语言智能发展现状思考

常宝宝
北京大学计算语言学研究所
chbb@pku.edu.cn

2022-11-11

现状—捷报频传

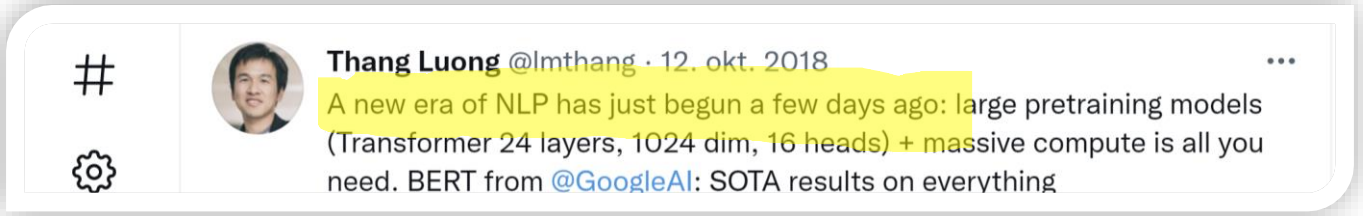
	GLUE	superGLUE
human	87.1	89.8
machine	**91.3**	**91.2**

- 情感分析超过人类水平
- 机器阅读理解超越人类水平
- 自然语言推理超越人类水平
  …

能知情、善解意、会推理的语言智能！

- 谷歌 AI 发布 BERT 模型，打破十一项 NLP 记录(2018年10月）
- 大规模预训练语言模型不断推陈出新…
  ELMo、GPT、BERT、RoBERTa、XLNet、BART、T5、GTP-2/3、PaLM
- 机器综合语言理解评测基准 GLUE 和 SuperGLUE，超过人类
  …

- 2018年 “NLP元年”



NLP元年 —— 1947年

Warren Weaver 和 Nobert Wiener的邮件往来

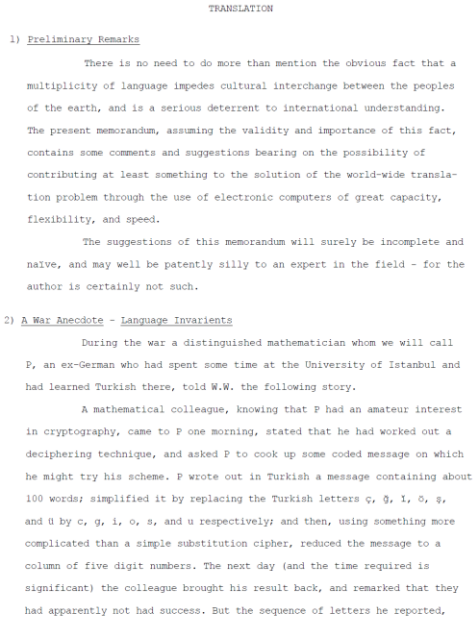
- 1947年3月4日，Warren Weaver 致 Nobert Wiener

one naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian, I say "This is really written in English, but it has been coded in some strange symbols.

"Have you ever thought about this? As a linguist and expert on computers, do you think it is worth thinking about?"

- 1947年4月30日，Nobert Wiener 致 Warren Weaver

"Second - as to the problem of mechanical translation, I frankly am afraid the boundaries of words in different languages are too vague and the emotional and international connotations are too extensive to make any quasi mechanical translation scheme very hopeful. I will admit that basic English seems to indicate that we can go further than we



TRANSLATION

1) Preliminary Remarks

There is no need to do more than mention the obvious fact that a multiplicity of language impedes cultural interchange between the peoples of the earth, and is a serious deterrent to international understanding. The present memorandum, assuming the validity and importance of this fact, contains some comments and suggestions bearing on the possibility of contributing at least something to the solution of the world-wide translation problem through the use of electronic computers of great capacity, flexibility, and speed.

The suggestions of this memorandum will surely be incomplete and naïve, and may well be patently silly to an expert in the field - for the author is certainly not such.

2) A War Anecdote - Language Invarients

During the war a distinguished mathematician whom we will call P, an ex-German who had spent some time at the University of Istanbul and had learned Turkish there, told W.W. the following story.

A mathematical colleague, knowing that P had an amateur interest in cryptography, came to P one morning, stated that he had worked out a deciphering technique, and asked P to cook up some coded message on which he might try his scheme. P wrote out in Turkish a message containing about 100 words; simplified it by replacing the Turkish letters ç, ğ, I, ö, ş, and ü by c, g, i, o, s, and u respectively; and then, using something more complicated than a simple substitution cipher, reduced the message to a column of five digit numbers. The next day (and the time required is significant) the colleague brought his result back, and remarked that they had apparently not had success. But the sequence of letters he reported,

Translation Memorandum
Weaver, 1949.7

NLP元年 —— 1947年

Warren Weaver 和 Nobert Wiener的邮件往来

- Nobert Wiener 致 Warren Weaver (续)

> By the way, I have been fascinated by McCulloch's work on such apparatus, and, as you probably know, he finds the wiring diagram of apparatus of this kind turns out to be surprisingly like the microscopic analogy of the visual cortex in the brain."

- Weaver提出用加密-解密思想建模翻译问题
  - 最早提出用经验方法建模NLP问题
  - 概念层面，目前流行的NLP方法仍是这一思想的延续和发展
- Wiener提出这一方法的局限性 emotional and international connotation
  - denotation vs. connotation，时至今日，对connotation建模乏力仍是NLP处理软肋
- Wiener提及神经网络的早期工作
  - 七十多年后，神经网络方法已经主宰了NLP领域

提纲

- 引言
- **自然语言处理研究范式变迁**
- 深度学习及其带来的进步
- 经验方法的数学基础和语言学基础
- 经验方法的局限性
- 机器理解和人类理解
- 结束语

自然语言处理研究范式的变迁

- 理性方法 与 经验方法
- 理性方法 / 符号方法 /基于规则的方法
  - 专家构建符号化的语言知识库、世界知识库、规则库
  - 机器利用规则和知识库完成自然语言处理任务
  - 专家 “教授” 机器
- 经验方法 / 数据驱动的方法/统计方法/机器学习方法
  - 专家标注自然语言数据(标注语料库)
  - 设计参数化语言处理模型，调整并确定参数让模型拟合语言数据
  - 机器利用拟合好的模型完成自然语言处理任务
  - 专家 “指导” + 机器 “学习”

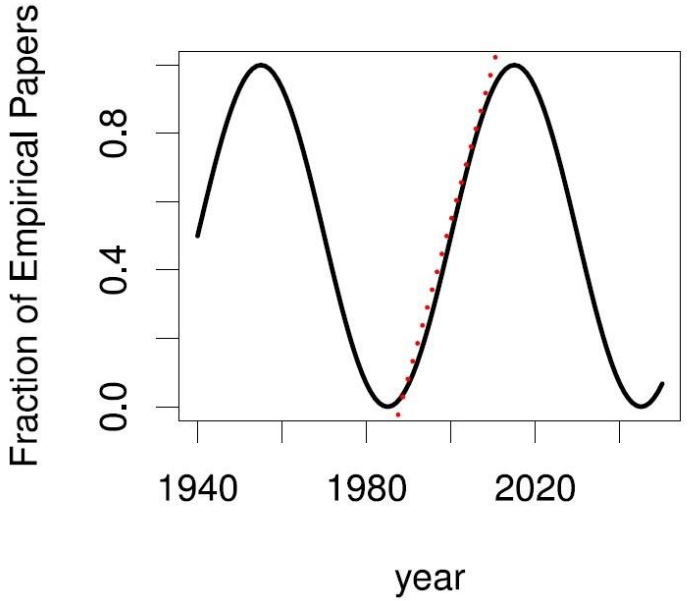
自然语言处理研究范式的变迁

- 1949年之后，NLP技术范式的演变
  - 经验方法 → 理性方法 → 经验方法
- 经验方法和理性方法周期性交替？

  1950年代 经验方法(Shannon、Skinner、Harris、Firth)
  1970年代 理性方法(Chomsky, Minsky) 经验主义出局
  1990年代 经验方法(IBM) 理性主义逐渐出局
  2010年代 重回理性方法？

  --- A PENDULUM SWUNG TOO FAR, Kenneth Church, 2007

- 经验方法在各种任务上都取得了更好的表现



提纲

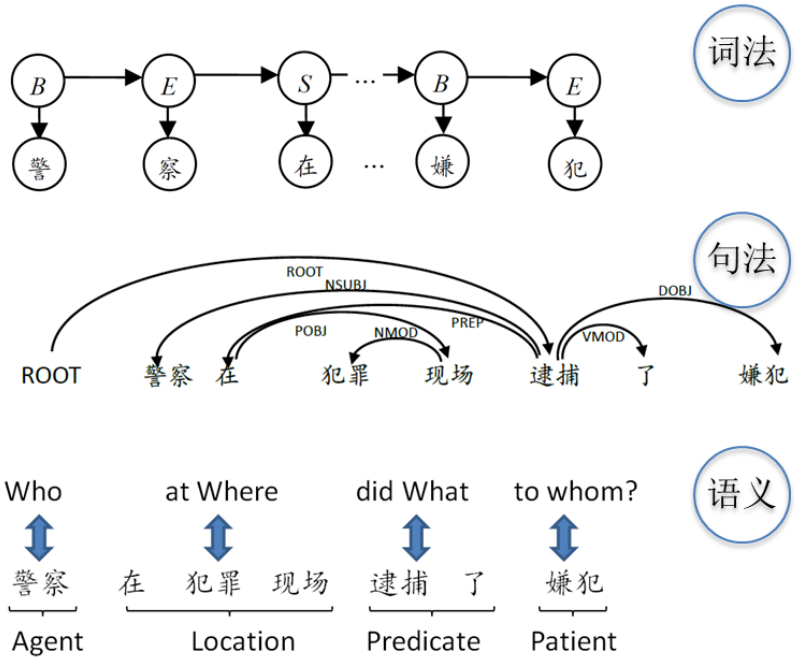
- 引言
- 自然语言处理研究范式变迁
- **深度学习及其带来的进步**
- 经验方法的数学基础和语言学基础
- 经验方法的局限性
- 机器理解和人类理解
- 结束语

深度学习及其带来的进步

- 2010年代　　深度学习 取代 传统浅层学习
  - 摧枯拉朽，几乎把所有的方法、模型都扫进了历史的垃圾堆中
  - 经验主义方法持续深化
- 2006年，在图像识别领域取得突破进展
  - Geoffrey E. Hinton，DBN(Deep Belief Nets)+预训练
- 2014年后，在NLP领域流行，给NLP带来了实实在在的进步
  - 把对语言处理的认识转换为网络结构的设计
  - FNN、CNN、RNN(LSTM)、Transformer、GCN、Attention…
- 2018年后，大规模预训练语言模型出现，
  - 逐渐形成 "预训练语言模型+微调"新型统一NLP范式
  - GPT、BERT、RoBERTa、XLNet、BART、T5、GTP-2/3…

我自己的体验

- 汉语分词
  - 精度与之前的模型可比，减少特征工程 (IJCNLP 2013，ACL 2014）)
- 句法分析
  - 91.6/90.39…**93.29**, 92.13(ACL 2015) → **93.51**,92.45(ACL 2016) →**96.35**,95.25(EMNLP 2018) [+**4.75**, +4.86]
- 汉语浅层句义分析
  - **…72.64**(ICCPOL 2009) → **77.21**(EMNLP 2015) → **79.67**(ACL 2017) → **79.92**(WSP4NLP 2017)[+**7.28**]



提纲

- 引言
- 自然语言处理研究范式变迁
- 深度学习及其带来的进步
- **经验方法的数学基础和语言学基础**
- 经验方法的局限性
- 机器理解和人类理解
- 结束语

数学基础

- 诺贝尔经济学奖得主Thomas Sargent，2018年8月在北京说：
  “人工智能其实就是统计学,……。”
- 统计学习理论(statistical learning theory)
- 语言被看作随机现象，是基于数据生成分布采样的结果

$$x_i \sim P(x_i)$$

- 试图根据语言实例，求解或部分求解数据生成分布
  - 假定语言生成符合某种随机过程
  - 假定存在X到Y的映射函数

$$p(x) = \sum_y p(x, y)$$

$$p(x, y) = p(y|x)p(x)$$

数学基础

- 模型： 参数化的随机过程或者将输入映射为输出的函数
  - bigram model 和 马尔可夫模型
  - 将文本(X)映射为情感标签(Y)的最大熵模型$P(Y|X)$
  - 用实际语言数据调整模型参数，让模型输出尽可能接近实际语言数据
  - 期望模型行为与人类语言行为一致
- 语言全体 和 语言样本
- 从语言样本求得模型，推广到全体(更重要的是推广到未来产生的语言数据)

数学基础

- 训练集/开发集
  - 训练集用于调整模型参数，开发集用于模型选择
- 测试集
  - 测试集是未来数据的样本
  - 用于衡量模型在未来数据上的处理能力
- 典型的研究范式
  - 确立任务和评价指标 → 建立训练集、验证集和测试集 → 提出参数化数学模型 → 训练模型 → 按照评价指标进行评价 →结果显著性检验
- 流程量化、严格、规范

语言学基础

- 经验方法也有语言学基础
- 分布语义学(distributional semantics)

  Words that occur in similar contexts tend to have similar meanings.

  ---Harris

  You shall know a word by the company it keeps.

  ---Firth
- 因为分布相似，所以意义相似
- 分布语义学是经验方法有效的基础保证。模型参数捕捉的是分布模式。

语言学基础

- 分布是有形的，是可观察的变量，机器可以看到并统计
- 意义是无形的，是隐藏变量，机器无法观察并统计。
- 经验方法通过观察有形的分布推断隐藏的意义。

- 分布信息是经验模型所依赖的所有信息
  - 两个词若出现在相似的语境中，意味着有相似的意义
  - 两个词若出现在相似的语篇中，意味着它们表达了相似的话题
  - 两句话因为分布相似，所以表达相似的情感
  - …
  - 因为是相似的分布，所以我们的模型给出相似的决策

深度学习

- 深度学习是经验方法的深化，数学基础还是**统计学习理论**
- 非线性的学习能力，更加强大的描述能力
  - 传统学习模型 与 (大致)线性可分问题
- 更好的表示机制，离散符号 → 连续向量
  - 苹果 → (0.26, 0.25, -0.39, …, -0.17) noun →(0.16, 0.13, 0.04,…)
  - 苹果、梨 表示成 两个不同的符号 vs 表示成两个向量
- 支持多层表示和表示学习
  - 通过多层变换，原始空间→目标空间 线性不可分→ (大致)线性可分
- 支持基于大数据的预训练(分布模式迁移)
  - word2vec(2013), Glove(2014), Elmo(2018), Bert(2018), Xlnet (2019)…

深度学习

- 端到端的训练(无需人工提取特征)
- 更少的前提假设(限制)，无需做过多的条件独立性假设
  - 形式上不限制观察语境的范围(没有马尔可夫假设、马尔可夫毯子)

- 深度模型容量大，需要更大规模的数据支持、需要算力
- 有时候，深度模型健壮性不足，对数据扰动比较敏感
- 大模型也带来许多问题，比如精调性能不稳定

- 总之，深度学习方法有更好的拟合能力，有更好的分布模式提取能力

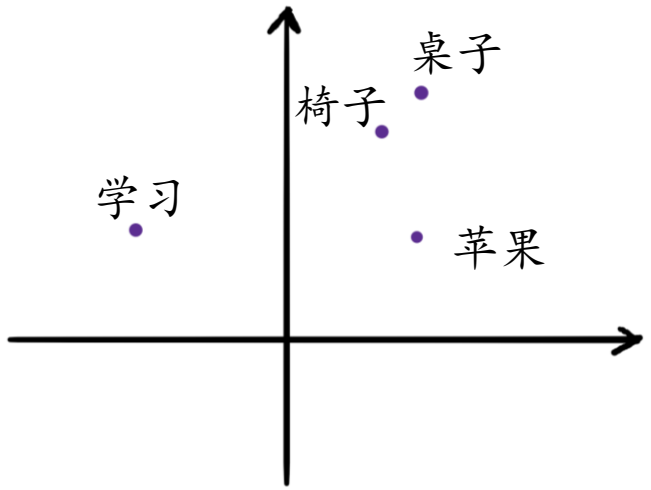
深度学习

- 基于深度学习的语言处理，语言基础还是分布语义学
- 词向量的预学习
- 词向量 因为两个词分布模式相似，所以有相似的向量表示

桌子-椅子-苹果-学习

木头桌子(√)、木头椅子(√) 、木头苹果(×)、木头学习(×)

一把椅子(√) 、一张桌子(√)、一个苹果(√)、一个学习(×)



深度学习

- 典型的预训练辅助任务，填空题
  - 自回归式重构任务

    他 坐 在 _____
    他 坐 在 椅子 _____
    他 坐 在 椅子 上 _____

  - 降噪自编码任务

    他 坐 在 ~~椅子~~ 上 看 书 → 椅子

- 通过辅助任务设计，迫使模型学习分布模式信息
- 自指导的任务设计实现了对超大规模数据的利用，NLP历史上从未有过

提纲

- 引言
- 自然语言处理研究范式变迁
- 深度学习及其带来的进步
- 经验方法的数学基础和语言学基础
- **经验方法的局限性**
- 机器理解和人类理解
- 结束语

经验方法的局限性

- 经验方法的基本思路是模型拟合：
  - 假设一个模型M
  - 用人工标注的数据去调整模型参数，让模型的输出和人类输出接近。

- 语言被视作随机现象或者随机过程的产物

- 模型运作原理和人类语言习得、理解和认知机制没有太大关系

经验方法的局限性

- 词袋模型(BOW)
  - 最简单的NLP模型
  - 很多时候很有效

- 按照词袋模型，如何讲话？

- 其他复杂的模型类似

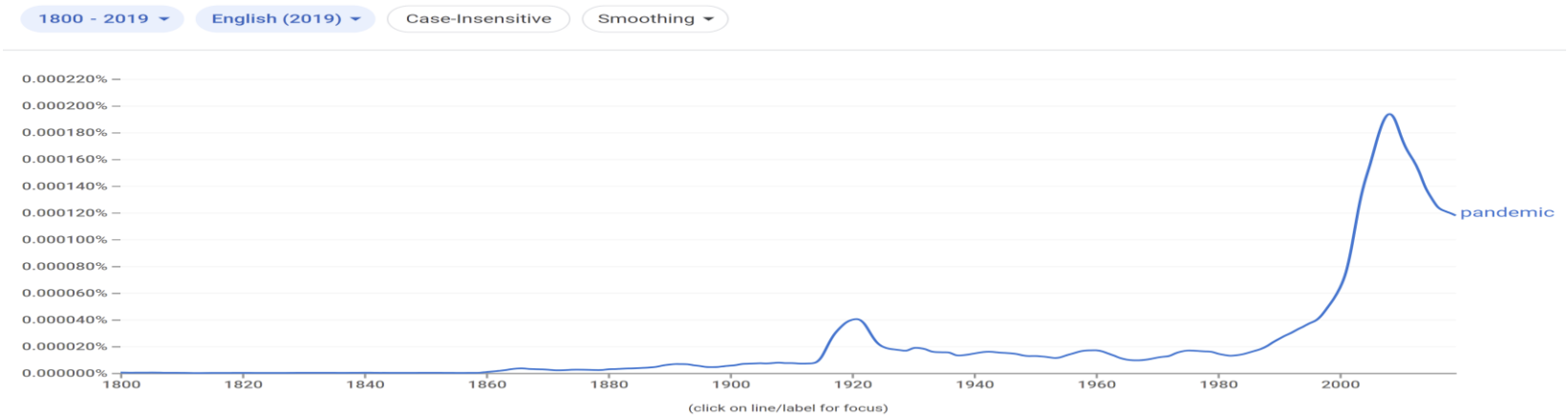
All models are wrong, some are useful.

---George Box

George Box (18 October 1919 – 28 March 2013) was a British statistician, who worked in the areas of quality control, time-series analysis, design of experiments, and Bayesian inference. He has been called "**one of the great statistical minds of the 20th century**".

经验方法的局限性

- 独立同分布假设(i.i.d distribution)
  - 同分布假设：所有数据(训练数据、未来数据)源自同一分布
  - 语言是动态变化的且领域多变，很难说遵从同一分布
  - 独立性假设：所有数据例子之间相互独立
  - 语言对象之间常常并不相互独立



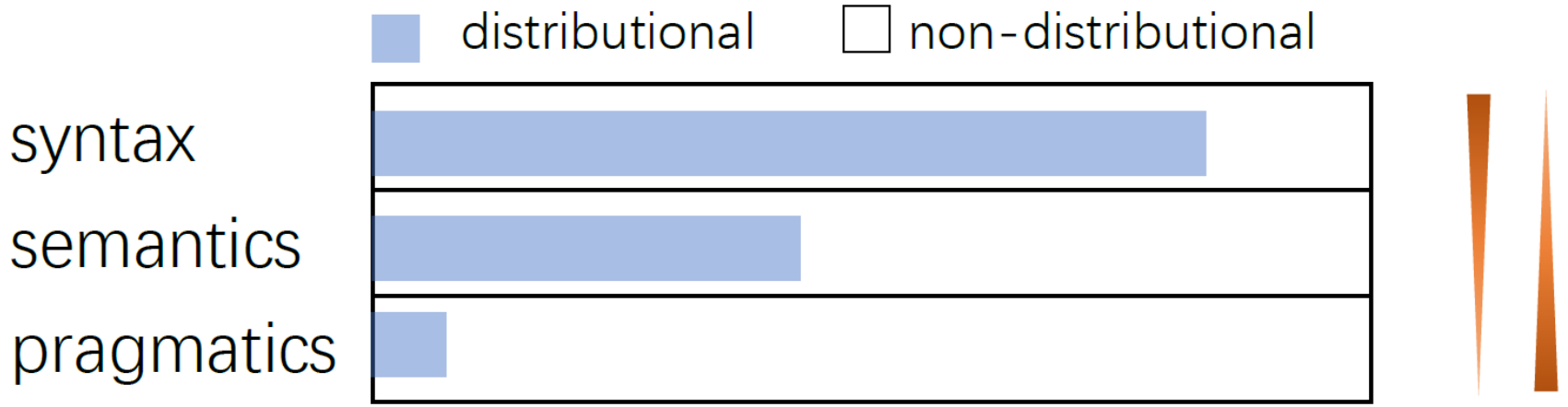
经验方法局限性

- 意义不同体现为分布不同，但并非所有意义不同都通过分布体现

- 经验方法的语言学基础
  - 基于可观察的形式分布推断分布背后的意义
  - 意义、情感、动机、交际功能

- 分布不能推导出所有的语义
  - 分布语义：　可由分布模式推导得到的语义
  - 非分布语义：不可由分布模式推导得到的语义

经验方法的局限性


distributional
non-distributional
syntax
semantics
pragmatics

提纲

- 引言
- 自然语言处理研究范式变迁
- 深度学习及其带来的进步
- 经验主义方法的数学基础和语言学基础
- 经验主义方法的局限性
- **机器理解和人类理解**
- 结束语

机器"理解" vs 人类理解

- 语言理解新任务
  - Machine Reading Comprehension　　机器阅读理解
  - Natural Language Inference (NLI)　　自然语言推理
- 过去很少做，现在很热门，效果很好

Premise	Hypothesis	Label
A man inspects the uniform of a figure in some East Asian country.	The man is sleeping.	contradiction
An older and younger man smiling.	Two men are smiling and laughing at the cats playing on the floor.	neutral
A soccer game with multiple males playing.	Some men are playing a sport.	entailment

机器"理解" vs 人类理解

Premise	Hypothesis	Label
A man inspects the uniform of a figure in some East Asian country.	The man is sleeping.	contradiction
An older and younger man smiling.	Two men are smiling and laughing at the cats playing on the floor.	neutral
A soccer game with multiple males playing.	Some men are playing a sport.	entailment

- 如果把premise去掉？
- 人类还能做出决策吗？ 0%
- 机器还可以做！！！ 67~69% (SNLI)

Unigram Patterns					
Entailment		Neutral		Contradiction	
outdoors	78.8	vacation	91.0	Nobody	99.7
sport	75.1	winning	89.9	No	95.8
instrument	74.4	favorite	88.7	cats	93.4
Several	54.7	addition	69.6	None	85.4
Yes	54.4	also	68.6	refused	80.5
various	53.7	locals	65.7	never	79.0

机器"理解" vs 人类理解

- 机器语言能力 超过 人类语言能力
- 评测的作用：寻找问题、改善模型
- 设置人类性能指标的本意(topline)
- 单个数据集上的超越
- 机器能力分数与人类能力分数的可比性

提纲

- 引言
- 自然语言处理研究范式变迁
- 深度学习及其带来的进步
- 经验主义方法的数学基础和语言学基础
- 经验主义方法的局限性
- 机器理解和人类理解
- **结束语**

结束语

- 自然语言理解很困难
- 使用传统机器学习方法，为自然语言处理性能带来显著进步
- 使用深度学习方法，为自然语言处理性能再次带来显著进步
- 不过，自然语言处理的数学基础和语言学基础没有根本性变化

"道" 和 "术"

- "道" 和 "术" 都很重要
- 有术无道，止于术。
- 由术入道，以道驭术。

Translation memorandum -- Weaver 1949

This approach brings into the foreground an aspect of the matter that probably is absolutely basic - namely, the statistical character of the problem. "Perfect" translation is almost surely unattainable. Processes, which at stated confidence levels will produce a translation which contains only X per cent "error," are almost surely attainable.

And it is one of the chief purposes of this memorandum to emphasize that statistical semantic studies should be undertaken, as a necessary preliminary step.

Translation memorandum -- Weaver 1949

Think, by analogy, of individuals living in a series of tall closed towers, all erected over a common foundation. When they try to communicate with one another they shout back and forth, each from his own closed tower. It is difficult to make the sound penetrate even the nearest towers, and communication proceeds very poorly indeed. But when an individual goes down his tower, he finds himself in a great open basement, common to all the towers. Here he establishes easy and useful communication with the persons who have also descended from their towers.

谢谢