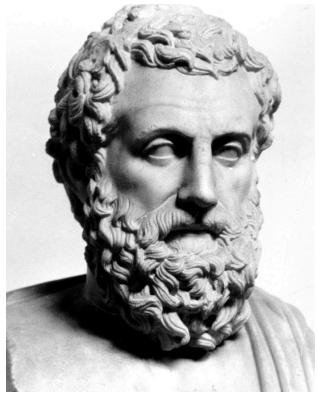
Artificial Intelligence For NLP Lesson-01



开课吧人工智能学院

2020.Jan.05

Outline

i. Course Background

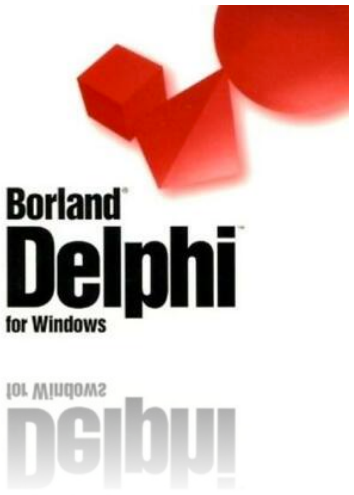
ii. AI Introduction

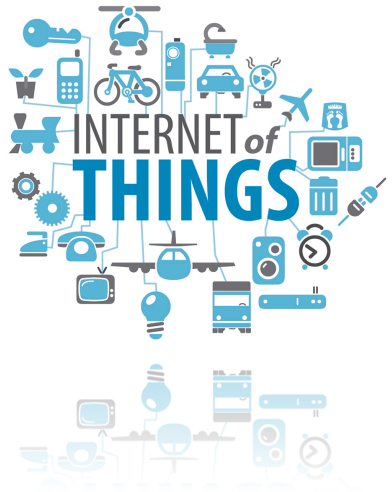
iii. Syntax Tree, Probability Language Model

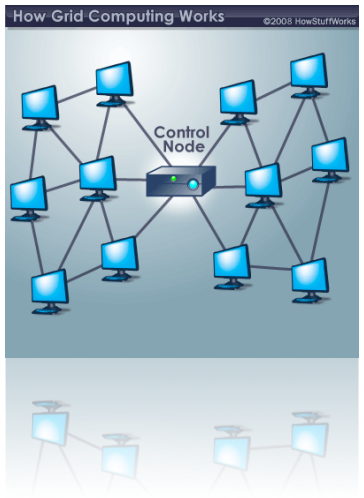
iv. 1st Assignment.

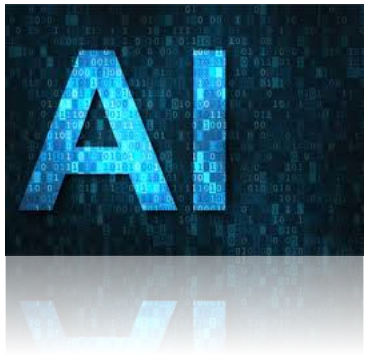
1/5 Background

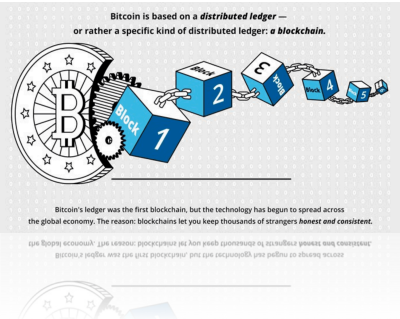
Things change, with time goes by.











Our mentors

- And plenty of references: https://github.com/Computing-Intelligence/References

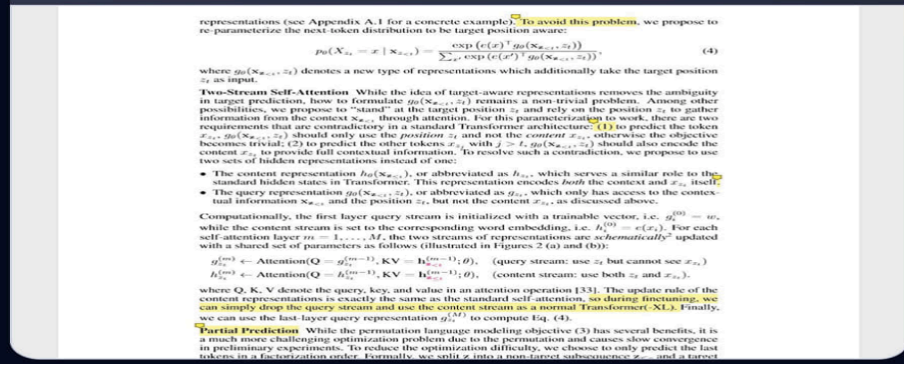
Our Project

项目1、文本自动摘要系统的构建讲解与

技术关键字：TextRank，Sentence Embedding, ex–tractive summarization

项目2、PDF，Word，Excel智能高亮标注

技术关键字：文档自动化读取， NER(命名实体识别)， 依存分析， 文档结构化解析与可视化



representations (see Appendix A.1 for a concrete example). To avoid this problem, we propose to re-parameterize the next-token distribution to be target position aware:

$$p_\theta(X_{z_t} = x \mid \mathbf{x}_{<z_t}) = \frac{\exp\left(e(x)^\top g_\theta(\mathbf{x}_{<z_t}, z_t)\right)}{\sum_{x'} \exp\left(e(x')^\top g_\theta(\mathbf{x}_{<z_t}, z_t)\right)}, \quad (4)$$

where $g_\theta(\mathbf{x}_{<z_t}, z_t)$ denotes a new type of representations which additionally take the target position z_t as input.

Two-Stream Self-Attention While the idea of target-aware representations removes the ambiguity in target prediction, how to formulate $g_\theta(\mathbf{x}_{<z_t}, z_t)$ remains a non-trivial problem. Among other possibilities, we propose to "stand" at the target position z_t and rely on the position z_t to gather information from the context $\mathbf{x}_{<z_t}$ through attention. For this parameterization to work, there are two requirements that are contradictory in a standard Transformer architecture: (1) to predict the token x_{z_t}, $g_\theta(\mathbf{x}_{<z_t}, z_t)$ should only use the *position* z_t and not the *content* x_{z_t}, otherwise the objective becomes trivial; (2) to predict the other tokens x_{z_j} with $j > t$, $g_\theta(\mathbf{x}_{<z_t}, z_t)$ should also encode the content x_{z_t} to provide full contextual information. To resolve such a contradiction, we propose to use two sets of hidden representations instead of one:

- The content representation $h_\theta(\mathbf{x}_{\leq z_t})$, or abbreviated as h_{z_t}, which serves a similar role to the standard hidden states in Transformer. This representation encodes *both* the context and x_{z_t} itself.
- The query representation $g_\theta(\mathbf{x}_{<z_t}, z_t)$, or abbreviated as g_{z_t}, which only has access to the contextual information $\mathbf{x}_{<z_t}$ and the position z_t, but not the content x_{z_t}, as discussed above.

Computationally, the first layer query stream is initialized with a trainable vector, i.e. $g_i^{(0)} = w$, while the content stream is set to the corresponding word embedding, i.e. $h_i^{(0)} = e(x_i)$. For each self-attention layer $m = 1, \ldots, M$, the two streams of representations are *schematically* updated with a shared set of parameters as follows (illustrated in Figures 2 (a) and (b)):

$$g_{z_t}^{(m)} \leftarrow \text{Attention}(\mathrm{Q} = g_{z_t}^{(m-1)}, \mathrm{KV} = \mathbf{h}_{<z_t}^{(m-1)}; \theta), \quad \text{(query stream: use } z_t \text{ but cannot see } x_{z_t})$$
$$h_{z_t}^{(m)} \leftarrow \text{Attention}(\mathrm{Q} = h_{z_t}^{(m-1)}, \mathrm{KV} = \mathbf{h}_{\leq z_t}^{(m-1)}; \theta), \quad \text{(content stream: use both } z_t \text{ and } x_{z_t}).$$

where Q, K, V denote the query, key, and value in an attention operation [33]. The update rule of the content representations is exactly the same as the standard self-attention, so during finetuning, we can simply drop the query stream and use the content stream as a normal Transformer(-XL). Finally, we can use the last-layer query representation $g_{z_t}^{(M)}$ to compute Eq. (4).

Partial Prediction While the permutation language modeling objective (3) has several benefits, it is a much more challenging optimization problem due to the permutation and causes slow convergence in preliminary experiments. To reduce the optimization difficulty, we choose to only predict the last tokens in a factorization order. Formally, we split $\mathbf{z}$ into a non-target subsequence $\mathbf{z}_{\leq c}$ and a target

项目3、面向服务的对话机器人的构建

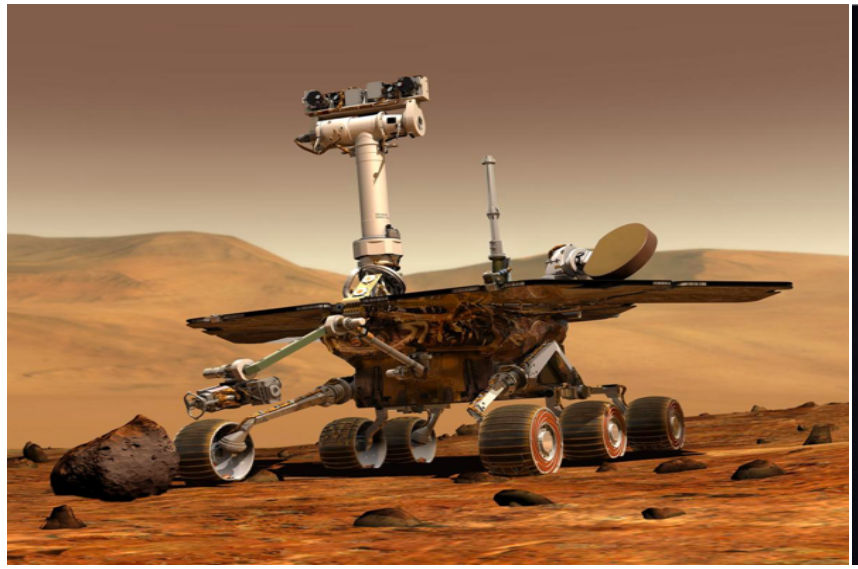
技术关键字: 意图识别 数据检索， 专家系统， 语言生成

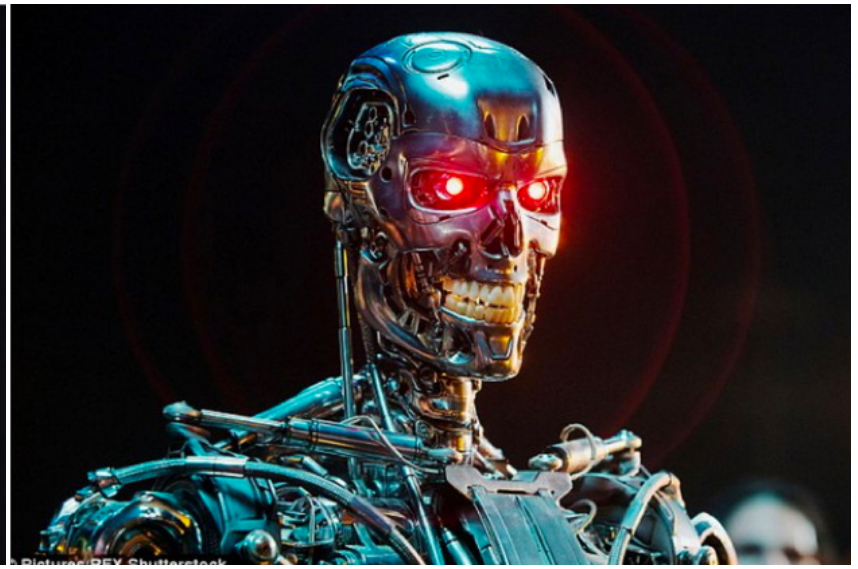
项目4、在线舆情自动检测系统

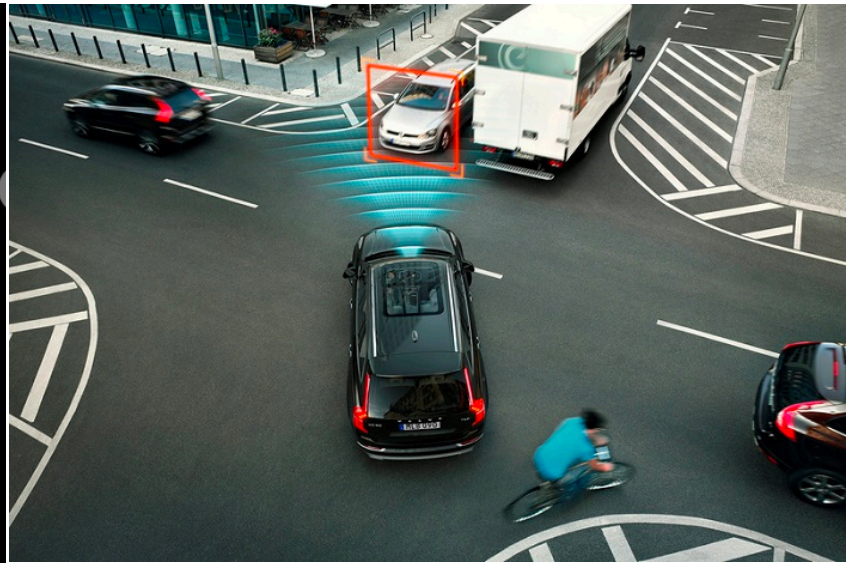
技术关键字： 情感分类，爬虫，文本结构化， 数据可视化

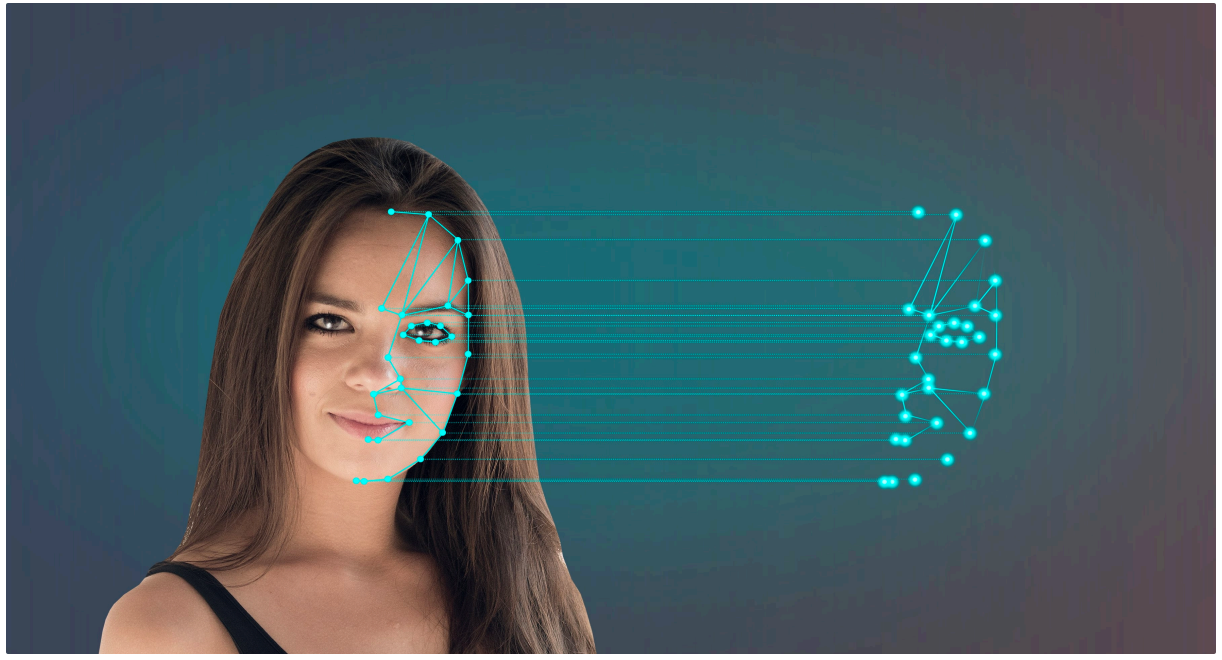
2/5 AI Introduction

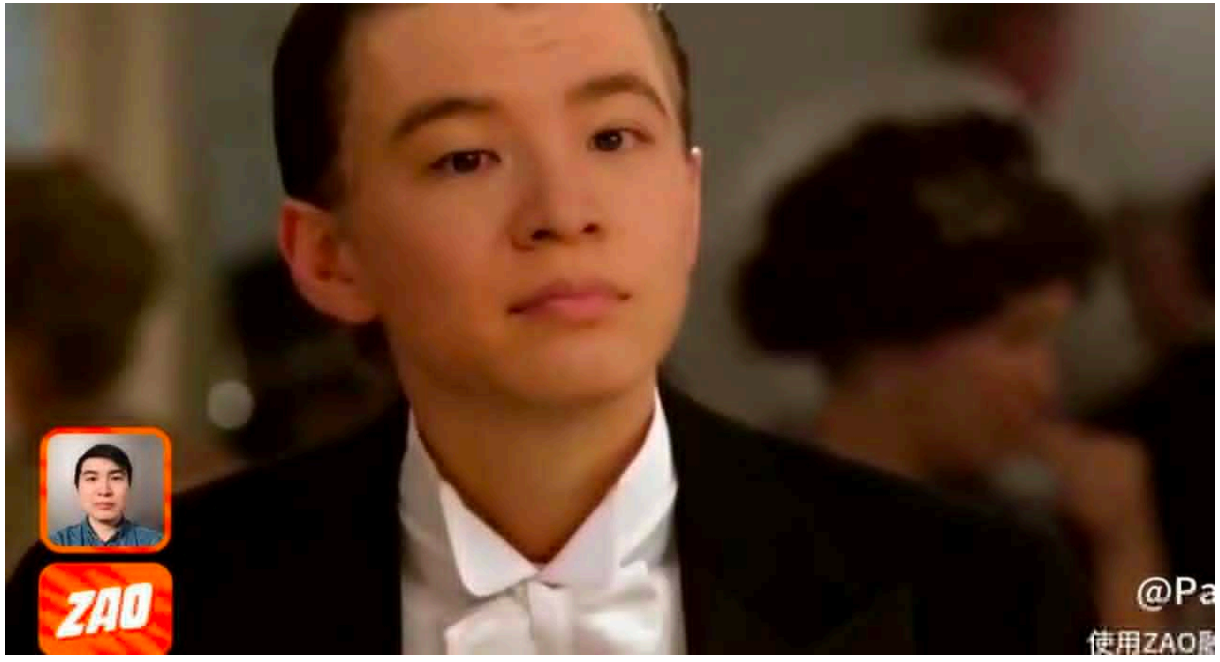
当我们提起AI的时候







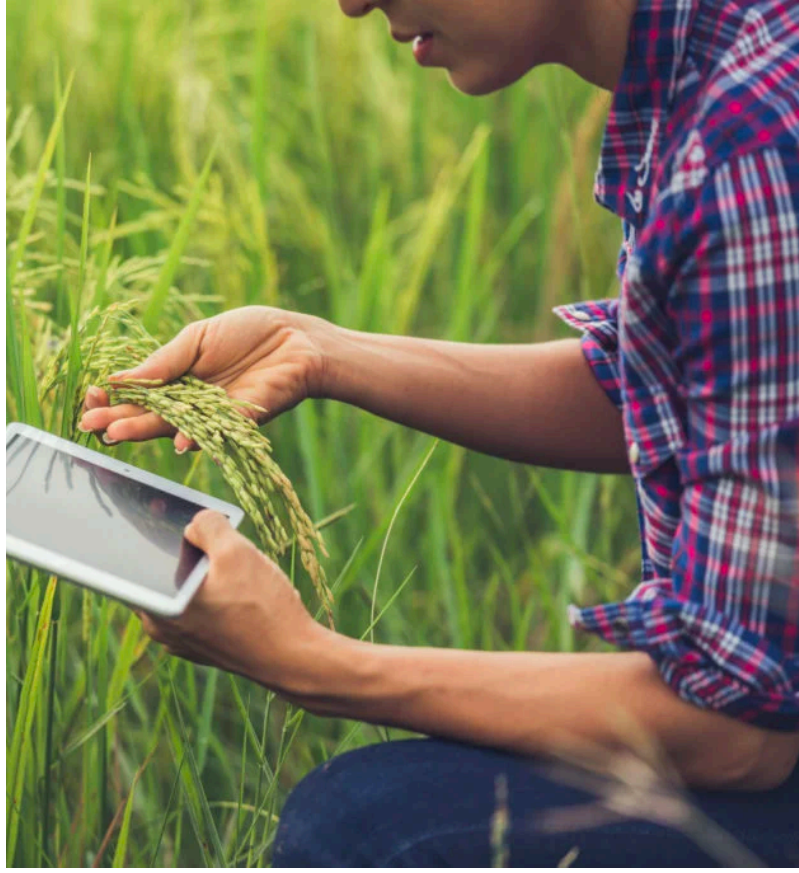


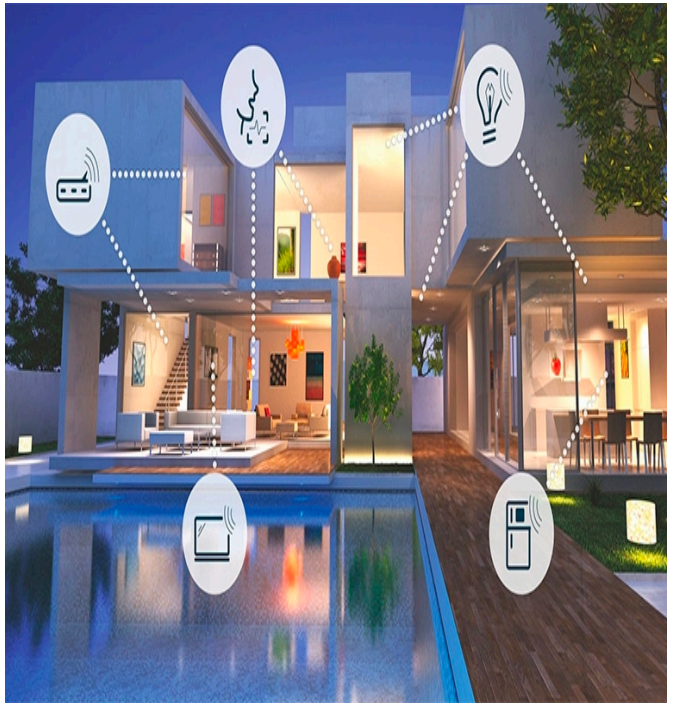

ZAO


Microsoft
微软五代小冰
“在亿万人之中，我只属于你。”
支持第三方平台
微软小冰是领先的跨平台人工智能机器人。目前，你已可以在以下平台使用她。技术对接需要逐步完成，敬请期待更多第三方平台。
关于小冰的常见问题
初次领养
已领养，登录
请输入中国大陆的手机号码
请输入右侧验证码
下一步
Microsoft

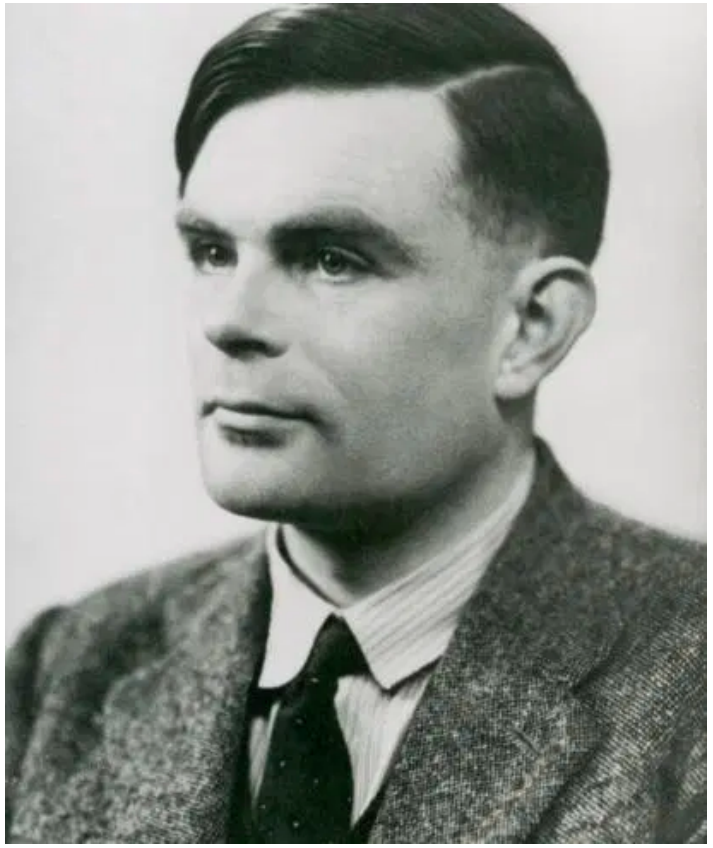
这些是不是AI呢？

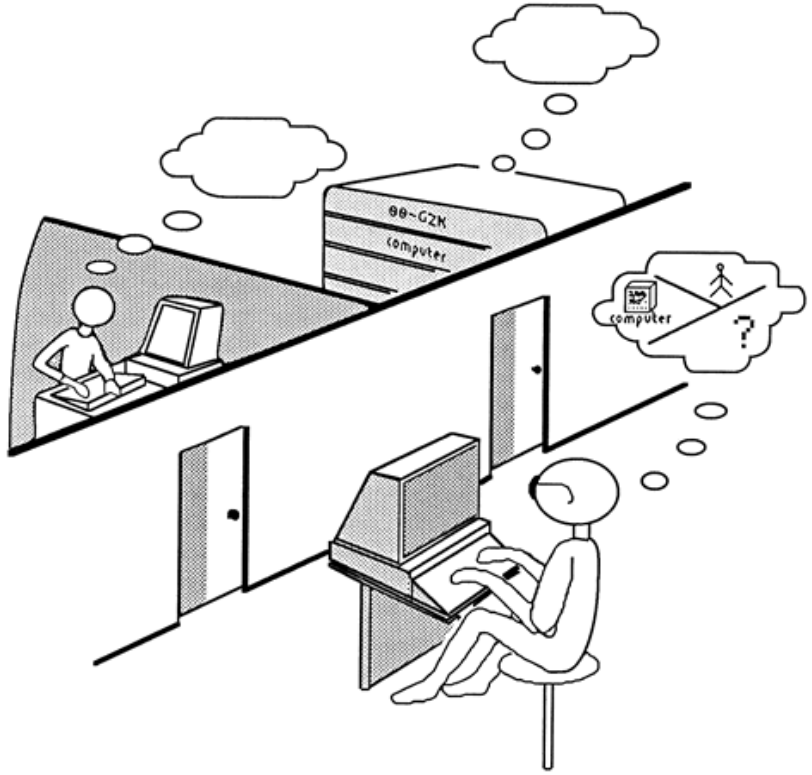
- 新德里、上海、杭州正在使用智慧系统降低车辆拥堵
- 农业学家使用更加先进的方式，跟踪农作物的生长
- 智慧家庭能够让我们远程控制自己的电器，甚至电器能够自己控制自己
- 使用新的能源控制系统，摩天大楼的能源消耗可以降低67%





Intelligence? 图灵测试





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

COMPUTING MACHINERY AND INTELLIGENCE

By A. M. Turing

1. The Imitation Game

I propose to consider the question, "Can machines think?" This should begin with definitions of the meaning of the terms "machine" and "think." The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous, If the meaning of the words "machine" and "think" are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, "Can machines think?" is to be sought in a statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words.

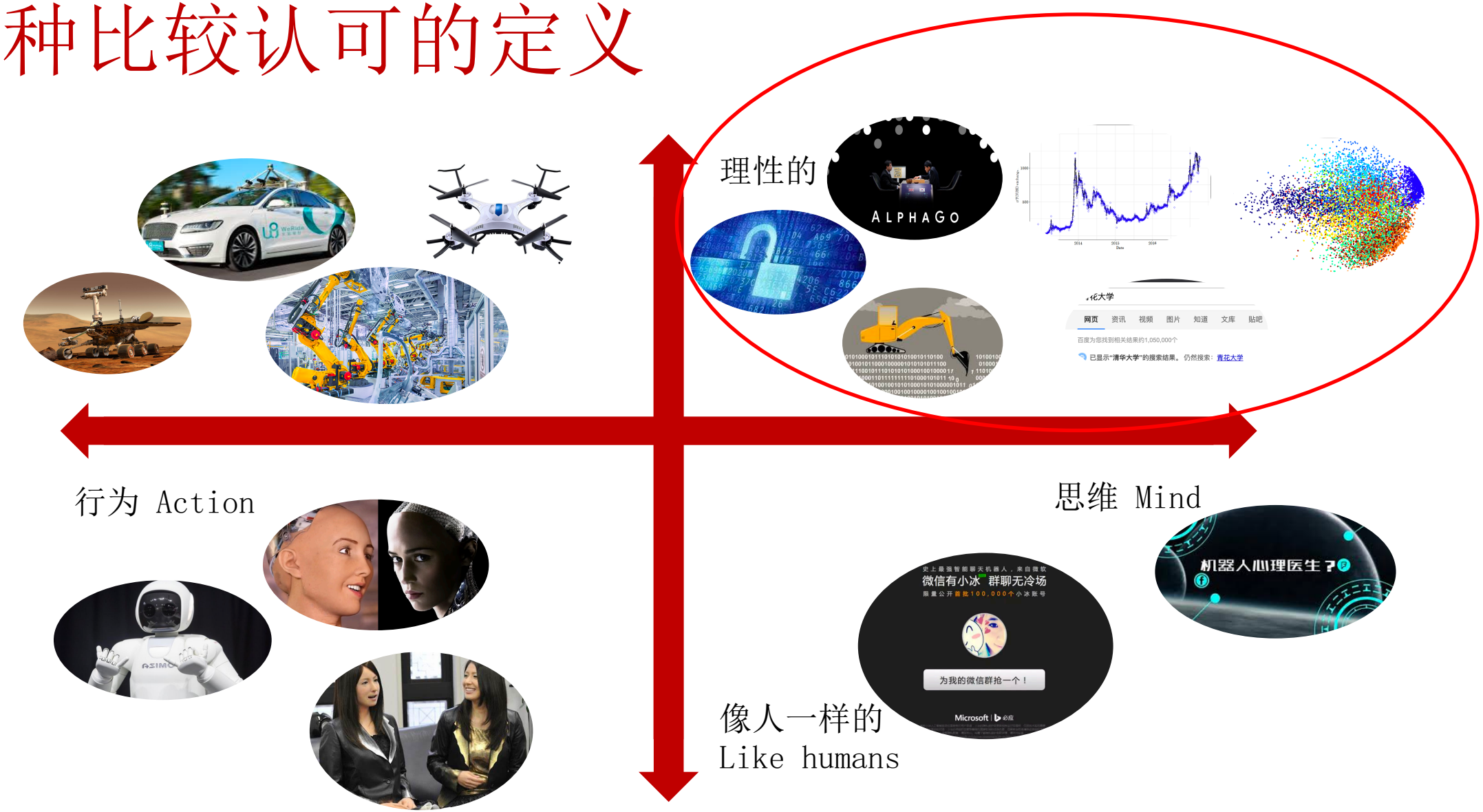
The new form of the problem can be described in terms of a game which we call the 'imitation game." It is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart front the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman. He knows them by labels X and Y, and at the end of the game he says either "X is A and Y is B" or "X is B and Y is A." The interrogator is allowed to put questions to A and B thus:

- People or Machine?
- Imitation Game

那AI到底是什么？

- Artificial Intelligence: Every aspect of learning or any other feature of intelligence can be so precisely described that a machine can be made to simulate it. (1956)

一种比较认可的定义


理性的
ALPHAGO
行为 Action
思维 Mind
机器人心理医生？
微信有小冰 群聊无冷场
为我的微信群抢一个！
Microsoft
像人一样的
Like humans

Question

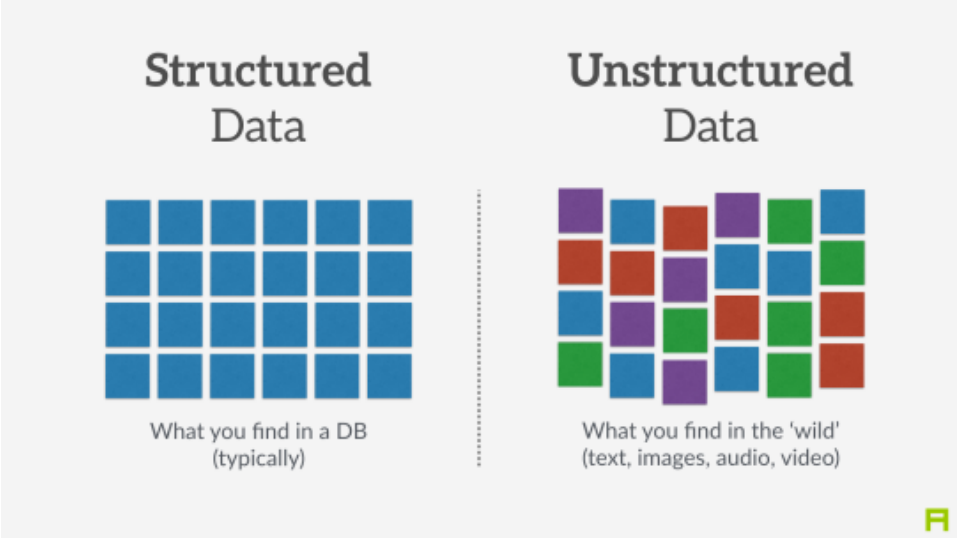
- Classify the applications or system based dimensions.

[ ] Auto Composition
[ ] Voice Recognition
[ ] Stock Predication
[ ] Service Robot
[ ] Anti Money Laundering

And, can you figure out some more?

Why Natural Language is so hard?

- 1. Text is Logic
- 2. Diversity
- 3. Unstructured
- 4. ..



AI Paradigm

- 1. Rule Based
- 2. Probability Based
- 3. Problem Solving: Search based
- 4. Mathematical or Analytic Based
- 5. Machine Learning (deep learning) Based

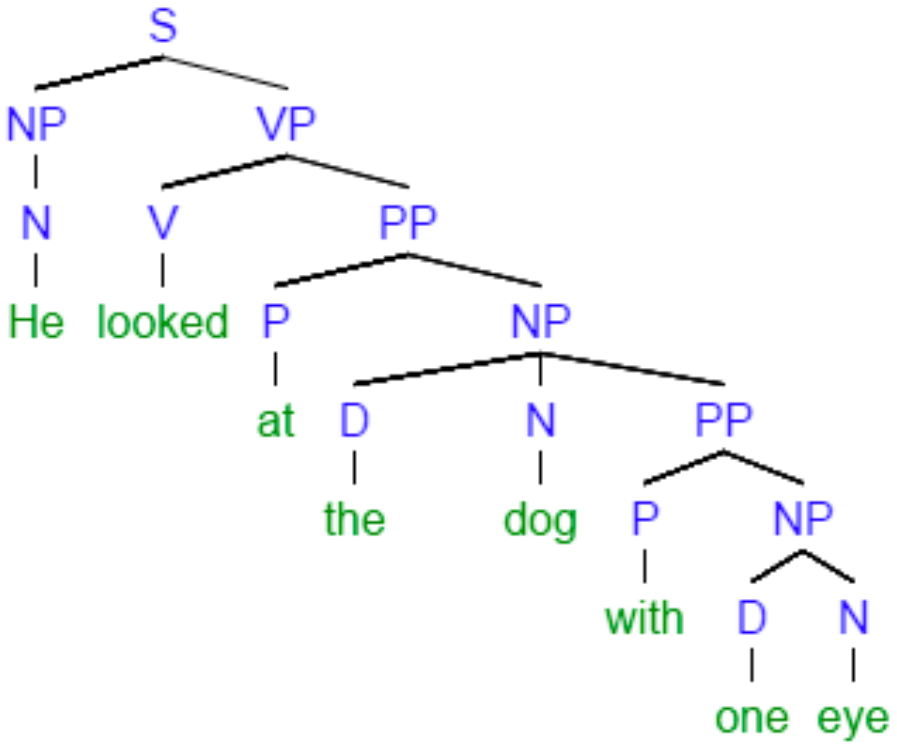
1. Rule Based

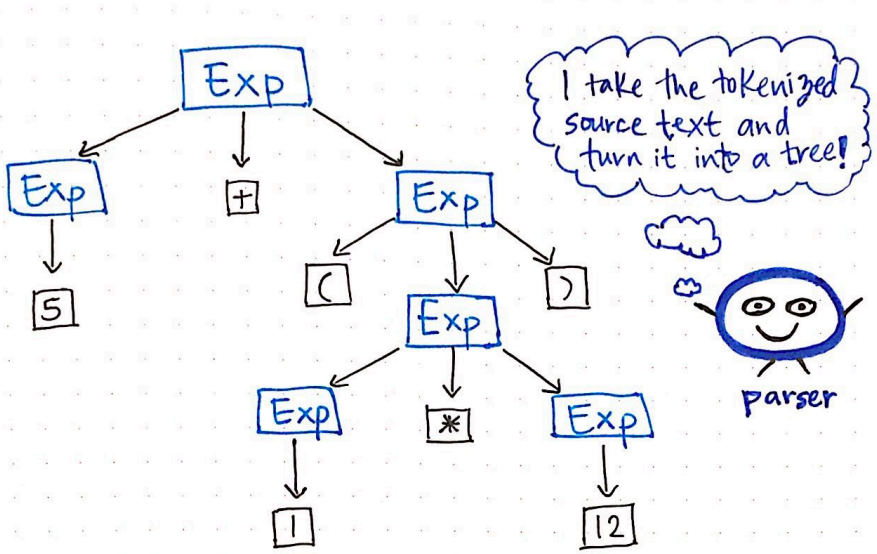
接待员 = """

句子 => 称呼 需要 活动 问题
称呼 => 您 | 你们 | 小伙子
需要 => 要不要 | 想不想 | 需不需要
活动 => 打牌 | 骑马 | 玩射击
问题 => 呢? | 吗？
"""

1. 小伙子想不想打牌呢？
2. 你们要不要骑马呢？
3. ...
4. 您需不需要玩射击呢?

1. Rule Based: Syntax Tree

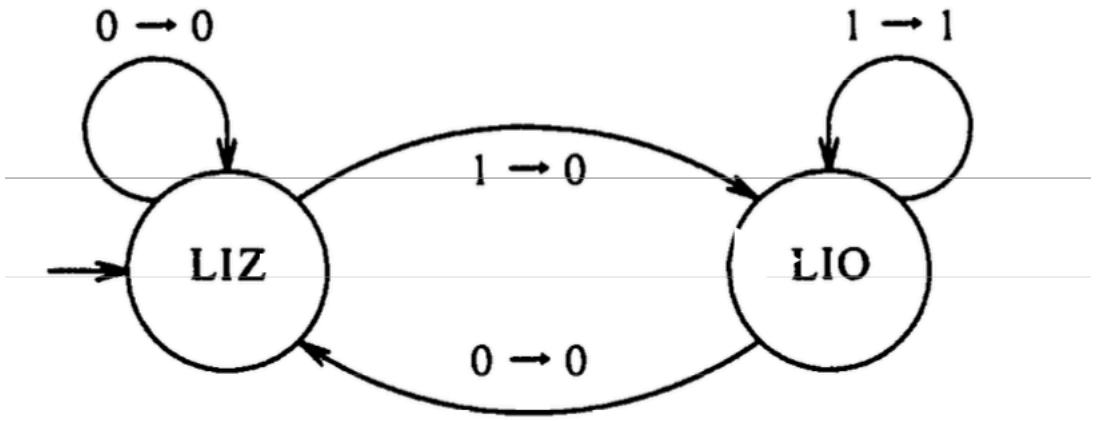


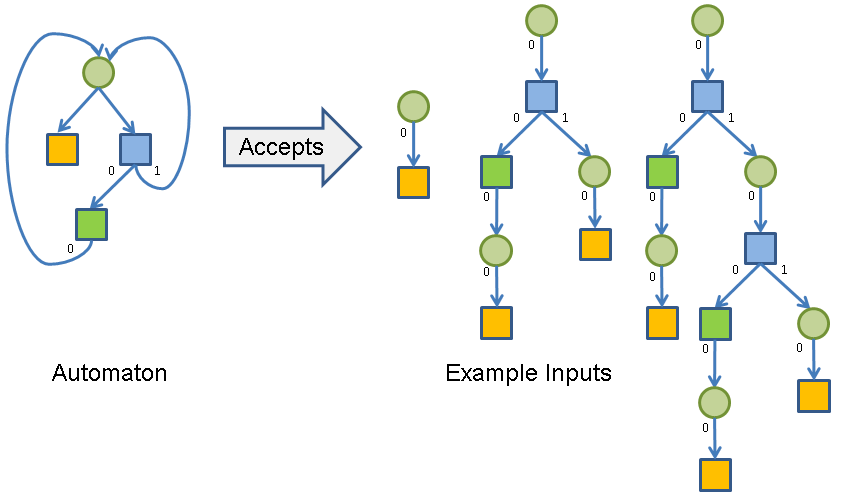


* First comes the lexical analysis phase, followed by the syntax analysis phase, which will generate a parse tree.

Automata

- Input: 011010111
- Output: 001000011





- LIZ: Last Input Zero, LIO: Last Input One

2. Probability Based

- A1. 前天早上吃晚饭的时候
- A2. 前天早上吃早饭的时候

- B1. 正是一个好看的小猫
- B2. 真实一个好看的小猫

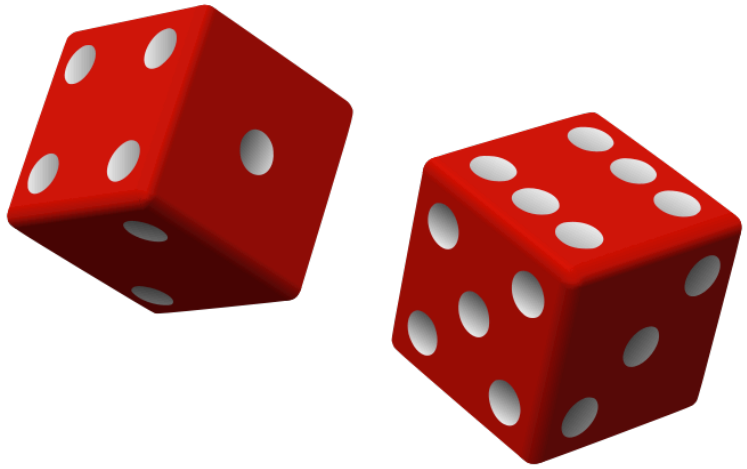
- C1. 我无言以对，简直
- C2. 我简直无言以对

名称	定　义	语法特点	类别	举　例
(一)名词	表示人或事物名称的词	①前面可以加数量词(一副对联) ②前面不能加不、很之类的副词(不搬联、很对联)。 ③后面不能加时态助词“了”(搬联了)	①具体名词	人、牛、山、水、对联
			②抽象名词	友谊、立场、观点、思想
(附)方位词	名词中表示方向位置的词	常用在名词或名词性短语的后面		东、西、南、北、前、后、中间、下边
(二)代词	具有替代或指示作用的词	①能够替代或指示各类实词。 ②一般不带修饰成分	①人称代词	我、你、他、我们、你们
			②指示代词	这、那、这里、那边
			③疑问代词	谁、什么、哪、多少
		①前面可以加副词(刚走、很想)。	①不及物动词	醒、病、游行、觉悟

2. Probability Based

- "Every time I fire a linguist, the performance of the speech recognizer goes up”

---------- **Frederick Jelinek** (18 November 1932 – 14 September 2010)



2. Probability Based: Language Model

- Language Model:

  The probability of a sentence.

N-gram Models

1-gram (Unigram)

$$P(w_i) = \frac{C(w_i)}{\sum_{\forall k} C(w_k)} = \frac{C(w_i)}{N}$$

# of tokens

2-gram (Bigram)

$$P(w_{i+1}|w_i) = \frac{C(w_i, w_{i+1})}{\sum_{\forall k} C(w_i, w_k)} = \frac{C(w_i, w_{i+1})}{C(w_i)}$$

Assignment-01

- 1. Self Review
- 2. Build Sentence Generation System Using Syntax Tree and Language Model
- 3. (Optional) Chat Bot Using Pattern
- 4. (Optional) Reading Turing’s Machine Intelligence Paper: See in our github