自然语言处理笔记

卢雨轩 19071125

2022 年 2 月 25 日

一、 绪论

平时成绩 40%
大作业 60%

1.1 NLP 的基本概念及发展历程

- 基本概念
  - 自然语言处理（Natural Language Processing, NLP）
    计算机通过可计算的方法对于自然语言的各级语言单位进行处理的方法
  - 语言
    人类特有的用来表达意思、交流思想的工具，是一种特殊的社会现象。
  - 自然语言
    指人类日常使用的语言，如汉语、英语等。
    强调人类发展过程中自然产生的，而不是人工编制的。
  - 语言学
    * 对语言的科学研究
    * 研究语言的本质、结构和发展规律
    * 两个基本属性：文字和声音
  - 语言的三种类型
    * 孤立语：词的形态变化少，语法关系用词序和虚词表示，如汉语
    * 曲折语：用词的形态变化表示语法关系，如英语
    * 黏着语：同时具有两者的特性：日语，汉语
  - 处理：指对信息的接收、存储、转化、传送和发送
    * 两个层次：字符处理和内容处理
    * 内容处理：输入、存储、输出等
    * 内容处理：词语切分、词性标注、结构分析、意义理解、推理、翻译...
  - 中文信息处理：针对中文的自然语言处理
- 相关提法
  - 自然语言处理
  - 自然语言理解
  - 计算机语言学
  - 统称为人类语言技术，Human Language Processing。

1.2 NLP 的研究内容与基本问题

- 研究内容：

  从应用角度：

  - 机器翻译，Machine Translation
  - 信息检索，Information Retrieval
  - 自动文摘，Automacic Summarization
  - 问答系统，Quastion Answering System
  - 信息过滤，Information Filtering
  - 信息抽取，Information Extraction
  - 文档分类，Document Categorization
  - 语音识别，Automatic Speech Recognition
  - 语音合成，Text To Speach
  - 说明：一般把语音相关独立出来，把文字相关作为自然语言处理的主题。

    从基础研究角度：
  - 分词
  - 词性标注
  - 短语切分
  - 语法分析
  - 语义分析
  - 篇章理解

    从资源建设角度：
  - 语料库资源建设
  - 语言学知识库建设

- 主要研究内容的分层：
  - 应用系统
  - 应用技术研究：自动问答等
  - 基础研究：词性标注、分词等
  - 资源建设：语料库建设

- 基本问题：
  - 形态学问题
    - 研究词由有意义的基本单位-词素的构成问题
  - 举发问题
    - 研究句子结构成分之间的相互关系
  - 语义问题
    - 研究如何从词的意思推断语句的意义
  - 语用学问题
    - 研究不同上下文中语句的 uzoyong
  - 语音学问题
    - 研究语音特性

1.3 现状、主要困难和基本方法

- 发展现状
  - 缺乏理论基础和完整体系
  - 浅层问题尚未解决，但是已经开始挑战深层次问题
  - 深度学习得到广泛关注
  - 开放域问题处理
- 主要困难
  - 大量的歧义现象
    - 词法歧义：
      - I'll see Prof. Zhang home.
        'll 是缩写，两个 ' 意义不同
      - 乒乓球拍卖完了
        乒乓球/拍卖/完了
        乒乓球拍/卖/完了
    - 结构歧义：
      今天中午吃馒头
      今天中午吃食堂
    - 语义歧义：
      同一个词有不同意义
    - 语音歧义
    - 多音字以及韵律歧义
  - 大量未知的语言现象
    - 新词、人名、属于
    - 新含义
    - 新用法，新句型
- 基本研究方法
  - 理性主义方法
    - 从理论为基础，解释歧义行为
  - 经验主义方法
    - 基于语料库应用数学方法进行统计分析，发现知识或抽取统计规律。
  - 区别：研究对象不同
  - 区别：理论基础不同
  - 区别：范围不同
  - 区别：方法不同

1.4 课程内容