**第1章 绪 论**

**1.1 基本概念**

**1.1.1 语言学与语音学**

**语言学 (linguistics)**：是指对语言的科学研究，研究语言的本质、结构和发展规律的科学。语音和文字是语言的两个基本属性。

**语音学 (phonetics)**：研究人类发音特点，特别是语音发音特点，并提出各种语音描述、分类和转写方法的科学。

**1.1.2 自然语言处理**

**计算语言学 (Computational Linguistics)**：通过建立形式化的计算模型来分析、理解和生成自然语言的学科，是人工智能和语言学的分支学科。 计算语言学是典型的交叉学科，其研究常常涉及计算机科学、语言学、数学等多个学科的知识。与内容接近的学科自然语言处理相比较，计算语言学更加侧重基础理论和方法的研究。

**自然语言理解 (Natural Language Understanding, NLU)**：自然语言理解是探索人类自身语言能力和语言思 维活动的本质，研究模仿人类语言认知过程的自然语言处理方法和实现技术的一门学科。它是人工智能早期研究的领域之一，是一门在语言学、计算机科学、认知科学、信息论和数学等多学科基础上形成的交叉学科。

**自然语言处理 (Natural Language Processing, NLP)**：自然语言处理是研究如何利用计算机技术对语言文本（句子、篇章或话语等）进行处理和加工的一门学科，研究内容包括对词法、句法、语义和语用等信息的识别、分类、提取、转换和生成等各种处理方法和实现技术。

**1.1.3 关于“理解”的标准**

图灵试图解决长久以来关于如何定义思考的哲学争论，他提出了一个虽然主观但可以操作的标准：如果一个计算机系统的表现（act）、反应（react）和互相作用（interact） 都和有意识的个体一样，那么，这个计算机系统就应该被认为是有意识 的。

**图灵测试**：测试人在一段规定的时间内，在无法看到反应来源的情况下，根据两个实体 （被测试的计算机系统和另外一个人）对他提出的各种问题的反应来判断做出反应的是人还是计算机。通过一系列这样的测试，从计算机被误判为人的几率就可以测出计算机系统所具有的智能程度。

**1.2 自然语言处理研究的内容和面临的困难**

**1.2.1 自然语言处理研究的内容**

（1）机器翻译（machine translation, MT）

（2）自动文摘（automatic summarizing或automatic abstracting）

（3）信息检索（information retrieval）

（4）文档分类（document categorization/classification）

（5）问答系统（question-answering system）

（6）信息过滤（information filtering）

（7）信息抽取（information extraction）

（8）文本挖掘（text mining）

（9）舆情分析（public opinion analysis）

（10）隐喻计算（metaphorical computation）

（11）文字编辑和自动校对（automatic proofreading）

（12）作文自动评分

（13）光读字符识别（optical character recognition, OCR）

（14）语音识别（speech recognition）

（15）文语转换（text-to-speech conversion）

（16）说话人识别／认证／验证（speaker recognition/identification/verification）

**1.2.2 自然语言处理涉及的几个层次**

形态学（morphology）：形态学（又称“词汇形态学”或“词法”）是语言学的一个分支，研究词的内部结构，包括屈折变化和构词法两个部分。

语法学（syntax）：研究句子结构成分之间的相互关系和组成句子序列的规则。其关注的中心是：为什么一句话可以这么说，也可以那么说？

语义学（semantics）：是一门研究意义，特别是语言意义的学科。

语用学（pragmatics）：是现代语言学用来指从使用者的角度研究语言，特别是使用者所作的选择、他们在社会互动中所受的制约、他们的语言使用对信递活动中其他参与者的影响。

**1.2.3 自然语言处理面临的困难**

最终需要解决的关键问题就是歧义消解（disambiguation）问题和未知语言现象的处理问题。

基本问题：

1. 形态学 (Morphology) 问题：研究词(word) 由有意义的基本单位－词素 (morphemes)的构成问题。单词的识别/ 汉语的分词问题。
2. 句法 (Syntax) 问题：研究句子结构成分之间的相互关系和组成句子序列的规则。
3. 语义 (Semantics) 问题：研究如何从一个语句中词的意义，以及这些词在该语句中句法结构中的作用来推导出该语句的意义。
4. 语用学(Pragmatics) 问题：研究在不同上下文中语句的应用，以及上下文对语句理解所产生的影响。从狭隘的语言学观点看，语用学处理的是语言结构中有形式体现的那些语境。相反，语用学最宽泛的定义是研究语义学未能涵盖的那些意义。
5. 语音学(Phonetics) 问题：研究语音特性、语音描述、分类及转写方法等。

**1.3 自然语言处理的基本方法及其发展**

**1.3.1 自然语言处理的基本方法**

一般认为，自然语言处理中存在着两种不同的研究方法，一种是理性主义（rationalist）方法，另一种是经验主义（empiricist）方法。

理性主义：通常通过一些特殊的语句或语言现象的研究来得到对人的语言能力的认识，而这些语句和语言现象在实际的应用中并不常见。问题求解的基本思路：基于规则的分析方法建立符号处理系统。

经验主义：偏重于对大规模语言数据中人们所实际使用的普通语句的统计。求解问题的思路：基于大规模真实语料(语言数据) 建立计算方法。

**1.3.2 自然语言处理的发展**

**1.4 自然语言处理的研究现状**