0. 写在前面

1 本课程总体结构

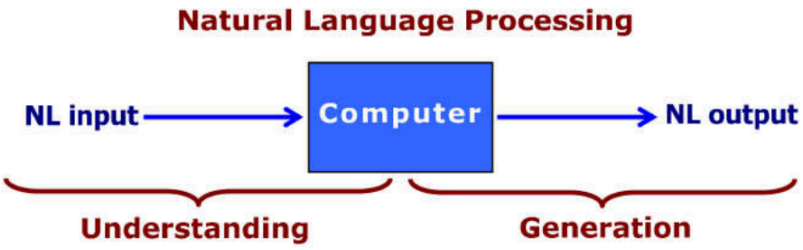
章节		教学内容
第一章 引言 (刘均, 2)		概念与研究背景；主要任务；挑战与研究方向；相关资源
第二章：自然语言的统计特性 (刘均, 1)		Zipf定律、Heaps定律、Benford 定律。
第三章：语言模型	词袋模型 (刘均, 3)	语言模型；词袋模型（BoW）；TF-IDF。 NLU任务：情感分析、文本聚类。
	概率语言模型 (李辰, 6)	概率语言模型；n-gram 模型；最大似然估计；平滑技术。 NLU任务：分词、语义关系抽取。
	主题模型 (刘均, 6)	生成模型；主题模型的图表示；LSA、PLSA、LDA；NMF等。 NLU任务：话题检测、推荐。
	神经网络语言模型 (李辰, 6)	分布式表示；C&W、CBOW、Skip-Gram、Glove等。 NLU任务：对话、实体消歧。
第四章：机器翻译	概述 (李辰, 1)	面临的挑战；发展历程；方法类别及特点；MT评估。
	统计机器翻译 (李辰, 3)	统计MT；Noisy Channel模型；IBM模型。
	神经网络机器翻译 与大语言模型 (刘均, 4)	RNN与LSTM简介；Encoder-Decoder框架；Attention模型；大语言模型。

- 这门课由于由两位老师授课，个人感觉结构比较混乱
- 由于时间紧任务重经费无，所以笔记还是按PPT内容和以上结构展开，即使有很多不合理的地方

2 考试有关事项

1. NLU的概念与背景

1 NLU与NLP



1. 自然语言理解：
    - 含义：让计算机理解人类语言的结构+语义
    - 应用：信息检索/情感识别/机器翻译/拼写检查/知识图谱构建
2. 自然语言处理：对自然语言的分析/理解/生成，即NLU+NLG(Generation)

2 AI-Hard问题

1. 含义：问题等同于AI核心的问题，即如何让计算机具有人类智能
2. 典型：NLU/NLP，CV

3 NLU面临的挑战

1. 计算机的特性：善于处理明确/结构化/无歧义的语言(如编程语言)
2. 自然语言特性：具有复杂的上下文以及歧义性(Ambiguity)

歧义类型	含义	示例
词汇歧义	词汇具有不同含义	Fuck可以是动词/语气词
句法歧义	一个句子被解析成不同的结构	南京市长江大桥
语义歧义	句中包含了不明确的词	John kissed his wife, and so did Sam
回指歧义	之前提到的词，在后面句子含义不同	小李告诉小王**他**生了
语用歧义	短语/句子不同语境下含义不同	可以站起来吗 (询问能力or请求)

2. NLU的语法任务

2.1. 词汇层面的任务

1 词干抽取(Stemming)

1. 含义：抽取词的词干(Stem)与词根(root)，比如Niubilitiness $\rightarrow$ $\begin{cases} \text{词干: Niubility} \\ \text{词根: Niubi} \end{cases}$

2. 处理方法：

方法	含义	限制
利用形态规则	机械地去处所有后缀，如-ing/-tion	不规则变形词不适用
基于词典	按照词典中的映射还原词性	受限于词典规模
高级方法	n-gram法/隐马可夫/机器学习	受限于语料库大小

2️⃣词形还原(Lemmatization)

1. 含义：将不同形式词汇还原为词目(Lemma)，如am, is, are, was, were→be
2. 对比：词干抽取完全不考虑上下文，词形还原考虑一定的上下文
    - 示例：

$$\text{We are meeting in the zoom meeting} \xrightarrow[\text{词形还原}]{\text{词干抽取}} \begin{cases} \text{词干抽取: meet} \\ \text{词形还原: meet+meeting} \end{cases}$$

3. 处理方法：
    - 基于规则：人工给予的语言学规则，或者机器学习训练出来的规则
    - 基于词典：受限于词典，只适用于简单语言

💡词形还原/词干抽取并非100%必要，比如细颗粒度情感分析就需要高精度文本(时态/复数等)

3️⃣词性标注(Part-of-speech tagging)

1. 含义：为文本中每个词标记词性(名词/动词/形容词)
2. 方法：基于规则(人工)，基于隐马可夫模型(HMM)，基于机器学习(SVM/神经网络)
3. 挑战：分词&词义多义性

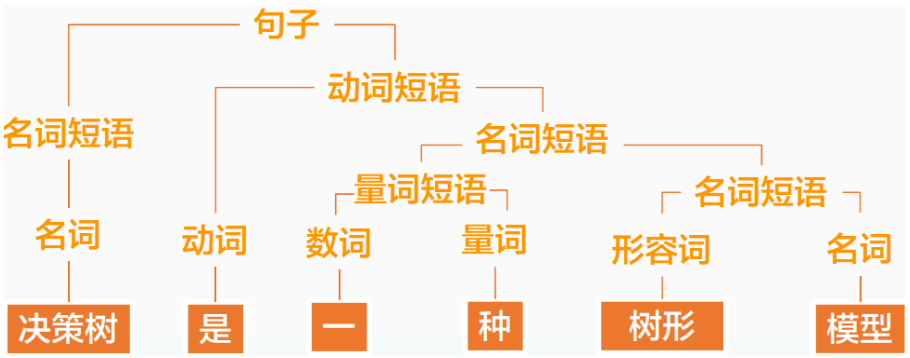
4️⃣术语抽取(Terminology extraction)

1. 含义：信息抽取的子任务，识别文本中特定领域的专门术语
2. 方法：机器学习，统计(TF/IDF)，外部知识库
3. 挑战：新术语/跨领域术语

2.2. 句法层面的任务

1️⃣句法分析(Parsing)

1. 含义：分析文本中单词/短语之间的句法关系



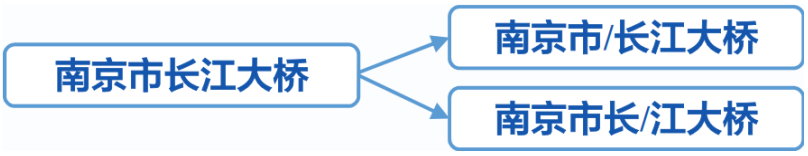
2. 方法：统计学(概率上下文无关文法/最大化信息熵的原则)，机器学习方法(RNN)

2️⃣ 断句(Sentence breaking)

1. 含义：句子边界消歧(例如．可表示短句/缩写/小数点等)+句子分割
2. 方法：基于最大熵，神经网络

3️⃣ 分词(Word segmentation)

1. 含义：仅对词汇间没有明显边界的语言(中文)而言，将连续字符分割为有意义单词



2. 方法：基于字典(正逆向匹配)，基于统计(HMM/SVM)
3. 难处：未登录词/切分歧义

3. NLU的语义任务

3.1. 文本生成/转换

1️⃣ 机器翻译

1. 含义：将文本/语音从一种语言翻译到另一种语言
2. 方法：基于规则(源/目标语言形态语法等)，基于统计(用大型语料库构建概率模型)，神经网络
3. 难题：词句歧义/对语料库大小强依赖/低频词句/长句子

2️⃣ 问答与对话

1. 含义：实现自然语言形式的人机交互
2. 分类：
   - 5W1H类问题：即Who/What/When/Where/Why & How
   - Open/Closed-domain类：回答可以没有限制 or 专注于某一领域
3. 方法：检索法(从库中抽取语料回答)，生成法(检索+推理)，Pipeline，Seq2Seq
4. 难题：多知识约束/多轮对话/多模态/可解释性.....

3️⃣ 自动文摘

1. 含义：为文档生成一段包含原文档要点的**压缩文档**，例如搜索引擎结果

2. 方法：对要点进行不修改的抽取，对要点概括(修改/复述)
3. 难题：难以评估，可理解性问题，需要背景知识

3.2. 文本信息提取

1️⃣命名实体识别(NER, Named entity recognition)

1. 命名实体：现实世界中的某个对象，如
   Obama is the president of the United States
2. NER：信息提取的子任务，识别文本中所有实体+分配到特定类别(名字/时间/数量)
3. 方法：语法规则(效果好但需要大量人工规则)，统计方法(需要标注大量数据)
4. 难题：领域依赖性(如医学实体/术语)，实体类型多样

2️⃣关系抽取

1. 含义：检测文本中实体的语义关系，并将各种关系分类
2. 方法：结合了领域知识的机器学习
3. 难点：训练集难以构建，自然语言的歧义

3.3. 文本内容分析

1️⃣文本分类

1. 含义：自动将文本划分到预定类中，比如垃圾邮件过滤/情感识别/黄色内容识别
2. 方法：特征提取→训练分类器(朴素贝叶斯/KNN/SVM)
3. 难题：特征难提取(需要大量标注)，数据非平衡问题

2️⃣情感分析

1. 含义：识别文本中的情感状态，主观评价等
2. 方法：情感词库(Happy/Fucking)，统计方法(SVM/潜在语义分析)
3. 难题：修辞的多样(反语/讽刺)，分面观点(即将复杂事物分解为不同方面)

3 主题分割

1. 含义：将单个长文本分为多个较短的，主题一致的片段
2. 方法：
   - 基于内容变化：同一主题内容有高度相似性→通过聚类
   - 基于边界特性：主题切换时会有边界(如过渡性/总结性文本)
3. 难点：任务目标模糊(主题多样)，无关信息干扰，歧义性

3.4. 文本歧义消解

1 词义消歧

1. 含义：确定一词多义词的含义
2. 方法：基于词典(叙词表/词汇知识库)，基于机器学习(小语料库的半监督学习/标注后的监督学习)
3. 难题：词义的离散型(一个词的不同含义可能完全不搭边)，需要常识

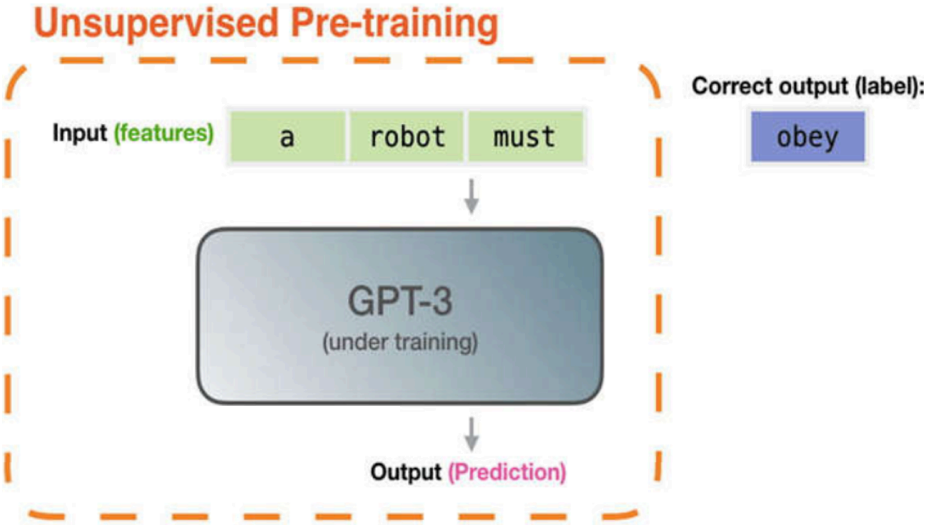
2 共指消解

1. 含义：识别文本中表示同一个事物的不同代称
2. 示例：甲队打败了乙队，他们更强 —消解后→ 虽然甲队打败了乙队，但他们更强
3. 方法：
   - 启发式：如最接近语法兼容词，即在代词前寻找最近的+语法匹配的词
   - 基于ML：如Mention-Pair Models/Mention-Ranking Models
4. 难题：如何应用背景知识，歧义(哪哪都有它，考试的万金油解答嘿嘿)

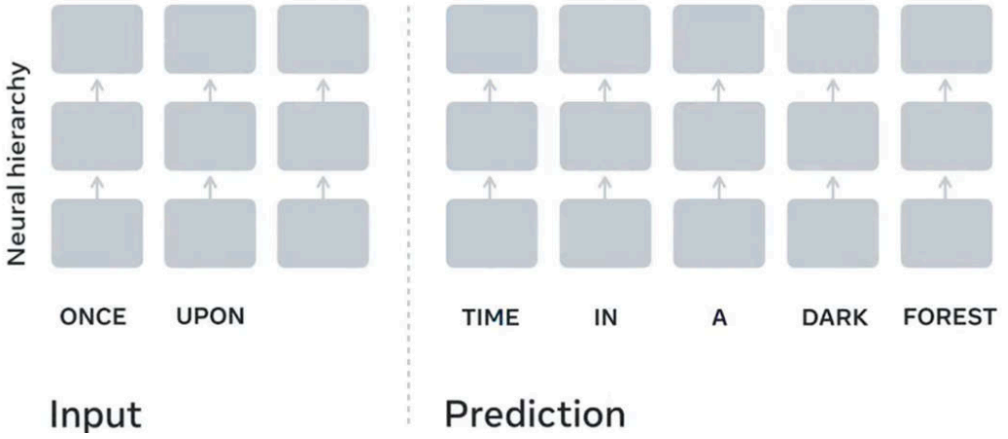
4. 关于大模型的概述

1 大规模预训练模型

1. 含义：用自监督方式在大规模数据集上训练的深度学习模型→后续可通过微调来适应特定任务
2. 原理：以GPT-3为例
   - 训练阶段：以"完形填空"形式训练一个序列预测器



   - 预测阶段：当前文本序列 —逐个预测每个Token→ 生成最有可能的文本序列



2 大模型的特点

特性	描述	备注
大模型	模型参数达到万亿级	模型性能$\propto$参数量$^{\alpha}$ (即标度律)
大数据	模型参数越多$\rightarrow$训练所需数据也更多	模型性能$\propto$数据量$^{\alpha}$
大算力	训练对算力要求高，预测时所有参数都被激活	算力需求$\propto 2^{\text{参数规模}}$ (即指数级)

3 大模型的局限

局限	含义
大模型幻觉	生成错误/不存在的信息，比如说埃菲尔铁塔在德国(其实也没错)
生成404内容	TG不太喜欢的，你懂
理解不了多模态	例如很多时候同时输入文本+图片，GPT往往会忽略图片内容
推理能力差	无法进行逻辑/尝试/数值推理，*但是好像GPT-o1解决了这一问题*
难以个性化	对特定场景/需求/认知水平的处理欠佳，或许可以通过参数调控(?)