自然语言处理研究范畴

一、自然语言处理的主要范畴

- 文本朗读([Text to speech)/语音合成(Speech synthesis)
- 语音识别(Speech recognition)
- 中文自动分词(Chinese word segmentation)
- 词性标注(Part-of-speech tagging)
- 句法分析(Parsing)
- 自然语言生成(Natural language generation)
- 文本分类(Text categorization)
- 信息检索(Information retrieval)
- 信息抽取(Information extraction)
- 文字校对(Text-proofing)
- 问答系统(Question answering)
- 机器翻译(Machine translation)
- 自动摘要(Automatic summarization)

二、自然语言处理研究的难点

- 单词的边界界定

在口语中，词与词之间通常是连贯的，而界定字词边界通常使用的办法是取用能让给定的上下文最为通顺且在文法上无误的一种最佳组合。在书写上，汉语也没有词与词之间的边界。

- 词义的消歧

许多字词不单只有一个意思，因而我们必须选出使句意最为通顺的解释。

- 句法的模糊性

自然语言的文法通常是模棱两可的，针对一个句子通常可能会剖析(Parse)出多棵剖析树(Parse Tree)，而我们必须要仰赖语意及前后文的资讯才能在其中选择一棵最为适合的剖析树。

- 有瑕疵的或不规范的输入

例如语音处理时遇到外国口音或地方口音,或者在文本的处理中处理拼写,语法或者光学字符识别(OCR)的错误。

- 语言行为与计划

句子常常并不只是字面上的意思；例如，“你能把盐递过来吗”，一个好的回答应当是把盐递过去；在大多数上下文环境中，“能”将是糟糕的回答，虽说 回答“不”或者“太远了我拿不到”也是可以接受的。再者，如果一门课程去年没开设，对于提问“这门课程去年有多少学生没通过？”回答“去年没开这门课”要 比回答“没人没通过”好。

三、当前自然语言处理研究的发展趋势

第一，传统的基于句法-语义规则的理性主义方法受到质疑，随着语料库建设和语料库语言学的崛起，大规模真实文本的处理成为自然语言处理的主要战略目标。

第二，统计数学方法越来越受到重视，自然语言处理中越来越多地使用机器自动学习的方法来获取语言知识。

第三，浅层处理与深层处理并重，统计与规则方法并重，形成混合式的系统。

第四，自然语言处理中越来越重视词汇的作用，出现了强烈的“词汇主义”的倾向。词汇知识库的建造成为了普遍关注的问题。

四、统计自然语言处理

统计自然语言处理运用了推测学、机率、统计的方法来解决上述，尤其是针对容易高度模糊的长串句子，当套用实际文法进行分析产生出成千上万笔可能性时所引发之难题。处理这些高度模糊句子所采用消歧的方法通常运用到语料库以及马可夫模型(Markov models)。统计自然语言处理的技术主要由同样自人工智能下与学习行为相关的子领域：机器学习及资料采掘所演进而成。

五、参考资料

Wikipedia – Natrual Language Processing : http://en.wikipedia.org/wiki/Natural_language_processing

李劼 Joker Lee
November, 2009
At BUPT