

自然语言处理与 文本数据挖掘

谢海华 助理研究员

北京雁栖湖应用数学研究院 (BIMSA)

大数据与人工智能 (BDAI)

2022.03



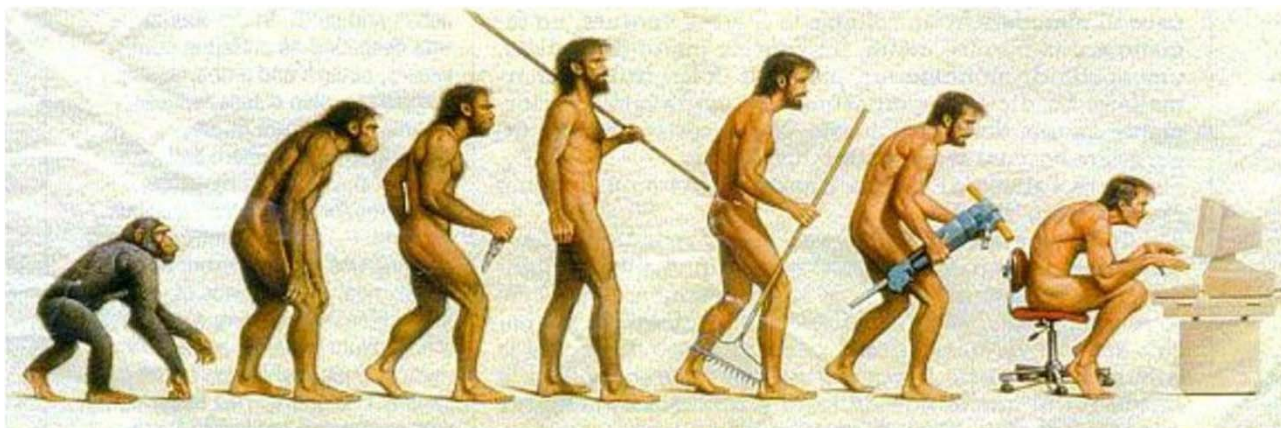
第一章 绪论

1.1 背景及问题

1.1 背景及问题

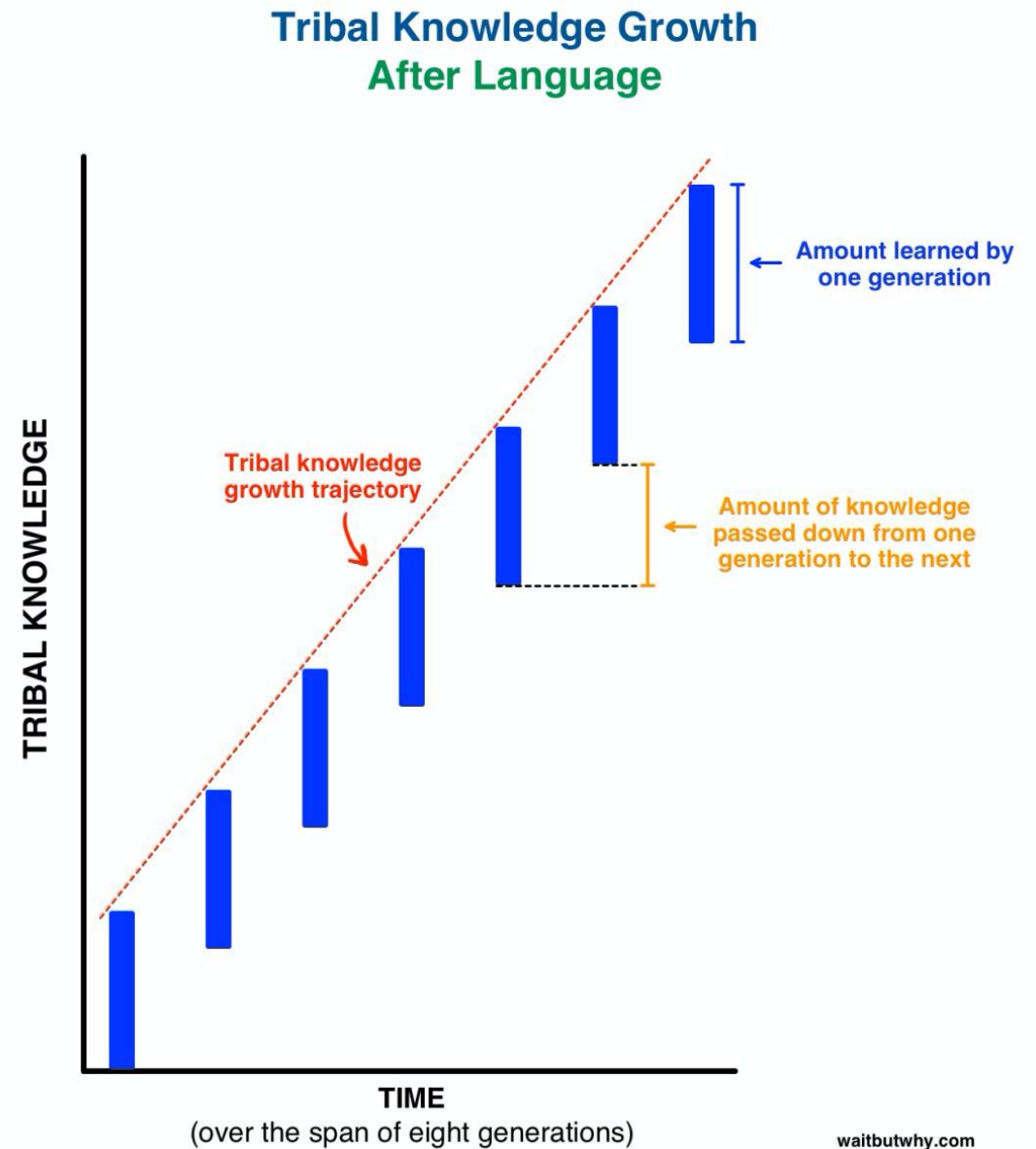
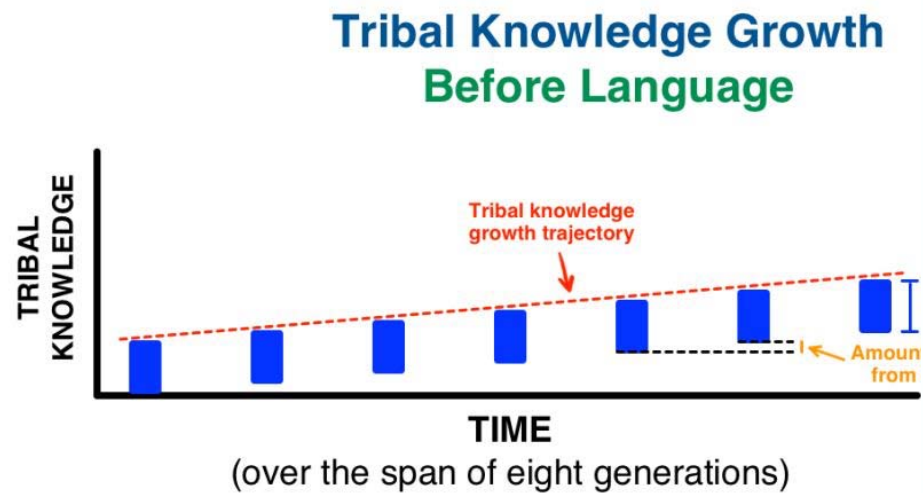
什么是自然语言？

- 自然语言是人类社会发展过程中自然产生的语言，是人类智慧和文明的产物。

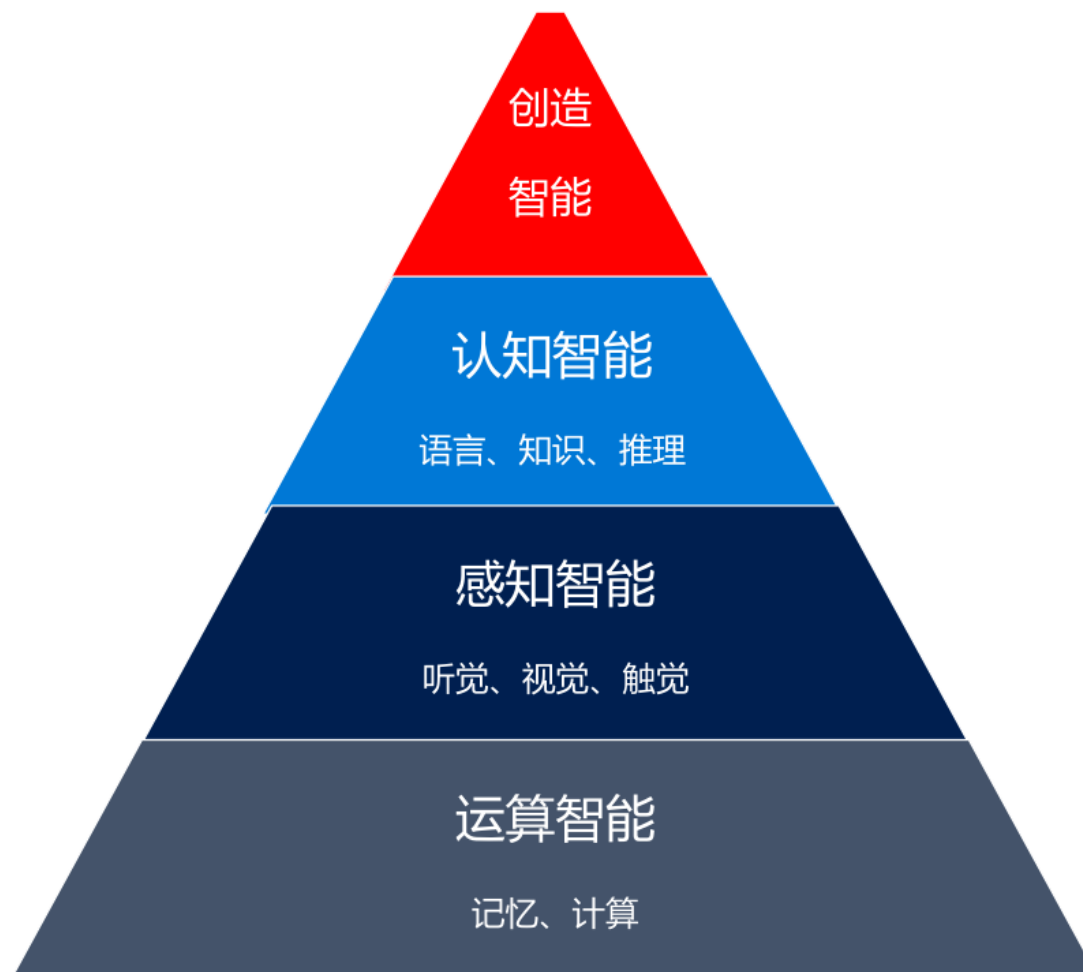


- 语言是思维的载体，是人类交流思想、表达情感的最自然、最直接、最方便的工具。
- 人类历史上以语言文字形式记载和流传的知识占知识总量的80%以上。

1.1 背景及问题



1.1 背景及问题



1.1 背景及问题



oo-form

Pedro Henrique Mar
YDorph
LUMJIS (Lisbon
pedrohenr
marinho@google.i

Abstract

Transformers struggle when it comes to parsing decisions and are much more expressive than vanilla PCFGs. However, they suffer from representational and inference complexities. For representation, the addition of textual information greatly increases the number of parameters to be estimated and exacerbates the data sparsity problem during learning, so the expectations maximization (EM) based estimation of L-PCFGs has to rely on sophisticated smoothing techniques and factorizations (Collins, 2003). As for inference, the CYK algorithm for L-PCFGs has a $O(l^3 G)$ complexity, where l is the sentence length and G is the grammar constant. Although Eisner and Satta (1999) manage to reduce the complexity to $O(l^2 G)$, inference with L-PCFGs is still relatively slow, making them less popular nowadays.

Recently, Zhu et al. (2020) combine the idea of factoring the binary rule probabilities (Collins, 2003) and neural parameterization (Kim et al., 2019) and propose neural L-PCFGs (NL-PCFGs), achieving good results in both unsupervised dependency and constituency parsing. Neural parameterization

Combining

Shaojing Jim Jia* and
1. Wangxuan Li
2. J
Feli
3. Postdoctoral W
(347

Abstract

Multi-turn conversational understanding of context question. It is a major challenge of while understanding context question, or an impression feature, information from context to each token turns to each token input sequence constructed to enhance the impact of the proposed idea

Neural Bi-Lexicalized PCFG Induction

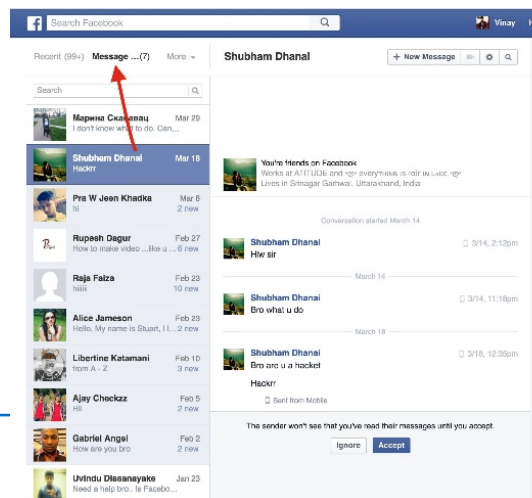
Songlin Yang*, Yanzheng Zhao*, Kewei Tu*
*School of Information Science and Technology, ShanghaiTech University
Shanghai Engineering Research Center of Intelligent Vision and Imaging
Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences
University of Chinese Academy of Sciences
910CC, University of Edinburgh
(yangs1, tukw1)@shanghaitech.edu.cn
yanzhao.ed@gmail.com

Abstract

Neural lexicalized PCFGs (L-PCFGs) (Zhu et al., 2020) have been shown effective in grammatical complexity. However, to reduce computational complexity, they make a strong independence assumption on the generation of the child word and thus bilateral dependencies are ignored. In this paper, we propose an approach to parameterize L-PCFGs without making implausible independence assumptions. Our approach directly models bilateral dependencies and meanwhile reduces both learning and representation complexities of L-PCFGs. Experimental results on the English WSJ dataset confirm the effectiveness of our approach in improving both training speed and unsupervised parsing performance.

1 Introduction

Probabilistic context-free grammars (PCFGs) has been an important probabilistic approach to syntactic analysis (Lari and Young, 1990; Jelinek et al., 1992). They assign a probability to each of the



1.1 背景及问题

- 如何让计算机能够自动或半自动地理解自然语言文本，懂得人的意图和心声？
- 如何让计算机实现海量语言文本的自动处理、挖掘和有效利用，满足不同用户的各种需求，实现个性化信息服务？

自然语言处理

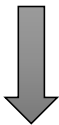
Natural Language Processing

1.2 学科产生与发展

1.2 学科的产生与发展



Warren
Weaver,
Rockefeller
Institute



Norbert
Wiener,
Professor at
MIT

[Reproduced by permission of the Rockefeller Foundation Archives]

March 4, 1947

Dear Norbert:

I was terribly sorry, when in Cambridge recently, that I got unavoidably held up by several unexpected jobs, and did not get a chance to see you.

One thing I wanted to ask you about is this. A most serious problem, for UNESCO and for the constructive and peaceful future of the planet, is the problem of translation, as it unavoidably affects the communication between peoples. Huxley has recently told me that they are appalled by the magnitude and the importance of the translation job.

I wondered if it were unthinkable to design a computer which would **TRANSLATE**

... inferred considerable about, powerful new mechanized methods in cryptography - methods which I believe succeed even when one does not know what language has been coded - 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. I will now proceed to decode."

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

Cordially,

Warren Weaver.

Professor Norbert Wiener
Massachusetts Institute of Technology
Cambridge 39, Massachusetts

WW:AEB

1.2 学科的产生与发展



摩尔，麦卡锡，明斯基，赛弗里奇、所罗门诺夫

Summer Research Project on **Artificial Intelligence** (Dartmouth Conference, 1956)

- **自然语言理解** (natural language understanding, NLU) 成为人工智能研究的核心问题之一。

1.2 学科的产生与发展

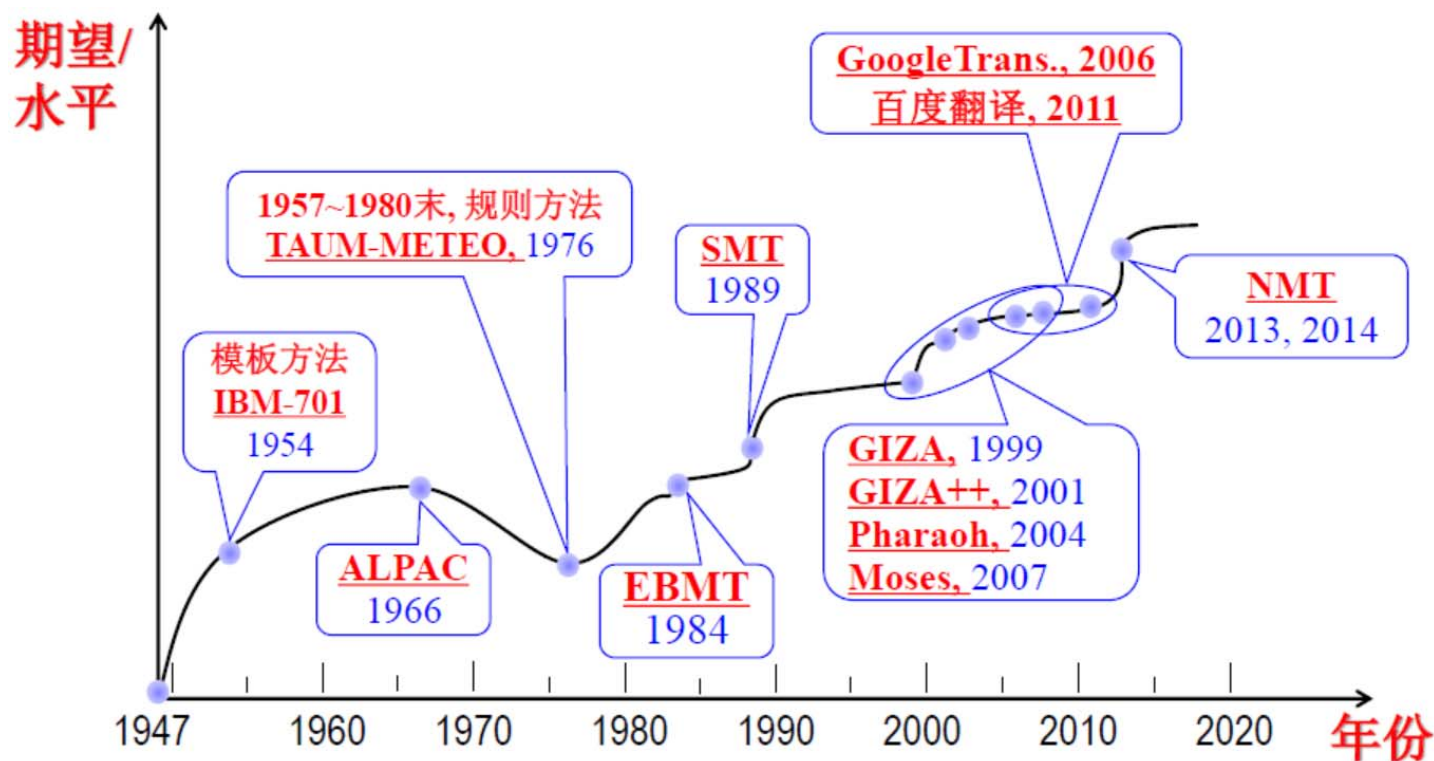
- 自 W. Weaver 和 A. D. Booth 提出机器翻译概念后，美国和英国的学术界对机器翻译 ([Machine Translation](#), MT) 产生了浓厚的兴趣，并得到了实业界的支持。
- 1954年 Georgetown 大学在 IBM 协助下，用 IBM-701 计算机实现了世界上第一个 MT 系统，实现俄译英翻译，1954年1月该系统在纽约公开演示。系统只有250条俄语词汇，6条语法规则，可以翻译简单的俄语句子。
- 随后10 多年里，MT研究在国际上出现热潮。

1.2 学科的产生与发展

- 1962年国际计算语言学学会 ([Association for Computational Linguistics, ACL](#)) 成立; 1965年国际计算语言学委员会 ([International Committee on Computational Linguistics, ICCL](#)) 成立。
- 1964年, 美国科学院成立语言自动处理咨询委员会 ([Automatic Language Processing Advisory Committee, ALPAC](#)) , 调查机器翻译的研究情况, 并于1966年11月公布了一个题为“语言与机器”的调查报告, 简称ALPAC 报告, 宣称: “在目前给机器翻译以大力支持还没有多少理由”, “机器翻译遇到了难以克服的语义障碍(semantic barrier)”。从此, 机器翻译研究在世界范围内进入低迷状态。计算语言学 ([Computational Linguistic, CL](#)) 术语首次以正式身份出现在这个报告里。
- 1980s, 随着计算机网络的快速发展和普及, 以开发实用自然语言处理系统为目标的语言工程技术应运而生, 自然语言处理 ([Natural Language Processing, NLP](#)) 术语由此诞生。

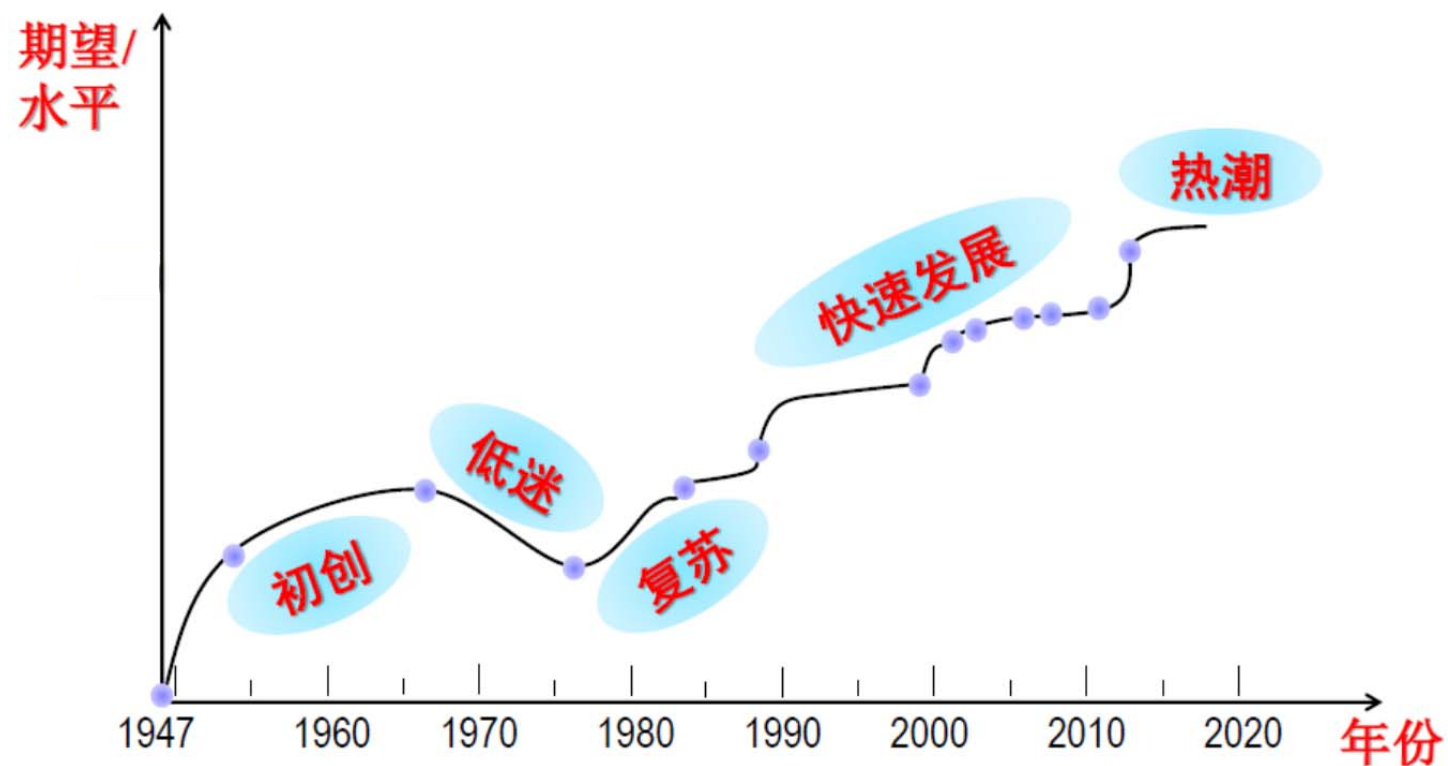
1.2 学科的产生与发展

机器翻译 技术的发展



1.2 学科的产生与发展

机器翻译
技术的发展



1.2 学科的产生与发展

◆ 定义-1：自然语言理解

自然语言理解是探索人类自身语言能力和语言思维活动的本质，研究模仿人类语言认知过程的自然语言处理方法和实现技术的一门学科。它是人工智能早期研究的领域之一，是一门在语言学、计算机科学、认知科学、信息论和数学等多学科基础上形成的交叉学科。

《计算机科学技术百科全书》（宗成庆）

清华大学出版社，2018.05

1.2 学科的产生与发展

◆ 定义-2: 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，是人工智能和语言学的分支学科。**计算语言学**是典型的交叉学科，其研究常常涉及计算机科学、语言学、数学等多个学科的知识。与内容接近的学科自然语言处理相比较，计算语言学更加侧重**基础理论和方法**的研究。

《计算机科学技术百科全书》（常宝宝）

清华大学出版社，2018.05

1.2 学科的产生与发展

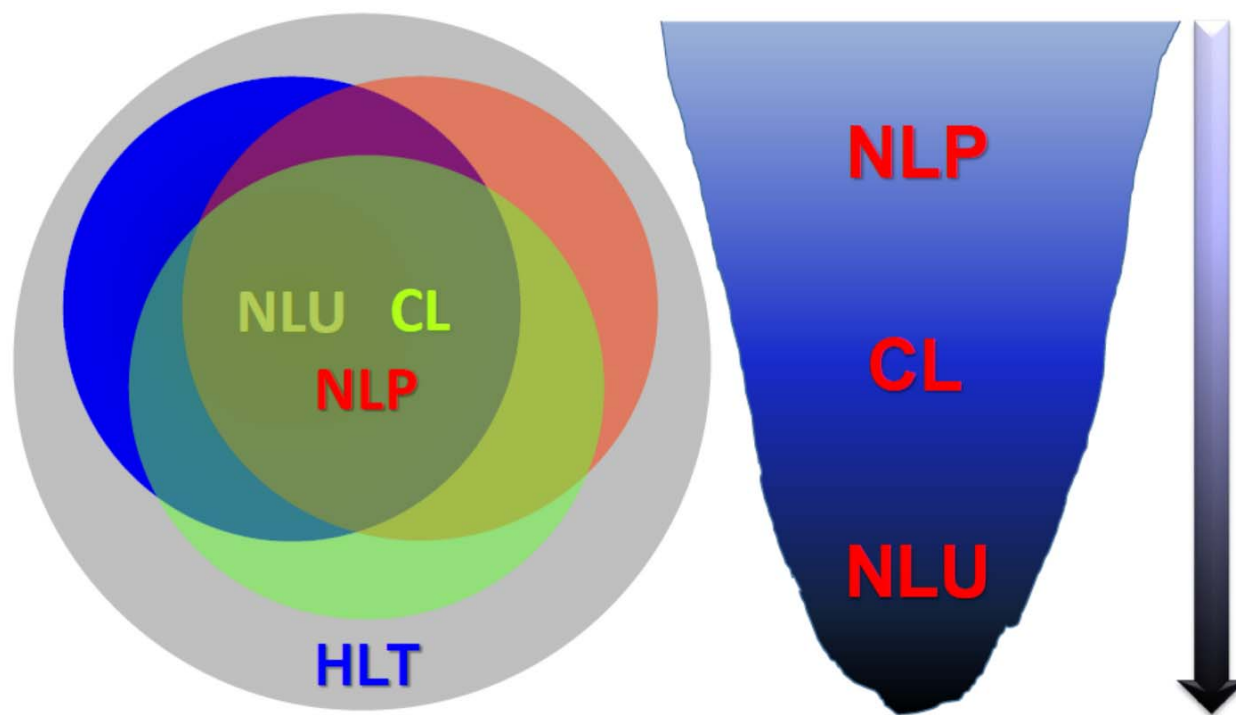
◆ 定义-3：自然语言处理

自然语言处理是研究如何利用计算机技术对语言文本（句子、篇章或话语等）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**和**语用**等信息的识别、分类、提取、转换和生成等各种处理方法和实现技术。

《计算机科学技术百科全书》（宗成庆）

清华大学出版社，2018.05

1.2 学科的产生与发展



统称：人类语言技术
(Human Language
Technique, HLT)

1.2 学科的产生与发展

◆ 定义-4：中文信息处理

以**中文为主要处理对象**的自然语言处理技术。

在不引起混淆的情况下，通常指汉语。但是从广义上讲，中文指中国境内使用的语言，可以包括汉语、蒙古语、藏语、维吾尔语等几十种语言。

中文信息处理不仅仅是中国人自己感兴趣的研究课题，任何一家国际互联网公司都从来不敢忽略或藐视它。

1.2 学科的产生与发展

■ 学科的理论意义

- 探索人脑语言理解的本质，揭示语言认知的奥秘
- 研究和建立计算语言学理论体系
- 推动相关学科产生与发展

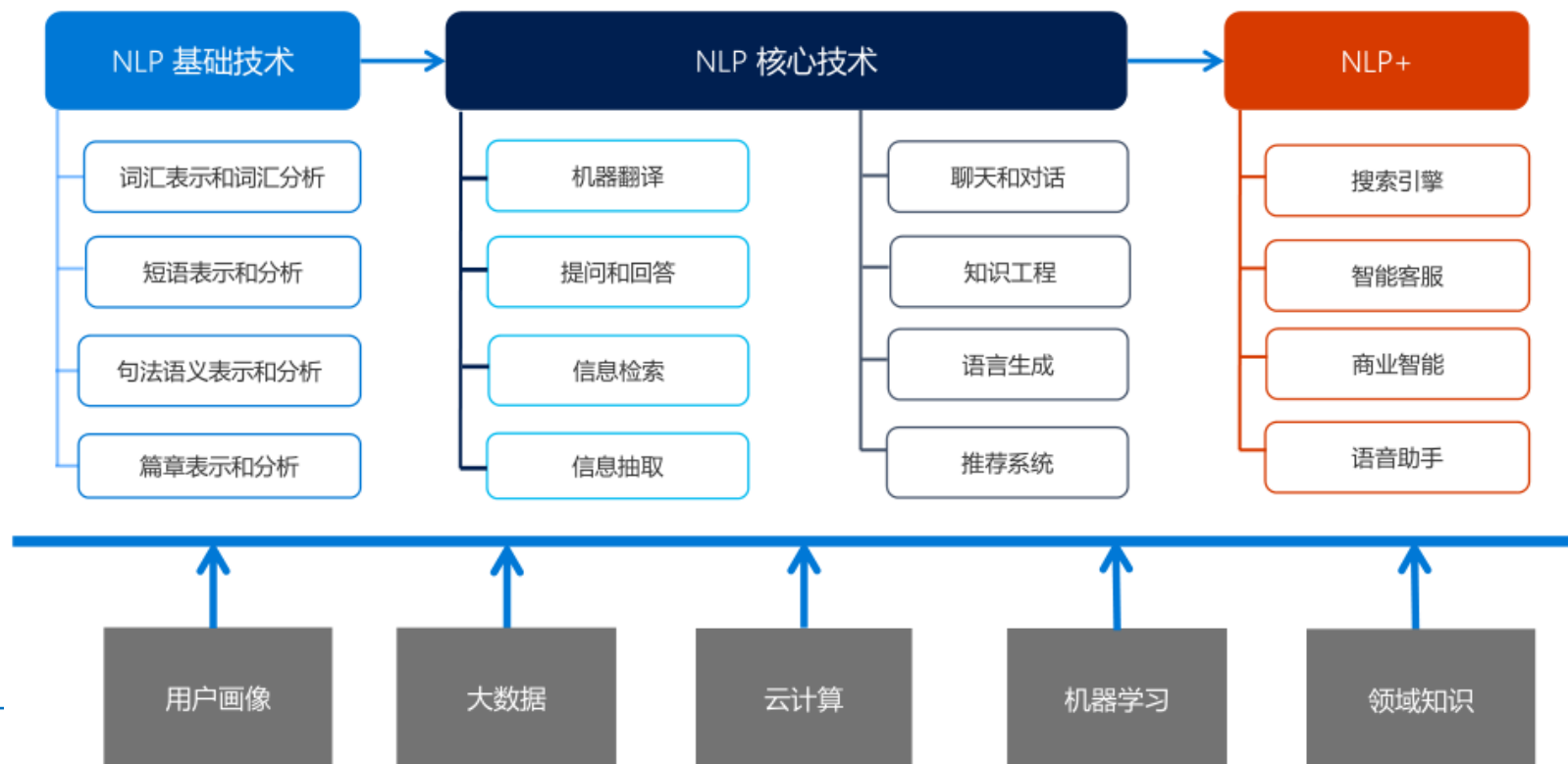
■ 学科的应用价值

- 打破人类语言的障碍，实现任意时间、任意地点、任意语言的无障碍自由通信
- 准确理解人的意图，提高个性化信息服务的质量
- 建立高智能人机交互系统
- 保障网络内容安全，维护国家和公共安全
- 保护民族语言文化，促进全球化社会发展

1.3 研究内容

1.3 研究内容

自然语言处理（NLP）是人工智能的一个分支，用于分析、理解和生成自然语言，以方便人和计算机设备进行交流，以及人与人之间的交流。



1.3 研究内容—NLP的基本任务

□ 自然语言分析：分析语言表达的结构和含义

- 词法分析：获取**单词**的语言学信息，具体任务包括：词性标注、命名实体识别、分词等。
- 句法分析：分析**语句**的短语结构，确定各个词语、短语之间的关系，以及它们在语句中的作用。
- 语义分析：结合上下文，获得词语和短语的**准确含义**。
- 语用分析：分析不同上下文中的语句应用，以及上下文对语句理解所产生的影响。关注的重点在于：为什么在特定的上下文中要使用该语句？
- 语音分析：根据音位规则，从语言流中区分出各个独立的音素，再根据音位形态规则，找出各个音节及对应的词素和词语。

□ 自然语言生成：自动生成自然语言，可应用于机器写作，自动对话等。

1.3 研究内容—NLP的基本任务1-分词

- ❑ 任务定义
 - 输入：句子，如：“请问苯环中碳碳键键能能否否定定论一？”
 - 输出：词语序列，如：“请/问/苯环/中/碳碳键/键能/能否/否定/定论一？”
- ❑ 交叉歧义：需要全局句法语义知识甚至外部知识
 - 乒乓球拍卖完了（球拍 vs. 拍卖）
- ❑ 新词识别：不断涌现
 - 人名、地名、机构名、公司名、产品名、歌名、电影名、缩略词、简称等
- ❑ 领域移植问题：法律、医药等不同领域的文本
- ❑ 分词粒度问题
 - 不同应用需求不同
 - 人工标注时粒度的把握很难，数据质量存疑

1.3 研究内容—NLP的基本任务2-实体识别

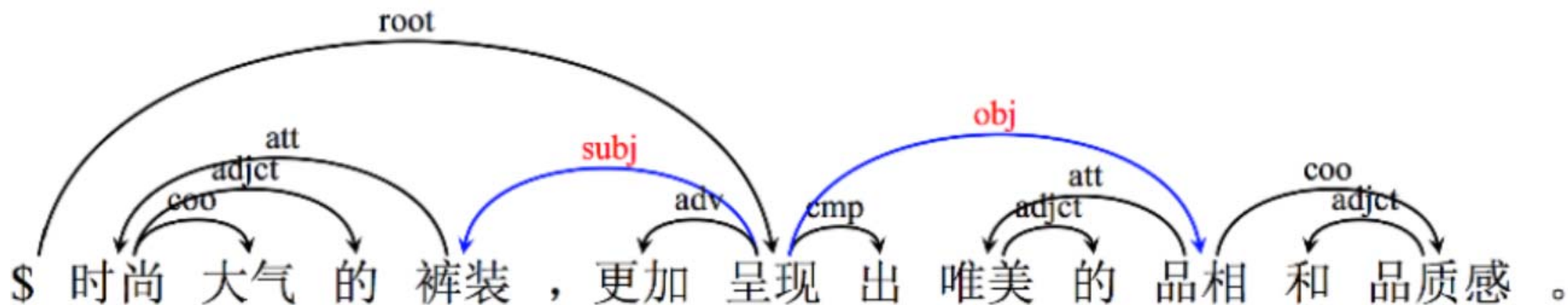
- 在给定文本中识别出实体，并标注给定类别
- 类别例子：人名、地名、组织机构名、产品名、时间

史蒂夫·乔布斯，1955年2月24日出生于美国加利福尼亚州旧金山，美国企业家、苹果公司联合创始人。他经历了苹果几十年的起落与兴衰，相继推出了Macintosh、iMac、iPod、iPhone、iPad等风靡全球的电子产品，深刻地改变了现代通讯、娱乐、生活方式。

1.3 研究内容—NLP的基本任务3-句法分析

□ 任务定义

- ✓ 输入：句子的词语序列
- ✓ 输出：依存句法树



1.3 研究内容—NLP的基本任务4-指代消解

红玉上来回道：“平姐姐说，**奶奶**刚出来了，**他**就把银子收了起来，才张材家的来讨，当面称了给**他**拿去了。”说着将荷包递了上去，又道：

“平姐姐教我回**奶奶**：才旺儿进来讨**奶奶**的示下，好往**那家**子去。平姐姐就把那话按着**奶奶**的主意打发**他**去了。”凤姐笑道：“**他**怎么按我的主意打发去了？”红玉道：“平姐姐说：**我们奶奶**问**这里奶奶**好。原是我们**二爷**不在家，虽然迟了两天，只管请**奶奶**放心。等**五奶奶**好些，**我们奶奶**还会了**五奶奶**来瞧**奶奶**呢。**五奶奶**前儿打发了人来说，**舅奶奶**带了信来了，问**奶奶**好，还要和**这里的姑奶奶**寻两丸延年神验万全丹。若有了，**奶奶**打发人来，只管送在**我们奶奶**这里。明儿有人去，就顺路给那边**舅奶奶**带去的。”

----- 《红楼梦》第二十七回

1.3 研究内容—NLP的基本任务5-语义分析

□ 语义分析指将文本转换为可计算的知识表示

- ✓ 输入：自然语言句子
- ✓ 输出：**结构化，机器可读的表示**，称为语义表示（或非正式地，“语义”）
- ✓ 例如：“找一本清代人写的唐代史书” →

$\text{find}(\text{book}): \text{label}(\text{book}) = \{\text{Tang \& History}\} \& \text{era}(\text{writer}(\text{book})) = \text{Qing}$

**不同的应用，语义表示方法会不一样，
分析方法也会不一样**

1.3 研究内容—NLP的基本任务6-篇章分析

篇章：一系列连续的语段或句子构成的语言整体单位

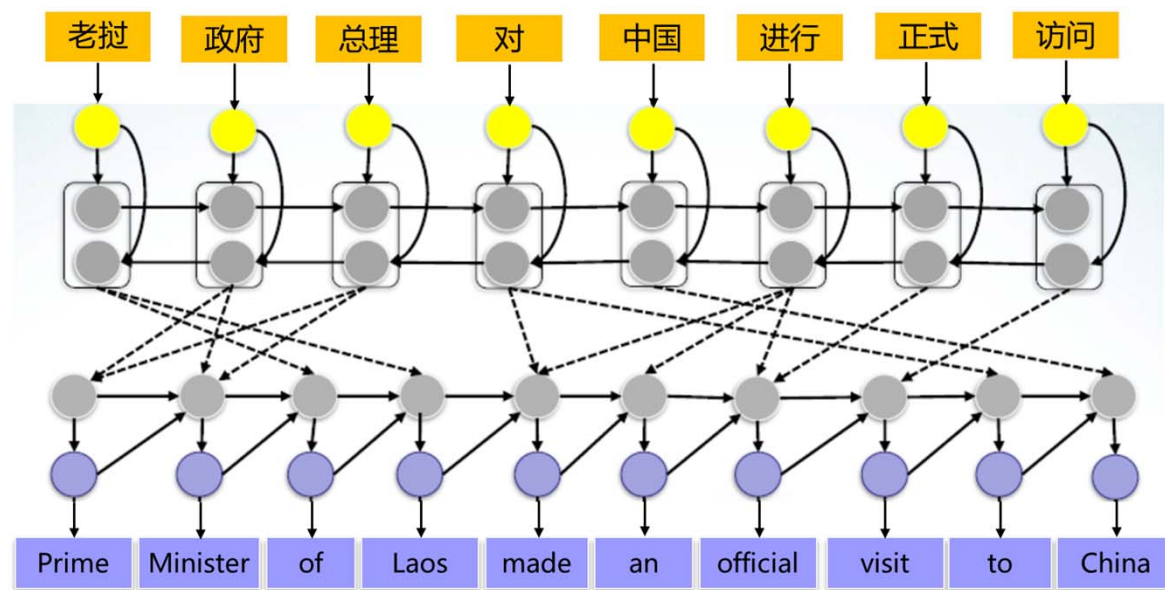
- 比尔来自于美国。今天交通非常拥挤。长江贯穿中国的多个省市。因此，自然语言处理是计算机科学与语言学的融合。
- 这里的交通非常拥挤。张先生早上6:40之前就得出发，【 】常常会提前半小时到办公室；如果【】稍晚一点，他很可能会迟到。

人们理解自然语言通常是在篇章级进行
“不能断章取义”

1.3 研究内容—NLP应用研究

■ 机器翻译 (Machine translation, MT)

- 实现一种语言到另一种语言的自动翻译。
- 应用：文献翻译、网页辅助浏览等。
- 代表系统：
 - <http://translate.google.cn>
 - <http://fanyi.baidu.com/>
 - <http://www.systransoft.com>
 - <https://www.deepl.com/translator>



1.3 研究内容—NLP应用研究1

- 机器翻译 (Machine translation, MT)
 - 基于词典的方法 (Bilingual Word)
 - 基于规则转换的方法 (Rule Transfer)
 - 基于中间语言的方法 (Interlingua)
 - 基于实例的方法 (Example-based)
 - 基于统计的方法 (Statistical MT)
 - 基于神经网络的方法 (Neural MT)

1.3 研究内容—NLP应用研究2

- **信息检索** (Information retrieval) : 也称情报检索, 就是利用计算机系统从大量文档中找到符合用户需要的相关信息。
 - 全球网站数量超过10亿, 网页的数量不计其数, 只有1%的信息被有效地利用。
 - 代表系统:
 - <http://www.google.com>
 - <http://www.baidu.com.cn>
 - <https://www.bing.com/>

1.3 研究内容—NLP应用研究3

■ 问答系统 (Question-answering System)

- 通过计算机系统对人提出的问题的理解，利用自动推理等手段，在有关知识资源中自动求解答案并做出相应的回答。
- 有时与语音技术和多模态输入/输出技术，以及人机交互技术等相结合，构成人机对话系统 (Man-computer Dialogue System) 。
- 应用：IBM Watson 自动问答系统、Siri、微软小冰。



1.3 研究内容—NLP应用研究3

■ 问答系统 (Question-answering System)

- 事实类(factoid): CNCC 2018在哪个城市召开?
- 描述类(descriptive): 新发布的iPhone手机有什么特点?
- 过程类(procedural): 个人护照怎么申请办理?
- 计算类(calculation): IJCAI 2018有多少华人作者?
- 因果类(casual): 为什么中国会发生疫苗事件?
- 观点类(opinion): 你对疫苗事件和中美贸易战有什么看法?

1.3 研究内容—NLP应用研究4

■ 情感和情绪分析 (Sentiment Analysis)

- 对于产品评论和新闻等文本中表达的**意见、情感、情绪、主客观性、评价对象**等方面的研究
- 情感分析在工业界和学术界都已经有着广泛的应用和研究



舆情监控
(面向政府)



企业征信
(面向企业)



聊天/服务机器人
(面向个人)

1.3 研究内容—NLP应用研究5

■ 信息抽取 (Information Extraction)

- 信息抽取是指从非结构化/半结构化文本中抽取指定类型的信息，如**实体**、**属性**、**关系**、**事件**、**商品记录**等。

出生地

史蒂夫·乔布斯，1955年2月24日出生于美国加利福尼亚州旧金山，美国企业家、苹果公司联合创始人。他经历了苹果公司几十年的起落与兴衰，先后领导和推出了Macintosh、iMac、iPod、iPhone、iPad等风靡全球电子产品，深刻地改变了现代通讯、娱乐、生活方式。

1.3 研究内容—NLP应用研究

- **自动文摘** (Automatic summarization / Automatic abstracting)
 - 将原文档的主要内容或某方面的信息自动提取出来，并形成原文档的摘要或缩写。
 - 应用：电子图书管理、情报获取、观点挖掘等。
- **文档分类** (Document Categorization)：利用计算机系统对大量的文档按照一定的分类标准（例如，根据主题或内容划分等）实现自动归类。

1.3 研究内容—NLP应用研究

- 文字编辑和自动校对 (Automatic proofreading)
 - 对文字拼写、用词、甚至语法、文档格式等进行自动检查、校对和编排。
 - 应用：排版、印刷和书籍编撰等。
- 文字识别 (Character Recognition)
- 语音识别 (Automatic Speech Recognition, ASR)
- 语音合成 (Text-to-speech Synthesis)

1.4 技术挑战

1.4 技术挑战—机器能理解自然语言吗？

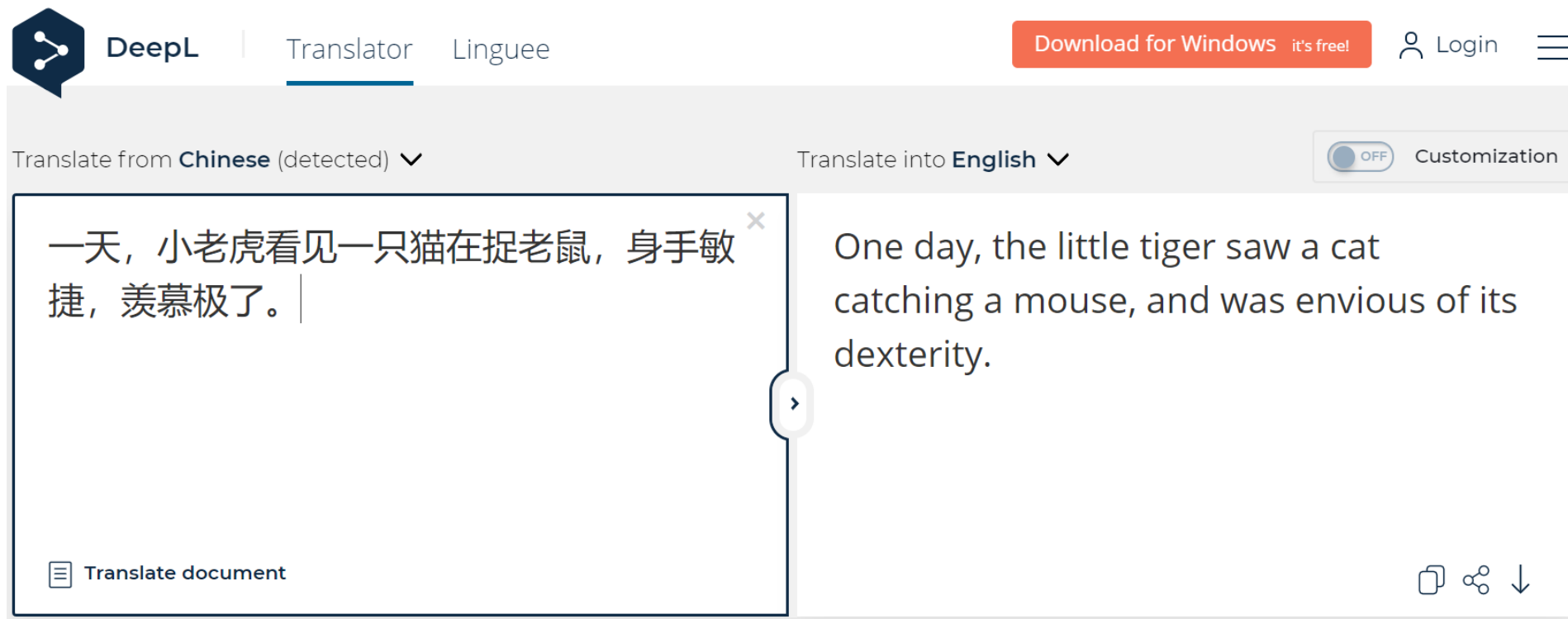
- ◆ 一天，小老虎看见一只猫在捉老鼠，身手敏捷，羡慕极了。
- ◆ 篮球放不进箱子里，（？）太大了（太小了/形状不对）。

1.4 技术挑战—机器能理解自然语言吗？

一天，小老虎看见一只猫在捉老鼠，身手敏捷，羡慕极了。

- One day, the little tiger saw a cat catching mice. He was quick and envious.
- One day, the little tiger saw a cat catching mice, quick-witted, envious.
- One day, the little tiger saw a cat catching mice. He was agile and envied him very much.
- One day, the little tiger saw a cat catching mice, and envied her very much.

1.4 技术挑战—机器能理解自然语言吗？



1.4 技术挑战—机器能理解自然语言吗？

王师傅是卖鱼的，每公斤鱼进价36元。现市场价卖48元一斤。顾客买了两公斤，给了王师傅200元假钱，王师傅没零钱，于是找邻居换了200元。事后邻居存钱过程中发现钱是假的，被银行没收了，王师傅又赔了邻居200，请问王师傅一共亏了多少？

1.4 技术挑战

Why NLP so Hard?

- 功能：语言是对世界的认知和理解。
- 知识：语言学知识、外部知识、领域知识、常识...
- 特点：组合性、开放的、动态的、长尾效应...
- 语用性：环境、上下文、信息、意图...
-

1.4 技术挑战

困难之一：大量歧义 (ambiguity) 现象

□ 词法歧义，例如：

(1) I'll see Prof. Zhang home.

(2) 自动化研究所取得的成就

自动化 / 研究所 / 取得 / 的 / 成就

自动化 / 研究 / 所 / 取得 / 的 / 成就

(3) 门把手弄坏了

门 / 把 / 手 / 弄 / 坏 / 了

门把手 / 弄 / 坏 / 了



1.4 技术挑战

□ 词性歧义：

1) 名词：苍蝇；2) 动词：飞行

(1) Time flies like an arrow.

1) 介词：像，好似；2) 动词：喜欢

- 时间像箭一样飞驰（光阴似箭）。
- 时间苍蝇喜欢箭（有一种苍蝇叫“时间”）。

(2) “动物保护警察” 明年上岗。

(3) 张某借高某人民币14000元，今还欠款4000元。

1.4 技术挑战

□ 结构歧义：

- (1) “喜欢乡下的孩子” , “关于鲁迅的文章”
- (2) “今天中午吃馒头” , “今天中午吃食堂”
“今天中午吃大碗” , “今天中午吃了闭门羹”
- (3) I saw a man with a telescope.
 - I saw [a man with a telescope].
 - I [saw a man] with a telescope.

英语句子歧义组合的开塔兰数 (Catalan Numbers) $C_n = \binom{2n}{n} \frac{1}{n+1}$

其中, n 为句子中介词短语的个数, $\binom{2n}{n} = \frac{(2n)!}{n! * n!}$

1.4 技术挑战

□ 语义歧义：

他说：“她这个人真有意思(interesting)”。她说：“他这个人怪有意思的(funny)”。于是人们以为他们有了意思(wish)，并让他向她意思意思(express)。他火了：“我根本没有那个意思(thought)”！她也生气了：“你们这么说是什么意思(intention)”？事后有人说：“真有意思(interesting)”。也有人说：“真没意思(nonsense)”。

人们的语言表达中大量地使用**缩略语**和**隐喻**的表达方式，例如：

要把权力装进制度的**笼子**；**老虎苍蝇**一起打。

破**四旧**，除**四害**；消灭一切**牛鬼蛇神**。

1.4 技术挑战

困难之二：大量未知语言现象

- ◆ 新词、人名、地名、术语等，例如：裸退、非典、新冠、大山、宁静、宁夏、失联、互联网+，共享经济。
- ◆ 新含义，例如：苹果、奔腾、同志、老虎、苍蝇等。
- ◆ 新用法和新句型等，尤其在口语中或部分网络语言中，不断出现一些“非规范的”新的语句结构。如：被长工资，很中国，百度一下。

1.4 技术挑战

归纳起来，NLU 所面临的挑战：

- 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。

1.4 技术挑战

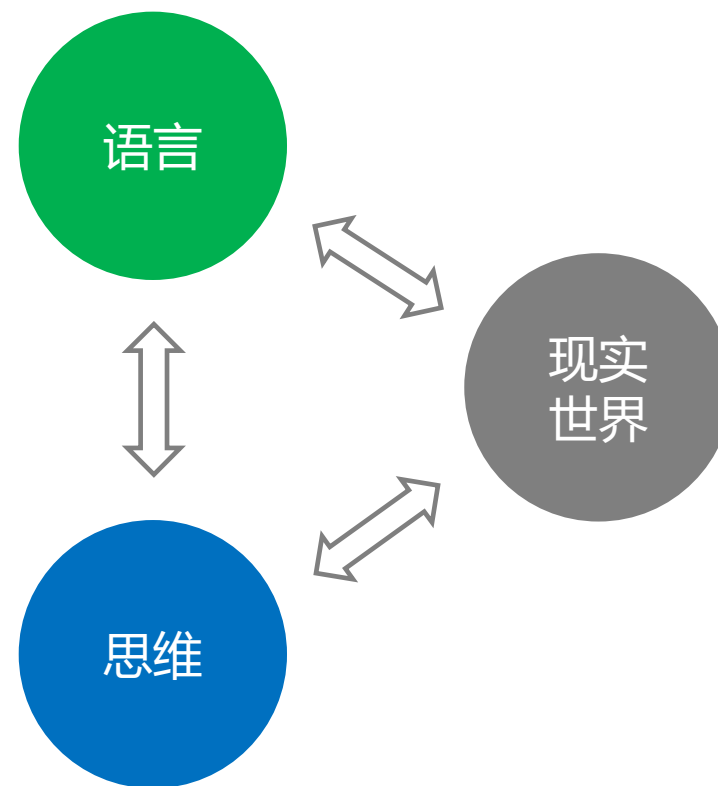
- ❖ 机器翻译中映射单元的不对等性：词法表达不相同、句法结构不一致、语义概念不对等：



1.4 技术挑战

■ 人脑理解语言是一个复杂的思维过程

- 语言学、心理学
- 逻辑学、认知科学
- 计算机科学
- 统计学、信息论
- 背景知识、常识等



1.4 技术挑战—研究现状

基本现状

- 部分问题得到了解决，可以为人们提供辅助性帮助，例如：专业领域文档翻译，电子词典，搜索引擎，文字录入等；
- 基础问题研究仍任重而道远，如：语义表示和计算、高质量的自动翻译等；
- 社会需求日益迫切：信息服务、通讯、网络内容管理、情报处理、国家安全等；
- 许多技术离真正实用的目标还有相当的距离，尚未建立起有效、完善的理论体系。

1.5 基本研究方法

1.5 基本研究方法

理性主义：通过一些特殊的语句或语言现象的研究来获取人的语言能力的认识，而这些语句和语言现象在实际的应用中可能并不常见。

□ 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。

- 规则库开发： $N + N \rightarrow NP$
- 词典标注：#工作：N(uc); V;
- 推导算法设计：归约、推导、歧义消解方法...

知识库 + 推理系统 $\rightarrow$ NLP系统

理论基础：Chomsky的文法理论

1.5 基本研究方法

经验主义：从大规模语言数据中统计学习出语言学规律。

- 求解问题的思路：基于大规模真实语料（语言数据）建立计算方法。
- 大规模真实数据的收集、标注：真实性、代表性、标注信息.....
- 统计模型建立：模型的复杂性、有效性、参数训练方法.....

语料库+统计模型 → NLP系统

理论基础：统计学、信息论、机器学习

1.5 基本研究方法

■ 以机器翻译为例

给定英语句子：

There is a book on the desk.

将其翻译成汉语。

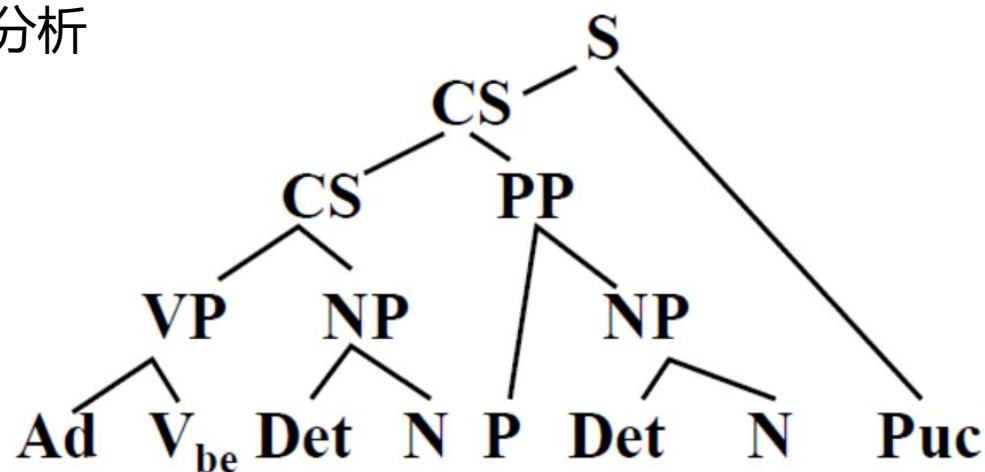
1.5 基本研究方法

□ 基于规则的方法

- 对英语句子进行词法分析

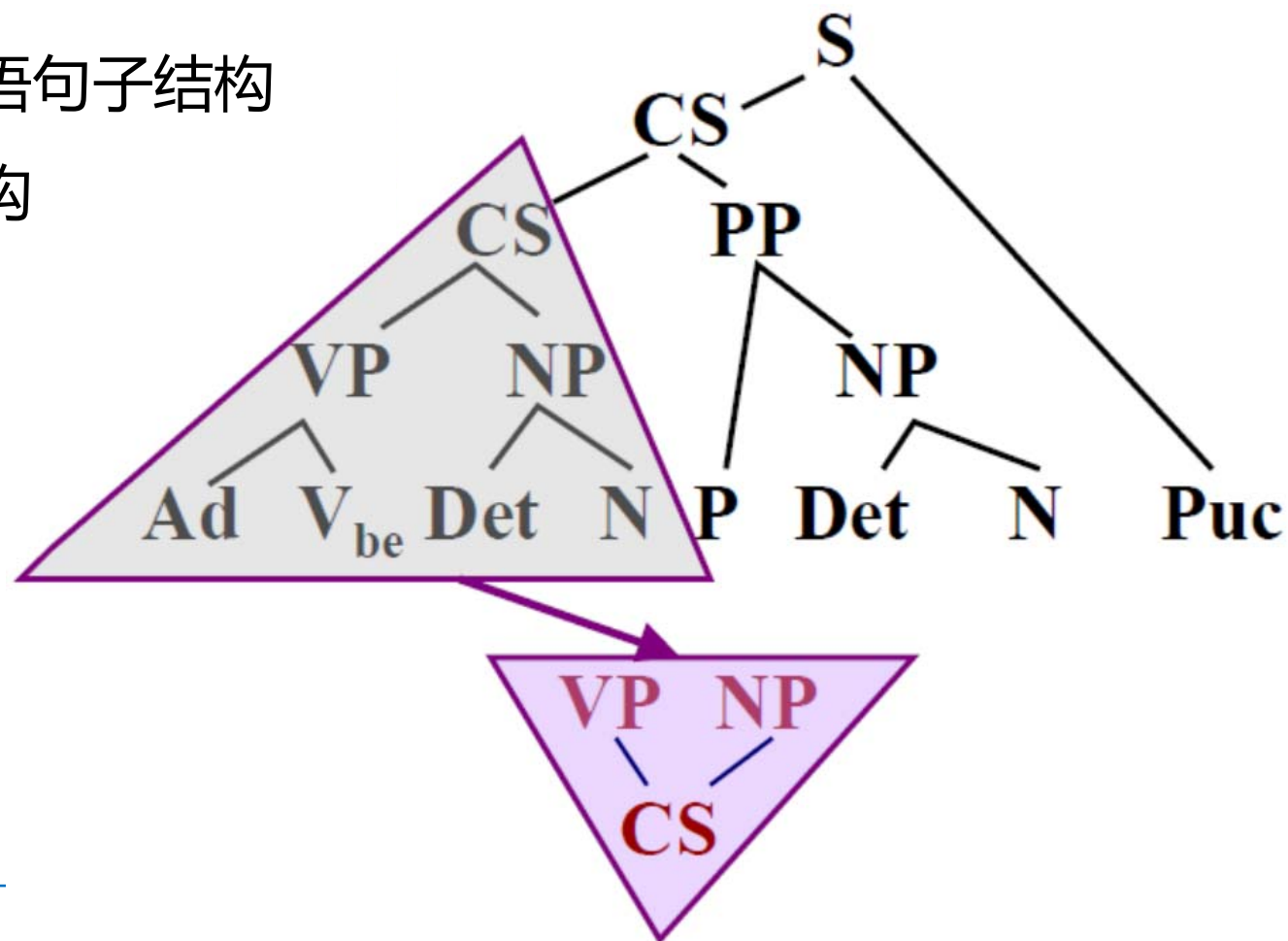
There/Ad is/V_{be} a/Det book/N on/P the/Det desk/N ./Puc

- 对英语句子进行句法结构分析



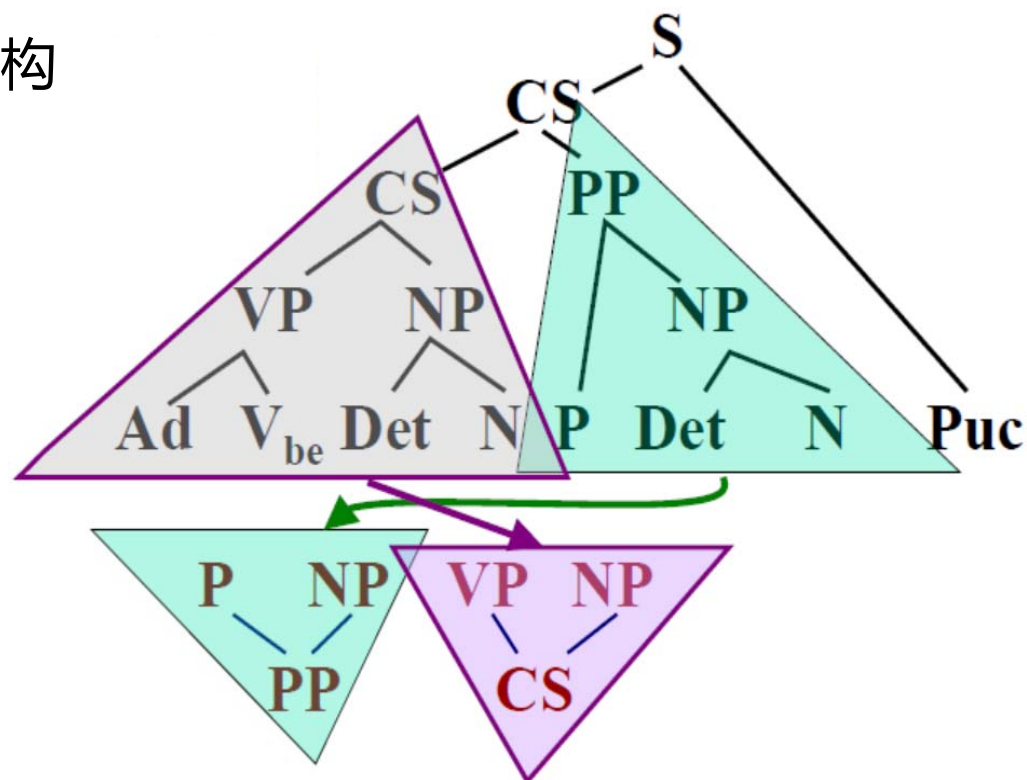
1.5 基本研究方法

利用转换规则将英语句子结构
转换成汉语句子结构



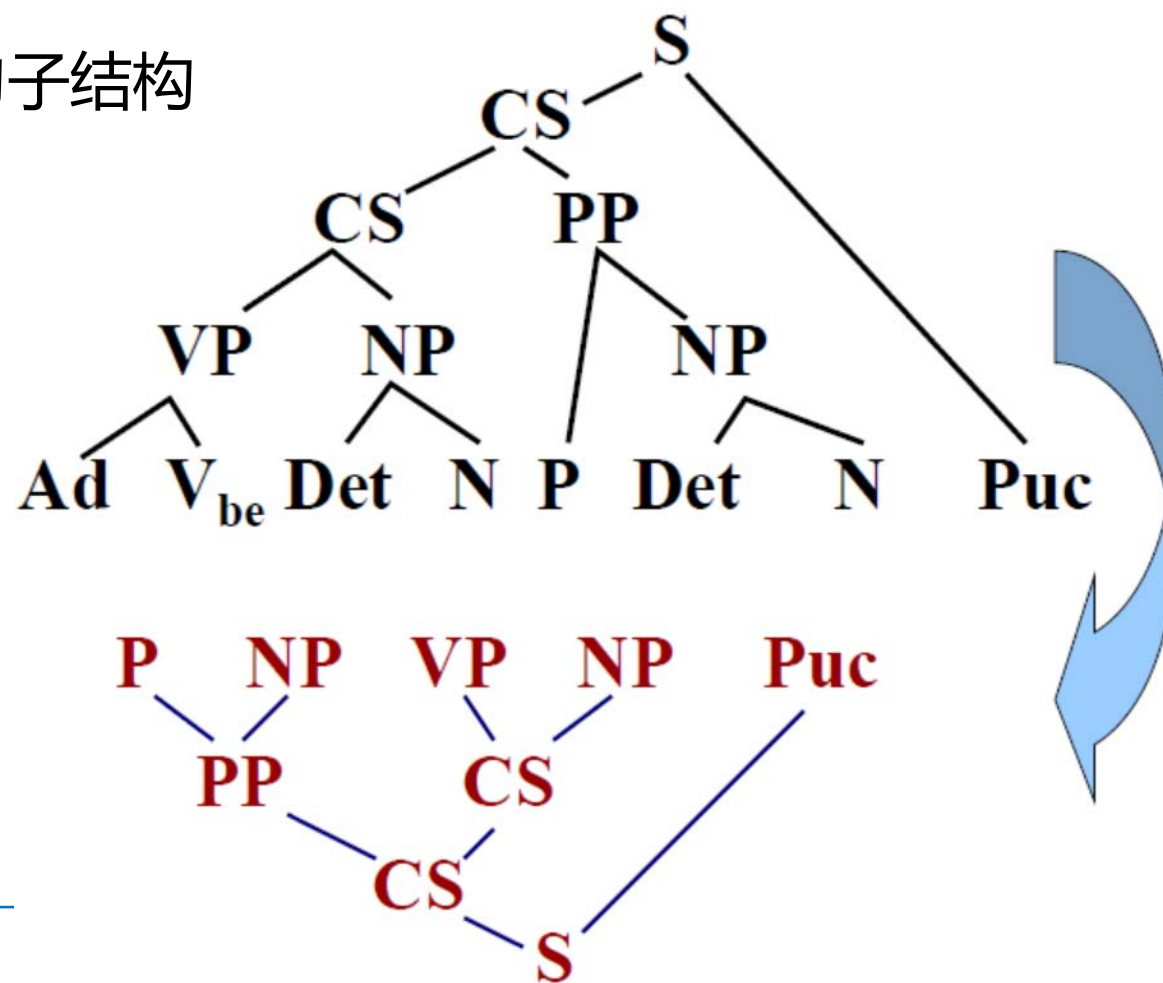
1.5 基本研究方法

利用转换规则将英语句子结构
转换成汉语句子结构



1.5 基本研究方法

利用转换规则将英语句子结构
转换成汉语句子结构



1.5 基本研究方法

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

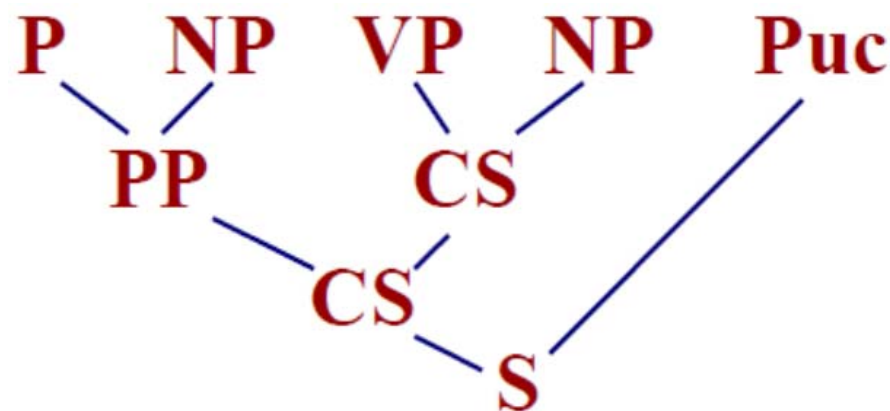
#a, Det, —

#book, N, 书; V, 预订

#desk, N, 桌子

#on, P, 在X 上

#There be, V, 有

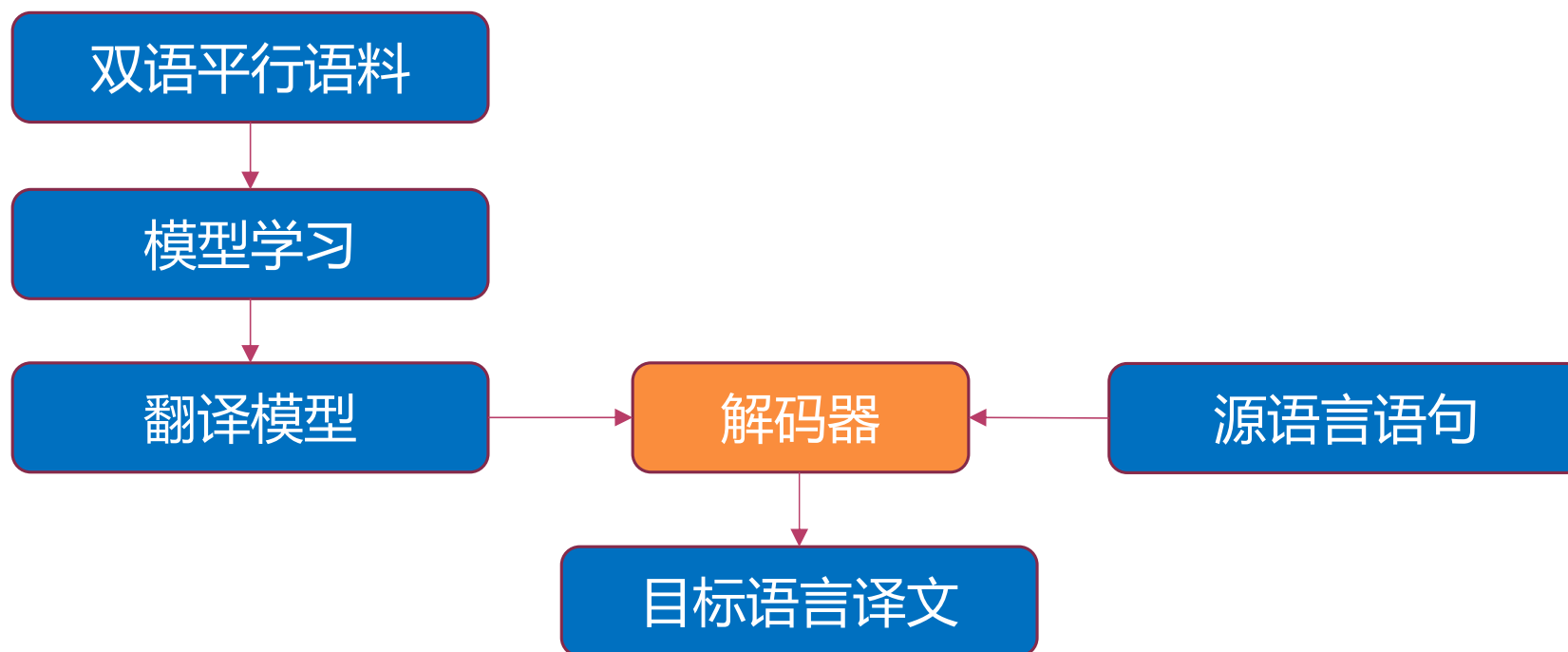


输出译文：

在桌子上有一本书。

1.5 基本研究方法

□ 数据驱动的翻译方法（如SMT和NMT）



1.5 基本研究方法

□ 数据驱动的翻译方法

给定源语言句子: $E = e_1e_2\dots e_m$

将其翻译成目标语言句子: $C = c_1c_2\dots c_l$

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

$$\hat{C} = \operatorname{argmax}(P(C|E)) = \operatorname{argmax}(P(C)P(E|C))$$

求解使 P 值
最大的 C

语言模型
Language Model

翻译模型
Translation Model

1.5 基本研究方法

构建解码器 (decoder) , 快速搜索最优翻译候选:



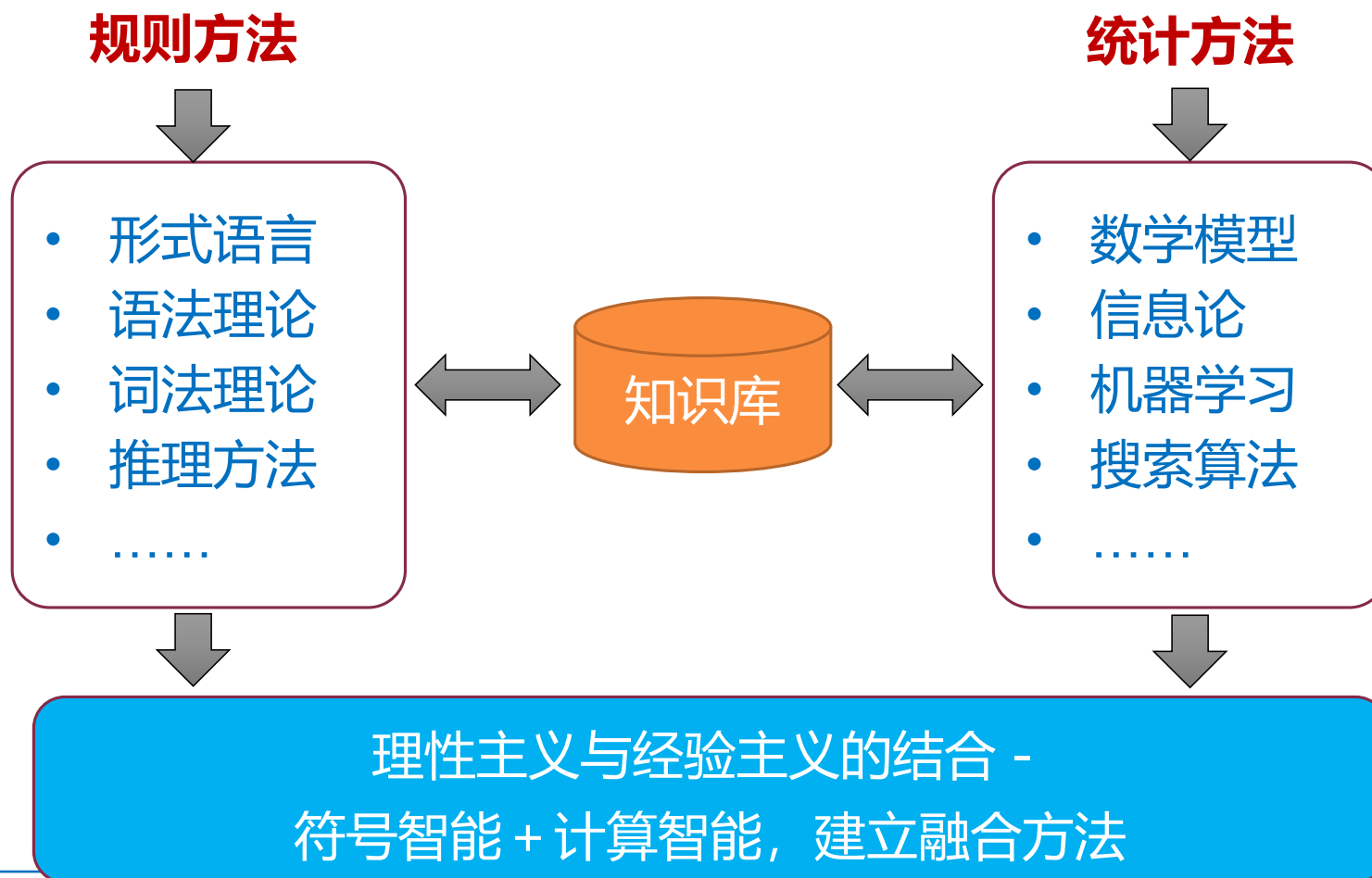
■ 三个关键问题:

- 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 , 使得 $P(C) * P(E|C)$ 有较大值

■ 主要任务:

- 收集大规模双语句子对、目标语言句子
- 参数训练与模型优化

1.5 基本研究方法



参考文献

专著:

1. C. D. Manning, H. Schütze. 1999. Foundations of Statistical Natural Language Processing. The MIT Press.
2. D. Jurafsky, and J. H. Martin. 2000. Speech and Language Processing, Prentice Hall, 2000 (冯志伟, 孙乐译, 自然语言处理综论, 电子工业出版社)
3. 冯志伟, 自然语言处理的形式模型, 中国科学技术大学出版社
4. 宗成庆, 统计自然语言处理(第2版), 清华大学出版社

参考文献

期刊:

1. Computational Linguistics
2. Natural Language Engineering
3. ACM TALLIP
4. Machine Translation
5. IEEE Trans. on Audio, Speech, and Language Processing
6. 中文信息学报/ 计算机学报/ 软件学报/ 计算机研究与发展

会议论文集:

1. Proceedings of ACL (Annual Meeting of the Association for Computational Linguistics)
2. Proceedings of NAACL, EMNLP
3. Proceedings of COLING (International Conference on Computational Linguistics)
4. Proceedings of IJCNLP (International Joint Conference on Natural Language Processing)
5. 国内相关会议论文集

参考文献

NLP课程:

大学	课程名	主讲老师	课程网站
Columbia University	NLP	Michael Collins	http://www.cs.columbia.edu/~cs4705/
CMU	NLP	Chris Dyer	http://demo.clab.cs.cmu.edu/NLP/
MIT	NLP	Robert C. Berwick	http://web.mit.edu/6.863/www/fall2012/
Stanford University	NLP	Dan Jurafsky and Christopher Manning	http://online.stanford.edu/course/natural-language-processing

思考题

1-1. 请说明如下句子有多少种不同的含义？

- ① He drew one card.
- ② 咬死猎人的狗。
- ③ 这座碑是为纪念反对分离主义者叛乱中牺牲的英雄而建立的。

1-2. 试比较汉英句子中地点状语位置的差异。

1-3. 下列语言中哪些为自然语言？

世界语、C语言、鸟语、甲骨文

问题及讨论

Q&A