自然语言处理内容要点

一、自然语言处理概述

1、理解什么是自然语言

2、自然语言与人工语言(编程语言)的区别

从词汇量、是否结构化、是否具有歧义性、容错性、发展易变性、简略性等方面分析比较理解。

3、什么是自然语言处理

自然语言处理（Natural Language Processing：NLP）是研究用机器处理人类语言的理论和技术以实现在人与人以及人与计算机之间利用自然语言进行交互的一门学科。

自然语言处理实现人与计算机的通信交互，包含两方面：自然语言理解和自然语言生成。要能区分两者并能举例说明。

4、了解自然语言处理的应用领域及发挥的作用，并能结合实际场景举例说明

如文本分类、信息检索、信息提取、自动摘要、文本生成、机器翻译、情感分析、信息过滤、个性化推荐等。

5、自然语言处理在人工智能发展中所处的阶段

人工智能三个层次，NLP 所处位置。

二、自然语言处理研究内容与现状

1、主要研究内容

由浅入深划分为几个层面：词法、句法、语义、语用。

2、词法分析的主要内容，包括哪些主要技术，词法分析的作用

词法分析是后续高级任务的基础。

中文分词：将文本分隔为有意义的词语。对于中文来说分词尤为重要。有的句子分词方式不同会产生歧义。

词性标注：确定每个词语的类别和浅层的歧义消除（如名词、动词、形容词等）。

命名实体识别：识别出句子中一些有意义的实体，如人物、地名、机构、术语、专用名词等。

去停用词：去掉对文本特征没有作用的字词，如标点符号、语气、人称等。

3、句法分析的主要内容，包括哪些主要技术，句法分析的作用

句法分析主要研究句子的结构与成分、词语之间的相互关系以及组成句子的规则。

句法结构分析:主要分析句子的成分结构,确定各个词在句子中起到的作用,如主谓宾定状补等。

依存关系分析：主要分析句子中各个词汇之间的相关关系，如并列、比较、从属、递进等，以便理解清楚句子逻辑与所包含的深层次含义。

4、语义分析的主要内容，包括哪些主要技术，语义分析的作用

语义分析主要研究如何根据文本中的句法结构、依存关系和句子中各个词语的意义，理解语句乃至一段文本所表示的语义。

词义表示：研究如何将自然语言中的词表示为向量，便于后续计算分析。

词义消歧：根据上下文确定一个词在语境中的含义。

语义角色标注：标注句子中的谓语和其他成分之间的关系。

语义关系分析:分析句子中词语之间的语义关系,包括利用上下文实现指代消歧。

5、语用分析的主要内容，包括哪些主要技术，语用分析的作用

语用分析主要研究词语、句子在不同上下文中的应用,根据上下文关系从整体理解、分析语句的含义。

相对于语义分析来说,语用分析增加了对上下文、语境等方面的分析与理解,可以提取出更多附加信息和深层次含义，是一种面向应用的高层次语言学分析。

包含文本分类与文本聚类、信息抽取、主题分析、意图识别、语境分析和情感分析等方面的问题。

6、常用的 NLP 工具包

7、NLP 的发展历程：发展阶段、主要特点、代表方法与模型等

大体可以分为三个阶段:基于规则的理性主义方法、基于统计学习的经验主义方法、基于深度神经网络的深度学习方法。

8、了解 NLP 研究面临的困难

一方面是语言场景中分词、歧义、多义性等带来的问题,另一方面是统计学习方法和深度学习方法在理论模型、效果和资源需求等方面仍有一定局限性。

三、自然语言处理主要过程与方法

1．理解什么是语料，如何获取语料

2、了解文本预处理的内容与方法

语料清洗、分词、词性标注、去停用词等。

3、特征工程

为什么要做特征工程？

掌握常用的特征，理解其含义、做法和特点:

统计特征（TF、IDF、TF-IDF）；

词向量（词袋模型、One-Hot 表示法、词嵌入），要理解为什么要用词向量，One-Hot 表示法和词嵌入的各自特点与区别；

实体抽取：命名实体识别、实体消岐等。

4、任务模型

统计学习模型；

常用的深度学习模型：CNN、RNN；

RNN 的特点与基本类型，为什么适合用于自然语言处理；

NLP 中常用的 RNN 变体模型：LSTM、GRU、Seq2Seq；

注意力机制。

5、应用分析