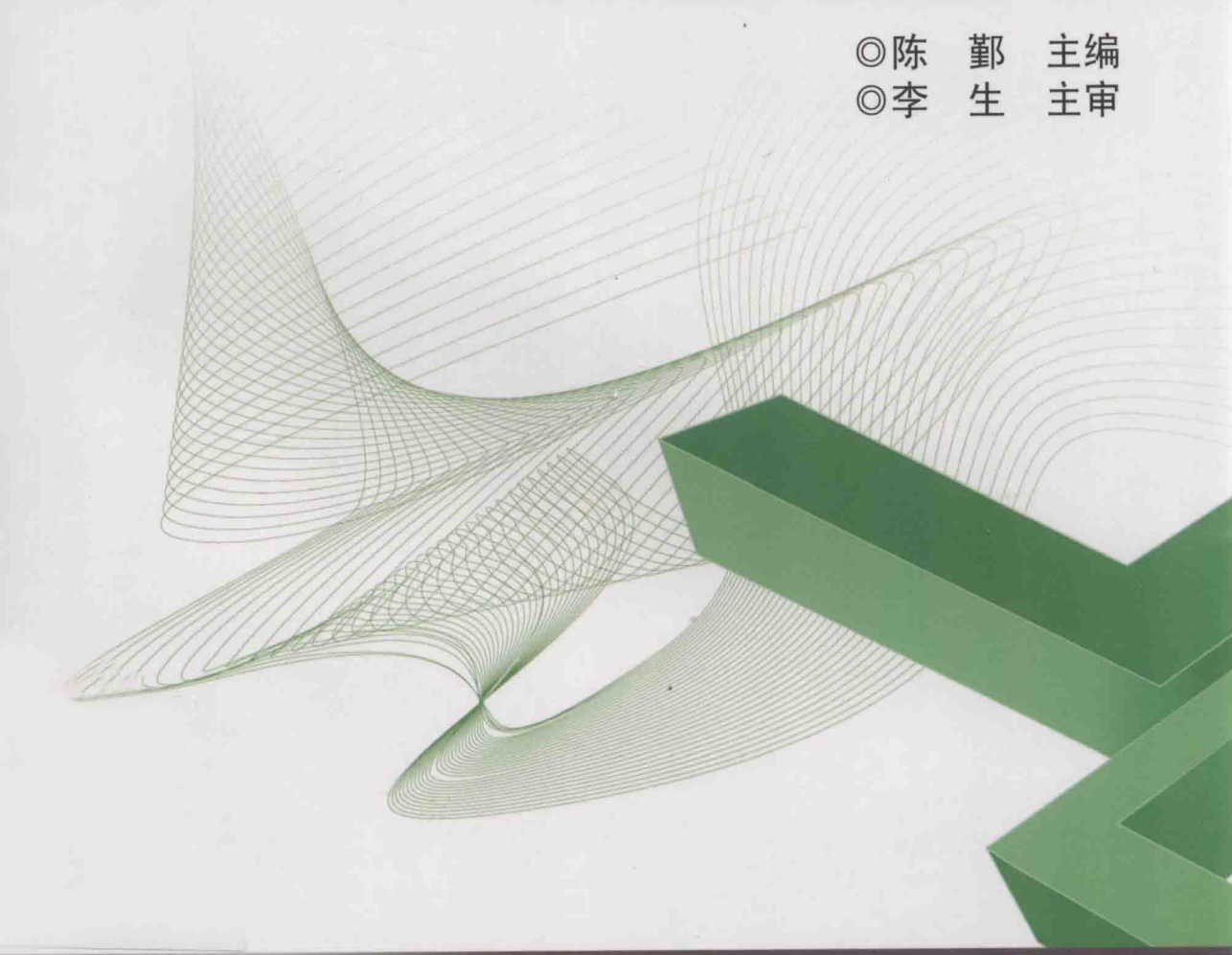
“十二五”国家重点图书出版规划项目
高等学校“十二五”规划教材·计算机软件工程系列

Basic Theories and Methods of Natural Language Processing

自然语言处理基本理论和方法

◎陈　鄞　主编
◎李　生　主审



注重夯实基础理论/简明生动/通俗易懂

系统讲解直观案例/培养能力/加强实践

哈爾濱工業大學出版社
HARBIN INSTITUTE OF TECHNOLOGY PRESS

“十二五”国家重点图书出版规划项目
高等学校“十二五”规划教材·计算机软件工程系列

自然语言处理基本理论和方法

陈　鄞　主编
李　生　主审

哈爾濱工業大學出版社

内容提要

本书对自然语言处理的基本理论和方法进行介绍。主要内容包括字符集的编码体系、语言计算模型、语言学资源建设、自然语言的词法分析、句法分析和语义分析等。本书内容源于作者多年的教学及科研心得，适合作为高等院校计算机相关专业本科生及研究生课程的教材。

图书在版编目(CIP)数据

自然语言处理基本理论和方法/陈鄞主编. —哈尔滨：哈尔滨工业大学出版社，2013.8

ISBN 978-7-5603-4126-2

Ⅰ.①自… Ⅱ.①陈… Ⅲ.①自然语言处理-高等学校-教材 Ⅳ.①TP391

中国版本图书馆 CIP 数据核字(2013)第 134250 号

策划编辑 王桂芝
责任编辑 李广鑫
出版发行 哈尔滨工业大学出版社
社　　址 哈尔滨市南岗区复华四道街 10 号　邮编 150006
传　　真 0451-86414749
网　　址 http://hitpress.hit.edu.cn
印　　刷 黑龙江省委党校印刷厂
开　　本 787mm×1092mm　1/16　印张　11.75　字数　288 千字
版　　次 2013 年 8 月第 1 版　2013 年 8 月第 1 次印刷
书　　号 ISBN 978-7-5603-4126-2
定　　价 28.00 元

高等学校“十二五”规划教材

计算机软件工程系列

编审委员会

◎序

Foreword

随着计算机软件工程的发展和社会对计算机软件工程人才需求的增长，软件工程专业的培养目标更加明确，特色更加突出。目前，国内多数高校软件工程专业的培养目标是以需求为导向，注重培养学生掌握软件工程基本理论、专业知识和基本技能，具备运用先进的工程化方法、技术和工具从事软件系统分析、设计、开发、维护和管理等工作能力，以及具备参与工程项目的实践能力、团队协作能力、技术创新能力和市场开拓能力，具有发展成软件行业高层次工程技术和企业管理人才的潜力，使学生成为适应社会市场经济和信息产业发展需要的“工程实用型”人才。

本系列教材针对软件工程专业“突出学生的软件开发能力和软件工程素质，培养从事软件项目开发和管理的高级工程技术人才”的培养目标，集9家软件学院（软件工程专业）的优秀作者和强势课程，本着“立足基础，注重实践应用；科学统筹，突出创新特色”的原则，精心策划编写。具体特色如下：

1. 紧密结合企业需求，多校优秀作者联合编写

本系列教材编写在充分进行企业需求、学生需要、教师授课方便等多方市场调研的基础上，采取了校企适度联合编写的做法，根据目前企业的普遍需要，结合在校学生的实际学习情况，校企作者共同研讨、确定课程的安排和相关教材内容，力求使学生在校学习过程中就能熟悉和掌握科学研究及工程实践中需要的理论知识和实践技能，以便适应就业及创业的需要，满足国家对软件工程人才的需要。

2. 多门课程系统规划，注重培养学生工程素质

本系列教材精心策划，从计算机基础课程→软件工程基础与主干课程→设计与实践课程，系统规划，统一编写。既考虑到每门课程的相对独立性、基础知识的完整性，又兼顾到相关课程之间的横向联系，避免知识点的简单重复，力求形成科学、完整的知识体系。

本系列教材中的《离散数学》、《数据库系统原理》、《算法设计与分析》等基础教材在引入概念和理论时，尽量使其贴近社会现实及软件工程等学科的技术和应用，力图将基本知识与软件工程学科的实际问题结合起来，在具备直观性的同时强调启发性，让学生理解所学的

知识。《软件工程导论》、《软件体系结构》、《软件质量保证与测试技术》、《软件项目管理》等软件工程主干课程以《软件工程导论》为线索，各课程间相辅相成，互相照应，系统地介绍了软件工程的整个学习过程。《数据结构应用设计》、《编译原理设计与实践》、《操作系统设计与实践》、《数据库系统设计与实践》等实践类教材以实验为主题，坚持理论内容以必需和够用为度，实验内容以新颖、实用为原则编写。通过一系列实验，培养学生的探究、分析问题的能力，激发学生的学习兴趣，充分调动学生的非智力因素，提高学生的实践能力。

相信本系列教材的出版，对于培养软件工程人才、推动我国计算机软件工程事业的发展必将起到积极作用。

2011 年 7 月

前　言

自然语言处理(Natural Language Processing,NLP)技术的产生可以追溯到20世纪50年代,它是一门集语言学、数学、计算机科学和认知科学等于一体的综合性交叉学科。经过半个多世纪的发展,自然语言处理的应用硕果累累,产生了很好社会效益和经济效益,在文字识别、语音合成等领域的技术已经达到了实用化的水平。目前,自然语言处理技术还进一步应用到网络内容管理、网络信息检控、不良信息的过滤和预警等方面,在现代信息科学的发展中,起着越来越重要的作用。

近年来,笔者在哈尔滨工业大学为软件工程专业的研究生讲授“自然语言处理”这门课程。软件工程专业学生的特点是实践能力较强,但在理论基础方面与计算机专业的学生相比稍显薄弱。而现有的专著和教材大多是面向专业技术人员或计算机专业的学生,因此在某些理论和方法的叙述上对于软件工程专业的学生来说过于抽象,理解起来比较困难。解决这一问题的最好办法就是多举例子,通过形象、直观、通俗易懂的实例来帮助学生进行理解。这种方法适合于讲解那些想象起来非常困难的理论或算法,因为想象起来很困难,这是难以理解的症结。因此在教材的编写上,既要吸取国内外教材的优点,广泛搜集恰当的例子,同时也要力图在较难的知识点或算法上精心设计大量简洁、直观的例子,把学生从抽象的理论讲解中解放出来。

笔者在授课过程中,参阅了大量的文献,通过比较,提取出最合适的定义和解释等,并大量采用形象举例法和模拟比喻法,形成自己的讲义。本书正是在笔者以往教案的基础上经过反复修改补充完成的。

由于学时所限(28学时),本书侧重于讲述NLP的基本理论与方法。本书的读者定位在对自然语言处理有一定兴趣的计算机相关专业的本科生或研究生。通过对本书的阅读,可以使读者对自然语言处理的相关知识有一个基本的了解,并可以起到将有志于从事此项研究的同学引入这一研究领域的作用,为将来开展研究工作打下坚实的基础。

本书一共9章,可以分成两个部分。第一部分是第1~6章,介绍自然语言处理的基础知识,包括字符集的编码体系、语言计算模型、语言学资源建设等;第二部分是第7~9章,介绍自然语言的基本技术,包括自然语言的词法分析、句法分析和语义分析等。

本书在编写过程中得到了哈尔滨工业大学赵铁军、李东、杨沐昀、徐冰等几位老师的大力支持,他们给出了很多宝贵的建议。笔者的老师李生教授在百忙中担任了本书的主审,使我深感荣幸。在此,谨向他们表示最诚挚的感谢!

由于作者水平有限,书中疏漏在所难免,敬请读者批评指正。

编者

2013年4月

于哈尔滨工业大学

目　录

第1章

绪　论

1.1　什么是自然语言处理

在 Stanley Kubrick(斯坦利·库布里克)2001 年执导的电影 *A Space Odyssey* 中,有一台称为 HAL 的机器人,它具有 20 世纪最受人们认可的一些特征。HAL 是一个具有高级语言处理能力,并且能够理解英语和说英语的智能计算机(Jurafsky et al,2000)。下面就是电影中的角色 Dave 先生和智能机器人 HAL 之间的一段对话:

Dave Bowman: Open the pod bay doors, HAL.

HAL: I'm sorry Dave, I'm afraid I can't do that.

HAL 的作者 Arthur 曾经乐观地预言,到了一定时期,我们就可以制造出像 HAL 这样的智能计算机。现在我们离这样的预言还有多远呢?

我们认为,像 HAL 这样的机器人,至少应该能够通过语言与人类进行交流。

首先,为了确定 Dave 先生讲什么,机器人 HAL 必须能够分析它从 Dave 那里接收到的声音信号,并把这些信号复原成文字的序列;接下来,HAL 需要分析这些文字序列所表达的含义,也就是说理解自然语言文本的意义;为了生成回答,HAL 必须把它要表达的意思组织成文字的序列,也就是说以自然语言文本来表达给定的意图、思想等;最后,HAL 需要把这些文字序列转化成 Dave 能够识别的声音信号。除了人机之间的自然语言通信之外,HAL 也应该具备一些与语言相关的智能处理行为,如查询资料、解答问题、摘录文献、翻译材料等。

事实上,像 HAL 这样的智能机器人是一种高级的计算机,与早期的计算机只可以处理计算机语言不同,它可以处理(理解、使用)人类的语言。让机器可以处理人类的语言,这样,人们就可以使用自然语言与计算机进行通信(交流),这是人们长期以来所追求的。人们可以用自己最习惯的语言来使用计算机,而无需再花大量的时间和精力去学习不很自然和习惯的各种计算机语言。实现人-机之间直接通过自然语言(声音、文字)或图形图像交换信息是下一代计算机(第五代计算机)的主要研制目标。

尽管 *A Space Odyssey* 只是一部科幻片,但是 HAL 所需要的一些与语言相关的技术现在已经研制出来了,并且有一部分技术已经商品化了。我们把这些统称为自然语言处理(Natural Language Processing,NLP)。自然语言处理是人工智能领域的重要内容,研究用电子计算机模拟人的语言交际过程,使计算机能理解和运用人类社会的自然语言,实现人机之间的自然语言通信,以代替人的部分脑力劳动,包括查询资料、解答问题、摘录文献、汇编资料及一切有关自然语言信息的加工处理。

与自然语言处理密切相关的另一个概念就是"计算语言学(Computational Linguistics)",它

是语言学的一个分支，专指利用电子计算机进行语言研究。

1.2 自然语言处理的研究内容

实现人机间自然语言通信意味着要使计算机既能理解它所接收到的自然语言信息的意义，也能以自然语言表达给定的意图、思想等。前者称为自然语言理解(Natural Language Understanding)，后者称为自然语言生成(Natural Language Generation)。因此，NLP 大体包括了自然语言理解和自然语言生成两个部分。自然语言理解把自然语言转化为计算机程序更易于处理的形式，自然语言生成把计算机数据转化为自然语言。历史上对自然语言理解研究得较多，而对自然语言生成研究得较少，但这种状况近年来已有所改变。

自然语言理解主要包括两个方面：

(1)语言信息的录入。

语言信息的录入具体包含两个部分：语音信息的录入和文字信息的录入。语音信息的录入是指将输入计算机的语音信号识别转换成书面语表示，这一过程称为语音识别(Speech Recognition)；文字信息的录入包括键盘输入、手写输入和印刷体输入。

(2)文本理解。

从某种意义上来说，NLP 的最终目的应该是在语义理解的基础上实现相应的操作。我们知道，一个句子的语义主要是由其核心谓语动词决定的。一般来说，一个句子中除了核心谓语动词以外，还包含一些名词性成分，我们称之为实体(Entity)。语义分析的本质就是确定这些名词性成分与核心谓语动词之间的关系。这些关系在语言学中称之为“格(Case)”(Fillmore，1966)，包括施事格、受事格、时间格、处所格、工具格、来源格、结果格、目标格等。因此，对于一个复杂的句子，要想分析其句子的含义，首先要识别出句子中的核心谓语动词以及各名词性成分；而要想识别出这些成分，需要分析句子的结构(分析句子结构的过程称为“句法分析”)；而要想分析句子的结构，又需要先确定句子中各个单词的词性/词类(确定句子中各单词类型的过程称为“词法分析”)。可见，文本分析(理解)的过程可以具体划分为词法分析、句法分析和语义分析三个步骤。

例如，对于英语句子“In the room, he broke a window with a hammer.”，通过词法分析，我们可以得到与这个句子对应的词性序列，如图 1.1 所示。

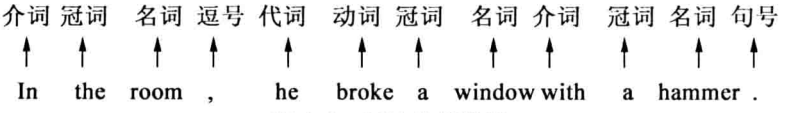


图 1.1 词法分析举例

接下来，根据英语的句法知识，我们可以分析这个句子的结构，如图 1.2 所示。根据这个句法分析结果，我们可以识别出句子中的核心谓语动词 break 以及各名词性成分。

接下来，我们需要根据语义规则来分析这个句子的语义。在语言学中，一般用“格语法(Case Grammar)”(Fillmore，1971，1975)来表示语法体系深层结构中的语义概念。格语法是 1966 年由美国语言学家菲尔摩(C. Fillmore)提出的一种语言学理论。所谓格语法，简单地说，就是带有格的语法规则。例如，如果英语的格语法中存在如图 1.3 所示的这样一条语义规则，那么，我们就可以分析出前面这个英语句子的含义。即，这个句子叙述的核心事件是“break

(打)"这件事,其中实施这件事情的施事者是"he",被"打"的受事者是"a window",实施这件事所用的工具是"a hammer",这件事发生的处所是"in the room"。

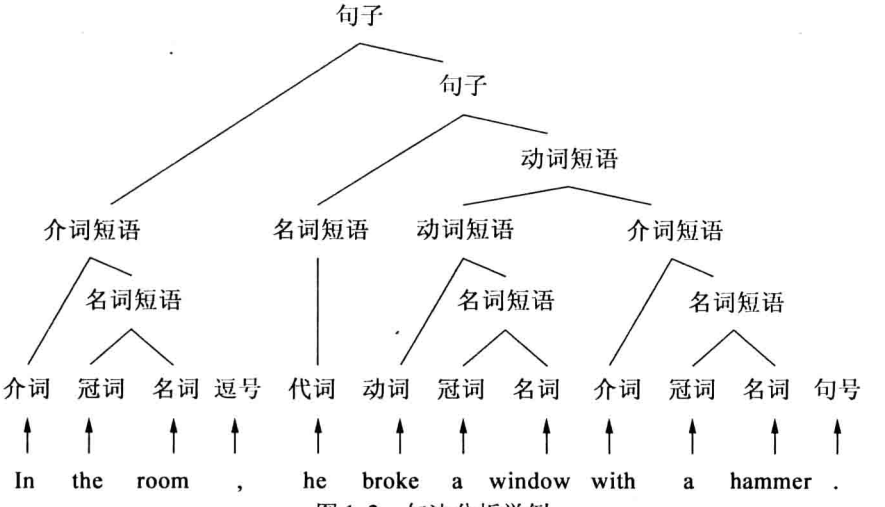


图 1.2 句法分析举例

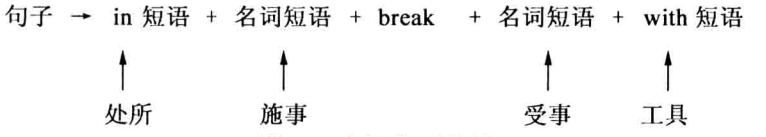


图 1.3 语义规则举例

应用格语法进行句子的语义分析,分析结果可以用"格框架(Case Frame)"来表示。例如,上面的分析结果可以表示为:

```
[ break
  [ case - frame
    agent:he
    object:a window
    locative:in the room
    instrument:a hammer
  ]
  [ modals
    time: past
    voice: active
  ]
]
```

自然语言生成的工作过程与自然语言理解相反,是从抽象的概念层次开始,通过选择并执行一定的语义和语法规则生成文本。如果是一个语音系统,那么还需要将书面文本自动转换成对应的语音表征,这一过程称为语音合成(Speech Synthesis)。

一般来说,NLP 研究的内容可以归纳为如图 1.4 所示的几个层次:

应用研究
机器翻译、信息检索、自动问答、文本校对、 自动摘要、辅助写作、语音识别、语音合成……
应用基础研究
英语形态分析、汉语自动分词、词性标注、短语切分、 名实体识别、句法分析、语义分析、篇章理解……
基础研究
字符集的编码体系、语言计算模型、 资源建设(语科库、知识库)……

图 1.4 NLP 研究内容的层次划分

(1)基础研究。基础研究包括字符集的编码体系、语言计算模型、资源建设(语料库、知识库)等。其中,字符集的编码体系解决了文字信息的存储和交换问题;语言计算模型把 NLP 作为语言学的分支来研究,它只研究语言及语言处理与计算相关的方面,而不管其在计算机上的具体实现,其最重要的基础是语法形式化理论和数学理论;NLP 系统离不开语料库和知识库等资源的支持,因此,资源建设也是 NLP 的主要研究内容之一。

(2)应用研究。应用研究包括键盘输入、文字识别、语音识别、语音合成、机器翻译、信息检索、信息抽取、信息挖掘、信息过滤、信息推送、信息标引、自动问答、自动摘要、文本分类、文本校对、辅助写作等。

(3)应用基础研究。尽管不同的 NLP 系统可能是千差万别的,但是它们在文本理解阶段所完成的任务是大致相同的,即词法分析(包括英语形态分析、汉语自动分词、词性标注、名实体识别等)、句法分析和语义分析等,以上内容称为应用基础研究。

1.3 自然语言处理的应用领域

NLP 的应用面非常广泛,涉及文化、教育、军事、医疗、商业、政务、社交等各个领域。尤其是 21 世纪以来,由于国际互联网的普及,网络已成为人们获取知识和信息的重要手段。生活在信息网络时代的现代人,几乎都要与网络打交道,都要或多或少地使用 NLP 的研究成果从广阔无边的互联网上获取或挖掘各种知识和信息。

以下举一些我们身边典型的例子。

1. 文化教育

数字图书馆:数字图书馆是一项全新的社会事业,它借鉴图书馆的资源组织模式,借助计算机和网络通信等高新技术,以普遍存取人类知识为目标,创造性地运用知识分类和精准检索手段,有效地进行信息整序,使人们获取信息时不受空间限制,很大程度上也不受时间限制。数字图书馆的建设过程中,会用到 NLP 中的文本分类、信息标引、自动文摘等技术。其中,文本分类技术根据图书分类法,对文献进行自动分类;信息标引自动给出文本的主题词,包括抽词标引和赋词标引两种;自动文摘根据不同比例及用户的不同需求自动编写文摘。

远程教育:远程教育的自动答疑系统会用到 NLP 中的自动问答技术,系统根据用户的问题收集教材中的相关内容,汇总后提供给用户。NLP 技术还可以帮助进行学生情况调查分析,

根据学生的提问情况，自动分析学生的主要问题所在，以便对症下药地改进教学内容（刘挺，2007）。

自动判卷系统：让计算机阅读数百篇典型的大学生论文并给这些论文打分，计算机的打分结果与人的打分结果几乎毫无区别，难以分辨（Landauer et al.，1997）。

自动阅读家庭教师：让计算机充当自动阅读家庭教师，帮助改善阅读能力。它能教小孩阅读故事，当阅读人出现阅读错误时，计算机能使用语音识别器来进行干预（Mostow and Aist，1999）。

智能解说与体育新闻实时解说：给计算机装上图像识别系统，它就可以观看一段足球比赛的录像，并用自然语言报告比赛的情况（Washlster，1989）。

新闻定制：根据用户的兴趣偏好，为用户定制新闻。

2. 医疗

聊天机器人：由系统工程师约瑟夫·魏泽堡和精神病学家肯尼斯·科尔比在20世纪60年代共同开发的Elisa系统是世界上第一个真正意义上的聊天机器人。许多心理学家和医生都想请它为人进行心理治疗，一些病人在与它谈话后，对它的信任甚至超过了人类医生（Weizenbaum，1966）。

残疾人智能帮助系统：对于有言语或交际障碍的残疾人，计算机能预见下面将要出现的词语，给他们作出提示，或者当他们说话时帮助在词语方面进行扩充，使残疾人能完整地说出简洁的话语（Newell et al.，1998；McCoy et al.，1998）.

3. 商务

自助呼叫中心：以自动问答的方式，从企业提供的大量技术支持资料中自动获取答案，满足用户的需求，减少呼叫中心的人力服务费用。

用户投诉信的自动分类和汇总系统：将用户的投诉信自动分发给企业的不同部门去处理，自动发现投诉信中的焦点问题，协助企业决策。

4. 政务

政务自动咨询系统：市民通过互联网，以问答的方式咨询政府的政策和办事流程等。

投诉自动汇总分析系统：将市民的投诉自动分类汇总，以供政府决策。

首长办公系统：自动汇总来自各下属部门的文件，并提取重要内容提供给领导阅读。

行政简报自动编写系统：定期自动编写简报，在政府部门内交流。

5. 公共设施

天气预报播报系统：早在1976年，加拿大就研制出了天气预报播报系统。计算机程序能够接受每天的天气预报数据，然后自动生成天气预报报告，不用经过进一步编辑就可以用英语和法语公布（Chandioux，1976）。

餐饮查询系统：到美国马萨诸塞州坎布里奇市的访问者可以用口语问计算机在什么地方可以吃饭，系统查询一个关于当地饭店的数据库之后，会给出相关信息作为回答（Zue et al. 1991）。

6. 内容安全

垃圾邮件（短信）过滤：包括广告、色情和反动邮件（短信）的过滤和分析。

企业商业秘密防泄露：监测从企业内部发出的邮件，封杀包含企业机密的邮件。

聊天室和BBS监控：过滤黄色话题或反动言论。

7. 移动计算

海量短信的自动化处理：电视台或广播电台常常提供在线的短信参与活动，大量短信发送到电视台需要及时地分类汇总，以便主持人作出反应，比如概括出大多数用户最关心的问题等。

8. 社交网络

微博数据挖掘；社交网络数据分析（影响力分析，偏好、兴趣建模，社区发现等）；舆情分析，观点挖掘，新事件发现等。

1.4 自然语言处理中用到的知识

自然语言处理是一门融语言学、计算机科学、数学、心理学、逻辑学、声学于一体的科学，而以语言学为基础。理解自然语言，需要关于外在世界的广泛知识以及运用操作这些知识的能力。自然语言理解的研究，综合应用了现代语音学、音系学、词法学、句法学、语义学、语用学的知识。

1. 语音学知识

为了确定 Dave 讲什么，HAL 必须能够分析它所接收的声音信号，并把 Dave 的这些信号复原成词的序列。与此相似，为了生成回答，HAL 必须把它的回答组织成词的序列，并且生成 Dave 能够识别的声音信号。要完成这两方面的任务，需要语音学（Phonetics）和音系学（Phonology）的知识，这样的知识可以帮助我们建立词如何在话语中发音的模型。

2. 词法学知识

值得注意的是，HAL 还能说出类似“I'm”和“can't”这样的缩略形式，并且还能识别并产生单词这样或那样的变体（例如，识别 doors 是复数）。这些与形态特征相关的知识都属于词法学知识。这些知识能够反映关于上下文中词的形态和行为的有关信息。

以英语为例，英语词汇由两部分构成：词干（stem）和词缀（affix），词干是单词中不可缺少的部分，有些词干可以独立成词。词缀分为前缀（prefix）和后缀（suffix）。从语素构成单词的方法可以分为两大类（可能部分地交叉）：屈折（inflection）和派生（derivation）。

屈折把词干和一个词缀结合起来，所形成的单词一般与原来的词干属于同一个词类。英语的屈折系统相对简单，只有名词、动词和部分形容词有屈折变化。

英语的名词只有两个屈折变化：一个词缀表示复数（plural），一个词缀表示领属（possessive）。大多数名词使用规则复数，拼写时在名词后面加 s；在以-s、-z、-sh、-ch、-x 结尾的名词词干后面加 es（例如，ibis/ibises、waltz/waltzes、thrush/thrushes、finch/finches、box/boxes）；以-y 结尾的名词，当-y 前面是一个辅音时，把-y 改为 i（例如，butterfly/butterflies）。对于领属后缀，规则单数名词和不以 -s 结尾的复数名词是通过加 's 实现的（例如，llama's、children's）；在规则复数名词后面以及某些以-s 或-z 结尾的人名后面通常只加'（例如，llamas'、Euripides' comedies）。

对于英语中的规则动词，只要知道了词干，就能预见到它的其他形式，在词干后面分别加上三个可预见的词尾 s，ing，ed，然后再进行某些有规律的拼写变化。在加后缀 ing 和 ed 时，前面的单独辅音字母要重叠（例如，beg/begged/begging）。如果最后一个字母是-c，则其重叠形式拼写为 ck（例如，picnic/picnicking）。正如在名词中那样，在以-s、-z、-sh、-ch、-x 结尾的动词