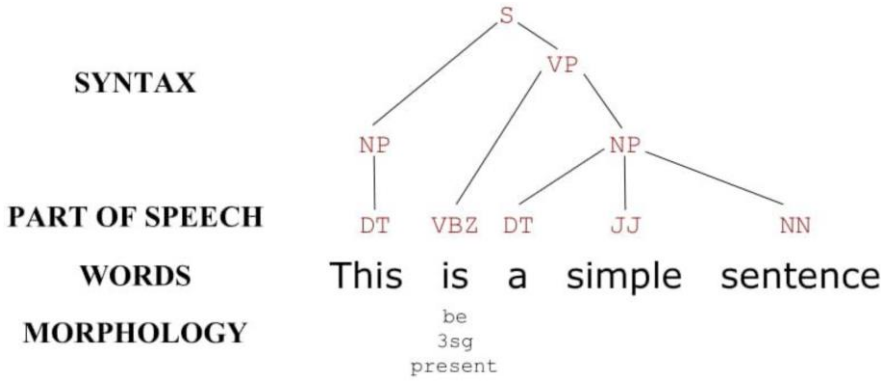
What dose an NLP system need to know a language?

- Phonetics (语音学)
  - Study how humans produce and perceive sounds
- Morphology (词形学)
  - Study how words are formed: such as stems, root words, prefixes, suffixes
- Part-of-Speech (词性)
  - Predict which category a word is assigned to in accordance with its syntactic functions

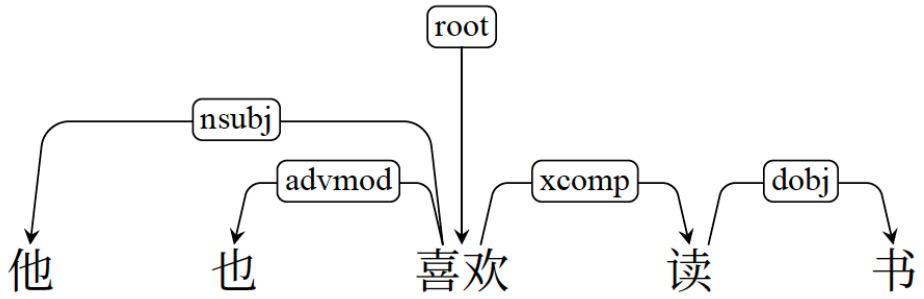


Levels of Linguistic Knowledge

- Syntax (句法)
  - Study how words and morphemes combine to form larger units such as phrases and sentences
  - Constituency parsing (成分分析)
  - Dependency parsing (依存分析)



Constituency parsing tree



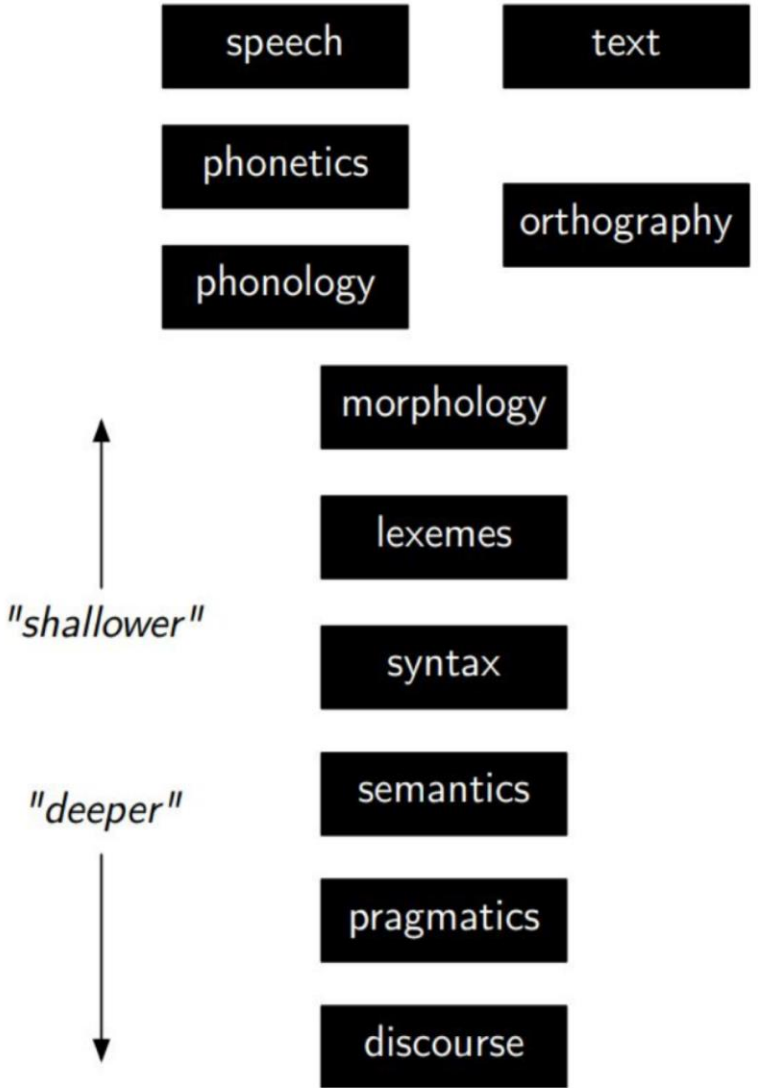
Dependency parsing tree

Constituency/Dependency Parsing with Stanza toolkit

Levels of Linguistic Knowledge

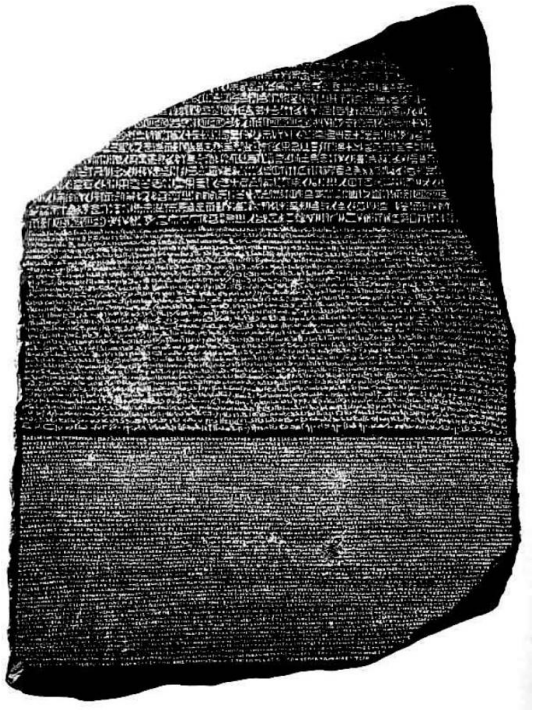
- Semantics (语义)
  - Study meaning of words, phrases, sentences, or discourse
  - Lexical relation
    - Synonymy(同义)/antonymy(反义)
    - Hypernymy(上位词)/hyponymy(下位词)
  - Named entity recognition
  - Word sense disambiguation
- Pragmatics (语用学)
  - Study how context contributes to meaning
  - Implicature (言外之意)

When Sebastian Thrun PERSON started working on self-driving cars at Google ORG in 2007 DATE, few people outside of the company took him seriously.



Corpora

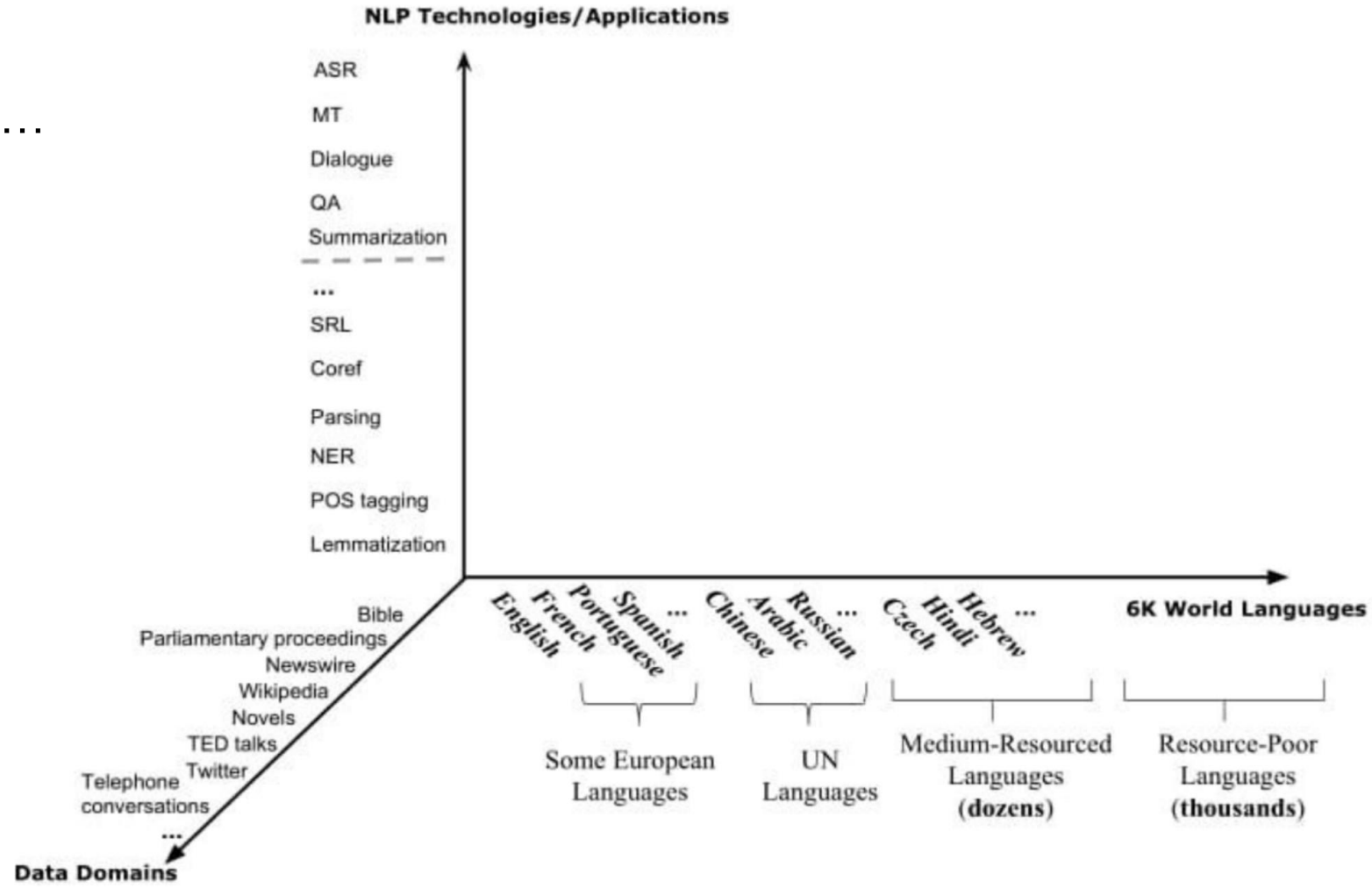
- A corpus is a collection of text
  - Often annotated in some way
  - Sometimes just lots of text
- Examples
  - Penn Treebank: 1M words of parsed WSJ
  - Canadian Hansards: 10M+ words of aligned French/English sentences
  - Web: billions/trillions of words
  - Amazon reviews



Rosetta Stone

Data Domains

- Diverse domains
- Language, task, source, style, …

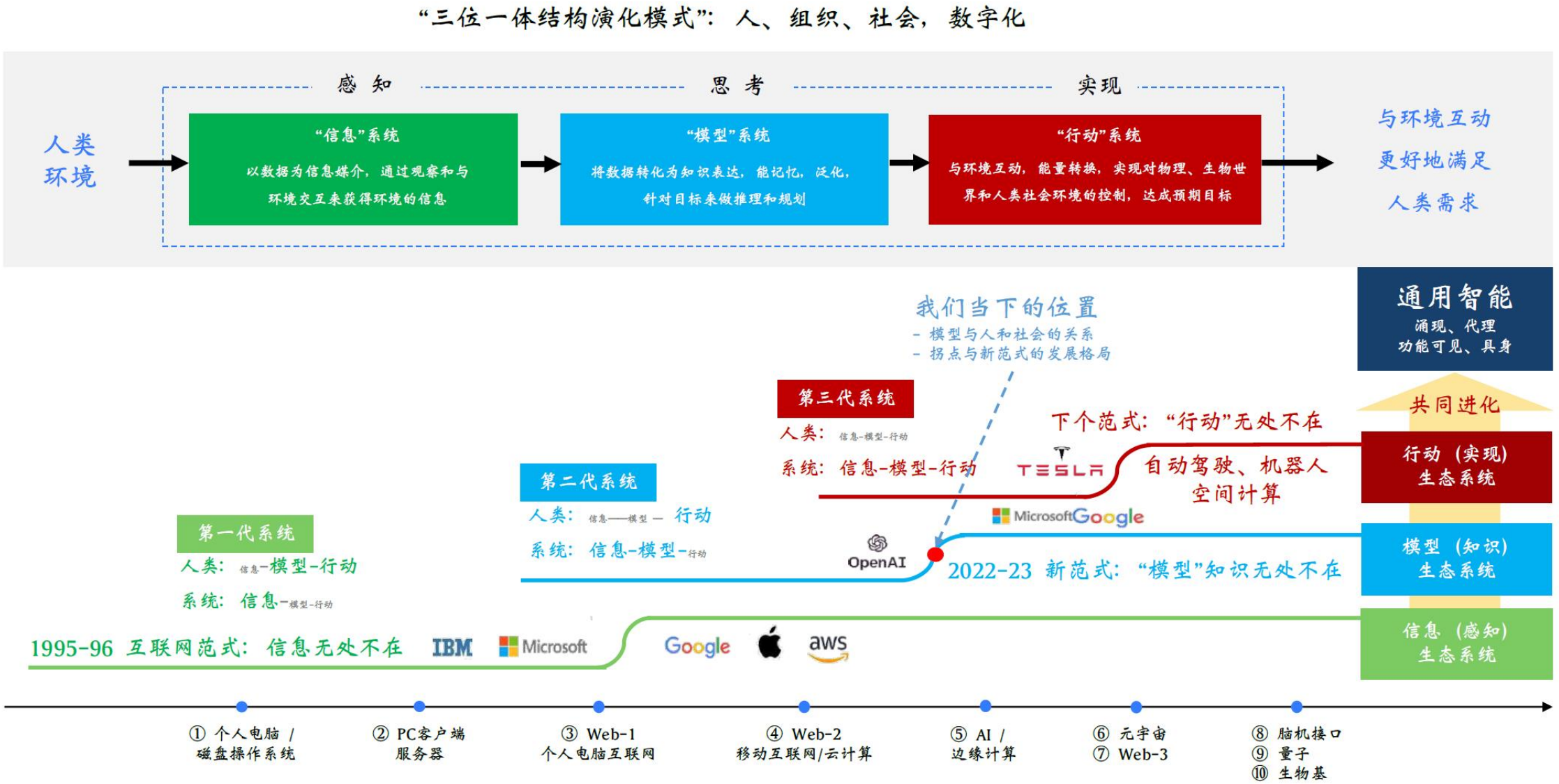


陆奇：我的大模型世界观

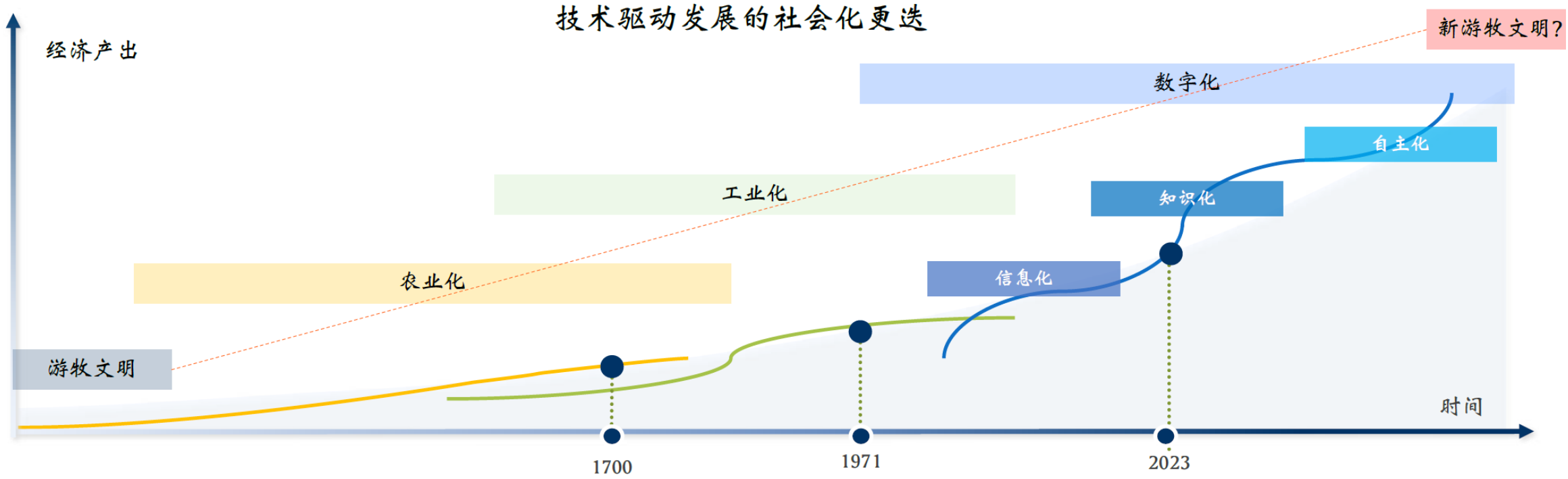
南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

- 任何复杂体系，包括一个人、一家公司、一个社会，甚至数字化本身的数字化体系，包括：
  - 信息系统，从环境当中获得信息；
  - 模型系统，对信息做一种表达，进行推理和规划；
  - 行动系统，我们最终和环境做交互，达到人类想达到的目的；
- 我们每个人都是模型的组合。人有三种模型：
  - 认知模型，我们能看、能听、能思考、能规划；
  - 任务模型，我们能爬楼梯、搬椅子剥鸡蛋；
  - 领域模型，我们有些人是医生，有些人是律师，有些人是码农。
- “这一次大模型拐点会让所有服务经济中的人、蓝领基本都受影响，因为他们是模型，除非有独到见解，否则你今天所从事的服务大模型都有。”
- 更多阅读：《陆奇：我的大模型世界观》

新范式的新拐点


“三位一体结构演化模式”：人、组织、社会，数字化
感 知
思 考
实现
人类
环境
“信息”系统
以数据为信息媒介，通过观察和与环境交互来获得环境的信息
“模型”系统
将数据转化为知识表达，能记忆，泛化，针对目标来做推理和规划
“行动”系统
与环境互动，能量转换，实现对物理、生物世界和人类社会环境的控制，达成预期目标
与环境互动
更好地满足
人类需求
我们当下的位置
- 模型与人和社会的关系
- 拐点与新范式的发展格局
通用智能
涌现、代理
功能可见、具身
共同进化
第三代系统
人类：信息-模型-行动
系统：信息-模型-行动
下个范式：“行动”无处不在
TESLA
自动驾驶、机器人
空间计算
行动（实现）
生态系统
第二代系统
人类：信息——模型 — 行动
系统：信息-模型-行动
Microsoft Google
OpenAI
2022-23 新范式：“模型”知识无处不在
模型（知识）
生态系统
第一代系统
人类：信息-模型-行动
系统：信息-模型-行动
1995-96 互联网范式：信息无处不在
IBM
Microsoft
Google
aws
信息（感知）
生态系统
① 个人电脑 / 磁盘操作系统
② PC客户端 服务器
③ Web-1 个人电脑互联网
④ Web-2 移动互联网/云计算
⑤ AI / 边缘计算
⑥ 元宇宙
⑦ Web-3
⑧ 脑机接口
⑨ 量子
⑩ 生物基

新范式的社会影响



时代	农业化	工业化	数字化：信息无处不在	数字化：模型无处不在	数字化：行动无处不在
人作为核心生产力	人：体力为主 简单工具 没有流动性	人：体力为主，脑力为辅 机械、电器、电子等自动化工具 地域流动性	人：脑力为主，体力为辅 电脑，手机等信息化工具 全球流动性	人：创新为主，其他为辅 新一代认知思考能力工具 逐步替代脑力劳动	人：创新探索 下一代自主行动工具 全面替代脑力和体力劳动
经济范式	农业经济	商品经济	服务经济	体验经济	（暂）AI经济
代表职业	农民	工人	码农，设计师，分析师	创业者，科学家，艺术家	人类新价值系统

数字化应用-技术推动+ 需求拉动：大模型为先

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

需求拉动

人类需求	消费者：需求层次 x 人群分类x 时间周期													创作者	企业：降本增效					
数字化满足人类需求六种能力	通讯	社交	内容	游戏	消费	旅游	健身	医疗	教育	金融	住房	睡眠	信息知识	食品饮料	开发者	法律	行政	OA办公	政府城市
I:信息(2D) 间接感知交互																			
II :体验 (3D) 直接具象交互																			
III: 关系 (抽象) 信任| 激励 | 所有权																			
IV:物理外部环境 就地 | 直接																			
V:生理内部感知 内部 | 直接 | 健康																			
IV:知识 | 模型 嵌入 | 启用																			

信息

自然语言为先
知识为主
行动导向
信息基础
三位一体为王

模型

行动

闭环壁垒

使用数据
模型使用
模型开发
自有数据
大模型基础

技术推动

● 科技驱动模式

-单一突破
-多点汇散
-垂直渗透
-水平扩散

模型覆盖宽度足够广

模型深度是决定因素

● 需求拉动模式

-狭窄的单点进入
-广泛的多点进入
-垂直渗透
-水平扩张

需求对知识的敏感度

需求对三位一体的敏感度

IV	⑤	知识 | 模型：嵌入和启用
V	⑧④⑤⑨	生理感受：内在且直接
IV	⑤④⑨	物理环境：就地和直接
III	⑦③④⑤⑥⑨	关系：信任| 激励 | 拥有权
II	⑥③④⑤⑦⑨	3D 体验：具体化的直接交互
I	③④①②⑤⑨	2D 信息：感知间接交互

① 个人电脑 / 磁盘操作系统
② PC客户端 服务器
③ Web-1 个人电脑互联网
④ Web-2 移动互联网/云计算
⑤ AI / 边缘计算
⑥ 元宇宙
⑦ Web-3
⑧ 脑机接口
⑨ 量子
⑩ 生物基

NLP on the Job Market

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

- Positions:
  - Research scientist
  - Machine learning engineer
  - Algorithm engineer, software engineer
  - Product manager, project manager, etc.

互联网/AI

后端开发	Java C/C++ PHP Python C# .NET Golang Node.js 语音/视频/图形开发 高性能计算工程师 GIS工程师 区块链工程师 全栈工程师 软件项目经理 其他后端开发
前端/移动开发	前端开发工程师 Android iOS U3D UE4 Cocos 技术美术 JavaScript 鸿蒙开发工程师
测试	测试工程师 软件测试 自动化测试 功能测试 测试开发 硬件测试 游戏测试 性能测试 渗透测试 测试经理
运维/技术支持	运维工程师 IT技术支持 网络工程师 网络安全 系统工程师 运维开发工程师 系统管理员 DBA 电脑/打印机维修 系统安全 技术文档工程师
人工智能	图像算法 自然语言处理算法 大模型算法 数据挖掘 规控算法 SLAM算法 推荐算法 搜索算法 语音算法 风控算法 高性能计算工程师 算法工程师 算法研究员 机器学习 深度学习 自动驾驶系统工程师 数据标注/AI训练师

NLP on the Job Market

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

岗位投递 | 腾讯校招

NLP on the Job Market

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

岗位描述

1、基于海量数据场景，参与业务数据分析及数据模型构建，驱动业务增长；

2、负责规划和搭建业务数据指标体系和分析体系，建设智能决策方法论，寻找业务收入增长点；

3、通过实验科学、因果推断等方法验证评估策略的效果和价值；

4、建设用户画像、构建特征工程实施方案，不断优化和改进数据模型，提升线上数据服务效果；

5、跟进业界最新的广告、推荐、NLP、复杂网络等领域进展，快速实现并应用于实际任务中。

岗位要求

1、本科以上学历，数学、统计、运筹学、计算机、经济学、通信等相关专业；

2、严谨的逻辑思维能力和扎实的数理专业知识基础；掌握常见的统计分析、实验设计、机器学习等

3、具备优秀的编码能力，扎实的数据结构和算法功底，熟练掌握至少一种编程语言，Java、Scala、

4、优秀的分析问题和解决问题的能力，对解决具有挑战性问题充满激情；

5、熟悉机器学习、自然语言处理、数据挖掘中一项或多项，有相关实践项目经验者优先；

6、沟通表达清晰；高效的协作能力；积极主动、自驱力强；思维严谨、系统；批判性思维。

加分项或注意事项

1、在相关领域的顶会或核心刊物上发表过高水平论文；

青云课题 | 顶尖师资 | 顶峰战场 | 定向培养 | 丰厚回报 | 招募对象

- AI大模型
- 基础架构
- 高性能计算
- 大数据
- 多媒体
- 安全
- 游戏引擎
- 机器人
- 量子
- 金融科技

TEG-新一代大模型分布式训练基础框架研究	WXG-大模型在搜索系统中的应用落地研究	CDG-广告全场景基础大模型
CSIG-音视频语音识别及视觉大模型研究	IEG-3D生成大模型	PCG-多模态大型语言模型的研发与应用
TEG-混元大模型-数据合成技术与数据价值衡量方法研究	WXG-基础大模型研究与应用	TEG-混元基座模型-大模型AI Coding及代码Base Agent研究

数据科学方向 | 腾讯校招

NLP on the Job Market

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

端侧AI训练算法工程师 40-70K·16薪

收藏 立即沟通

深圳 经验不限 硕士

职位描述 微信扫码分享 举报

工作职责
1. 负责LLM和多模态大模型在移动端平台(CPU/GPU/NPU)训练和推理优化
2. 探索端侧大模型轻量化训练方法，设计量化感知训练算法
3. 分析和解决在异构硬件上训练时遇到的技术难题

任职要求
1. 硕士及以上学历，人工智能、计算机科学、电子、信息工程、机器人等专业
2. 熟练使用C++/python
3. 对大模型加速优化方案有深入了解，对投机采样，GQA，MOE，Lora量化等技术有开发与优化经验者优先
4. 有TFlite (Micro)，NCNN，SNPE/QNN，EAI，NeuroPilot等移动端深度学习框架开发经验者优先
5. 熟悉1/4/8bit量化，掌握蒸馏、剪枝、NAS等模型优化手段
6. 熟悉CPU/DSP/GPU等硬件平台的算子加速

叶先生 刚刚活跃
小米 · 算法工程师

去App
与BOSS随时沟通

大模型算法工程师-AI Coding-（北... 50-80K·15薪

收藏 立即沟通

深圳 3-5年 本科

职位描述 微信扫码分享 举报

职位描述
1、负责研发提效场景所需要的大模型能力的研发和应用，研究高质量数据的挖掘和合成、大模型的对齐效率、Agent的设计&端到端训练等等，不断思考和跟进AI的最新进展对我们的价值，探索下一代的研发模式；
2、高质量数据挖掘清洗使用，数据自动、半自动合成方案研究探索，设计针对代码场景的原子任务、全链路任务的评测方法；
3、研究LLM/VLM训练与优化技术，包括微调、强化学习（RLHF）、知识蒸馏等，提高大模型在代码场景下的能力；
4、尝试落地到各种应用场景，比如：IDE代码补全、代码能力QA、场景化Agent、代码自动修复、漏洞检测等等；
5、持续跟踪LLM/VLM领域的最新技术动态，并将其应用于实际业务场景中，推动技术落地。
职位要求
1、优秀的代码能力、数据结构和基础算法功底，熟练掌握至少一门语言，ACM/ICPC、NOI/IOI、Top Coder、Kaggle等比赛获奖者优先；
2、熟悉NLP、CV、ML等相关的技术，深入理解大模型或图片视频生成等相关技术栈（如RLHF、SFT、Dit/Vit/Transformer、Stable Diffusion等）；
3、在大模型领域，主导过有影响力的项目或论文者优先；在ACL/EMNLP/ECCV/CVPR等顶会发表论文者优先；
4、有代码基座经验，有强化学习结合大模型落地经验，有Multi-Agent、Tool-Use等有相关经验优先；
5、出色的问题分析和解决能力，有自主探索解决方案的能力；
6、良好的沟通协作能力，能和团队一起探索新技术，推动技术进步。

祝先生 刚刚活跃
字节跳动 · 招聘专家

去App
与BOSS随时沟通

NLP on the Job Market

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

- Will NLP become the next civil engineering?
  - More talents, salary reduction, lay offs, …
- Emerging opportunities
  - Better LLMs
  - Applications of LLMs in various domains
  - Machine learning operation/DevOps for machine learning (MLOps)
- What is your niche?
  - Research, engineering (technology is not everything)
  - Business, marketing, product
  - Hybrid

How to Learn

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

- Try with official documents and examples
  - Don't be afraid. Software is designed for easier use
  - Start from basic examples, step towards the real applications
- Open courses and tutorials
  - Don't wait to be taught
  - Search and learn by yourself
  - MOOC/Wechat/Zhihu/Bilibili, etc.
- Learn from those who are willing to share

•智源社区每日分享，https://hub.baai.ac.cn/

•微博账号：爱可可-爱生活，宝玉XP，蚁工厂，

•微信公众号：PaperWeekly, 夕小瑶科技说，机器之心，李rumor

How to Learn

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

- Two possible pattern
  - Learning process: what -> why -> how
    - Technician vs. engineer vs. researcher/scientist
  - Research process (sometimes): what -> how -> why
- Another two possible pattern
  - Deep-first search
  - Width-first search
  - Lean thinking (精益开发)

How to Learn

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

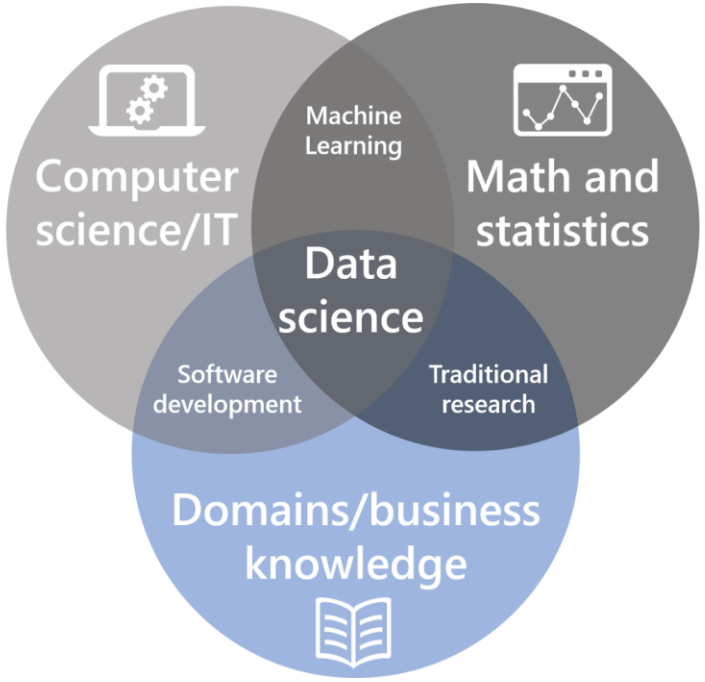
- What should we pay more attention to in the LLM era?
  - “认知卸载”
  - “无分数学习”
  - “人工智能能为人类做什么” vs. “人工智能正在对人类做什么”

Data Science vs. Computer Science

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

- Computer science
  - Theory and practice of computing and coding
  - Everything from software to the operating systems they run on, and to the base hardware that interacts with the OS
- Data science
  - Aim at real-world problems
  - Work for understanding of how business works and how to improve
  - Emphasis more on using instead of designing

CS自学指南



More reading: https://www.indeed.com/career-advice/finding-a-job/data-science-vs-computer-science

GPU Resources

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

- Google Colab
  - Free for 12 hours, buy pro/pro+
- 阿里天池
  - Free for 60 hours，8 hours/each
  - 天池大赛
- Kaggle GPU
  - Running Kaggle Kernels with a GPU
- AutoDL

Learning Resources

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

- CS224N: Natural Language Processing with Deep Learning
- COS 484: Natural Language Processing
- Huggingface NLP course

Thank you