绪论

问题的提出

1. （1）跨语言通信与信息获取；（2）机器翻译市场需求大；（3）舆情监测。
2. （1）如何让计算机实现自动地或人机互助地语言处理功能？（2）如何让计算机实现海量语言信息的自动处理、知识挖掘和有效利用？

基本概念

1. **语言学**：（1）历时~ / 历史~；（2）共时~；（3）描述~；（4）对比~；（5）结构~。
2. **语音学**：
    - （1）发音~；（2）声学~；（3）听觉~。
    - （1）一般~：与语言学无关；（2）实验~：使语言学研究的一部分。
3. **自然语言处理 / 自然语言理解 / 计算语言学。**
4. 语系：
    - **屈折语**：词的形态变化表示语法，如英语、法语；
    - **黏着语**：专门的附加成分表示语法，如日语、汉语、土耳其语；
    - **孤立语/分析语**：词序和虚词表示语法，如汉语。

NLP的产生与发展

1. 源于**机器学习。**
2. 发展历程：1960s（萌芽期）-> 1960s-1970s（步履维艰）-> 1966 ALPAC -> 1970s-1980s（复苏）-> 1980s至今（蓬勃发展）。

研究内容

1. 自然语言处理：以**文字**为处理对象，主要内容有：机器翻译、文本摘要、文本理解等。
2. 语音技术：语音识别、语音合成，主要内容有：语音翻译、人机对话、多媒体检索。

基本问题和主要困难

1. 基本问题：形态学问题、语法学问题、语义学问题、语用学问题、语音学问题。
2. 困难：**大量歧义现象、大量未知语言现象。**
3. NLU所面临的挑战：
    - **普遍存在的不确定性**：词法、句法、语义、语用和语音；
    - **未知语言现象的不可预测性**：新的词汇、术语、语义、语法；
    - **始终面临的数据不充分性**：有限语言集合无法涵盖开放的语言现象；

- **语言知识表达的复杂性**：语义知识的模糊性和复杂关联性；
- **机器翻译中映射单元的不对等性**：词法表达、句法结构、语义概念不对等。

基本研究方法

1. 理性主义与经验主义的差异：
    - **认识差异**：内在语言官能 vs. 感官输入、简单联想与通用化操作；
    - **研究差异**：语言知识结构 vs. 实际语言数据；
    - **理论差异**：文法理论 vs. 概率论、信息论、机器学习；
    - **方法差异**：基于规则 vs. 基于统计；
    - **NLP系统**：知识库+推理模型 vs. 语料库+统计模型。

国内外研究机构