NLP（自然语言处理）基础概念

1. NLP任务：理解文本，"思考"并回复

- 中文分词Chinese Word Segmentation，CWS
  - 不同于英文简单空格分词
- 子词切分Subword Segmentation
  - 适用于处理词汇稀疏问题，通过已知的子词来分析新词
  - 方法：Byte Pair Encoding BPE，WordPiece，Unigram，SentencePiece等
- 词性标注Part-of-Speech Tagging，POS Tagging
  - 名词，动词，形容词等
  - 方法：隐马尔可夫模型（Hidden Markov Model，HMM）、条件随机场（Conditional Random Field，CRF）或者基于深度学习的循环神经网络 RNN 和长短时记忆网络 LSTM 等
- 文本分类Text Classification
  - 理解文本含义和上下文，对文本进行分类
  - 方法：深度学习，神经网络
- 实体识别Named Entity Recognition, NER
  - 将文本中的具有特定意义的实体识别出来：人名，地点时间等，帮助理解关键因素
- 关系抽取Relation Extraction
  - 抽取实体之间的关系：因果，拥有，亲属，位置等
- 文本摘要Text Summarization
  - 概括原文
    - 抽取式摘要：从原文中选词句，准确性高，但略显生硬
    - 生成式摘要：理解并概括
      - 基于注意力机制的序列到序列模型（Seq2Seq）
- 机器翻译Machine Translation, MT
  - 转换语言但不丢意思
- 自动问答Automatic Question Answering, QA
  - 理解问题之后根据数据回答问题
    - 需要上述任务的完成

2. 文本表示的发展

- 词向量Vector Space Model, VSM
  - 将文本转换为向量，（维度1，维度2，……）
    - 维度：特征项，字、词、短语
    - 维度值：特征项在文本中的权重
      - 计算公式：词频TF、逆文档频率TF-IDF等
  - 应用：文本相似度计算、文本分类、信息检索等
  - 问题：数据稀疏性问题，维数灾难问题，忽略了文本上下文的结构关系
  - 改进：改进特征表示方法，改进和优化特征项权重的计算方法
- 语言模型N-gram
  - 基于统计，基于马尔可夫假设：一个词的出现概率仅依赖于它前面的N-1个词
    - 计算句子出现的条件概率
      - N-gram模型通过条件概率链式规则来估计整个句子的概率。具体而言，对于给定的一个句子，模型会计算每个N-gram出现的条件概率。并将这些概率相乘以得到整个句子的概率。例如，对于句子"The quick brown fox"，作为trigram模型，我们会计算 $P(the|<s>,<s>)$、$P(quick|<s>,the)$、$P(brown|the,quick)$、$P(fox|quick,brown)$等概率。并将它们相乘。
  - 问题：当N较大时，会出现数据稀疏性问题；此外，N-gram模型忽略了词之间的范围依赖关系，无法捕捉到句子中的复杂结构和语义信息。
    - 模型的参数空间会急剧增大，相同的N-gram序列出现的概率变得非常低，导致模型无法有效学习，模型泛化能力下降。
- Word2Vec
  - 词嵌入技术，基于神经网络NNLM，利用词的上下文捕捉词之间关系，让词义相似的词在向量空间中距离较近
    - Skip-Gram模型
      - 利用目标词的向量表示计算上下文中的词向量
        - 适用于大型语料
      - 连续词袋模型CBOW(Continuous Bag of Words)
        - 是根据目标词上下文中的词对应的词向量，计算并输出目标词的向量表示
        - 适用于小型数据集
  - 生成低维密集向量，减少计算复杂度和存储需求
- ELMo（Embeddings from Language Models）
  - 能够捕捉到词汇的多义性和上下文信息，生成的词向量更加丰富和准确
    - 先利用语言模型进行预训练
    - 再在特定任务中微调