基本概念

语言学：是指对语言的科学研究

- 语音和文字是语言的两个基本属性

语音学：研究人类发音特点，特别是语音发音特点，并提出各种语音描述，分类和转写方法的科学

计算语言学：通过建立形式化的计算模型来分析，理解和生成自然语言的学科。与自然语言处理相比，更侧重基础理论和方法的研究

自然语言理解：是探索人类自然语言能力和语言思维活动的本质

自然语言处理：研究如何利用计算机技术对语言文本进行处理和加工的一门学科，研究内容包括对词法，句法，语义等信息进行处理的方法

屈折语：用词的形态变化表示语法关系

黏着语：词内有专门表示语法意义的附加成分

孤立语：形态变化少，语法关系靠次序和虚词表示

研究内容

- 信息过滤：通过计算机系统自动识别和过滤那些满足特定条件的文档信息
- 信息抽取：从指定文档中或海量文本中抽取出用户感兴趣的信息

基本问题和主要困难

形态学问题

研究词由词素的构成问题

- 词素：词根，前缀，后缀，词尾

句法问题

研究句子结构成分之间的相互关系和组成句子序列的规则

语义问题

如何从一个语句中词的意义，以及这些词在该语句中句法结构中的作用，来推导出该语句的意义

语用学问题

研究在不同上下文语句的应用，以及上下文对语句理解所产生的影响。

大量歧义现象

- 词法歧义
- 词性歧义
- 结构歧义
- 语义歧义
- 语音歧义
- 多音字及韵律等歧义

大量未知语言现象

新词，新含义，新用法和新句型等

总结

- **普遍存在的不确定性**：词法、句法、语义、语用和语音各个层面
- **未知语言现象的不可预测性**：新的词汇、新的术语、新的语义和语法无处不在
- **始终面临的数据不充分性**：有限的语言集合永远无法涵盖开放的语言现象
- **语言知识表达的复杂性**：语义知识的模糊性和错综复杂的关联性难以用常规方法有效地描述，为语义计算带来了极大的困难

- **机器翻译中映射单元的不对等性：** 词法表达不相同、句法结构不一致、语义概念不对等

基本研究方法

理性主义：通常通过一些特殊的语句或语言现象的研究来得到对人的语言能力的认识。

- 问题求解的基本思路：基于规则的分析方法建立符号处理系统
  - 规则开发：N + N -> NP
  - 词典标注
  - 推导算法设计
  - 知识库 + 推理系统 -> nlp 系统

经验主义：偏向于对大规模语言数据中人们所实际使用的普通语句的统计

- 问题求解的思路：基于大规模真实语料建立计算方法
  - 大规模真实数据的收集，标注
  - 统计模型的建立
  - 语料库 + 统计模型 -> nlp 系统

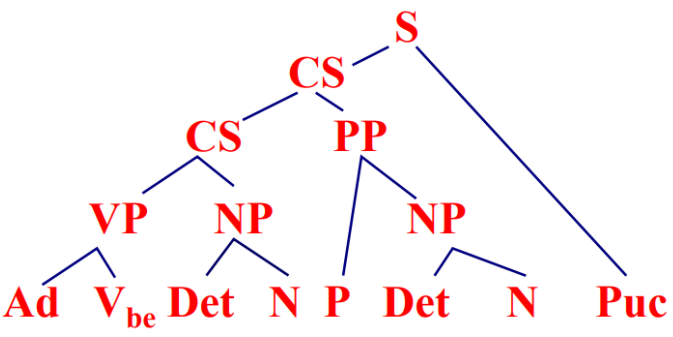
以机器翻译为例

基于规则的方法

- 词法分析：

$$\text{There/Ad is/}V_{be}\text{ a/Det book/N on/P the/Det desk/N ./Puc}$$

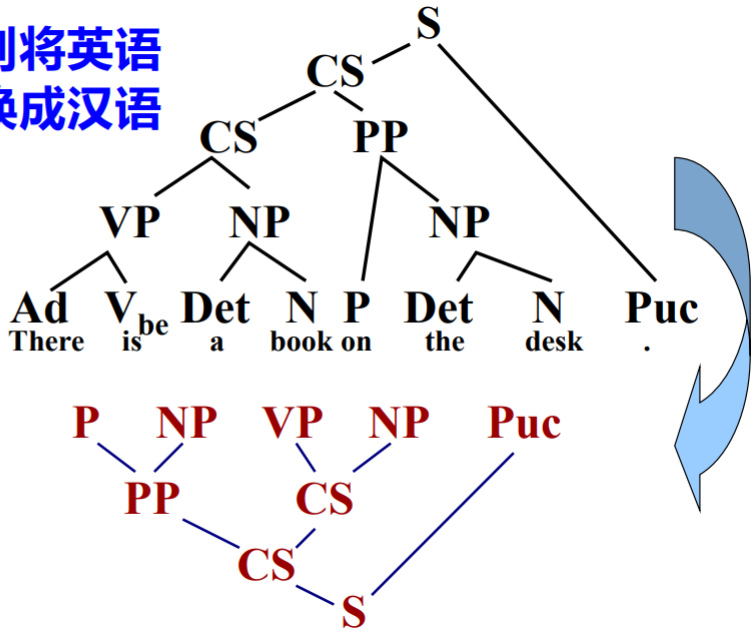
- 利用句法规则进行句法结构分析



动词短语（verb phrase, VP）
名词短语（noun phrase, NP）
介词短语（preposition, PP）
2类连词（conjunction, 分别记作： CC, CS）

- 利用转换规则将英语句子结构转换成汉语句子结构

✧利用转换规则将英语句子结构转换成汉语句子结构

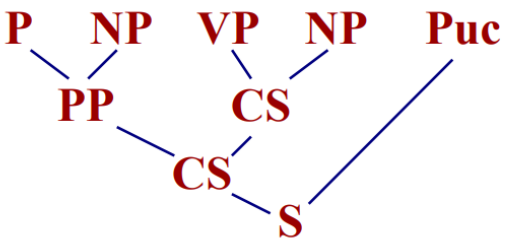


- 根据转换后的句子结构，利用词典和生成规则生成翻译的结果句子

✧根据转换后的句子结构，利用词典和生成规则生成翻译的结果句子

词对应的中文

#a, Det, 一
#book, N, 书; V, 预订
#desk, N, 桌子
#on, P, 在 X 上
#There be, V, 有

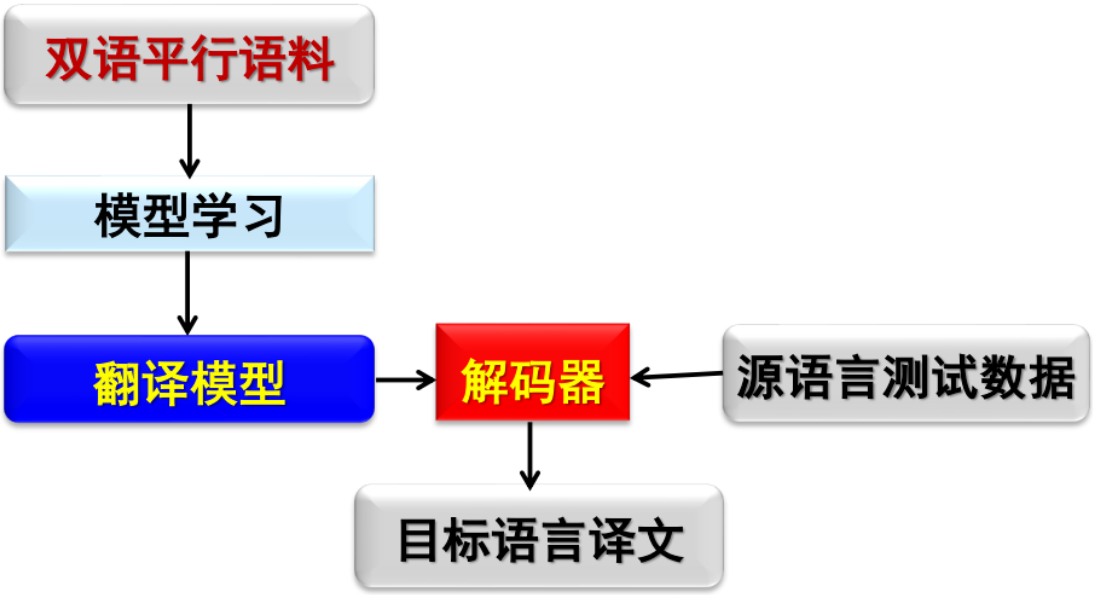


利用规则将这些中文放到对应的位置

输出译文：

在桌子上有一本书。

数据驱动的翻译方法



给定源语言句子：$E = e_1^m \equiv e_1 e_2 \cdots e_m$

将其翻译成目标语言句子：$C = c_1^l \equiv c_1 c_2 \cdots c_l$

根据贝叶斯公式：$P(C \mid E) = \dfrac{P(C)P(E \mid C)}{P(E)}$

求解使P值最大的C：$\hat{C} = \arg\max_C P(C)P(E \mid C)$

语言模型
(Language model, LM)

翻译模型
(Translation model, TM)

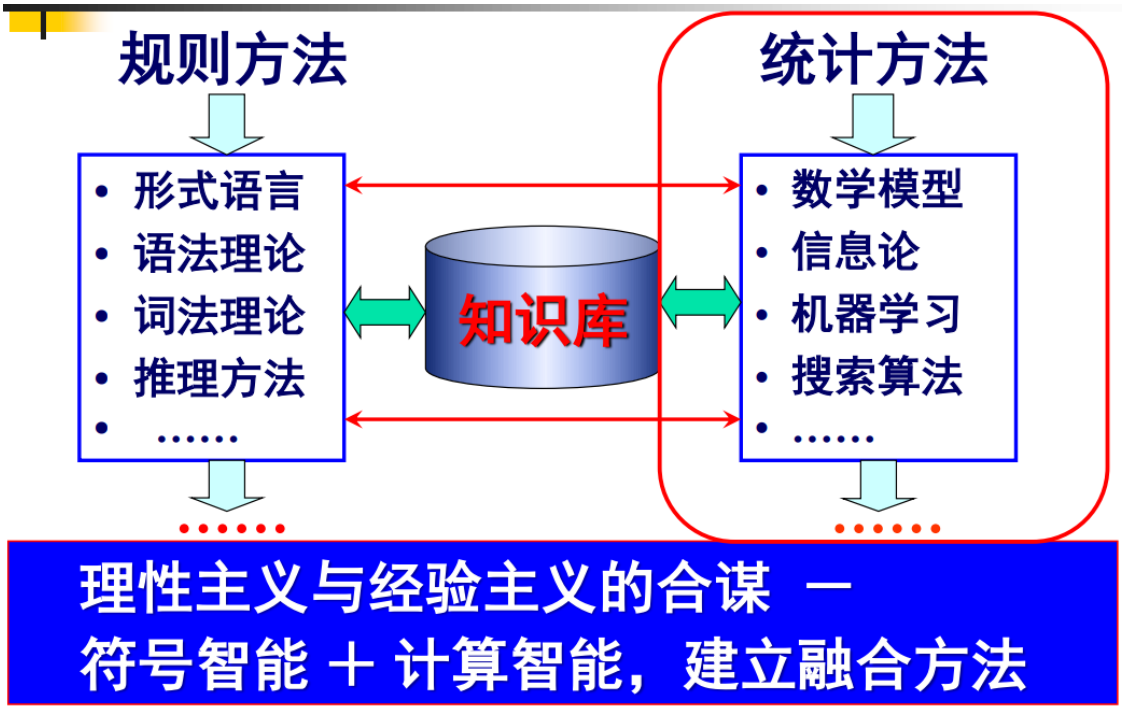
构建解码器(decoder)，快速搜索最优翻译候选：

三个关键问题

- 估计语言模型概率 $P(c)$
- 估计翻译模型概率 $P(E|C)$
- 快速有效地搜索候选译文 C，使 $P(C) * P(E|C)$ 最大

主要任务

- 收集大规模双语句子对（求 P(E|C)），目标语言句子（求 P(C)）
- 参数训练与模型优化


规则方法
统计方法
• 形式语言
• 语法理论
• 词法理论
• 推理方法
• ……
知识库
• 数学模型
• 信息论
• 机器学习
• 搜索算法
• ……
理性主义与经验主义的合谋 —
符号智能 + 计算智能，建立融合方法