



# 自然语言处理

通用人工智能与数字经济创新团队

西南财经大学

# 课程简介

- 本课程旨在介绍大模型与通用人工智能的基本概念、原理和应用。通过理论授课、实践案例分析和讨论，学生将获得对大模型和通用人工智能领域的全面了解，培养相关技能和知识，为未来在该领域的研究和应用工作打下坚实基础。

# 课程信息

- 课程类型：专业方向课
- 学分：3
- 课时：17周，1次课/周，3学时/1次课，共51学时
- 教材：自编教材，《自然语言处理：大模型理论与实践》
- 先修课程：概率论与数理统计、机器学习等

# 参考教材



- 教材：**自然语言处理：大模型理论与实践**
- 适用人群：**本科生、研究生、科研人员**
- 适用学科专业：**计算机、人工智能学科**
- 主要编者：**赵宇，任福继等**

教材官网：<https://nlp-book.swufenlp.group/>

# 授课安排

一·绪论	3	十·提示工程	3
二·词向量	3	十一·涌现	3
三·统计语言模型	3	十二·大模型评估	3
四·神经语言模型	3	十三·探讨	3
五·预训练语言模型	3	十四·大模型本地开发	3
六·大语言模型架构	3	十五·基于大模型的应用开发	3
七·多模态大模型框架	3	<ul style="list-style-type: none"><li>■ 教学时间：1周—17周（17周）</li><li>■ 教学地点：周五 1-3节：H301</li><li>■ 教学方式：概念讲解 + 实践 + 答疑</li></ul>	
八·大模型预训练	3		
九·大模型微调	3		

# 考核方式

- 考核方式 (暂定) : 平时成绩 (50%) + 期末闭卷考试 (50%)
- 平时成绩  
考勤、课堂表现和平时作业 (20%) + 课程大实验 (30%)

# 预备知识

## A.1 概率论基本概念

A.1.1 概率和概率分布

A.1.2 贝叶斯法则

## A.2 信息论基本概念

A.2.1 熵

A.2.2 联合熵和条件熵

A.2.3 互信息

A.2.4 相对熵与交叉熵

## A.3 机器学习的基本概念

A.3.1 训练方式

A.3.2 常用算法和模型

## A.4 强化学习基本概念

A.4.1 马尔可夫过程决策

A.4.2 策略迭代

A.4.3 重要性采样

A.4.4 近端策略优化算法

# 相关学术会议与学术组织

## NLP领域主要学术会议：

- Association for Computational Linguistics (ACL) **CCF A**
- Conference on Empirical Methods in Natural Language Processing (EMNLP) **CCF B**
- North American Chapter of the Association for Computational Linguistics (NAACL) **CCF B**
- International Conference on Computational Linguistics (COLING) **CCF B**
- The China National Conference on Computational Linguistics (CCL)
- Conference of the European Chapter of the Association for Computational Linguistics (EACL)

## NLP领域学术组织：

ACL、国际计算语言学委员会（ICCL）、中国人工智能学会自然语言理解专委会、中国中文信息处理学会、中国计算机学会自然语言处理专委会



金融智能与金融工程四川省重点实验室  
Financial Intelligence and Financial Engineering  
Key Laboratory of Sichuan Province 通用人工智能与金融创新团队

# 第一章 绪论

# 目录

1.1自然语言处理概述

1.2自然语言处理简史

1.3自然语言处理传统研究内容

1.3.1传统基础技术

1.3.2实际应用

1.4自然语言处理与大模型发展现状

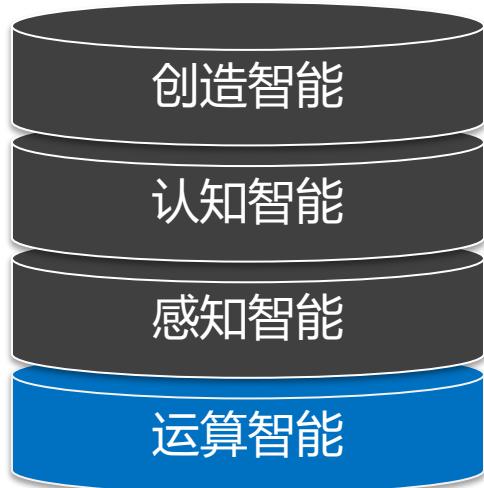
# 1.1 自然语言处理概述



金融智能与金融工程四川省重点实验室  
Financial Intelligence and Financial Engineering  
Key Laboratory of Sichuan Province

自然语言处理被誉为人工智能皇冠上的明珠。

—— 无名氏



记忆存储、计算能力，  
机器早已经超过人类。  
(超级计算机)



# 1.1 自然语言处理概述

自然语言处理被誉为人工智能皇冠上的明珠。

——无名氏



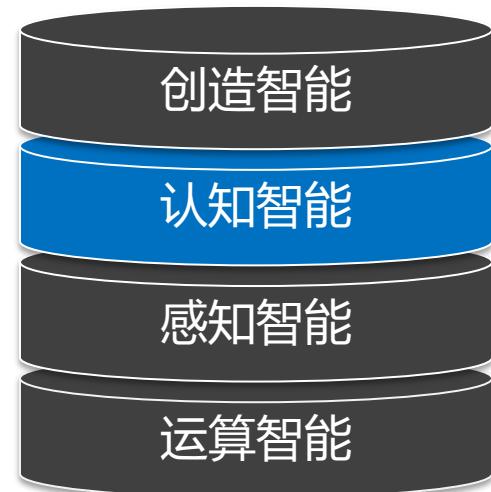
包括听觉、视觉、触觉；  
随着深度学习的引入，大幅度提高语音识别和图像识别的识别率，计算机在感知智能层面已经做得相当不错，在一些典型的测试题下，达到或者超过了人类的平均水平  
(语音识别、图像识别)



# 1.1 自然语言处理概述

自然语言处理被誉为人工智能皇冠上的明珠。

——无名氏



理解、运用语言的能力，掌握知识、运用知识的能力，以及在语言和知识基础上的推理能力。认知智能主要集中在语言智能，即自然语言处理，它简单理解了句子、篇章，实现了帮助搜索引擎、仿照系统提供一些基本的功能、提供一些简单的对话翻译。

(自然语言处理)

# 1.1 自然语言处理概述



金融智能与金融工程四川省重点实验室  
Financial Intelligence and Financial Engineering  
Key Laboratory of Sichuan Province

自然语言处理被誉为人工智能皇冠上的明珠。

——无名氏

创造智能

认知智能

感知智能

运算智能



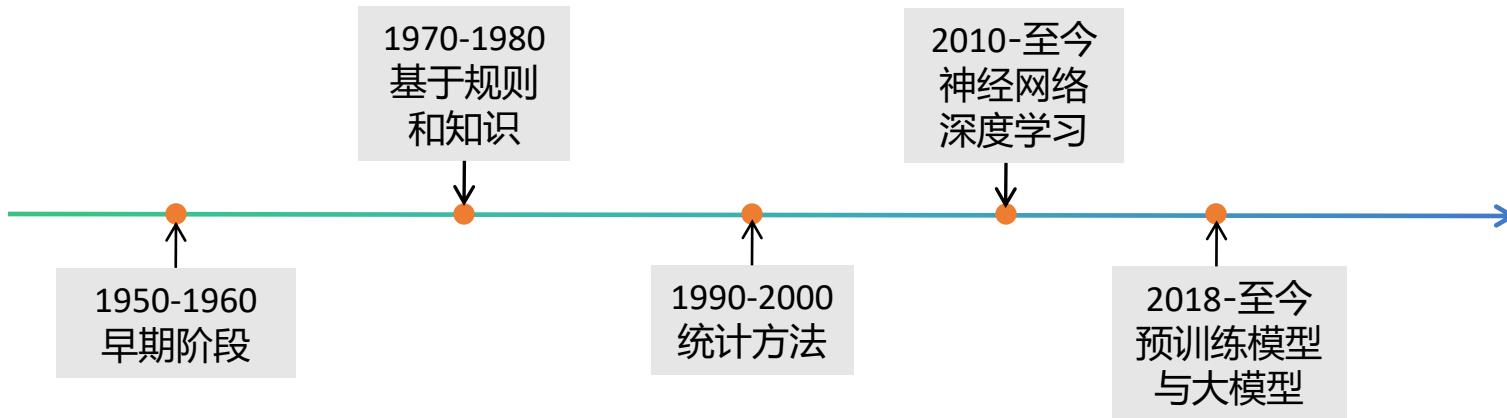
人们利用已有的条件，利用一些想象力甚至有一些是臆断、梦想，想象一些不存在的事情包括理论、方法、技术，通过实验加以验证，然后提出新的理论，指导更多实践，最后产生很好的作品或产品。

(创造力)



## 1.2 自然语言处理简史

自然语言处理技术的发展可以追溯到20世纪50年代初，经历了若干重要阶段和关键突破。其发展历程如下：





# 1.3 传统研究内容

## 1.3.1 传统基础技术

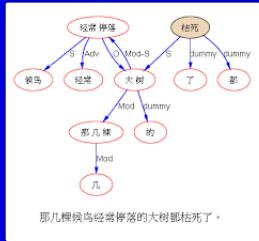
- 1 自动分词
- 2 命名实体识别
- 3 词性标注
- 4 句法分析
- 5 语义分析
- 6 篇章分析

## 1.3.2 实际应用

- 1 机器翻译
- 2 文本分类与情感分类
- 3 信息抽取与自动文摘
- 4 信息检索与问答系统
- 5 对话系统
- 6 阅读理解

## 1.3 传统研究内容

# 自动分词



- 一般领域

- 人名
  - 组织结构名
  - 地名
  - 数学表达式

- 旅游领域

- 景点名
  - 酒店名
  - 目的地名
  - 机场名

那几棵候鸟经常停落的大树都枯死了。

## 自动实体识别



在兴义市区期间，入住的是富康国际酒店。从兴义机场到市中心的富康国际酒店打车不过十来分钟。富丽堂皇的大堂、穿着民族服饰的前台服务员、雅致的房间，都让我对兴义有着不错的印象。刘氏庄园位于兴义城南。

词性标注

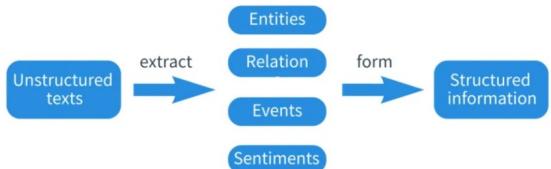


# 机器翻译

# “Machine learning...”

機器學習

# 信息抽取



## 文本分类 （情感分类）

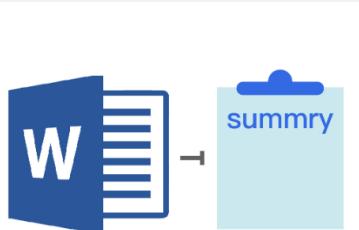


# 句法分析

语义分析



对话系统



## （摘要生成） 阅读理解

- 自然语言处理有哪些主要任务?**

自然语言处理(NLP)是计算机科学和人工智能的一个重要分支。其主要任务包括但不限于以下几类:

  - 语义分析:** 分析句子的结构, 指出句子中各个词汇的依赖关系和句法功能。
  - 语用分析:** 理解句子或文本的意义, 包括语义分类和句子意义结构分析。
  - 情感分析:** 通过对文本中的情感表达模式或主观信息进行识别、提取、负面或正面情感倾向的分析。
  - 命名实体识别:** 识别文本中的特定实体, 如人名、地名或组织名。
  - 机器翻译:** 将文本从一种语言翻译成另一种语言, 同时保持该语言的语体性和完整度。
  - 文本生成:** 基于给定的语境信息生成新的符合语境要求的文本。
  - 对话系统:** 通过文本或语音交互, 在对话过程中理解用户的意图并做出响应。
  - 语义角色标注:** 将文本中的词语与它们在语句中的角色进行关联。
  - 语义指代消解:** 将文本中的代词与它们所指代的名词进行匹配。
  - 语义相似度计算:** 以词汇为基础进行一步或多步的语义分析。
  - 语义网:** 为本领域的各个方面构建一个语义网络, 例如, 动物学知识库。

自然语言处理通常需要利用机器学习和深度学习技术来实现, 以便更好地处理语义方面的语言现象。

信息检索  
问答系统

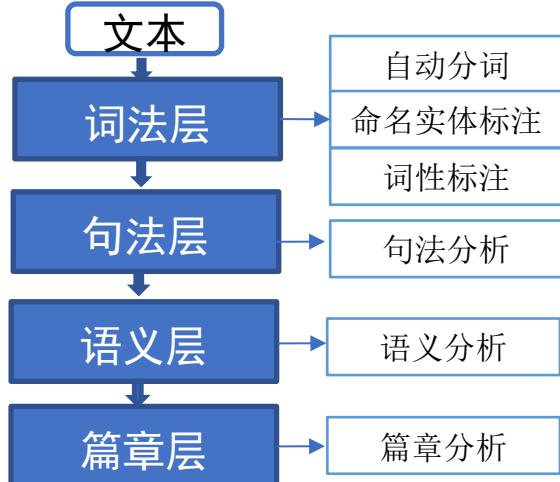
篇章分析



## 1.3 传统研究内容

### 自然语言处理

#### 传统基础技术



#### 实际应用





## 1.3.1传统基础技术：自动分词

词是自然语言处理的基本单位。

屈折语系：**We are good friends.**(英语)

**Somos buenos amigos.** (西班牙语)

有空格分隔和词形变化

孤立语系：**我们是好朋友。** (汉语)

**私たちはいい友達だよ。** (日语)

无明显分隔标记

无词形变化，用词语来  
表示时态

- 我们是好朋友。
- 我们曾经是好朋友。
- 我们现在是好朋友。



## 1.3.1 传统基础技术：自动分词

从断句看分词，信息“无损”传递：

明日逢春好不晦气 终年倒运少有余财

**断句1：明日逢春，好不晦气；终年倒运，少有余财。**



ChatGPT

这两句话是用来描述一个人非常不顺利和不幸的情况。"明日逢春，好不晦气" 这句话是在说，尽管即将迎来新的春天（通常是个好兆头或新的开始），但是却非常不顺利和倒霉。"终年倒运，少有余财" 这句话则是在描述这个人一整年都很不顺利，而且财务状况也不好，很少有多余的财富积累。整体来看，这段话描绘了一个人在新的一年中可能会遭遇连续的不幸和财运不佳的情况。

**断句2：明日逢春好，不晦气；终年倒运少，有余财。**

这两句话展示了一个积极乐观的情景。"明日逢春好，不晦气" 这句话表明，即将到来的春天是美好和顺利的，没有任何不吉利或不顺的事情发生。"终年倒运少，有余财" 这句话则表示，整年中不顺的事情很少，而且财务状况良好，有多余的财富积累。整体来看，这段话描绘了一个人在新的一年中将会遇到好运和财运顺畅的情况。



## 1.3.1 传统基础技术：自动分词

- **自动分词 (Automatic word segmentation)** : 是指让计算机系统将连续的文本流切分成独立的、有意义的词语单元的过程。

示例：人们/穿行在/重庆市/解放碑/步行街

- **自动分词面临的问题：**

1. **分词规范**: 确定词语边界的规则或指导原则。

不规范划分：火车 / 穿过 / 南京 / 市长/ 江大桥

符合规范划分：火车 / 穿过 / 南京市 / 长江大桥

2. **歧义切分**: 不同切分方式带来的语义的曲解。

示例：拍 / 电影的人

拍电影的 / 人

3. **未登录词识别**: 文本中没有出现在预定义词典或训练语料库中的词语，即一些新出现的词汇，如 “栓Q” 、 “凡尔赛” 。



## 1.3.1 传统基础技术：自动分词

- 自动分词方法：

### 1. 最大匹配法 (Maximum Matching, MM) : 分为FMM(正向)、BMM(逆向)

**核心原理**: 通过从左到右 (FMM) 或从右到左 (BMM) 扫描待分词文本，配合预定义的词典<sup>1</sup>来找出最长的匹配词。



1. 词典是指经过处理和统计的词库（语料库）。



## 1.3.1 传统基础技术：自动分词

例

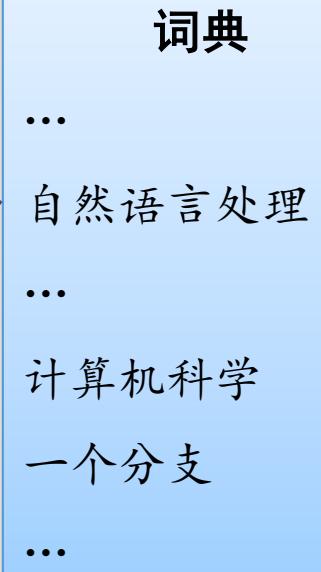
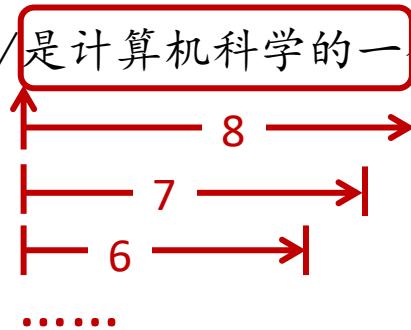
假设词典中包含单词有“自然语言处理”、“计算机科学”和“一个分支”，且最长单词的字数为8。

输入字串：自然语言处理是计算机科学的一个分支。

切分过程：



自然语言处理/是计算机科学的一个分支。



FMM最终切分结果：自然语言处理/是/计算机科学/的/一个分支。



## 1.3.1 传统基础技术：自动分词

- 自动分词方法：

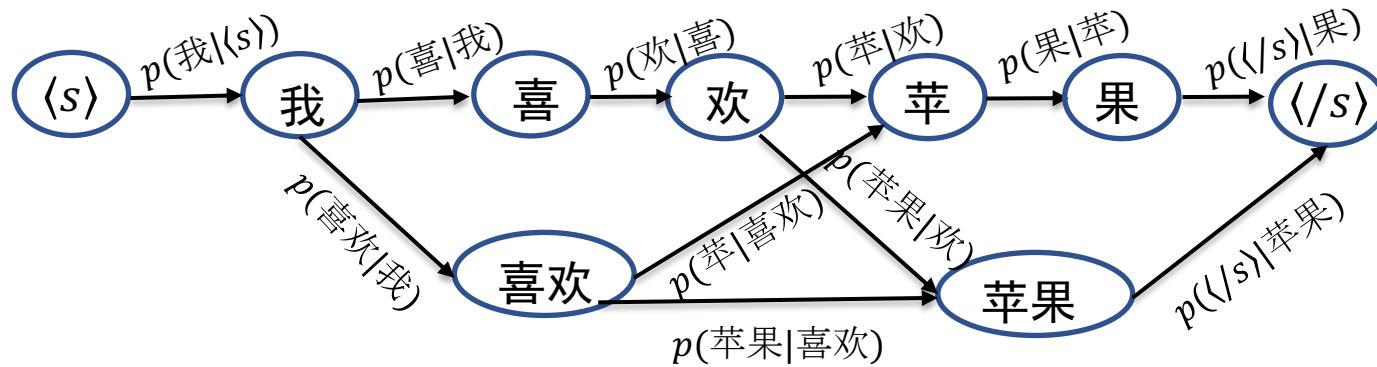
### 2. 基于语言模型的分词方法：

- a) 建立  $n$  元语言模型，计算每个词在给定上下文中出现的**条件概率**；
- b) 构造  $n$  元的**切分词图**，图中的结点表示可能的词候选，边表示路径，计算词在当前位置的前  $n - 1$  个词的下的条件概率表示代价；
- c) 利用相关搜索算法找到图中代价最小的路径作为分词结果输出。

## 1.3.1 传统基础技术：自动分词

例

假设有一个简短的中文文本：“我喜欢苹果”，并且有一个包含“我”、“喜欢”和“苹果”的词典。使用二元语言模型进行分词。



建立切分词图后，应用相关搜索算法（如动态规划）来找到图中代价最小的路径。

最终切分结果：我/喜欢/苹果。



## 1.3.1 传统基础技术：自动分词

- 自动分词方法：

### 3. 基于最短路径的分词方法

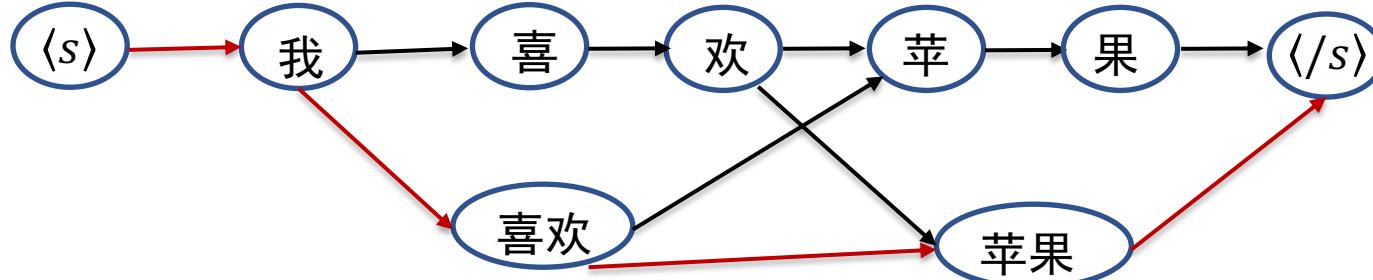
通过构建分词图，利用最短路径算法来找到最佳的分词序列，从而实现自动分词。能够充分考虑词语之间的连贯性和上下文信息。

- 与基于语言模型的分词技术相似。
- 不同之处在于，**节点间的权重**并非基于条件概率，而是根据词频或词语概率确定，有时为简化处理，所有权重也可能统一设定为 1。



## 1.3.1 传统基础技术：自动分词

例 假设权重统一设定为1，则最短路径为节点最少的分词路径。



最终切分结果：我/喜欢/苹果。



# 1.3.1传统基础技术：自动分词

## 扩展 自动分词方法（大语言模型—提示学习法）

E

You

请给下面的文段分词：所以标点放在不同的地方文章就会有不同的意思正如韩愈在师说中所说的彼童子之师授之书而习其句读者因此古代的断句大多是老师教的老师专门讲解文章告诉学生在哪个地方应该断教会之后学生再读书时完全就可以凭借自己以前学过的断句来理解了



ChatGPT

以下是文段的分词结果：

所以 / 标点 / 放在 / 不同 / 的 / 地方 / 文章 / 就 / 会 / 有 / 不同 / 的 / 意思 / 正如 / 韩愈 / 在 / 师说 / 中 / 所说 / 的 / 彼 / 童子 / 之 / 师 / 授之 / 书 / 而 / 习 / 其 / 句 / 读者 / 因此 / 古代 / 的 / 断句 / 大多 / 是 / 老师 / 教 / 的 / 老师 / 专门 / 讲解 / 文章 / 告诉 / 学生 / 在 / 哪个 / 地方 / 应该 / 断 / 教会 / 之后 / 学生 / 再 / 读书 / 时 / 完全 / 就 / 可以 / 凭借 / 自己 / 以前 / 学过 / 的 / 断句 / 来 / 理解 / 了 /。

这样的分词方式可以更好地理解文段的结构和意义。



## 1.3.1 传统基础技术：自动分词

### 扩展 自动分词方法（大语言模型—提示学习法）

所以标点放在不同的地方文章就会有不同的意思！正如韩愈在《师说》中所说的：“彼童子之师，授之书而习其句读者。”因此古代的断句，大多是老师教的，老师专门讲解文章，告诉学生在哪个地方应该断，教会之后学生再读书时完全就可以凭借自己以前学过的断句来理解了。

E

You

请给下面的文段补充标点符号：所以标点放在不同的地方文章就会有不同的意思正如韩愈在师说中所说的彼童子之师授之书而习其句读者因此古代的断句大多是老师教的老师专门讲解文章告诉学生在哪个地方应该断教会之后学生再读书时完全就可以凭借自己以前学过的断句来理解了



ChatGPT

所以，标点放在不同的地方，文章就会有不同的意思。正如韩愈在《师说》中所说的：“彼童子之师，授之书而习其句。”读者因此可知，古代的断句大多是老师教的。老师专门讲解文章，告诉学生在哪个地方应该断。教会之后，学生再读书时，完全就可以凭借自己以前学过的断句来理解了。



# 1.3.1传统基础技术：自动分词

## 扩展 自动分词方法（大语言模型—提示学习法）

SIGHAN

Second International Ch

The Second International Chinese Word Segmentation Bakeoff took place over the summer of 2005 and the results were presented at the [4th SIGHAN Workshop](#), held at [IJCNLP'05](#), October 14-15.

Corpora from the following organizations were used:

- CKIP, Academia Sinica, Taiwan
- City University of Hong Kong, Hong Kong SAR
- Beijing University, China
- Microsoft Research, China

The complete training, testing, and gold-standard data sets, as well as the scoring script, are available for research use:

- [icwb2-data.rar](#) [40 MB, [md5](#)]
- [icwb2-data.zip](#) [50 MB, [md5](#)]
- [icwb2-data.tar.bz2](#) [37MB, [md5](#)]

The [Detailed Instructions](#) for the bakeoff are available. Please read them carefully.

Segmentation guidelines for the following corpora are available. These were supplied to SIGHAN by each data provider, and converted into PDF by the organizer:

Corpus	MS Word	PDF
Academia Sinica	<a href="#">516 KB</a>	<a href="#">336 KB</a>
City University of Hong Kong	<a href="#">154 KB</a>	<a href="#">237 KB</a>
Peking University	<a href="#">177 KB</a>	<a href="#">294 KB</a>
Microsoft Research	<a href="#">41 KB</a>	<a href="#">70 KB</a>

The [collected results](#) of all participating systems are also available.

Many thanks to the data providers and the bakeoff participants!

## Contact Information

The bakeoff was organized by Tom EMERSON of Basis Technology Corp.

Questions on the bakeoff should be addressed to [Tom Emerson](#).



## 1.3.1 传统基础技术：自动分词

- 结果评估：

准确率 :  $P = \text{系统输出中正确的结果个数} / \text{系统所有输出结果个数}$

召回率 :  $R = \text{系统输出中正确的结果个数} / \text{测试集中正确的答案个数}$

测度值 :  $F = 2 * PR / (P + R)$



## 1.3.1 传统基础技术：自动分词

例

假设有一个汉语分词系统，该系统在一个测试集上有2230个分词结果，而标准答案是1980个词语。对比发现，系统切分出来的结果中有1872个是正确的。那么它的评价指标分别是多少？

$$P = \frac{1872}{2230} \times 100\% = 83.85\%$$

$$R = \frac{1872}{1980} \times 100\% = 94.55\%$$

$$\begin{aligned} F1 &= \frac{2 \times P \times R}{P + R} \times 100\% \\ &= \frac{2 \times 78.33 \times 91.35}{78.33 + 91.35} \times 100\% = 88.88\% \end{aligned}$$



## 1.3.1 传统基础技术：命名实体识别

- **命名实体识别 (Named Entity Tagging, NER)**：是指识别文本中具有特定意义的实体并将其分类为预定义的类别。

**核心目标：**是从非结构化文本中准确识别和分类具有特定语义价值的实体，并将它们归类到预先定义的类别中。

- **命名实体**

- 人名（中国人名和外国译名）
- 地名
- 组织机构名
- 数字日期
- 货币数量
- 其他

示例： 故宫 始建于公元 1406 年。  
地名                  时间



## 1.3.1 传统基础技术：命名实体识别

- 命名实体识别方法

1. 基于规则的方法：根据人工编写的基于模式或基于上下文的规则进行识别。

示例：包含“省”、“市”、“县”等关键字的字符串一般是地名

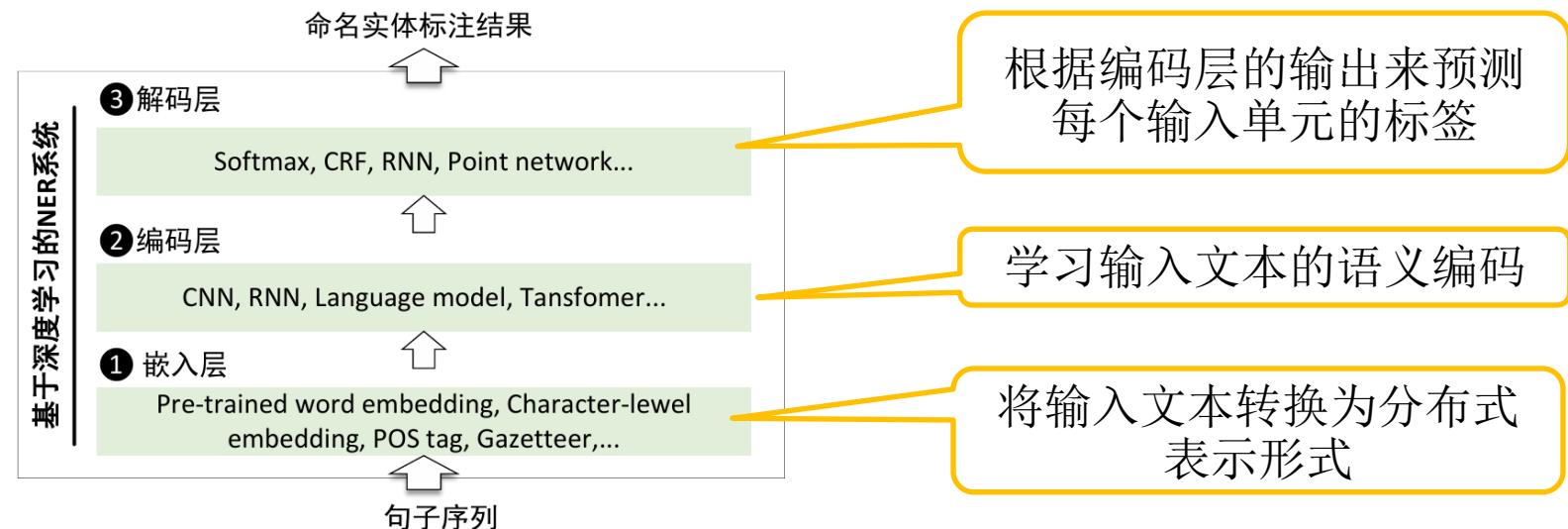
2. 基于统计的方法：其思想是利用人工标注的语料进行训练，通过对训练语料所包含的语言信息进行统计和分析，从训练语料中挖掘出特征。



## 1.3.1 传统基础技术：命名实体识别

### ● 命名实体识别方法

3. 基于深度学习的方法：其思想是利用深度学习非线性的特点，建立从输入到输出非线性的映射，学习得到更加复杂精致的特征，从而选择概率最大的实体。





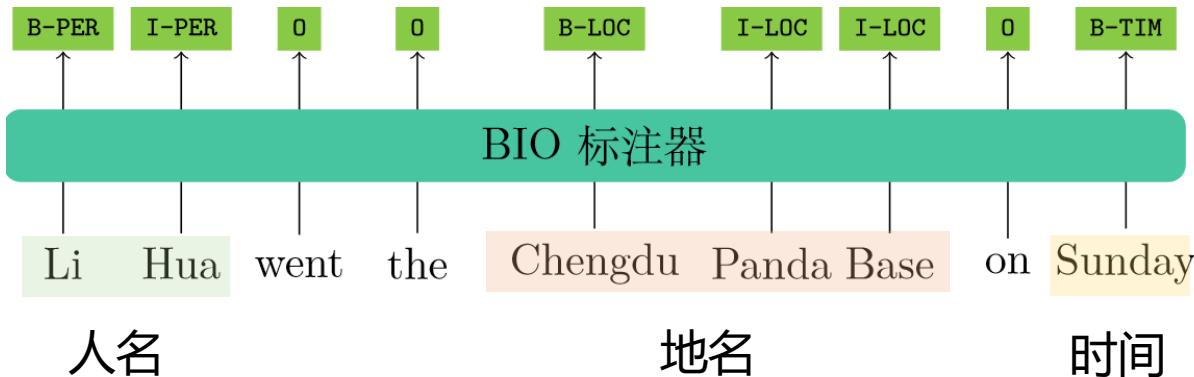
## 1.3.1 传统基础技术：命名实体识别

### ● 实体标注方法：BIO标注

BIO标注是序列标注的常见方法。通过标记来捕捉边界和命名实体类型，从而将命名实体识别看作一个逐字逐句序列标注的任务。

**B-X:** X类型实体的开头。    **I-X:** X类型实体的中后部分。

**O:**不属于命名实体的词。





## 1.3.1 传统基础技术：命名实体识别

### 💡 扩展 命名实体识别方法（大语言模型—提示学习法）

F

You

请给这段话进行命名实体标注：在过去的一场足球比赛中，巴黎圣日耳曼对阵利物浦。两队的球星梅西、内马尔和萨拉赫在比赛中展现出了出色的发挥。巴黎圣日耳曼的内格雷多在下半场取得了一粒关键的进球，最终帮助球队赢得了比赛。整个比赛在巴黎的巴尔德球场进行，吸引了来自世界各地的球迷。



ChatGPT

以下是对文段中命名实体的标注：

在过去的一场足球比赛中，[巴黎圣日耳曼]{组织}对阵[利物浦]{组织}。两队的球星[梅西]{人名}、[内马尔]{人名}和[萨拉赫]{人名}在比赛中展现出了出色的发挥。[巴黎圣日耳曼]{组织}的[内格雷多]{人名}在下半场取得了一粒关键的进球，最终帮助[球队]{组织}赢得了比赛。整个比赛在[巴黎]{地点}的[巴尔德球场]{地点}进行，吸引了来自世界各地的[球迷]{人名}。



## 1.3.1传统基础技术：词性标注

词性是词汇基本的语法属性。

Your answer is exactly right.  
你的回答完全正确。

“right”是一个形容词，用来描述  
“answer”

Everyone has a right to a fair trial.  
每个人都有权获得公正的审判。

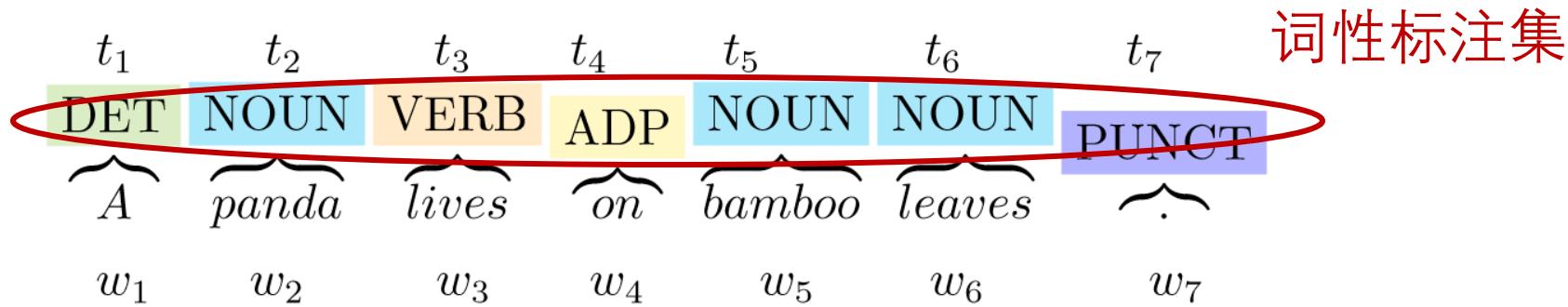
“right”是一个名词，指的是一种  
权利

- 每个词性都有一些通用的特点和用法，这些特点和用法决定了它们在句子中的位置、作用以及在语言理解和表达中的重要性。
- 可以根据词性猜测用法。



## 1.3.1 传统基础技术：词性标注

- **词性标注 (Parts-Of-Speech tagging)**：目标是为给定句子中的每个单词分配一个适当的词性标签，从而明确其在句子中的语法角色和功能。
- 词性标注是为了消除歧义，确定词语在句中所表达的含义，如确定单词 “book” 是动词还是名词。



- **词性标注集**：描述词性的集合，用于在自然语言处理中对单词进行标记和分类，如UPenn Treebank、ICTCLAS。



## 1.3.1 传统基础技术：词性标注

### ● 词性标注方法：

1. 基于规则的方法：根据已有的语法知识和语言规律人工设计规则，完成词性标注。

例

that 具有多种词性，比如可以做从属连词 (SCONJ)，副词 (ADV)

(SCONJ) : We know **that** it is bad.

(ADV) : It is not **that** bad.

编写消歧规则选择词性进行标注

*if* 后一个单词是 *ADJ*、*ADV*

*and* 后一个单词是句子边界

*and* 前一个单词不是动词

*then* 消除 *SCONJ*

*else* 消除 *ADV*



## 1.3.1 传统基础技术：词性标注

- 词性标注方法：

**2. 基于统计模型的词性标注方法：**从大量已标注文本中学习词语和词性之间的关联，常见模型如HMM、CRF。

**3. 基于规则与统计结合的方法：**将人工设计的规则和统计方法相结合，利用规则对一些特殊情况进行标注，然后利用统计方法对其他情况进行标注。特殊情况采用基于规则的方法，一般情况采用基于统计模型的方法。



## 1.3.1 传统基础技术：词性标注

### 💡 扩展 词性标注方法（大语言模型—提示学习法）

F

You

对这段话进行词性标注：“The sun set behind the horizon, casting a warm glow across the sky. The waves gently lapped the shore, creating a soothing rhythm. I sat on the sand, feeling the gentle breeze brush against my skin.



ChatGPT

以下是对文段中的词性进行标注：

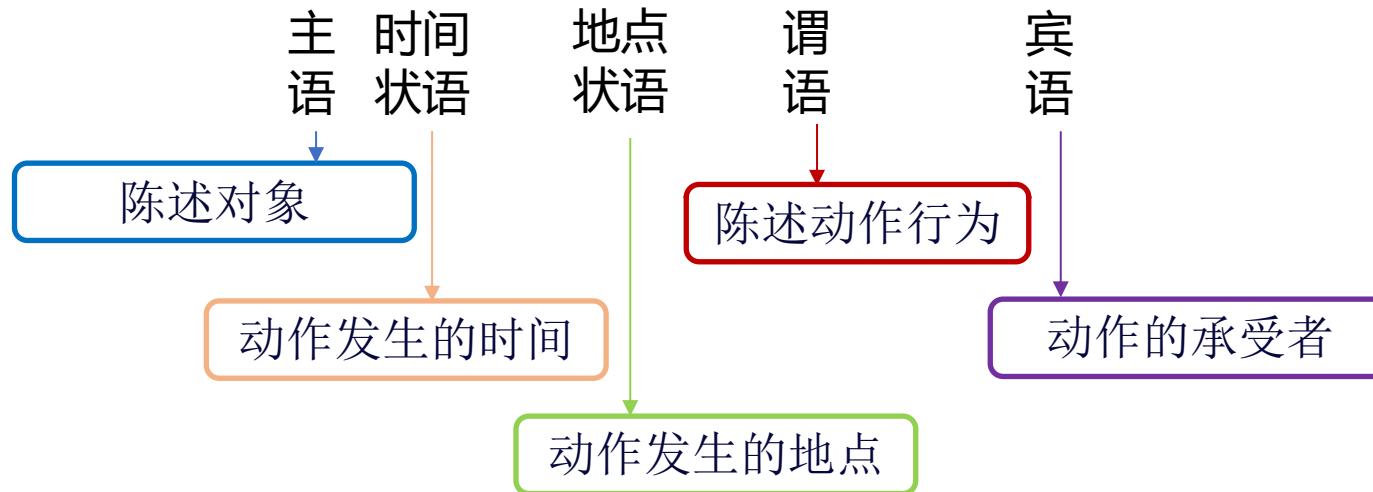
- **The** (DT) sun (NN) set (VBD) behind (IN) the (DT) horizon (NN), casting (VBG) a (DT) warm (JJ) glow (NN) across (IN) the (DT) sky (NN).
- **The** (DT) waves (NNS) gently (RB) lapped (VBD) the (DT) shore (NN), creating (VBG) a (DT) soothing (JJ) rhythm (NN).
- I (PRP) sat (VBD) on (IN) the (DT) sand (NN), feeling (VBG) the (DT) gentle (JJ) breeze (NN) brush (VB) against (IN) my (PRP\$) skin (NN).



## 1.3.1 传统基础技术：句法分析

语法结构是语言意义的重要组成部分。

他 昨天 在图书馆里 读了 一本有趣的书。



通过分析语法结构，可以清晰地理解这个句子的意思。



## 1.3.1 传统基础技术：句法分析

- 句法分析 (Syntax Parsing) : 过对句子的语法结构进行分析，确定句子中各个词语之间的语法关系，例如主谓关系、宾语关系、定语关系等，进而推断句子的含义。
- 两种类型
  - 短语结构分析 (又称句法结构分析)
  - 依存句法分析



## 1.3.1 传统基础技术：句法分析

- **短语结构分析：**识别句子中的短语、子句等，并表示出它们之间的语法关系，如主谓关系等，通常使用上下文无关文法 (Context-Free Grammar, CFG) 来描述句子的语法结构。

### 1. 上下文无关文法 (Context-Free Grammar, CFG)

四元组  $G = \{N, \Sigma, S, P\}$

- N: 非终结符的有限集合，如NP、VP
- $\Sigma$ : 终结符的有限集合：单词，符号
- S: 句子符或初始符
- P: 是一组产生式规则的有限集合，如 $NP \rightarrow Det\ Nominal$

产生式规则将句子分成分句或短语，如名词短语和动词短语，并定义了如何将非终结符替换为终结符或者其他非终结符。



## 1.3.1 传统基础技术：句法分析

- 如何理解文法？

成都是四川的省会。 → <名词短语><动词><名词短语><句号>  
→ <名词短语><动词短语><句号>

四川位于中国西南地区。 → <名词短语><动词><名词短语><句号>  
→ <名词短语><动词短语><句号>

熊猫是非常可爱的。 → <名词短语><动词><形容词短语><句号>  
→ <名词短语><动词短语><句号>

归纳以上3个句子，可以得到它们的主体结构

句子 → <名词短语><动词短语><句号>

产生式规则

此外，<动词短语>=<动词><形容词短语> | <动词><名词短语>



## 1.3.1 传统基础技术：句法分析

### ● 如何理解文法？

此外，产生式还有

- <动词短语>=<动词><形容词短语> | <动词><名词短语>
- <动词>= {是，位于}
- <形容词短语>= {非常可爱的}
- <名词短语>= {成都，四川，熊猫，四川的省会，中国西南地区}
- <句号>={。 }

代表一个语法范畴，是非终结符

集合中的词语仅表示自身，称为终结符



## 1.3.1 传统基础技术：句法分析

例

分析句子 “The cat chased the mouse.（猫追逐老鼠）”。使用 CFG 来表示句子的语法结构，并生成相应的句法树。

首先，需要定义一些基本的产生式规则：

$$\begin{aligned}S &\rightarrow NP \ VP \\NP &\rightarrow Det \ N \\VP &\rightarrow V \ NP \\Det &\rightarrow "The" \\N &\rightarrow "cat" \mid "mouse" \\V &\rightarrow "chased"\end{aligned}$$



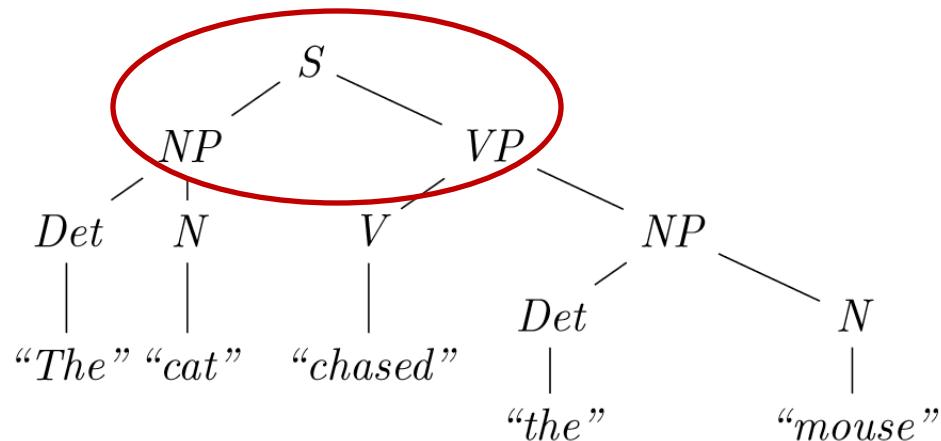
## 1.3.1 传统基础技术：句法分析

例

分析句子 “The cat chased the mouse.（猫追逐老鼠）”。使用 CFG 来表示句子的语法结构，并生成相应的句法树。

- 从开始符号S出发，应用规则“ $S \rightarrow NP VP$ ”

$S \rightarrow NP VP$



$S \rightarrow NP VP$   
 $NP \rightarrow Det N$   
 $VP \rightarrow V NP$   
 $Det \rightarrow "The"$   
 $N \rightarrow "cat" | "mouse"$   
 $V \rightarrow "chased"$



## 1.3.1 传统基础技术：句法分析

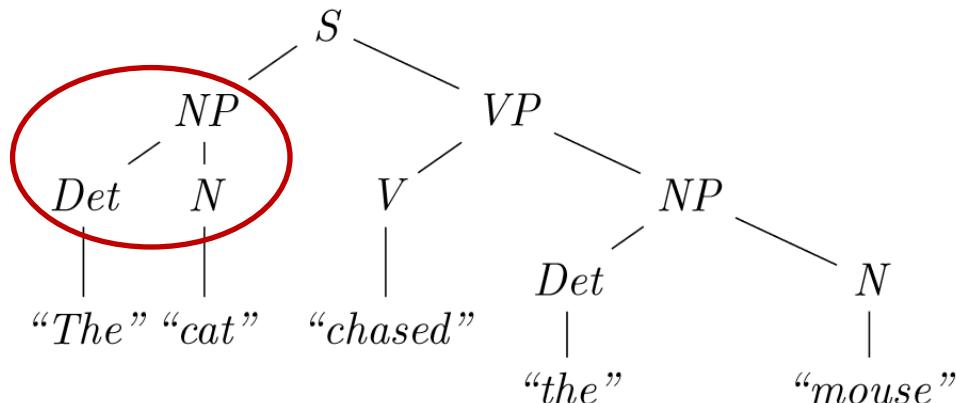
例

分析句子 “The cat chased the mouse.（猫追逐老鼠）”。使用 CFG 来表示句子的语法结构，并生成相应的句法树。

- 继续应用规则

$S \rightarrow NP\ VP$   $\rightarrow Det\ N\ VP$

$S \rightarrow NP\ VP$   
 $NP \rightarrow Det\ N$   
 $VP \rightarrow V\ NP$   
 $Det \rightarrow "The"$   
 $N \rightarrow "cat" \mid "mouse"$   
 $V \rightarrow "chased"$





## 1.3.1 传统基础技术：句法分析

例

分析句子 “The cat chased the mouse.（猫追逐老鼠）”。使用 CFG 来表示句子的语法结构，并生成相应的句法树。

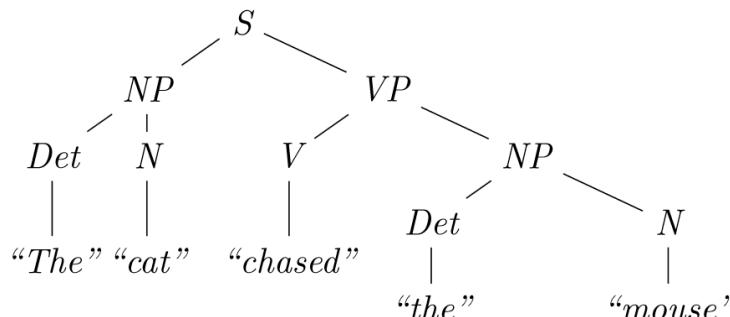
$S \rightarrow NP\ VP \rightarrow Det\ N\ VP \rightarrow Det\ N\ V\ NP \rightarrow Det\ N\ V\ Det\ N$

$\rightarrow "the"\ N\ V\ Det\ N \rightarrow "the"\ "cat"\ V\ Det\ N$

$\rightarrow "the"\ "cat"\ "chased"\ Det\ N$

$\rightarrow "the"\ "cat"\ "chased"\ "the"\ N$

$\rightarrow "the"\ "cat"\ "chased"\ "the"\ "mouse"$





## 1.3.1 传统基础技术：句法分析

2. 结构歧义问题：结构歧义是指一个句子可能有不同的语法解析或句法结构，产生不同的分析结果。

I saw the man with the telescope: 附着歧义

old men and women: 并列歧义

- I saw the **man** (who had) **with the telescope**.
- I **saw** the man (using) **with the telescope**.

- **old** (men and women).
- **(old men)** and women.



## 1.3.1 传统基础技术：句法分析

**3. 句法分析方法：CKY (Cocke-Kasami-Younger) 算法：**基于动态规划的思想，用一个二维表格来存储可能的短语结构，最后生成句法树。

- 对乔姆斯基文法进行范式化，产生式符合以下形式：

$$A \rightarrow w \text{ 或 } A \rightarrow BC$$

$$A, B, C \in N, w \in \Sigma, G = (N, \Sigma, P, S)$$



## 1.3.1 传统基础技术：句法分析

例

采用CKY算法分析句子 “The cat chased the mouse. (猫追逐老鼠) ”。

输入句子: The cat chased the mouse.



词性标注: Det N V Det N.

$S \rightarrow NP VP$   
 $NP \rightarrow Det N$   
 $VP \rightarrow V NP$   
 $Det \rightarrow "The"$   
 $N \rightarrow "cat" | "mouse"$   
 $V \rightarrow "chased"$

0					
	The				
		cat			
			chased		
				the	
					mouse

0	Det				
	The	N			
		cat	V		
			chased	Det	
				the	N
					mouse

0	Det	NP			S
	The	N			
		cat	V		VP
			chased	Det	NP
				the	N
					mouse

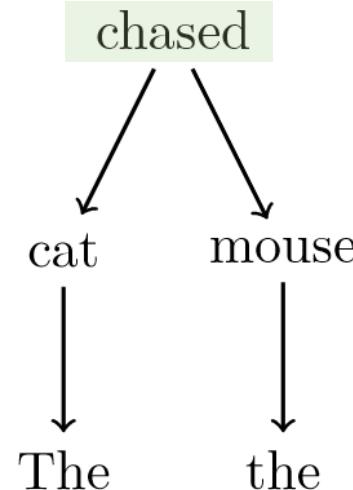


## 1.3.1 传统基础技术：句法分析

- **依存句法分析：**旨在分析句子中词汇之间的依存关系，句法结构完全通过单词或词语之间的有向二元语法关系（依赖词→中心词）来描述。

The cat chased the mouse.

带有方向的弧来表示两个成分之间的依存关系，支配者在有向弧的发出端，被支配者在箭头端，通常说被支配者依存于支配者。



依存树



## 1.3.1 传统基础技术：句法分析

### 3. 依存句法分析方法：

- **决策式依存句法分析**：基于移进-归约 (shift-reduce) 算法。该算法使用分析栈和输入缓冲区来管理分析过程，并定义一组预定义的转移操作：移动 (Shift)、左弧 (Left-Arc) 和右弧 (Right-Arc) 等。
- **基于图的依存句法分析**：基于图的分析器构建一个图结构来表示句子中的依存关系，它对长句子的解析更加准确。分析器在句子的所有可能树结构中寻找一个（或多个）得分最高的树结构，通过将搜索空间编码为有向图，采用图论的方法来寻找空间中的最优解。



## 1.3.1 传统基础技术：句法分析

例

采用决策式依存句法分析方法分析句子 “The cat chased the mouse.  
(猫追逐老鼠)”。

编号	操作	分析栈	输入缓冲区	关系集
0	初始化	[]	[The cat chased the mouse]	{ }
1	Shift	[ The ]	[cat chased the mouse]	{ }
2	Left-Arc	[]	[cat chased the mouse]	{The ← cat}
3	Shift	[ cat ]	[chased the mouse]	{The ← cat}
4	Left-Arc	[]	[chased the mouse]	{The ← cat, cat ← chased}
5	Shift	[ chased ]	[ the mouse]	{The ← cat, cat ← chased}
6	Shift	[ chased the ]	[ mouse]	{The ← cat, cat ← chased}
7	Left-Arc	[ chased ]	[ mouse]	{The ← cat, cat ← chased, the ← mouse}
8	Right-Arc	[ chased mouse]	[]	{The ← cat, cat ← chased, the ← mouse, mouse ← chased}
9	Reduce	[ chased]	[]	{The ← cat, cat ← chased, the ← mouse, mouse ← chased}
10	Reduce	[]	[]	{The ← cat, cat ← chased, the ← mouse, mouse ← chased}



## 1.3.1 传统基础技术：句法分析

- 两种句法分析的关系和区别

	表示方式	表示	目标	应用
短语结构分析	树状结构	嵌套和组合关系	句法关系	语法分析、句法分析、句法树生
依存句法分析	有向图	直接依存关系	依存关系	关系抽取、语义分析、机器翻译



## 1.3.1 传统基础技术：句法分析



### 扩展 句法分析方法（大语言模型—提示学习法）



You

对这句话进行句法分析：“The diligent student completed his assignment before the deadline.”



ChatGPT

以下是对句子 "The diligent student completed his assignment before the deadline." 的简单句法分析：

- "The diligent student" 是主语短语，其中 "The" 是冠词 (DT)， "diligent" 是形容词 (JJ)， "student" 是名词 (NN)。
- "completed" 是动词 (VBD)。
- "his assignment" 是宾语短语，其中 "his" 是形容词性物主代词 (PRP\$)， "assignment" 是名词 (NN)。
- "before the deadline" 是时间状语从句，其中 "before" 是介词 (IN)， "the" 是冠词 (DT)， "deadline" 是名词 (NN)。

整个句子是一个简单的陈述句，主语是 "The diligent student"，谓语是 "completed"，而宾语是 "his assignment"。时间状语从句 "before the deadline" 提供了更多关于动作完成时间的信息。



Regenerate



## 1.3.1 传统基础技术：语义分析

- **语义分析**: 对句子所表达的语义进行分析和解释，让计算机理解文本所传达的真正含义、推断用户的意图、理解文本中实体之间的关系。
- **语义表示**: 用形式化的结构表示自然语言表达的含义

意义表示语言：一阶逻辑 (First-Order Logic, FOL)

常量：指代某个特定对象，如Car、Restaurant

函数：所属格概念，如 “Chengdu’ s location” 表示为  
LocationOf(Chengdu)

变量：表示特定未知事物，或某个分类中的一切事物

示例：“I like apple and I don’t like pear.”

Like(Speaker,Apple)  $\wedge$   $\neg$ Like(Speaker,Pear)



## 1.3.1 传统基础技术：语义分析

- **词义消歧：**为单词选择符合上下文语境的正确词义。

词义消歧算法：

### 1. 简单最近邻算法

计算待消歧的单词和词汇库中的每个已知含义单词的相似性得分。通过得分比较，选择最佳的词义，完成词义消歧。

### 2. 基于特征的 WSD 算法

将待消歧的单词放置在上下文中，并提取与该单词相关的特征，然后通过训练一个分类器（如SVM分类器）来确定最适合的词义。

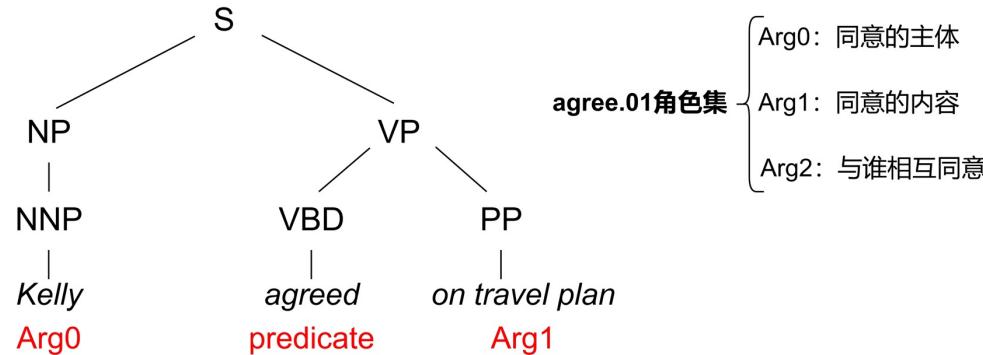
### 3. Lesk 算法

计算目标单词的字典注释与邻近单词的交集，并选交集最大的词义作为目标单词的正确词义。



## 1.3.1 传统基础技术：语义分析

- **语义角色标注：**是指给句子的成分分配语义角色标签的任务。以谓词为核心，分析句子中的其他成分与谓词之间的关系。





## 1.3.1 传统基础技术：语义分析

### 💡 扩展 语义分析方法（大语言模型—提示学习法）

F

You

为这句话生成一阶逻辑语言：The boy is playing with a ball in the park.



ChatGPT

将句子 "The boy is playing with a ball in the park." 转化为一阶逻辑语言可以表示为：

- Playing(Boy, Ball)  $\wedge$  InPark(Boy)



## 1.3.1传统基础技术：篇章分析

**篇章是由一组有结构、有组织的相邻句子构成的自然语言文本。**

悟空撩衣上前，摸了一把，乃是一根铁柱子，约有斗来粗，二丈有余长。他尽力两手挝过道：“忒粗忒长些！再短细些方可用。”说毕，那宝贝就短了几尺，细了一围。悟空又颠一颠道：“再细些更好！”那宝贝真个又细了几分。悟空十分欢喜，拿出海藏看时，原来两头是两个金箍，中间乃一段乌铁；紧挨箍有镌成的一行字，唤做“如意金箍棒”，重一万三千五百斤。心中暗喜道：“想必这宝贝如人意！”一边走，一边心思口念，手颠着道：“再短细些更妙！”拿出外面，只有二丈长短，碗口粗细。

-----摘自《西游记》



## 1.3.1 传统基础技术：篇章分析

- 篇章分析：衡量文本质量，分析语篇结构，监测语篇的连贯性。
- 语篇连贯性：



关系不连贯：晓晴从上海坐飞机到成都。她每天锻炼一个小时。

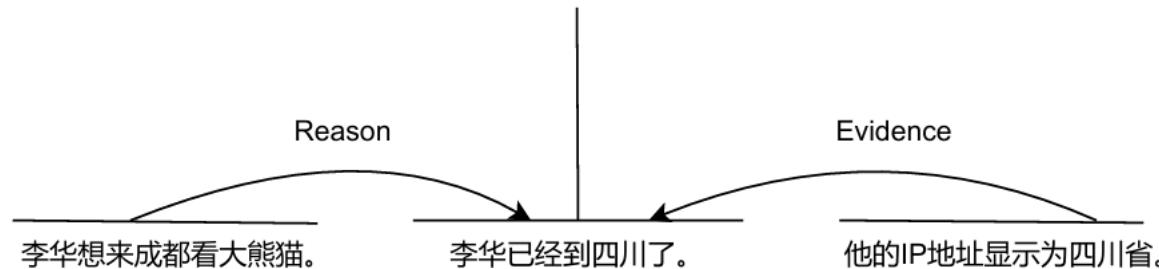
关系连贯：晓晴从上海坐飞机到成都。她要参加一项运动比赛。



## 1.3.1 传统基础技术：篇章分析

- 修辞结构理论：(Rhetorical Structure Theory, RST) 用核心 (Nucleus) 和卫星 (Satellite) 定义两个文本区间之间的关系。

- (i) 李华想来成都看大熊猫。
- (ii) 李华已经到四川了。
- (iii) 他的 IP 地址显示为四川省。





## 1.3.1 传统基础技术：篇章分析

- **指代消解：**就是将文本中指代同一话语实体的用语链接起来，形成一组指代链。



**共指**   示例：鲁迅原名周树人，字豫才。他是浙江绍兴人。  
“他”和“鲁迅”指代同一个话语实体“鲁迅”。

示例：Mary has a dog. It is very friendly.



示例：If you want some, there's coffee in the pot.





## 1.3.1 传统基础技术：篇章分析

- 指代消解需要进行指代识别，即找出句子中所有的指代。然后进行指代消解，确定这些指代与其先行语之间的关系，从而实现对文本的深层理解。

1. 指代识别
2. 上下文建模
3. 特征提取
4. 指代消解决策
5. 生成或更新指代关系



## 1.3.1 传统基础技术：篇章分析

### 💡 扩展 篇章分析方法（大语言模型—提示学习法）

为这句话生成RST关系树：“环保是我们每个人的责任，保护环境就是保护未来。减少废物、节约能源、推广可再生能源是我们应该积极采取的行动。只有共同努力，才能保护地球的美丽，让我们的后代享有一个可持续发展的世界。”



ChatGPT

对于给定的文本，以下是可能的RST (Rhetorical Structure Theory) 关系树的一种表示方式：

```
SCSS
Copy code

ROOT (主要论点): 环保是我们每个人的责任，保护环境就是保护未来。
|
| -- Cause (原因): 减少废物、节约能源、推广可再生能源是我们应该积极采取的行动。
|   |
|   | -- Explanation (解释): 这些行动有助于环境保护，从而保护未来。
|
| -- Condition (条件): 只有共同努力，才能保护地球的美丽，让我们的后代享有一个可持续发展的世界。
|
|   |
|   | -- Elaboration (详细说明): 共同努力是实现环境保护和可持续发展的条件。
|
| -- Elaboration (详细说明): 保护环境就是保护未来。
```



## 1.3.2 实际应用：机器翻译

- **机器翻译 (Machine Translation, MT)** : 机器翻译是利用计算机将一种语言的文本翻译成另一种语言的技术



- **语言差异:** SVO (中、英、法) 、 SOV (日语) VSO (阿拉伯语)
- **编码器-解码器网络:**  
编码器接受输入序列并创建其上下文  $h$ 。然后，将该上下文  $h$  传递给解码器，解码器生成目标语言的输出序列。
- **结果测评：准确性、流畅性**  
基于字符或词汇重叠、基于嵌入相似性



## 1.3.2 实际应用：机器翻译

### ● 机器翻译方法（大语言模型—提示学习法）

F

You

将这句话翻译成法文：

"It is a truth universally acknowledged that a single man in possession of a good fortune must be in want of a wife."



ChatGPT

La phrase peut être traduite en français comme suit :

"Il est universellement reconnu qu'un homme célibataire en possession d'une bonne fortune doit être en quête d'une femme."

Cette phrase est l'ouverture célèbre du roman "Orgueil et Préjugés" de Jane Austen.



## 1.3.2 实际应用：文本分类与情感分类

- **文本分类 (text classification)**：旨在将给定的文本分类为预定义的类别中的一个，常见类别包括情感、主题、语言和其他自定义类别。文本分类可形式化表示为如下数学函数：

$$f(\text{text}) \rightarrow \text{category}$$

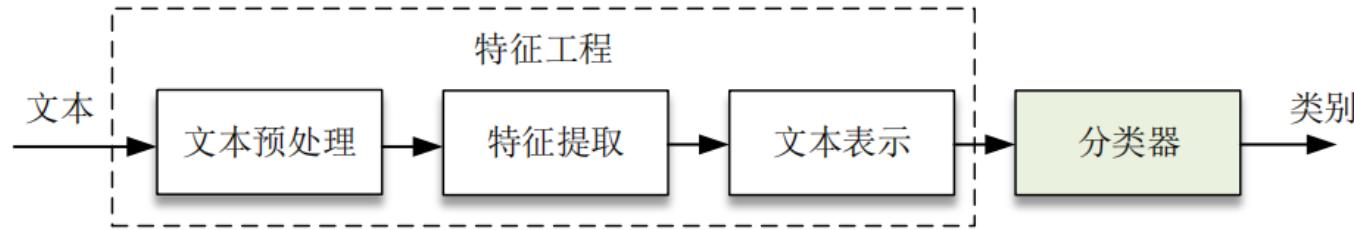
其中， $f$ 表示分类函数，接受文本作为输入， $\text{text}$ 是待分类的文本数据， $\text{category}$ 是将文本被分类到的一个或多个类别标签。

- **朴素贝叶斯分类 (Naive Bayes Classifiers)**：给定一个未标注的文本，朴素贝叶斯方法计算该文本属于每个类别的后验概率，并选择具有最高概率的类别作为分类结果。



## 1.3.2 实际应用：文本分类与情感分类

- **文本分类方法：**经典的文本分类方法包括基于**特征工程**和**浅层学习**。在不考虑训练集的情况下，一个文本分类系统可以被拆分为特征工程和分类器两个主要部分。系统流程如下所示：



其中，文本表示的任务是将非结构化信息转化为计算机可以理解的结构化信息。



## 1.3.2 实际应用：文本分类与情感分类

- **情感分析：**情感分析是一种常见的文本分类任务，即分析文本表达的情感倾向，是积极还是消极的态度。举例以大模型进行情感分析：



You

请判断这段影评的所表达的情感是积极还是消极的：“这部电影剧情拖沓，角色发展薄弱。配乐平庸，缺乏情绪的引导力。整个电影给人一种匆忙制作的感觉，缺乏精心策划和执行。”



ChatGPT

这段影评表达的情感是消极的。评论者对电影的剧情、角色发展、配乐和制作质量都提出了负面的评价，指出了这些方面的问题和不足之处。

- **其他分类任务：**垃圾邮件过滤、新闻分类、法律文书归档、医学文献识别以及在线广告定位等。

## 1.3.2 实际应用：信息抽取与自动文摘

- **信息抽取 (Information Extraction)** : 从非结构化的文本数据中抽取特定的实体、关系和事件等重要信息，并将其转化为结构化的形式（如数据库）

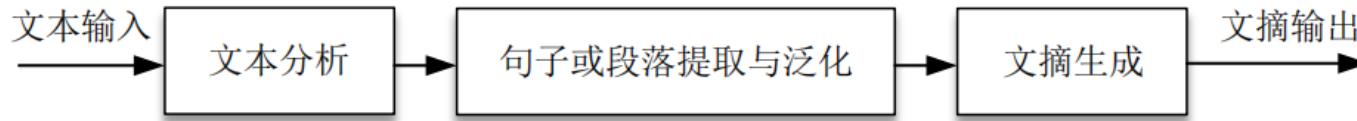


- **关系抽取 (Relation Extraction)** : 识别文本中实体之间的关系或关联，如雇佣关系。
- **事件抽取 (Event Extraction)** : 识别文本中的事件、触发词、参与者、时间等信息。



## 1.3.2 实际应用：信息抽取与自动文摘

- **自动文摘技术：**自动文摘技术是自然语言处理领域中的一个重要研究方向，其核心目标是通过算法和技术，从大量文本中提取关键信息和主要观点，生成简洁、凝练且信息丰富的摘要。其系统流程如下图所示：



- ✓ 文本分析：分词、词性标注、命名实体识别；
- ✓ 句子或段落提取与泛化：代表性句子、段落或信息块



## 1.3.2 实际应用：信息抽取与自动文摘

### ● 信息抽取方法（大语言模型—提示学习法）

F

You

请对这段话进行事件抽取：“昨天发生了一场车祸。一辆黑色轿车与一辆红色卡车在十字路口发生了碰撞。事故发生后，紧急救护车赶到现场，将轿车司机送往医院接受治疗。交通警察到场调查事故原因，并对现场进行了交通管制。”



ChatGPT

以下是对该段话进行事件抽取的结果：

1. 事件：车祸发生
  - 时间：昨天
  - 描述：一辆黑色轿车与一辆红色卡车在十字路口发生了碰撞。
2. 事件：紧急救护
  - 描述：事故发生后，紧急救护车赶到现场。
3. 事件：送医治疗
  - 描述：紧急救护车将轿车司机送往医院接受治疗。
4. 事件：交通警察调查
  - 描述：交通警察到场调查事故原因。
5. 事件：交通管制
  - 描述：交通警察对现场进行了交通管制。



Regenerate



## 1.3.2 实际应用：信息检索与问答系统

- **信息检索 (Information Retrieval, IR)** : 信息检索是一门研究如何从大量文本集合中有效获取与用户查询相关信息的技术。





## 1.3.2 实际应用：信息检索与问答系统

### ● 关键技术





## 1.3.2 实际应用：信息检索与问答系统

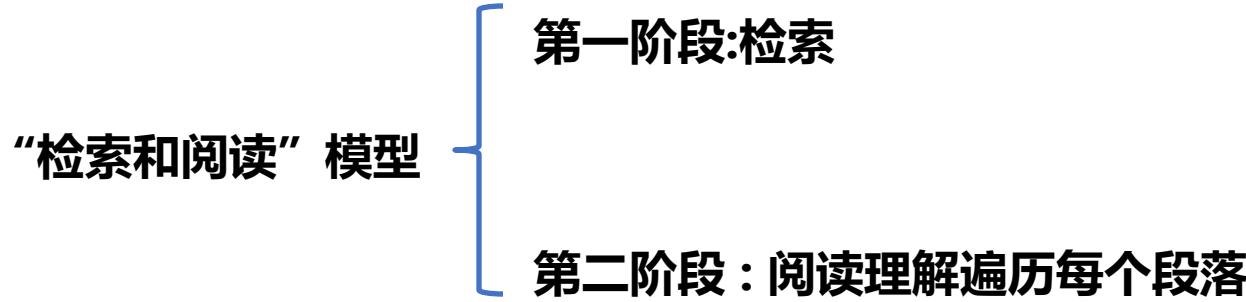
- **问答系统 (Question Answering System) :** 问答系统是一种旨在自动回答用户提出的自然语言问题人工智能应用。主要以事实型问答为主。





## 1.3.2 实际应用：信息检索与问答系统

- 基于信息检索的问答系统



- 基于生成的问答系统

不仅从已知信息中检索答案，还能根据问题的语义和上下文生成全新的答案。



## 1.3.2 实际应用：信息检索与问答系统

- **评估：**平均倒数排名 (mean reciprocal rank, MRR) , 其计算公式如下:

$$MRR = \frac{1}{Q} \sum_{i=1}^Q \frac{1}{rank_i}$$

其中, Q是测试集中问题的总数, rank<sub>i</sub>是第i个问题的正确答案的排名。

- **问答系统 (大语言模型—提示学习法)**

F You

请问《红楼梦》的作者是谁



ChatGPT

《红楼梦》的作者是曹雪芹。这部小说是中国文学史上的经典之一，也被认为是中国古典小说中的巅峰之作。



## 1.3.2 实际应用：对话系统

- **对话系统 (Dialogue Systems)** : 对话系统旨在模拟人类的对话行为并与用户进行自然语言交流。

嘿，你好，我想订一张明天飞往上海的机票。

好的，已为您找到9月15日成都到上海的航班。

有上午九点的吗？

航班号MU5406，明天早上九点起飞。

什么时候到上海？

抵达上海的时间是十一点45分

那帮我订这个航班。

好的，已为您打开XX软件进行预定，请您稍后



## 1.3.2 实际应用：对话系统

### ● 对话系统分类





## 1.3.2 实际应用：对话系统

### ● 对话系统（大语言模型—提示学习法）

F

You

我想从昆明去成都游玩。有什么推荐的交通方式吗？



ChatGPT

从昆明到成都有几种不同的交通方式供选择，具体取决于你的预算和时间安排。以下是一些常见的交通方式：

1. 飞机：最快捷的方式是搭乘飞机。你可以从昆明长水国际机场飞往成都双流国际机场。这是最快的选择，飞行时间通常在2小时左右。
2. 火车：如果你更喜欢悠闲的旅行体验，可以考虑搭乘高铁或普通火车。有多趟列车从昆明到成都，旅行时间因列车类型而异，通常在8到12小时之间。
3. 长途汽车：长途汽车也是一个选项，但旅行时间较长，通常在20小时以上。这适合那些预算有限且有耐心的旅行者。
4. 自驾车：如果你有自驾车的能力和许可，也可以考虑自驾前往成都。这将提供更大的灵活性，但需要更多的时间和计划。

现在轮到你了！你是否已经决定了前往成都的交通方式，或者还有其他问题需要我回答？



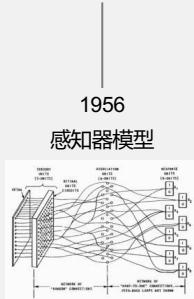
# 1.4 NLP和大模型发展现状

前神经网络 专家系统 概率推论

神经网络 基础模型 大规模模型

2011年之前

## 决策式/分析式 NLP 系统



1956

感知器模型

1980s  
Prolog &  
Lisp

专家系统  
解答特定领域  
的问题



1990s

机器学习算法



AI学科成立

Robot

机器语言翻译



Google  
Translate

1994  
GroupLens  
首个自动化  
推荐系统

amazon

ByteDance  
字节跳动

推荐算法

amazon

ByteDance  
字节跳动

2011  
大数据  
IBM沃森  
益智问答  
打败人类



2011

IBM沃森  
益智问答  
打败人类



2012

DNNs  
物体识别  
自动驾驶



2014

GAN  
生成图像但  
分辨率不高



2017

Transformer  
网络结构进化  
为后续大型模  
型打下基础



2020

GPT-3  
能够完成  
大多数  
NLP任务



2021

CLIP  
以40亿个文本-图像对  
为训练数据实现文本  
到图像跨模态

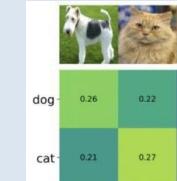


2022

DALL-E2  
由文本生成  
的高清具有  
独创性图像

2012年至今

## 生成式 NLP 时代

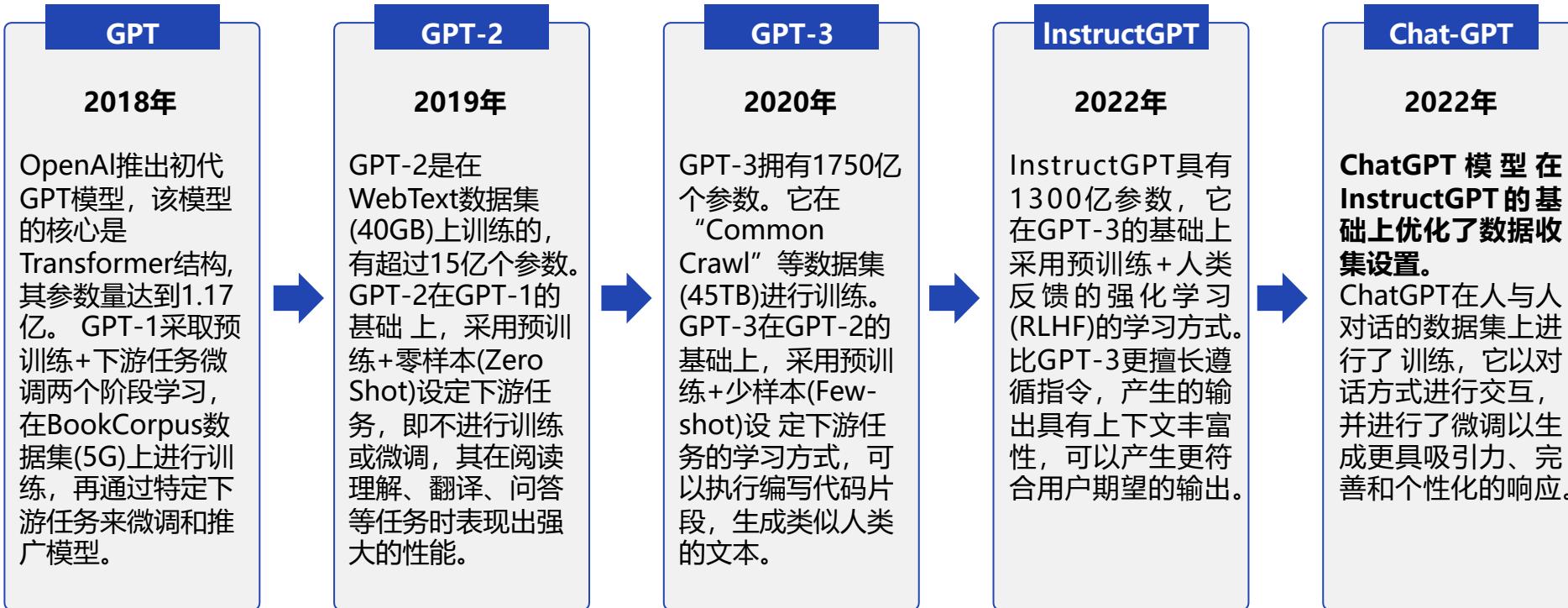


AI的急速发展使得AI模型在手写识别、语音识别、  
图像识别、阅读理解、语言理解等领域的表现加速  
超越人类平均水平

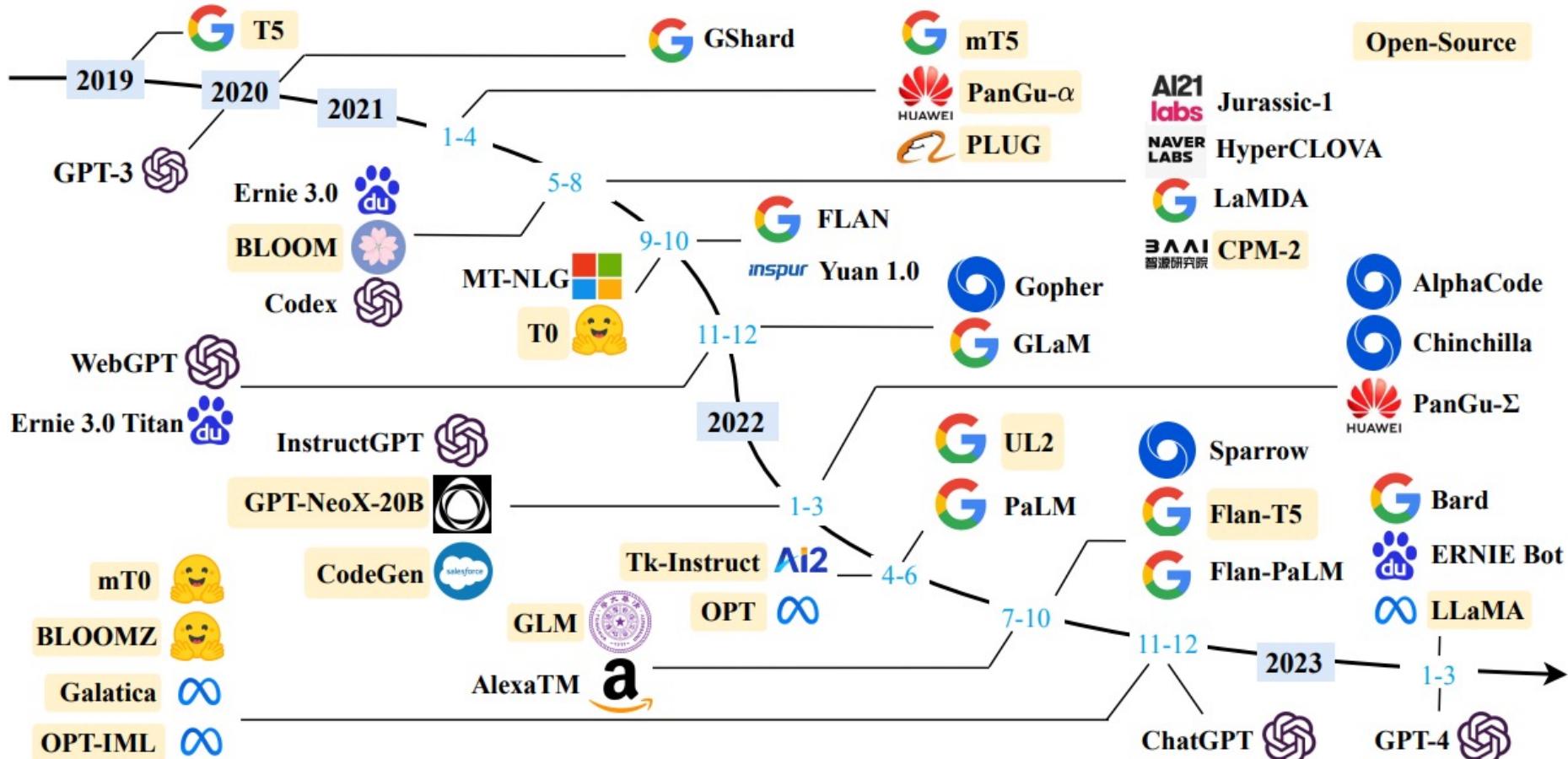


# 1.4 NLP和大模型发展现状

- ChatGPT在海量的文本数据上进行了预训练，可以对自然语言输入产生类似人类的回答。
- ChatGPT其核心是建立在GPT模型（Generative Pre-trained Transformer）的发展基础之上。



# 1.4 NLP和大模型发展现状





# 1.4 NLP和大模型发展现状

表 6.1: 大规模文本预训练模型对比表

发布者	模型	参数量	输入长度限制	访问方式	微调方式						
OpenAI	davinci	175B?	2048	API	None	Meta	OPT	175B	2048	开源	None
	text-davinci-001	175B?	2048	API	指令微调		OPT-IML	175B	2048	开源	指令微调
	code-davinci-002	175B?	8000	API			Galactica	120B	2048	开源	指令微调
	text-davinci-002	175B?	4000	API	代码微调 + 指令微调		LLaMA	65B	8192	开源	
	text-davinci-003	175B?	4000	API	代码微调 + 指令微调 + RLHF	微软/英伟达	TNLG v2	530B	2048	未开放	
	text-curie-001	6.7B	2048	API		Eleuther	GPT-J	6B	2048	开源	
	text-babbage-001	1.3B	2048	API			GPT-NeoX	20B	2048	开源	
	text-ada-001	350M	2048	API							
Google	LaMDA	137B		未开放	对话微调	Cohere	xlarge	52.4B	2048	API	
	Flan-LaMDA	137B		未开放	指令微调		large v20220720	13.1B	2048	API	
	T5	11B	512	开源			medium v20220720	6.1B	2048	API	
	UL2	20B	512	开源			small v20220720	410M	2048	API	
	Flan-T5	11B	512	开源	指令微调	AnthropicAI	Claude	52B	8192	未开放	
	PaLM	540B		未开放							
	U-PaLM	540B		未开放							
	Flan-PaLM	540B		未开放	指令微调	清华大学	GLM-130B	130B	2048	开源	指令微调
	Flan-U-PaLM	540B		未开放	指令微调						
DeepMind	Sparrow			未开放	指令微调 + RLHF	华盛顿大学	J1-Jumbo v1	178B	未知	API	
	Chinchilla	70B		未开放			J1-Grande v1	17B	未知	API	
	Gopher	280B	2048	未开放		AI21 Labs	J1-Large v1	7.5B	未知	API	



# 1.4 NLP和大模型发展现状

Model	Release Time	Size (B)	Base Model	Adaptation IT	Adaptation RLHF	Pre-train Data Scale	Latest Data Timestamp	Hardware (GPUs / TPUs)	Training Time	Evaluation ICL	Evaluation CoT
Open Source	T5 [71]	Oct-2019	11	-	-	1T tokens	Apr-2019	1024 TPU v3	-	✓	-
	mT5 [72]	Mar-2021	13	-	-	1T tokens	Apr-2019	-	-	✓	-
	PanGu- $\alpha$ [73]	May-2021	13*	-	-	1.1TB	-	2048 Ascend 910	-	✓	-
	CPM-2 [74]	May-2021	198	-	-	2.6TB	-	-	-	-	-
	T0 [28]	Oct-2021	11	T5	✓	-	-	512 TPU v3	27 h	✓	-
	GPT-NeoX-20B [75]	Feb-2022	20	-	-	825GB	Dec-2022	96 40G A100	-	✓	-
	CodeGen [76]	Mar-2022	16	-	-	577B tokens	-	-	-	✓	-
	Tk-Instruct [77]	Apr-2022	11	T5	✓	-	-	256 TPU v3	4 h	✓	-
	UL2 [78]	Apr-2022	20	-	✓	-	1T tokens	Apr-2019	512 TPU v4	-	✓
	OPT [79]	May-2022	175	-	-	180B tokens	-	992 80G A100	-	✓	-
	BLOOM [66]	Jul-2022	176	-	-	366B	-	384 80G A100	105 d	✓	-
	GLM [80]	Aug-2022	130	-	-	400B tokens	-	768 40G A100	60 d	✓	-
	Flan-T5 [81]	Oct-2022	11	T5	✓	-	-	-	-	✓	✓
	mT0 [82]	Nov-2022	13	mT5	✓	-	-	-	-	✓	-
	Galactica [35]	Nov-2022	120	-	-	106B tokens	-	-	-	✓	✓
	BLOOMZ [82]	Nov-2022	176	BLOOM	✓	-	-	-	-	✓	-
	OPT-IML [83]	Dec-2022	175	OPT	✓	-	-	128 40G A100	-	✓	✓
	LLaMA [57]	Feb-2023	65	-	-	1.4T tokens	-	2048 80G A100	21 d	✓	-



# 1.4 NLP和大模型发展现状

Model	Release Time	Size (B)	Base Model	Adaptation IT	Adaptation RLHF	Pre-train Data Scale	Latest Data Timestamp	Hardware (GPUs / TPUs)	Training Time	Evaluation ICL	Evaluation CoT
Closed Source	GShard [84]	Jan-2020	600	-	-	1T tokens	-	2048 TPU v3	4 d	-	-
	GPT-3 [55]	May-2020	175	-	-	300B tokens	-	-	-	✓	-
	LaMDA [85]	May-2021	137	-	-	2.81T tokens	-	1024 TPU v3	57.7 d	-	-
	HyperCLOVA [86]	Jun-2021	82	-	-	300B tokens	-	1024 A100	13.4 d	✓	-
	Codex [87]	Jul-2021	12	GPT-3	-	100B tokens	May-2020	-	-	✓	-
	ERNIE 3.0 [88]	Jul-2021	10	-	-	375B tokens	-	384 V100	-	✓	-
	Jurassic-1 [89]	Aug-2021	178	-	-	300B tokens	-	800 GPU	-	✓	-
	FLAN [62]	Oct-2021	137	LaMDA	✓	-	-	128 TPU v3	60 h	✓	-
	MT-NLG [90]	Oct-2021	530	-	-	270B tokens	-	4480 80G A100	-	✓	-
	Yuan 1.0 [91]	Oct-2021	245	-	-	180B tokens	-	2128 GPU	-	✓	-
	WebGPT [70]	Dec-2021	175	GPT-3	-	✓	-	-	-	✓	-
	Gopher [59]	Dec-2021	280	-	-	300B tokens	-	4096 TPU v3	920 h	✓	-
	ERNIE 3.0 Titan [92]	Dec-2021	260	-	-	300B tokens	-	2048 V100	28 d	✓	-
	GLaM [93]	Dec-2021	1200	-	-	280B tokens	-	1024 TPU v4	574 h	✓	-
	InstructGPT [61]	Jan-2022	175	GPT-3	✓	✓	-	-	-	✓	-
	AlphaCode [94]	Feb-2022	41	-	-	967B tokens	Jul-2021	-	-	-	-
	Chinchilla [34]	Mar-2022	70	-	-	1.4T tokens	-	-	-	✓	-
	PaLM [56]	Apr-2022	540	-	-	780B tokens	-	6144 TPU v4	-	✓	✓
	AlexaTM [95]	Aug-2022	20	-	-	1.3T tokens	-	128 A100	120 d	✓	✓
	Sparrow [96]	Sep-2022	70	-	-	✓	-	64 TPU v3	-	✓	-
	U-PaLM [97]	Oct-2022	540	PaLM	-	-	-	512 TPU v4	5 d	✓	✓
	Flan-PaLM [81]	Oct-2022	540	PaLM	✓	-	-	512 TPU v4	37 h	✓	✓
	Flan-U-PaLM [81]	Oct-2022	540	U-PaLM	✓	-	-	-	-	✓	✓
	GPT-4 [46]	Mar-2023	-	-	✓	✓	-	-	-	✓	✓
	PanGu- $\Sigma$ [98]	Mar-2023	1085	PanGu- $\alpha$	-	329B tokens	-	512 Ascend 910	100 d	✓	-



# 1.4 NLP和大模型发展现状

- **人工智能能够赋能经济社会发展各领域，下游各个领域的产业升级对大模型的需求持续走高。**例如，在泛交通领域，人工智能能够在自动驾驶、智能座舱、运行管理优化等多个方面为汽车行业创造价值；在医疗方面，人工智能可以提高疾病检测的效率以及医学影像领域的智能化分析。据测算，2022年中国人工智能行业市场规模为3,716亿人民币，预计2027年将达到15,372亿人民币，有望在下游制造、交通、金融、医疗等多领域不断渗透，实现大规模落地应用。
- 下游行业对人工智能需求呈现出碎片化、多样化的特点，从开发、精调、优化、迭代到应用都需要投入巨大的人力和财力，成本极高。而**大模型能够向外赋能，包括通过开放API的形式，降低AI应用开发门槛，提高落地部署效率和精度等，进而降低AI规模化部署的成本，满足各行业场景的应用需求，进一步推动人工智能进入工业化发展阶段。**





0

感谢聆听  
THANKS