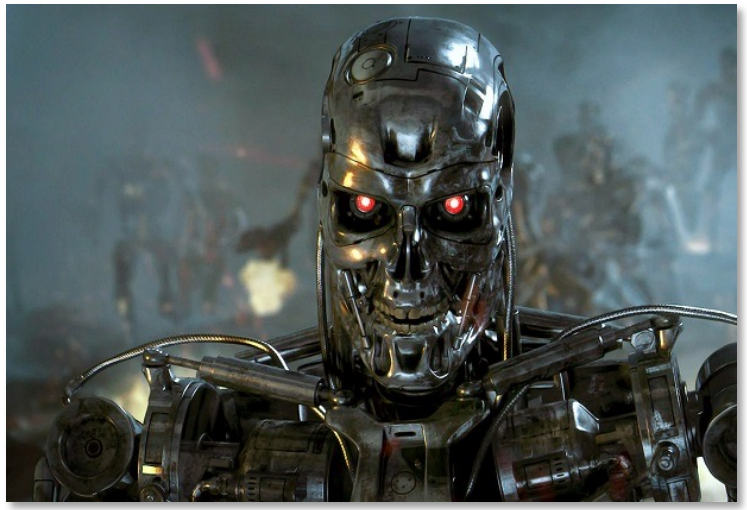
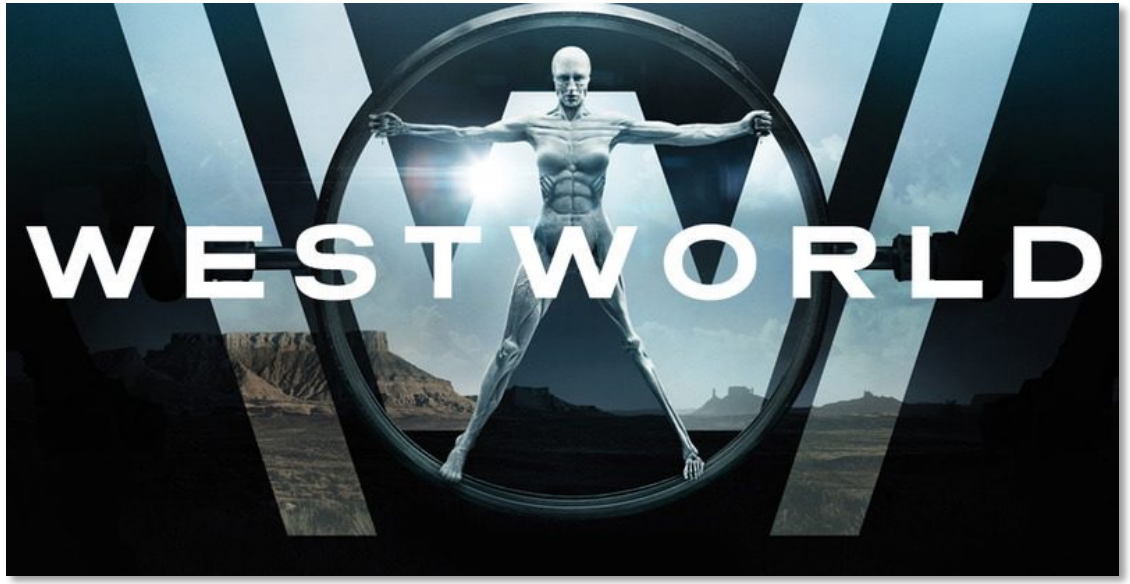
机器读心术—自然语言处理

复旦大学计算机科学技术学院

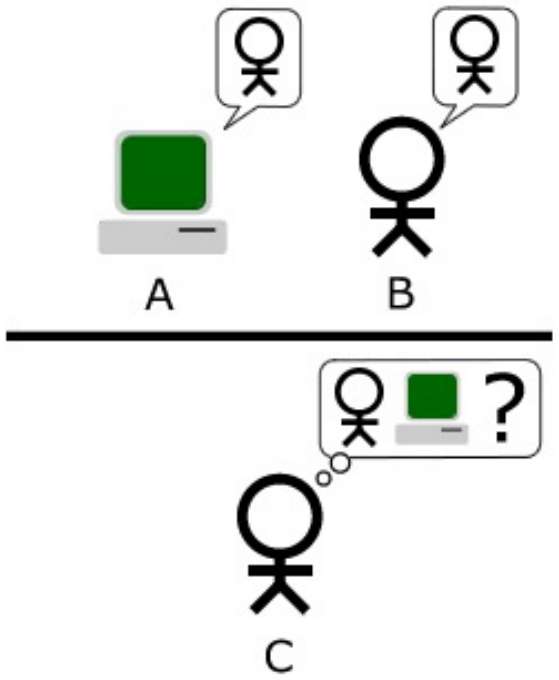
张奇

那些年，电影里的人工智能





图灵测试——屏幕后和你聊天的是人类还是计算机？



模仿游戏

机器能思考吗?

艾伦·麦席森·图灵

(1912 - 1954)

提问：给我写一首歌颂爱国的诗歌吧？

回答：别开我玩笑了，我哪有这个能耐啊。

提问: 34,957 加上 70,764等于多少？

回答: (停顿了大概10秒) 105,621.

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

自然语言处理是目前以及未来 AI 领域最重要的基础技术之一

自然语言处理技术

- 自然语言理解
  - 自然语言转化为计算机程序更易于处理的形式
- 自然语言生成
  - 把计算机数据转化为自然语言

自然语言处理是目前以及未来 AI 领域最重要的基础技术之一

对话系统

机器翻译

阅读理解

智能金融

智能医疗

智能司法

自然语言处理的应用

机器翻译——语言无障碍

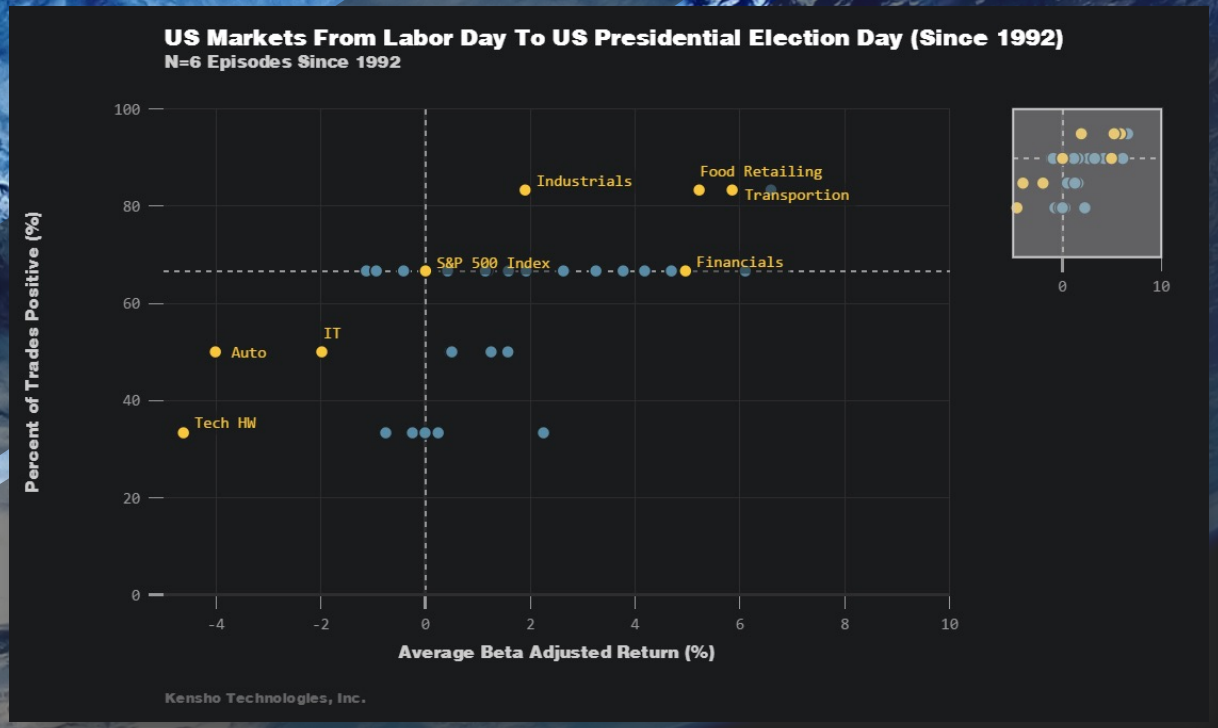
谷歌翻译每天服务2亿次

1960　　1990 – 2000　　2015　　Future

语言翻译的岗位会消失吗?

自然语言处理的应用

Kensho：比阿尔法狗还残暴的华尔街之狼



Kensho：1992年起，总统竞选阶段行业情况概览

■ Kensho如何影响传统金融

Kensho的软件Warren主要能实现两种功能：寻找事件和资产之间的相关性及对于其价格的影响，以及基于这些事件对资产未来价格走势做预测。

■ 如何用好Kensho这类智能金融产品？

——“你仍然需要问正确的问题”

Warren只能做到变量延展，但却无法替用户去“逻辑推理”事件可能的影响因素。

——“相关性不代表因果性”

当影响资产的相关因素越来越多的时候，如何识别事件背后的相关性和因果性就变得更加困难。

自然语言的应用

QUANTITATIVE SOCIAL SCIENCE

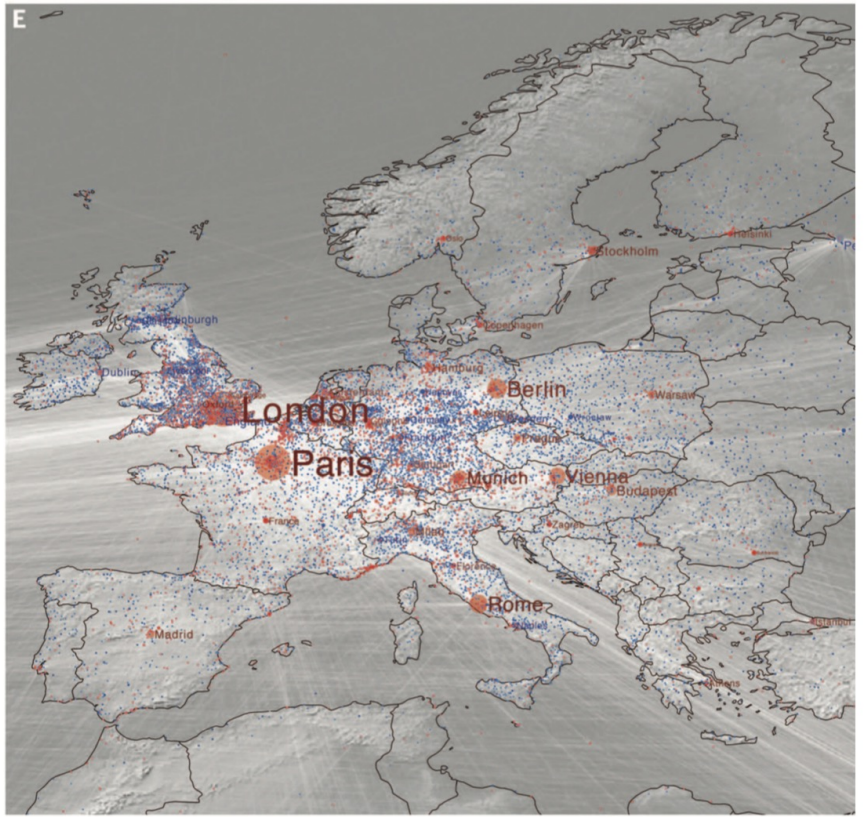
A network framework of cultural history

Maximilian Schich,[1,2,3*] Chaoming Song,[4] Yong-Yeol Ahn,[5] Alexander Mirsky,[2] Mauro Martino,[3] Albert-László Barabási,[3,6,7] Dirk Helbing[2]

Winckelmann Corpus, 18世纪人物

名人

出生地点 ---> 死亡地点

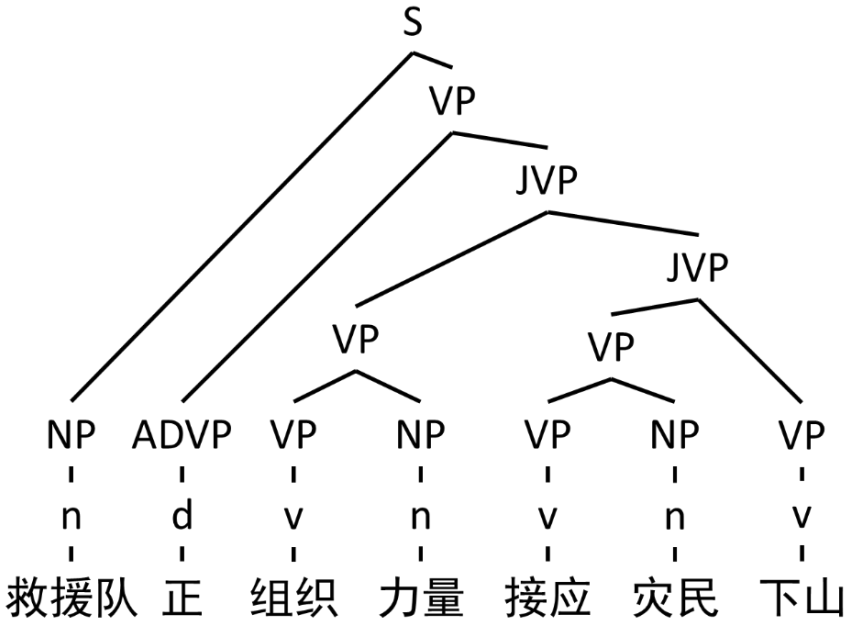


Freebase

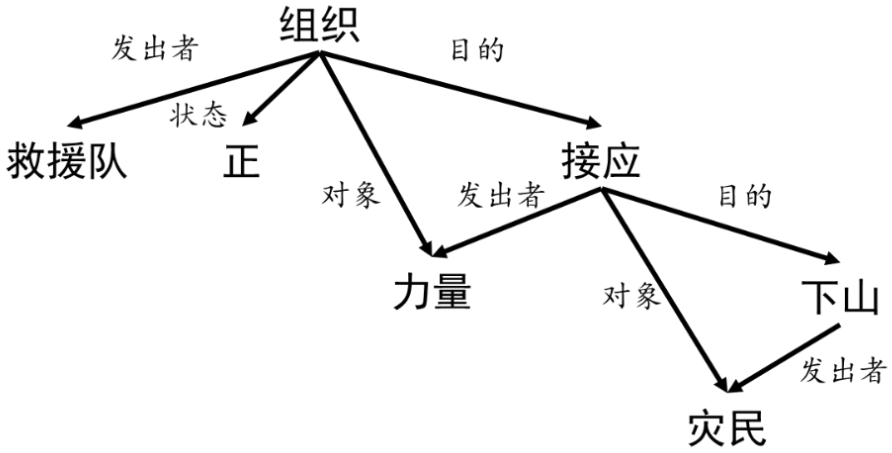
自然语言理解

输入： 救援队正组织力量接应灾民下山

输出：



句法结构



语义结构

自然语言理解的本质是结构预测

《自然语言处理》 刘知远

自然语言理解复杂程度

中文6763个常用字（GB2312），新闻文档平均句子长度20+

6763^20 = 4.00e+76

6763^40 = 1.61e+153

利用句法语义等知识**约束**求解

特点	例句
口语化	亲，请点赞哦！
缩略语	中石油
中英混合	我今天很Happy
新词	累觉不爱

自然语言理解复杂程度

分词

- 白天鹅在水里游泳
- 该研究所获得的成果

指代消解

- 重庆队得88分，客场负于台湾队2分。（台湾队和重庆队各得多少分？比赛地点？）

隐喻、幽默、夸张、双关、映射

- 冬天，能穿多少穿多少；夏天，能穿多少穿多少。
- 单身的来由：原来是喜欢一个人，现在是喜欢一个人。

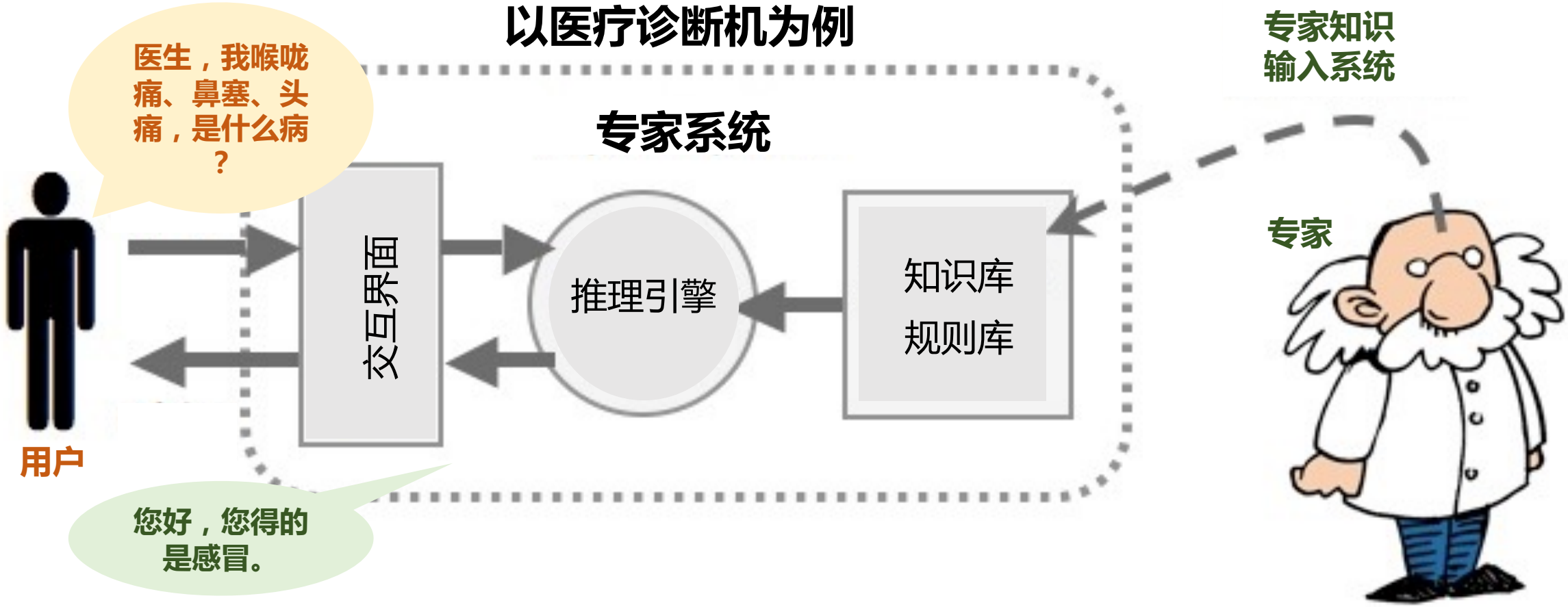
1970 — 80: 把全世界的知识都记录下来吧！

将知识基于规则表达

- 狗都有四条腿。
- 卡拉是条狗。

→ 所以卡拉有四条腿。

以医疗诊断机为例



1990 –至今: 上帝也会掷骰子

基于统计的方法

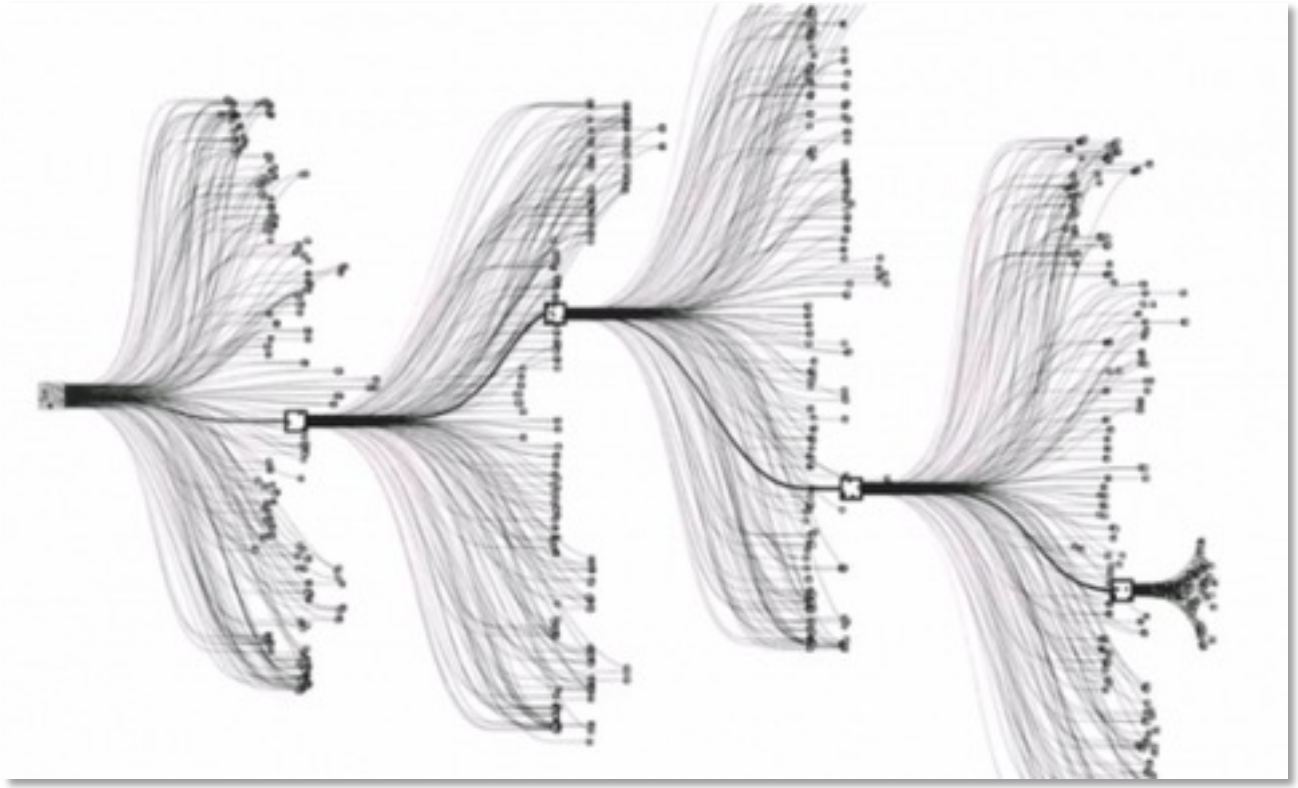
- 概率图模型
- 统计机器学习

买早餐要不要排队？

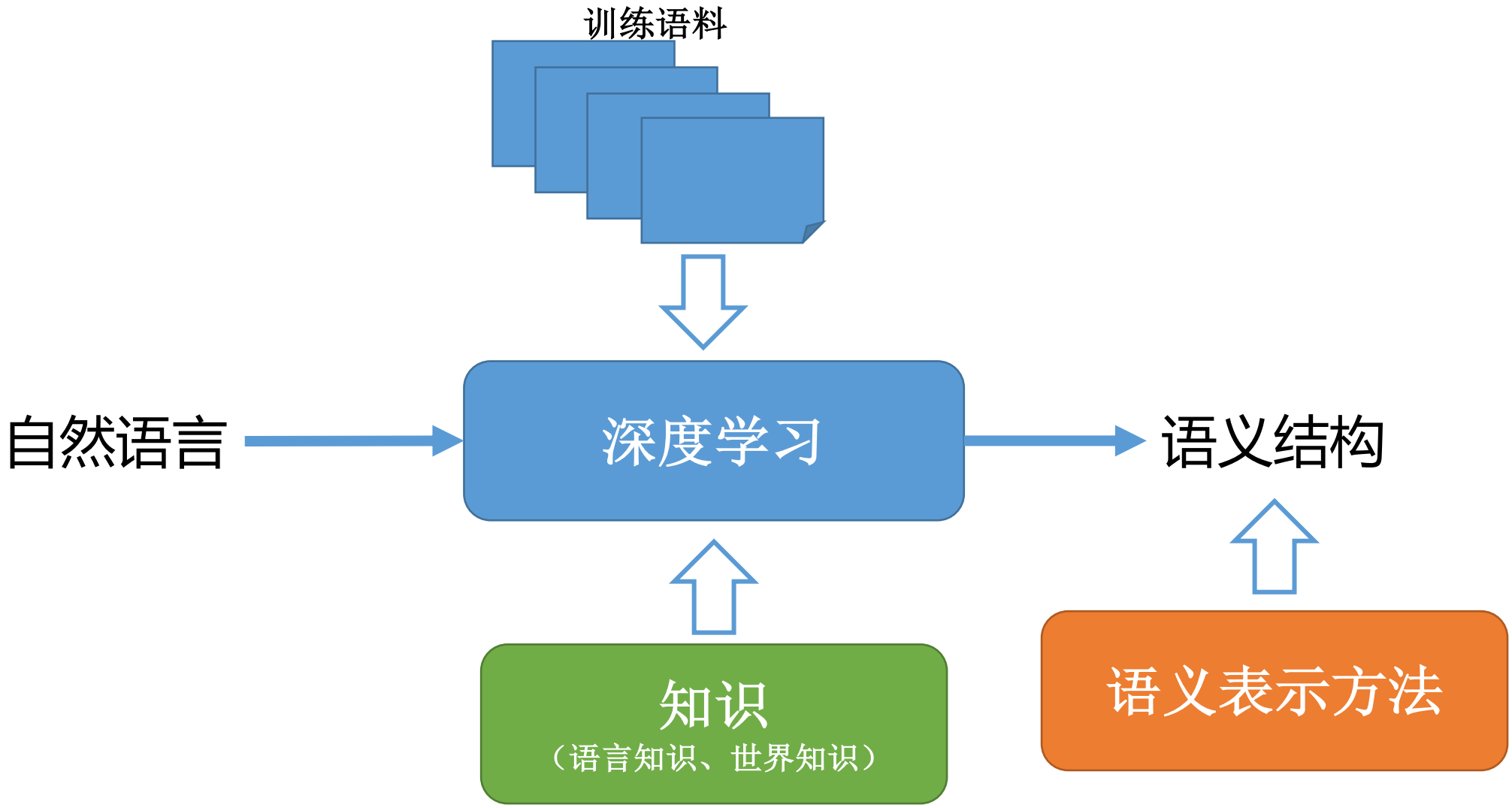
堵车吗？

今天天气怎么样？

公交上有没有位置？



自然语言理解技术发展趋势



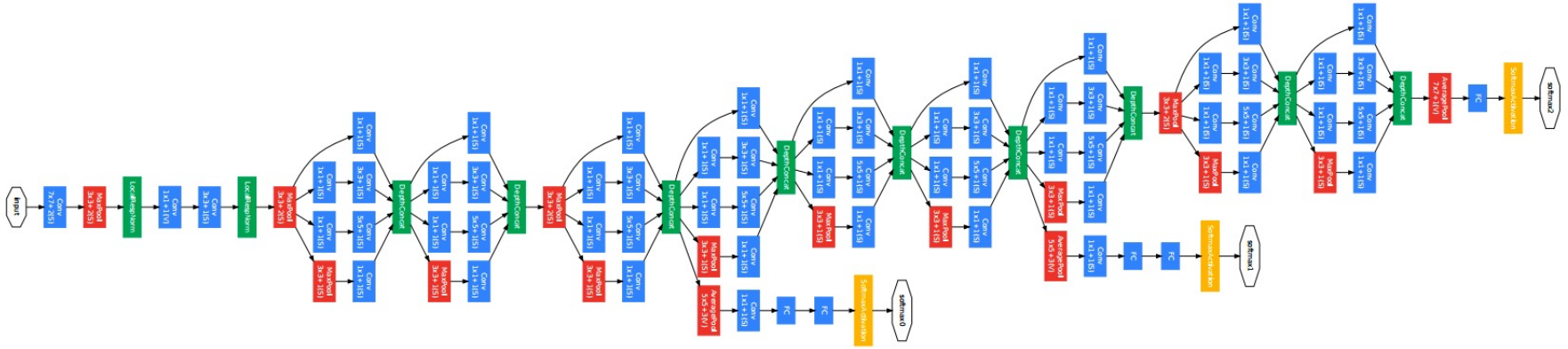
深度学习

深度学习是机器学习中的一个子领域

传统的机器学习方法通常需要**人工设计的表示以及抽取特征**

深度学习方法与传统机器学习方法不同，它可以直接进行**自动的表示学习**

Feature	NER
Current Word	✓
Previous Word	✓
Next Word	✓
Current Word Character n-gram	all
Current POS Tag	✓
Surrounding POS Tag Sequence	✓
Current Word Shape	✓
Surrounding Word Shape Sequence	✓
Presence of Word in Left Window	size 4
Presence of Word in Right Window	size 4



深度学习

深度学习使得自然语言处理众多任务取得了重大进展

- **基础技术**：分词、词性标注、实体识别、成分分析、句法分析、语义表示、语义匹配、情感倾向分析等
- **应用系统**：问题回答、对话系统、机器翻译、阅读理解等

单词向量表示 Word Embedding

expect =

$$\begin{pmatrix} 0.286 \\ 0.792 \\ -0.177 \\ -0.107 \\ 0.109 \\ -0.542 \\ 0.349 \\ 0.271 \\ 0.487 \end{pmatrix}$$

litoria

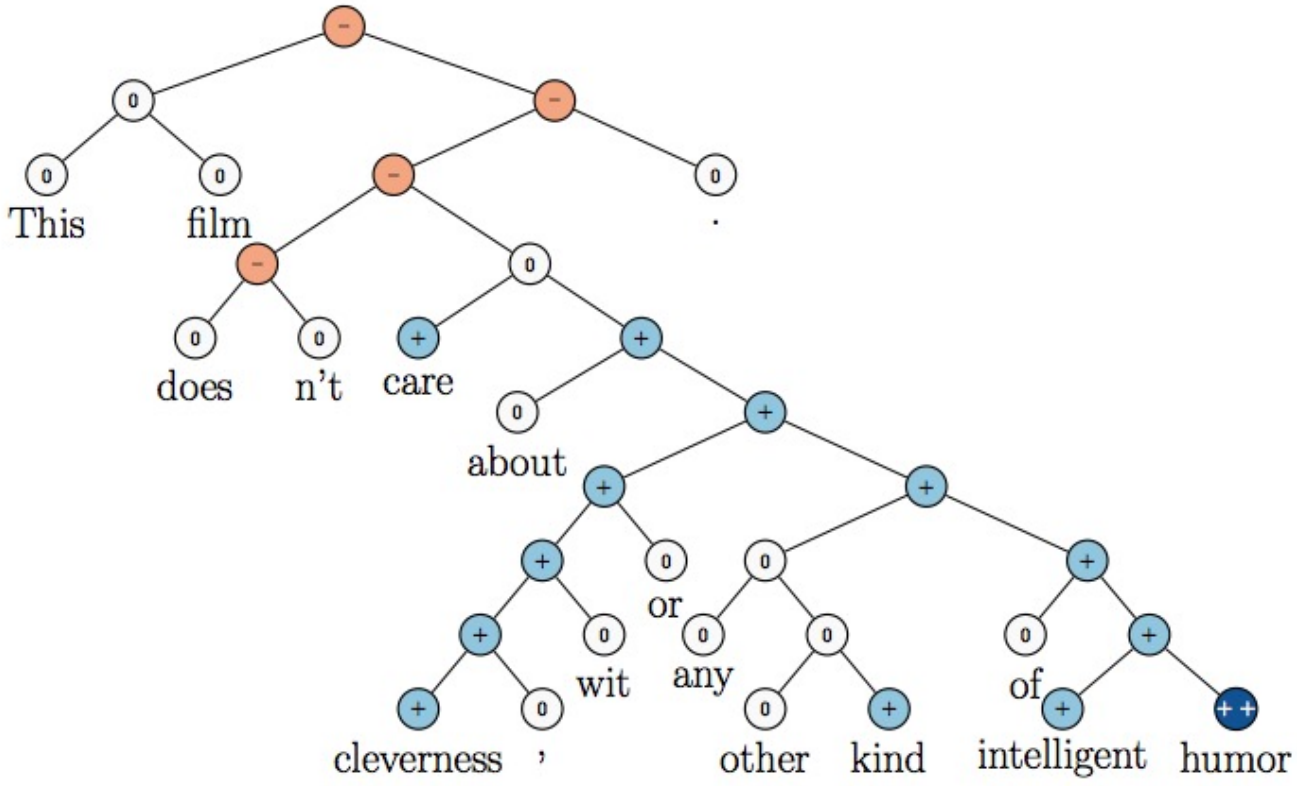
leptodactylidae

rana

eleutherodactylus

King - Man + Woman = Queen

情感倾向分析

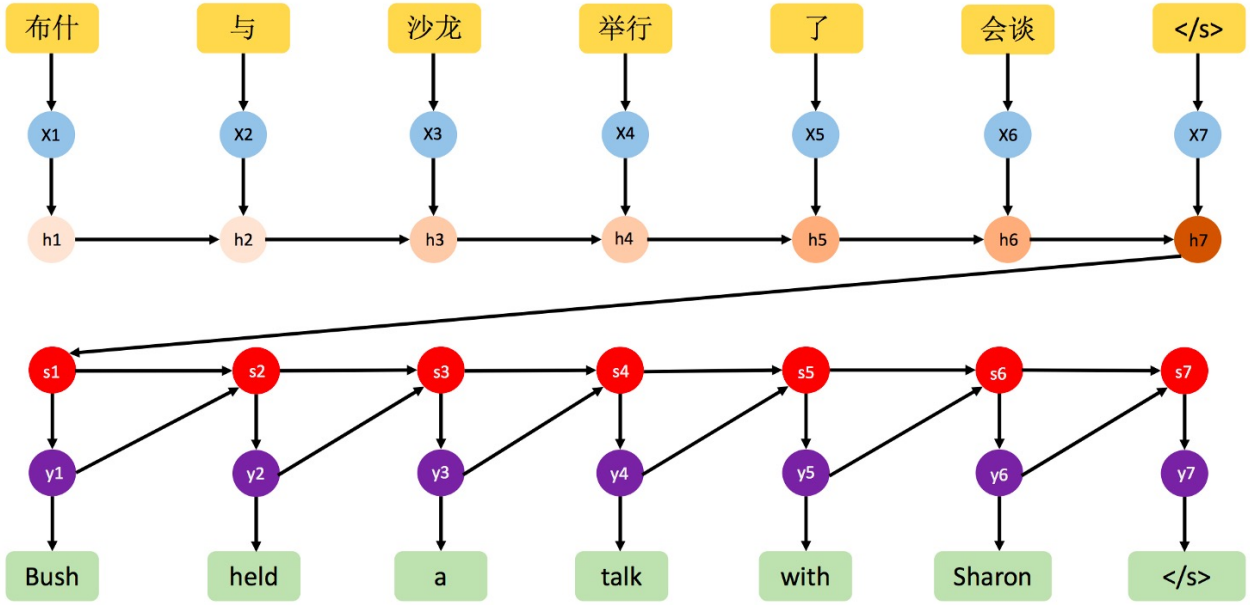


Socher et al. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank, EMNLP 2013

机器翻译

将源语言转换为一个向量表示，之后将该向量作为输入生成目标语言

[Sutskever et al. 2014, Bahdanau et al. 2014, Luong and Manning 2016]



(Sutskever et al., 2014)

生成式文本摘要

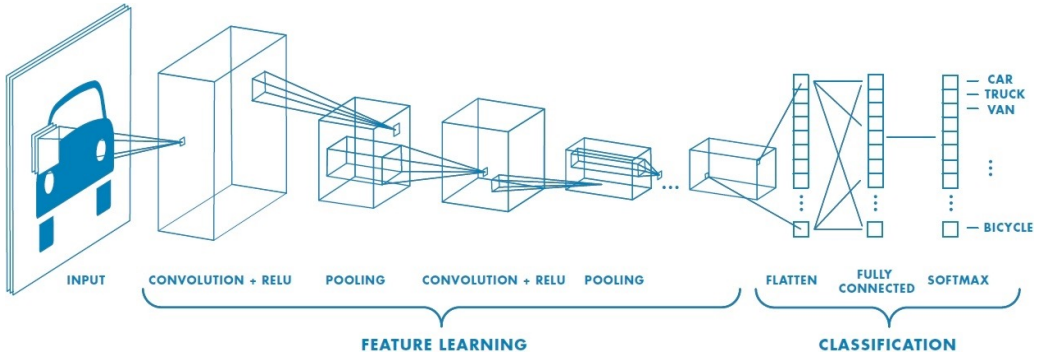
The bottleneck is no longer access to information; now it's our ability to keep up.
AI can be trained on a variety of different types of texts and summary lengths.
A model that can generate long, coherent, and meaningful summaries remains an open research problem.

The last few decades have witnessed a fundamental change in the challenge of taking in new information. The bottleneck is no longer access to information; now it's our ability to keep up. We all have to read more and more to keep up-to-date with our jobs, the news, and social media. We've looked at how AI can improve people's work by helping with this information deluge and one potential answer is to have algorithms automatically summarize longer texts. Training a model that can generate long, coherent, and meaningful summaries remains an open research problem. In fact, generating any kind of longer text is hard for even the most advanced deep learning algorithms. In order to make summarization successful, we introduce two separate improvements: a more contextual word generation model and a new way of training summarization models via reinforcement learning (RL). The combination of the two training methods enables the system to create relevant and highly readable multi-sentence summaries of long text, such as news articles, significantly improving on previous results. Our algorithm can be trained on a variety of different types of texts and summary lengths. In this blog post, we present the main contributions of our model and an overview of the natural language challenges specific to text summarization.

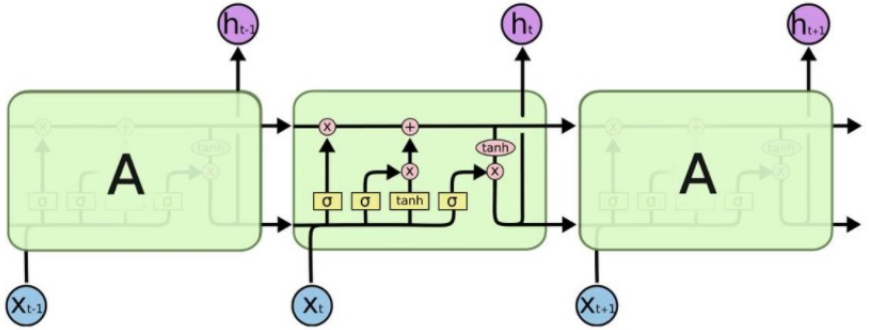
Paulus, Romain, Caiming Xiong, and Richard Socher. "A deep reinforced model for abstractive summarization." 2017

https://tryolabs.com/blog/2017/12/12/deep-learning-for-nlp-advancements-and-trends-in-2017/

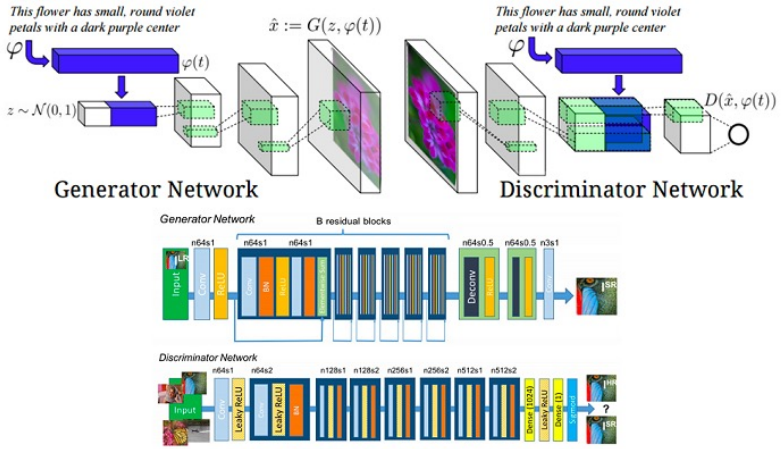
深度学习



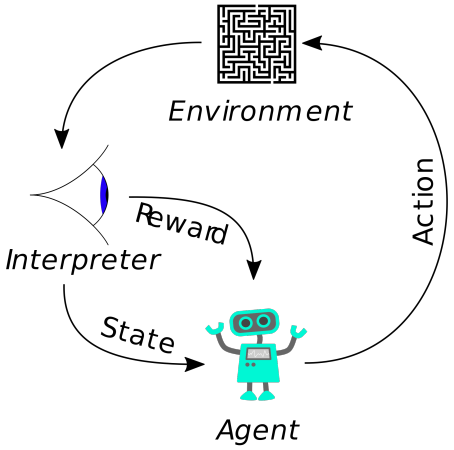
Convolutional Neural Networks



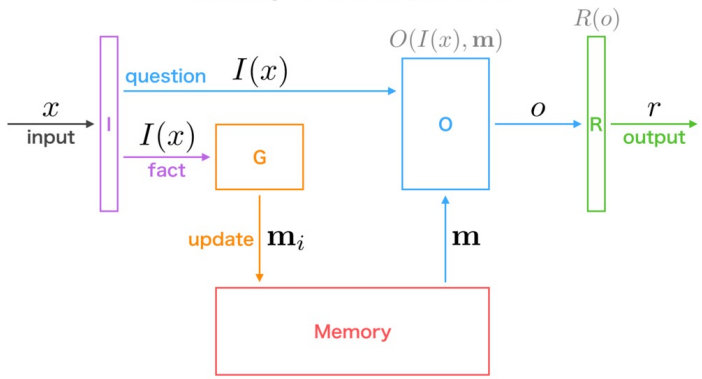
Recurrent Neural Networks



Generative Adversarial Networks



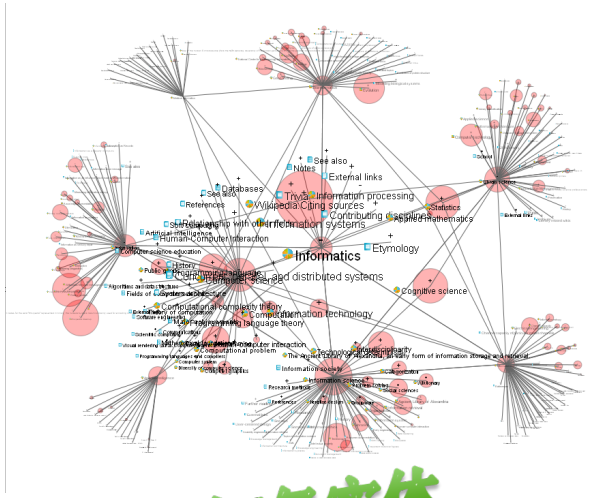
Reinforcement Learning



Memory Network

知识图谱

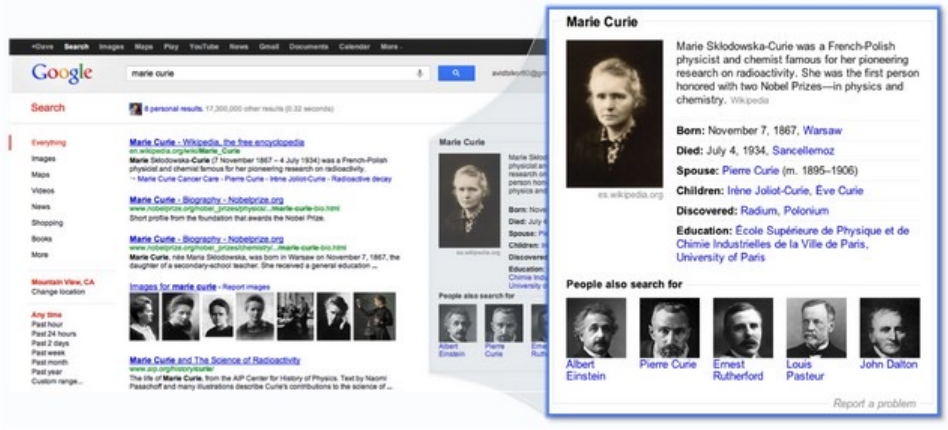
真正理解自然语言需要大规模、高覆盖率的知识资源



超过5亿实体
超过35亿条关系

ProBase

Google



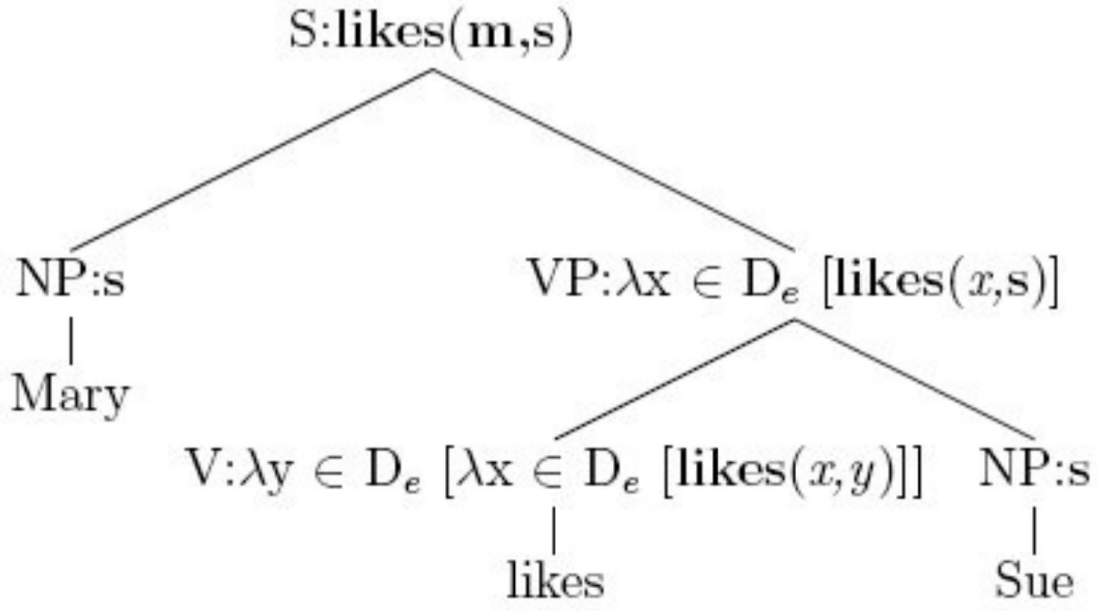
百度知心

搜狗知立方

目前的知识资源难以满足中文理解的需求

语义表示

真正理解自然语言需要大规模、高覆盖率的知识资源



Lambda Calculus

∀ = For all ; [e.g : every one, every body, any time, etc]

∃ = There exists ; [e.g : some one, some time, etc]

⇒ = Implication ; [if … then ….]

⇔ = Equivalent ; biconditional [if … and … only … if …]

¬ = Not ; negation

∨ = OR ; disjunction

∧ = AND ; conjunction

First Order Logic

自然语言处理发展趋势

论文写作　　百科写作

报告生成

新闻写作

语言生成　　自动作诗

专利写作

语言理解　　自动问答

机器翻译

文本分类

阅读理解

自动摘要

对话系统

目前自然语言处理方法的局限

自然语言处理到底发展到什么水平？

- **封闭环境VS开放环境**
  每一个任务的性能都是在封闭测试的环境下得到的，如果在开放环境下，性能会大大下降。

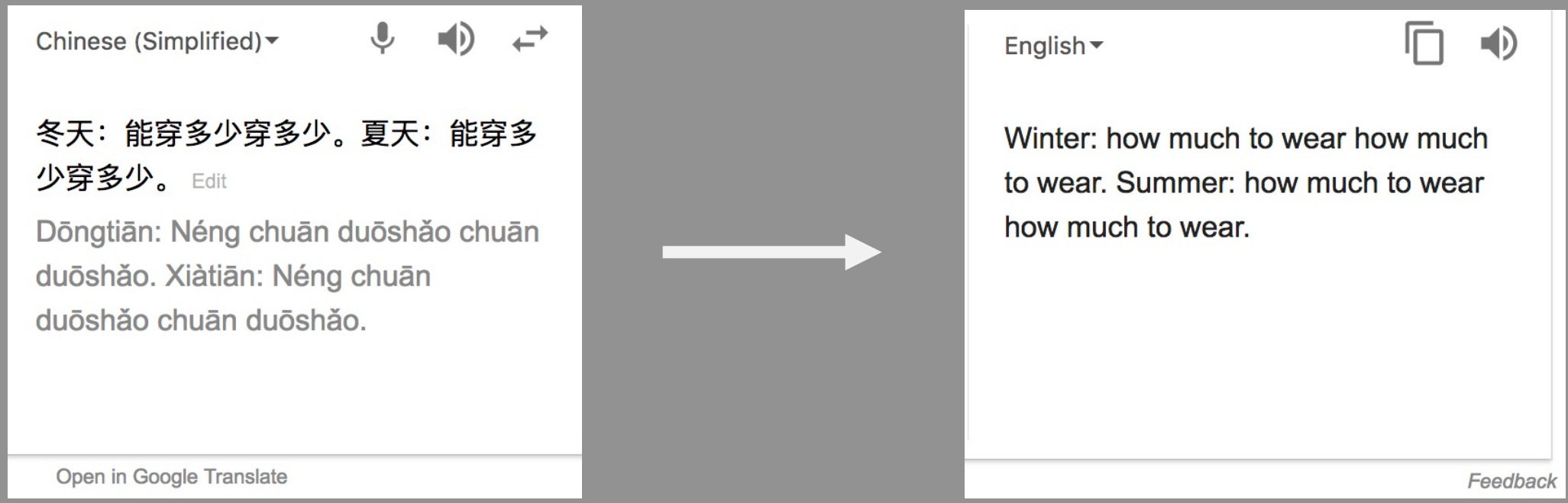
- **噪音敏感问题**
  对于噪音问题的敏感度极大，比如语音识别任务。

- **模式识别VS真正理解**
  所有智能的本质都是模式识别，而不是真正意义上的理解。

- **严重依赖数据**
  成功的方法大都严重依赖标记语料集合，没有数据就没有智能。

目前自然语言处理方法的局限

机器翻译真的明白你想说的意思吗？



自然语言处理应用

自然语言处理技术应用的5个基础条件

单一清晰的领域

自然语言处理算法研究员

海量数据

超大计算量

自动标注数据


自然语言处理是目前以及未来 AI 领域最重要的基础技术之一
THANK YOU
LUSITANIA ECLIPSE