## 简介

自然语言处理（Natural Language Processing，简称NLP）可分成“自然语言”和“处理”两部分来理解。

自然语言：它是人类发展过程中形成的一种信息交流的方式，包括口语及书面语，反映了人类的思维。现在世界上所有的语种语言，都属于自然语言，包括汉语、英语、法语等。

处理：如果只是人工处理的话，那我们原本就有专门的语言学来研究，也没必要特地强调“自然”。因此，这个“处理”必须是计算机处理的。但计算机毕竟不是人，无法像人一样处理文本，需要有自己的处理方式。

因此自然语言处理，简单来说即是**计算机接受用户自然语言形式的输入，并在内部通过人类所定义的算法进行加工、计算等系列操作，以模拟人类对自然语言的理解，并返回用户所期望的结果**。它是人工智能、计算机科学、信息工程的交叉领域，涉及统计学、语言学等的知识。

自然语言处理的目的在于**用计算机代替人工来处理大规模的自然语言信息，以此解放人类的双手**。由于语言是人类思维的证明，故自然语言处理是人工智能的最高境界，被誉为“人工智能皇冠上的明珠”

## NLP与深度学习

**Deep NLP = Deep Learning + NLP**

将自然语言处理的思想与表示学习结合起来，用深度学习的手法解决NLP目标。

1. 层次：语音、词汇、语法、语义等；
2. 工具：词性标注、命名实体识别、句法/语义分析等（技术）；
3. 应用：机器翻译、情感分析、客服系统、问答系统等（任务）。

深度学习的一个魅力之处是，它提供了一套“宇宙通用”的框架解决了各种问题。虽然工具就那么几个，但在各行各业都适用。

## NLP的层次分解

NLP的层次分解见图1：

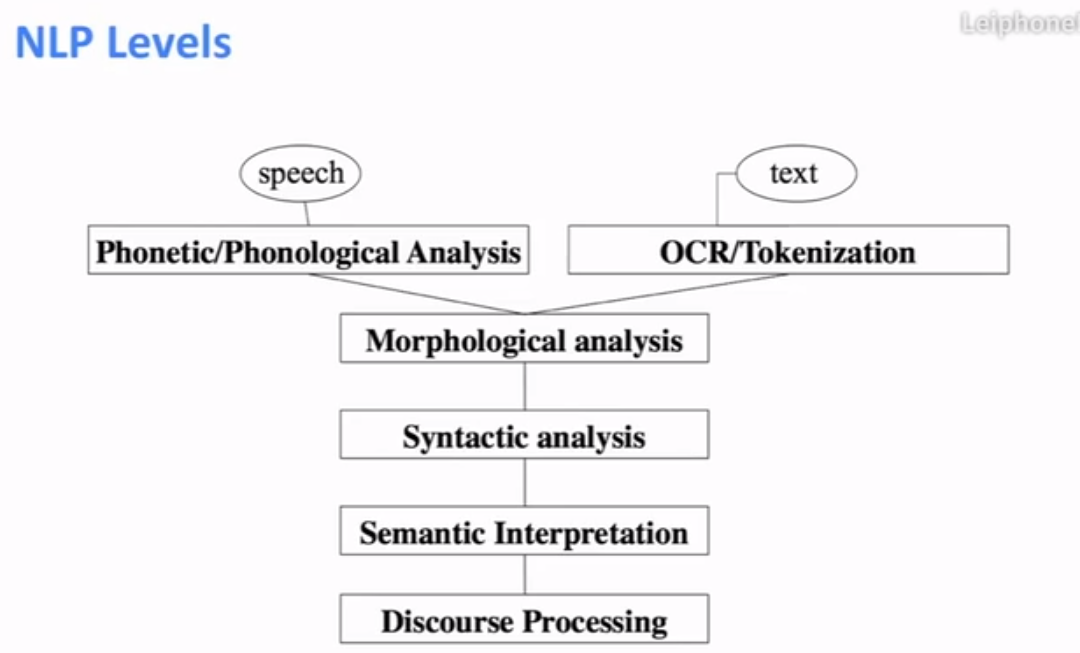


图1：NLP的层次

其中：

Speech：语音；

Phonetic/Phonological Analysis ：语音的/语音学分析；

Text：文字；

OCR/Tokenization ：光学字符识别/词语切分；

Morphological Analysis：词态学分析；

Syntactic Analysis：句法分析；

Semantic Interpretation：语义解释；

Discourse Processing：篇章分析。

## NLP的内容和技术

NLP的内容和技术见图2：

|  |  |  |
| --- | --- | --- |
| 内容：四大任务 | | 技术 |
| 词 | 词性标注 | 分词/POS Tag/NER/语义标注...... |
| 句 | 句子关系判断 | Entailment/QA/自然语言推理...... |
| 文 | 文本分类 | 文本分类/情感计算...... |
| 生 | 生成式任务 | 机器翻译/文本摘要...... |

图2：NLP的内容和技术

由上可知，NLP的内容分为词、句、文、生四大任务，蕴含“道生一，一生二，二生三，三生万物”的道理。

## NLP的技术路线

现在NLP的技术路线就是预训练+微调，见图3：

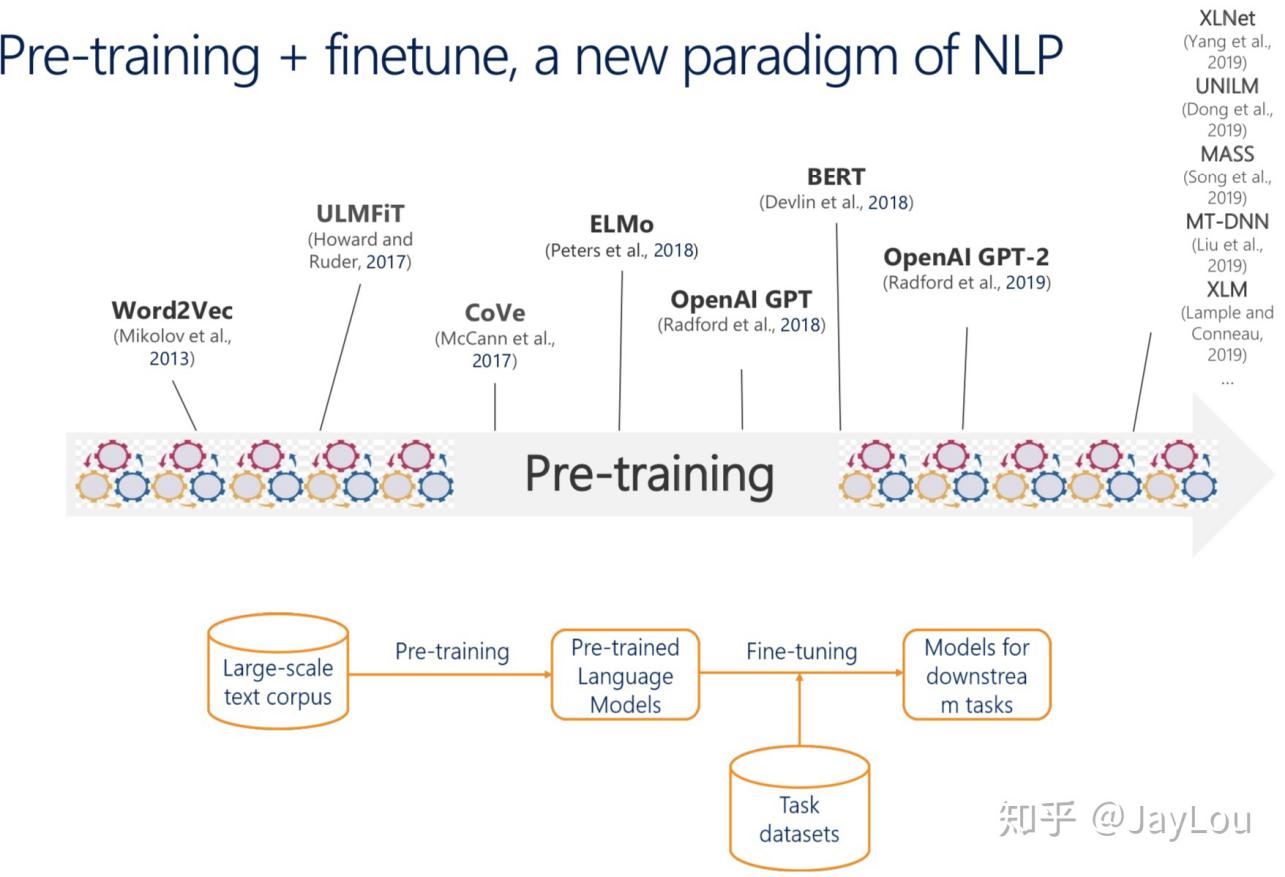


图3：NLP的技术路线

其中，按照训练方法和数据类型的不同，各种模型的分类见图4：

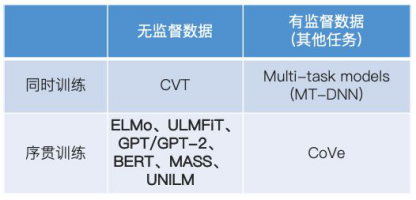


图4：NLP模型的分类

## NLP的技术演变

NLP的技术框架和模型演变逻辑如图5：

CNN

RNN

图5：NLP的技术框架和模型演变逻辑

XLM、MT-DNN

MASS、WWM

SpBERT、RoBERT

ALBERT、K-BERT

ERNIE、ERNIE2

T5、UNILM

Universal Transformer

XLNet

Transformer-XL

Decode

Encode

Transformer

GPT-1

GPT-2

BERT

Attention

(Q、K、V)

Seq2Seq

Encode-Decode

## NLP的技术体系

NLP的技术是一个庞大的体系，它分为以下四个部分：

### 概率&统计

包括5个分支：基础 (Basic) 、采样 (Sampling) 、信息理论 (Information Theory) 、模型 (Model) 、贝叶斯 (Bayesian) ，见图6：

图6：NLP的技术体系-概率统计部分

### 机器学习

包括7个分支：线性回归 (Linear Regression) 、逻辑回归 (Logistic Regression) 、正则化 (Regularization) 、非概率 (Non-Probabilistic) 、聚类 (Clustering) 、降维 (Dimensionality Reduction) 、训练 (Training) ，见图7：

图7：NLP的技术体系-机器学习部分

### 文本挖掘

包括6个分支：基本流程 (Basic Procedure) 、图 (Graph) 、文档 (Document) 、词嵌入 (Word Embedding)、序列标注 (Sequential Labeling) 、NLP基本假设 (NLP Basic Hypothesis) ，见图8：

图8：NLP的技术体系-文本挖掘部分

### **深度学习**

包括4个分支（见图9）：

一是基础 (Basic) ，详细梳理了NLP常用的几类网络：循环模型、卷积模型和递归模型；

二是语言模型 (Language Model) ，包含了编码器-解码器模型，以及词表征到上下文表征(Word Representation to Contextual Representation) 这两部分，许多著名模型，比如BERT和XLNet，都是在这里得到了充分拆解，也是需要努力学习的内容；

三是分布式表征 (Distributed Representation) ，许多常用的词嵌入方法都在这里，包括GloVe和Word2Vec。

四是任务 (Task) ，机器翻译、问答、阅读理解、情绪分析……

图9：NLP的技术体系-深度学习部分

