第1章

绪　论

自然语言作为人类思想情感最基本、最直接、最方便的表达工具，无时无刻不充斥在人类社会的各个角落。人们从出生后的第一声啼哭开始，就企图用语言（声音）来表达自己的情感和意图。随着信息时代的到来，人们使用自然语言进行通信和交流的形式也越来越多地体现出它的多样性、灵活性和广泛性。然而，人脑是如何实现自然语言理解这一认知过程的？我们应该如何建立语言、知识与客观世界之间的对应关系，并实现有效的概念区分和语义理解？从数学的角度讲，语义是否可计算？如果可计算，其计算模型和方法以及复杂度又如何？为什么世界上不同种族的人在拥有几乎相同的大脑结构和语声工作机理的情况下，却无法实现不同语言之间的相互理解？众多科学问题困扰着我们，目前计算机处理自然语言的能力在大多数情况下都不能满足人类社会信息化时代的要求。有关专家已经指出，语言障碍已经成为制约21世纪社会全球化发展的一个重要因素。因此，如何尽早实现自然语言的有效理解，打破不同语言之间的固有壁垒，为人际之间和人机之间的信息交流提供更便捷、自然、有效和人性化的帮助与服务，已经成为备受人们关注的极具挑战性的国际前沿研究课题，也是全球社会共同追求的目标和梦想。

从研究内容和方法上来看，自然语言处理研究集认知科学、计算机科学、语言学、数学与逻辑学、心理学等多种学科于一身，其研究范畴不仅涉及对人脑语言认知机理、语言习得与生成能力的探索，而且，包括对语言知识的表达方式及其与现实世界之间的关系，语言自身的结构、现象、运用规律和演变过程，大量存在的不确定性和未知语言现象以及不同语言之间的语义关系等各方面问题的研究。因此，自然语言处理是现代信息科学和技术研究不可或缺的重要内容，从事这项研究不仅具有重要的科学意义，而且具有巨大的应用价值。

1.1　基本概念

1.1.1　语言学与语音学

我们知道，语言作为人类特有的用来表达情感、交流思想的工具，是一种特殊的社会现象，由语音、词汇和语法构成。语音和文字是构成语言的两个基本属性，语音是语言的物质外壳，文字则是记录语言的书写符号系统[黄伯荣等，1991]。

根据《现代语言学词典》[克里斯特尔，2002]的定义，语言学(linguistics)是指对语言的科学研究。作为一门纯理论的学科，语言学在近期获得了快速发展，尤其从20世纪60年代起，已经成为一门知晓度很高的广泛教授的学科。

根据语言学家的注意中心和兴趣范围，语言学可以区分为一些不同的分支，例如，历时语言学(diachronic linguistics)或称历史语言学(historical linguistics)、共时语言学(synchronic linguistics)、一般语言学(general linguistics)、理论语言学(theoretical linguistics)、描述语言学(descriptive linguistics)、对比语言学(contrastive linguistics)或类型语言学(typological linguistics)、结构语言学(structural linguistics)等。

语音学(phonetics)是研究人类发音特点，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。语音学一般有三个分支：①发音语音学(articulatory phonetics)，研究发音器官是如何产生语音的；②声学语音学(acoustic phonetics)，研究口耳之间传递语音的物理属性；③听觉语音学(auditory phonetics)，研究人通过耳、听觉神经和大脑对语音的知觉反应。仪器语音学(instrumental phonetics)则是利用各种物理设备，如测量气流或分析声波的仪器等，来研究上述三个问题的任一方面[克里斯特尔，2002]。

由于语音学家研究的目标通常是发现支配语音性质和使用的普世原则，因此，语音学又常称作一般语音学或通用语音学(general phonetics)。实验语音学(experimental phonetics)具有相同的含义。

从研究方法上来看，如果研究者关心的只是语音发音、声学或知觉的一般性规律和特点，那么，语音学研究与语言学的关系不大。但是，如果研究者关注的重点是具体语言或方言(或语言、方言群)的语音特点时，我们往往很难说清楚语音学到底是一门独立的学科还是应看作语言学的一个分支。而在有些大学里，有的相关的系称为“语言学系”，有的则称为“语言学和语音学系”，但实际上“语言学系”也同样教授语音学。因此，为了避免这种名称上的差异可能给人们造成的错觉，一些聪明的外国人采用一种折中的办法，用复数的“语言科学(linguistic sciences)”来作为整个学科的统称，既包括语言学，也包括语音学。在本书中，我们愿意沿用这种复数的语言科学名称。

1.1.2 自然语言处理

自然语言处理(natural language processing，NLP)也称自然语言理解(natural language understanding，NLU)，从人工智能研究的一开始，它就作为这一学科的重要研究内容探索人类理解自然语言这一智能行为的基本方法。在最近二三十年中，随着计算机技术，特别是网络技术的迅速发展和普及，自然语言处理研究得到了前所未有的重视和长足的进展，并逐渐发展成为一门相对独立的学科，备受关注。

语言学家刘涌泉在《大百科全书》(2002)中对自然语言处理的定义为：“自然语言处理是人工智能领域的主要内容，即利用电子计算机等工具对人类所特有的语言信息(包括口语信息和文字信息)进行各种加工，并建立各种类型的人-机-人系统。自然语言理解是其核心，其中包括语音和语符的自动识别以及语音的自动合成。”

冯志伟对“自然语言处理”的解释为：自然语言处理就是利用计算机为工具对人类特

有的书面形式和口头形式的自然语言的信息进行各种类型处理和加工的技术[冯志伟，1996]。

美国计算机科学家马纳瑞斯(Bill Manaris)给自然语言处理的定义为："自然语言处理是研究人与人交际中以及人与计算机交际中的语言问题的一门学科。自然语言处理要研制表示语言能力(linguistic competence)和语言应用(linguistic performance)的模型，建立计算框架来实现这样的语言模型，提出相应的方法来不断地完善这样的语言模型，根据这样的语言模型设计各种实用系统，并探讨这些实用系统的评测技术。"[Manaris，1999]

通常认为，"计算语言学(computational linguistics)"这一术语是在美国科学院于 20 世纪 60 年代设立的自动语言处理咨询委员会(Automatic Language Processing Advisory Committee，ALPAC)于 1966 年宣布的对机器翻译技术的评估报告中首次提出来的，但实际上在 ALPAC 报告发布之前这一术语就已经出现了，如 1962 年美国成立了"机器翻译和计算语言学学会(Association for Machine Translation and Computational Linguistics)"[①②]，1965 年期刊 *MT*：*Mechanical Translation*[③] 更名为 *Mechanical Translation and Computational Linguistics*[④]。几乎同一时间，国际计算语言学委员会(The International Committee on Computational Linguistics，ICCL)成立，并于 1965 年组织召开了第一届国际计算语言学大会(The International Conference on Computational Linguistics，COLING)[⑤]。只不过在那一时期人们对于"计算语言学"是否能够真正成为一门独立的学科还没有足够的把握，因此，这一术语的出现还带着"犹抱琵琶半遮面"的羞涩味道，而 1966 年美国科学院公布的 ALPAC 报告作为一份正规、严谨的科学文献使"计算语言学"这一术语正式得到了学术界的承认。

目前对"计算语言学"这一术语并没有统一的严格定义，我们能够看到的定义基本上都是解释性的，但这并不影响我们对它的理解。根据英国《大不列颠百科全书》的解释，计算语言学是利用电子数字计算机进行的语言分析。虽然其他类型的语言分析也可以运用计算机，但计算机分析最常用于处理基本的语言数据，例如建立语音、词、词元素的搭配以及统计它们的频率[翁富良等，1998]。当然，从目前情况来看，这种解释似乎有点过时，因为它仅仅强调的是计算机作为辅助工具用于对自然语言进行一些相关的分析和统计，而没有把计算机作为一种可以提供主动服务，能够帮助人类达到对话、翻译、检索等若干目的的智能工具。

《现代语言学词典》[克里斯特尔，2002]中对计算语言学的定义为：语言学的一个分支，用计算技术和概念来阐述语言学和语音学问题。已开发的领域包括自然语言处理、言

---

① http://en.wikipedia.org/wiki/Computational_linguistics

② http://en.wikipedia.org/wiki/Association_for_Computational_Linguistics

③ 该期刊由 Vic Yngve 博士创建于 1954 年。

④ 该期刊于 1970 年停刊。1974 年，David Hays 延续之前的工作，创立了期刊 *American Journal of Computational Linguistics*（AJCL）。1984 年 AJCL 改名为 *Computational Linguistics*。目前该期刊已经成为国际计算语言学和自然语言处理领域的顶级学术期刊。详细情况请见网页 ACL Anthology（http://www.aclweb.org/anthology-new/docs/cl.html）。

⑤ http://nlp.shef.ac.uk/iccl/

语合成、言语识别、自动翻译、编制语词索引、语法的检测，以及许多需要统计分析的领域（如文本考释）。

语言学家刘涌泉在我国的《大百科全书》(2002)中给出的解释是：计算语言学是语言学的一个分支，专指利用电子计算机进行语言研究。

根据这些定义和上述(复数)语言科学的概念可以看出，计算语言学实际上包括以语音为主要研究对象的语音学基础及其语音处理技术研究和以词汇、句子、话语或语篇(discourse)及其词法、句法、语义和语用等相关信息为主要研究对象的处理技术研究。不难看出，早期对“计算语言学”和“自然语言处理”的解释带有一定的局限性。例如，在《现代语言学词典》给出的定义中，自然语言处理属于计算语言学研究的范畴。但实际上，近几年来由于自然语言处理技术的迅速发展，相关技术不断与语音识别(speech recognition)、语音合成(speech synthesis)等技术相互渗透和结合已经形成了若干新的研究分支，例如，基于语音输入输出的人机对话系统，语音翻译(speech-to-speech translation)，语音文档摘要(speech document summarization)和语音文档检索(speech document retrieval)等。因此，从目前情况来看，自然语言处理一般不再被看作是计算语言学范畴内的一个研究分支，而两者基本上是处于同一层次上的概念。

从术语的字面上来看，似乎“计算语言学”更侧重于计算方法和语言学理论等方面的研究，而“自然语言理解”更偏向于对语言认知和理解过程等方面问题的研究，相对而言，“自然语言处理”包含的语言工程和应用系统实现方面的含义似乎更多一些，但是，在很多情况下我们很难绝对地区分开“计算语言学”、“自然语言理解”与“自然语言处理”三个术语之间到底存在怎样的包含或重叠关系以及各自不同的内涵和外延。因此，很多人在谈到“计算语言学”、“自然语言理解”或“自然语言处理”这些术语时，往往默认为它们是同一个概念，至少在其外延上不再细究其差异。甚至有些专著中干脆直接这样解释：计算语言学也称自然语言处理或自然语言理解[刘颖，2002]。

本书主要介绍以词汇、语句、篇章和对话等为主要处理对象的自然语言处理技术的基本理论和实现方法，不涉及语音技术的细节。

值得说明的是，中文信息处理(Chinese information processing)作为专门以中文为研究对象的自然语言处理技术已经在世界范围内得到广泛关注，并取得了快速进展。顾名思义，“中文”就是中国的语言文字。从广义上理解，它可以是中国各民族使用的所有语言文字的总称。长期以来，中国境外(如新加坡、马来西亚等)华人使用的汉语文字被称为华文或中文，由于汉族在人口数量和地域分布上都占有绝对优势，因此，在不引起混淆的情况下，我们认为“中文”与“汉语”指同一概念。根据国家标准 GB12200.1—90“汉语信息处理词汇 01 部分：基本术语”的解释，“中文(Chinese)”特指汉语[宗成庆等，2009]。

“中文信息处理”又可划分为“汉字信息处理”和“汉语信息处理”两个分支，汉字信息处理主要指以汉字为处理对象的相关技术，包括汉字字符集的确定、编码、字形描述与生成、存储、输入、输出、编辑、排版以及字频统计和汉字属性库构造等[俞士汶，2006]。一般而言，汉字信息处理关注的是文字(一种特殊的图形)本身，而不是其承载的语义或相互之间的语言学关系，而“汉语信息处理”则是指对传递信息、表达概念和知识的词、短语、句子、篇章乃至语料库和网页等各类语言单位及其不同表达形式的处理技术[宗成庆等，

2009]。本书中提到的“中文信息处理”一般指后者。

1.1.3　关于“理解”的标准

当人们提到关于“理解”的标准时，总是不会忘记著名的英国数学家图灵(Turing)1950 年提出的测试标准。当时图灵提出这个测试的目的是用来判断计算机是否可以被认为“能思考”。后来这个测试被称为图灵测试(Turing test)，现已被多数人承认。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。为此，图灵设计了一种“模仿游戏”，即现在所说的图灵测试：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。

在自然语言处理领域中，人们采用图灵实验来判断计算机系统是否“理解”了某种自然语言的具体准则可以有很多，例如：通过问答(question-answering)系统测试计算机系统是否能够正确地回答输入文本中的有关问题；通过文摘生成(summarizing)系统测试计算机系统是否有能力自动产生输入文本的摘要；通过机器翻译(machine translation，MT)系统测试计算机系统是否具有把一种语言翻译成另一种语言的能力；通过文本释义(paraphrase)系统测试计算机系统是否能够用不同的词汇和句型来复述其输入文本，等等[石纯一等，1993]。

实际上，人们在自然语言处理领域研究的任何一个应用系统都可以拿来做图灵测试。按照人的标准对这些系统的输出结果进行评价，从而判断计算机系统是否达到了“理解”的效果。显然，被测试系统所表现出来的性能反映了计算机系统的“理解”能力。因此，我们从事自然语言理解研究的任务也就是研究和探索针对具体应用目的的新方法和新技术，使实现系统的性能表现尽量符合人类理解的标准和要求。

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

自然语言处理研究的内容十分广泛，根据其应用目的不同，我们可以大致列举如下一些研究方向：

(1) 机器翻译(machine translation，MT)：实现一种语言到另一种语言的自动翻译。

(2) 自动文摘(automatic summarizing 或 automatic abstracting)：将原文档的主要内容和含义自动归纳、提炼出来，形成摘要或缩写。

(3) 信息检索(information retrieval)：信息检索也称情报检索，就是利用计算机系统从海量文档中找到符合用户需要的相关文档。面向两种或两种以上语言的信息检索叫做跨语言信息检索(cross-language/trans-lingual information retrieval)。

（4）文档分类（document categorization/classification）：文档分类也称文本分类（text categorization/classification）或信息分类（information categorization/classification），其目的就是利用计算机系统对大量的文档按照一定的分类标准（例如，根据主题或内容划分等）实现自动归类。近年来，情感分类（sentiment classification）或称文本倾向性识别（text orientation identification）成为本领域研究的热点。该项技术拥有广泛的用途，公司可以利用该技术了解用户对产品的评价，政府部门可以通过分析网民对某一事件、政策法规或社会现象的评论，实时了解百姓的态度。因此，情感分类已经成为支撑舆情分析（public opinion analysis）的基本技术。

（5）问答系统（question-answering system）：通过计算机系统对用户提出的问题的理解，利用自动推理等手段，在有关知识资源中自动求解答案并做出相应的回答。问答技术有时与语音技术和多模态输入、输出技术，以及人-机交互技术等相结合，构成人-机对话系统（human-computer dialogue system）。

（6）信息过滤（information filtering）：通过计算机系统自动识别和过滤那些满足特定条件的文档信息。通常指网络有害信息的自动识别和过滤，主要用于信息安全和防护、网络内容管理等。

（7）信息抽取（information extraction）：指从文本中抽取出特定的事件（event）或事实信息，有时候又称事件抽取（event extraction）。例如，从时事新闻报道中抽取出某一恐怖事件的基本信息：时间、地点、事件制造者、受害人、袭击目标、伤亡人数等；从经济新闻中抽取出某些公司发布的产品信息：公司名称、产品名称、开发时间、某些性能指标等。前一种事件一般是过程性的，有一定的因果关系，而后一类事件则是静态事实性的。信息抽取与信息检索不同，信息抽取直接从自然语言文本中抽取信息框架，一般是用户感兴趣的事实信息，而信息检索主要是从海量文档集合中找到与用户需求（一般通过关键词表达）相关的文档列表，而信息抽取则是希望直接从文本中获得用户感兴趣的事实信息。当然，信息抽取与信息检索也有密切的关系，信息抽取系统通常以信息检索系统（如文本过滤）的输出作为输入，而信息抽取技术又可以用来提高信息检索系统的性能[李保利等，2003]。

信息抽取与问答系统也有密切的联系。一般而言，信息抽取系统要抽取的信息是明定的、事先规定好的，系统只是将抽取出来的事实信息填充在给定的框架槽里，而问答系统面对的用户问题往往是随机的、不确定的，而且系统需要将问题的答案生成自然语言句子，通过自然、规范的语句准确地表达出来，使系统与用户之间形成一问一答的交互过程。

（8）文本挖掘（text mining）：有时又称数据挖掘（data mining），是指从文本（多指网络文本）中获取高质量信息的过程。文本挖掘技术一般涉及文本分类、文本聚类（text clustering）、概念或实体抽取（concept/entity extraction）、粒度分类、情感分析（sentiment analysis）、自动文摘和实体关系建模（entity relation modeling）等多种技术[①]。当然，数据挖掘有时具有更广泛的含义，可以包括音视频数据、图像数据和统计数据等。

① http://en.wikipedia.org/wiki/Text_mining

(9) 舆情分析(public opinion analysis)：舆情是指在一定的社会空间内，围绕中介性社会事件的发生、发展和变化，民众对社会管理者产生和持有的社会政治态度。它是较多群众关于社会中各种现象、问题所表达的信念、态度、意见和情绪等等表现的总和[①]。网络环境下舆情信息的主要来源有：新闻评论、网络论坛(bulletin board system，BBS)、聊天室、博客(Blog)、新浪微博、聚合新闻(或称“简易供稿”(really simple syndication，RSS))、Facebook、QQ、Twitter 等社交网站。由于网上的信息量十分巨大，仅仅依靠人工的方法难以应对海量信息的收集和处理，需要加强相关信息技术的研究，形成一套自动化的网络舆情分析系统，及时应对网络舆情，由被动防堵变为主动梳理、引导。显然，舆情分析是一项十分复杂、涉及问题众多的综合性技术，它涉及网络文本挖掘、观点(意见)挖掘(opinion mining)等各方面的问题。

(10) 隐喻计算(metaphorical computation)：“隐喻”就是用乙事物或其某些特征来描述甲事物的语言现象[周昌乐，2009]。简要地讲，隐喻计算就是研究自然语言语句或篇章中隐喻修辞的理解方法。

(11) 文字编辑和自动校对(automatic proofreading)：对文字拼写、用词，甚至语法、文档格式等进行自动检查、校对和编排。

(12) 作文自动评分：对作文质量和写作水平进行自动评价和打分。

(13) 光读字符识别(optical character recognition，OCR)：通过计算机系统对印刷体或手写体等文字进行自动识别，将其转换成计算机可以处理的电子文本，简称字符识别或文字识别。相对而言，文字识别研究的主要内容更多地属于字符(汉字)图像识别问题，通常被看作是一个模式识别问题，但作者认为，对于一个高性能的文字识别系统而言，如果没有任何自然语言理解技术的参与是不可想像的。

(14) 语音识别(speech recognition)：将输入计算机的语音信号识别转换成书面语表示。语音识别也称自动语音识别(automatic speech recognition，ASR)。

(15) 文语转换(text-to-speech conversion)：将书面文本自动转换成对应的语音表征，又称语音合成(speech synthesis)。

(16) 说话人识别/认证/验证(speaker recognition/identification/verification)：对一说话人的言语样本做声学分析，依此推断(确定或验证)说话人的身份。

综上所述，涉及人类语言的任何应用技术几乎都隐含着自然语言处理的问题。当然，上面所列举的这些研究内容覆盖面较广，有很多内容不仅仅是自然语言处理的问题，例如信息检索、舆情分析、文字识别，甚至社交网络(social network)、社会计算(social computing)等，除此之外，还有情感计算(affective computing)、语言教学(language teaching)、口语考试自动评分等等，这些研究往往包含很多其他技术。本书不想陷入关于这些内容归属问题的争论，只是由于这些研究与自然语言处理密切相关，而简单地将其划归为自然语言处理研究的范畴，这也算是作者对自然语言处理学科的“偏心”吧。另外需要指出的是，语音识别、语音合成和说话人识别这三项内容常常被单独看作“语音技术”，本书不涉及对这三项内容的具体介绍。

---

① 参阅《光明日报》2007 年 1 月 21 日马海兵：《网络舆情及其分析技术》。

1.2.2 自然语言处理涉及的几个层次

如果撇开语音学研究的层面，自然语言处理研究的问题一般会涉及自然语言的形态学、语法学、语义学和语用学等几个层次。

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。由于词具有语音特征、句法特征和语义特征，形态学处于音位学、句法学和语义学的结合部位，所以形态学是每个语言学家都要关注的一门学科［Matthews，2000］。

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［毛茂臣，1988］。语义学的研究对象是语言的各级单位（词素、词、词组、句子、句子群、整段整篇的话语和文章，乃至整个著作）的意义，以及语义与语音、语法、修辞、文字、语境、哲学思想、社会环境、个人修养的关系，等等［陆善采，1993］。其重点在探明符号与符号所指的对象之间的关系，从而指导人们的言语活动。它所关注的重点是：这个语言单位到底说了什么？

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。目前还缺乏一种连贯的语用学理论，主要是因为它必须说明的问题是多方面的，包括直指、会话隐含、预设、言语行为、话语结构等。部分原因是由于这一学科的范围太宽泛，因此出现多种不一致的定义。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［克里斯特尔，2002］。因此，语用学可以是集中在句子层次上的语用研究，也可以是超出句子，对语言的实际使用情况的调查研究，甚至与会话分析、语篇分析相结合，研究在不同上下文中的语句应用，以及上下文对语句理解所产生的影响。其关注的重点在于：为什么在特定的上下文中要说这句话？

在实际问题的研究中，上述几方面的问题，尤其是语义学和语用学的问题往往是相互交织在一起的。语法结构的研究离不开对词汇形态的分析，句子语义的分析也离不开对词汇语义的分析、语法结构和语用的分析，它们之间往往互为前提。

1.2.3 自然语言处理面临的困难

根据上面的介绍，自然语言处理涉及形态学、语法学、语义学和语用学等几个层面的问题，其最终应用目标包括机器翻译、信息检索、问答系统等非常广泛的应用领域。其实，如果进一步归结，实现所有这些应用目标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。一方面，自然语言中大量存在的歧义现象，无论在词法层次、句法层次，还是在语义层次和语用层次，无论哪类语言单位，其歧义性始终都是困扰人们实现应用目标的一个根本问题。因此，如何面向不同的应用目标，针对不同语言单位的特点，研究歧义消解和未知语言现象的处理策略及实现方法，就成了自然语言处理面临的核心问题。

词汇形态歧义消解是自然语言处理需要解决的基本问题。请看如下例句：

例句 1　I'll see Prof. Zhang home.

例句 2　He books two tickets.

对于例句 1，系统需要正确地识别“I'll”是单词 I 和 will 的缩写，而“Prof.”中的“.”只是表明“Prof.”是“Professor”的缩写，并非句子的结束。

例句 3　自动化研究所取得的成就。

对于汉语而言，尽管不存在形态变化的问题，但如何划分词的边界始终是中文信息处理中面临的一个难题。例句 3 可以有两种划分：

(1)自动化 研究所 取得 的 成就 。

(2)自动化 研究 所 取得 的 成就 。

显然，“所”一旦被切分为介词，整个句子的结构就完全不一样了。

请看如下典型的结构歧义例句：

例句 4　Put the block in the box on the table.

在例句 4 中，“on the table”既可以修饰“box”，也可以限定“block”。于是，我们可以得到两种不同的句法结构：

(1) Put the block [in the box on the table].

(2) Put [the block in the box] on the table.

如果在这个句子中再增加一个介词短语(...in the kitchen)，我们可以得到 5 种可能的分析结果，另外再增加一个的话，就可以得到 14 种可能的分析结构[Samuelsson and Wiren，2000]。

类似地，见例句 5：

例句 5　I saw a man in the park with a telescope.

可以得到 5 种不同的分析结构[冯志伟，1996]，而 W. A. Martin 曾报道他们的系统对于以下句子可以给出 455 个不同的句法分析结果[Martin *et al.*，1987]：

例句 6　List the sales of the products produced in 1973 with the products produced in 1972.

实际上，这种歧义结构分析结果的数量是随介词短语数目的增加呈指数上升的，其歧义组合的复杂程度随着介词短语个数的增加而不断加深，这个歧义结构的组合数称为开塔兰数(Catalan numbers，记作 C_n)，即如果句子中存在这样 n(n 为自然数)个介词短语，C_n 可以由下式获得[Samuelsson and Wiren，2000]：

$$C_n = \binom{2n}{n}\frac{1}{n+1}$$

由此，歧义结构数目的急剧增加，使得句法分析算法面临的困难迅速增大，句法分析算法不得不消耗大量的时间在这样一个组合爆炸的候选结构中搜索可能的路径，以实现局部歧义和全局歧义的有效消解。

在现代汉语中，尽管一般不会出现像上述英语例句那样由于多个介词结构的挂靠成分不同而引起句子歧义结构数目大量存在的现象，但是，汉语中的各类歧义现象却也是普遍存在的。请看如下例句：

例句 7 喜欢乡下的孩子。

这个句子可以理解为“[喜欢/乡下]的孩子。”也可以理解为“喜欢[乡下/的/孩子]。”而句子：

例句 8 关于鲁迅的著作。

可以解析为“关于[鲁迅/的/著作]”，也可以解析为“[关于/鲁迅]的著作”。

句法结构歧义固然是自然语言处理中典型的问题，而词汇的词类（part-of-speech）歧义、词义歧义和句子的语义歧义等也同样是自然语言处理中普遍存在的问题。例如，英语动词“swallow”通常需要有生命的动物作为主语，客观存在的有形的东西（被吞咽的对象）作为宾语，但在实际运用中，当用于隐喻时就出现了例外。例如[Manning and Schütze，1999]：

例句 9 I swallowed his story，hook，line，and sinker.

例句 10 The supernova swallowed the planet.

在汉语中，似是而非、模棱两可的句子更是司空见惯。句子“咬死猎人的狗”既可以指“那只狗是咬死了猎人的狗”，也可以指“把那只猎人的狗咬死”；我们说“今天中午吃食堂”绝不意味着今天中午要把食堂吃下去，而是要在食堂吃午饭。而“今天中午吃馒头”和“今天中午吃大碗”与这句话有相同的表达形式，却有完全不同的含义；我们夸奖一个人说“这个人真牛”时，并不是说这个人是真正的牛，而是夸奖他真有能耐；说一个人嘴很硬，也不是指这个人的嘴长得坚硬，而是指他（她）守口如瓶，或坚决不承认、不改变自己说过的话；“火烧圆明园”与“火烧驴肉”也绝非同一种结构和含义。在《现代汉语词典》（1999，商务印书馆）里“打”字做实词使用时就有25种含义，在“打鼓、打架、打球、打酒、打电话、打毛衣”等用法中，“打”字的含义各有不同。除此之外，“打”字还可以用作介词（如：自打今天起）和量词（如：一打铅笔）。如何根据特定的上下文让计算机自动断定“打”字的确切含义恐怕不是一件容易的事情。

作为一个例子，请看如下这段幽默小片段：

他说：“她这个人真有意思（funny）。”她说：“他这个人怪有意思的（funny）。”于是人们以为他们有了意思（wish），并让他向她意思意思（express）。他火了：“我根本没有那个意思（thought）！”她也生气了：“你们这么说是什么意思（intention）？”事后有人说：“真有意思（funny）。”也有人说：“真没意思（nonsense）”。（原文见《生活报》1994.11.13.第六版）[吴尉天，1999]