NLP，让人类与智能机器的交互不再遥远；
深度学习，让语言解析不再是智能系统的瓶颈。

NLP汉语自然语言处理
原理与实践

郑捷◎著

围绕三个部分展开

自然语言理论、人工智能算法、算法实现和案例部分

多达15个算法的讲解

包含NShort、HMM、朴素贝叶斯、CRF、BP神经网络等

多达9个程序库的剖析

包括Python NLTK、Ltp3.X、HanLP、Word2Vec、CRF++、Keras等

中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

NLP 汉语自然语言处理原理与实践

郑　捷◎著

電子工業出版社
Publishing House of Electronics Industry
北京 · BEIJING

内容简介

本书是一本研究汉语自然语言处理方面的基础性、综合性书籍，涉及 NLP 的语言理论、算法和工程实践的方方面面，内容繁杂。

本书包括 NLP 的语言理论部分、算法部分、案例部分，涉及汉语的发展历史、传统的句法理论、认知语言学理论。需要指出的是，本书是一本系统介绍认知语言学和算法设计相结合的中文 NLP 书籍，并从认知语言学的视角重新认识和分析了 NLP 的句法和语义相结合的数据结构。这也是本书的创新之处。

本书适用于所有想学习 NLP 的技术人员，包括各大人工智能实验室、软件学院等专业机构。

图书在版编目（CIP）数据

NLP 汉语自然语言处理原理与实践 / 郑捷著. —北京：电子工业出版社，2017.1

ISBN 978-7-121-30765-2

Ⅰ. ①N… Ⅱ. ①郑… Ⅲ. ①汉语－自然语言处理－研究 Ⅳ. ①TP391

中国版本图书馆 CIP 数据核字（2016）第 321878 号

策划编辑：李　冰
责任编辑：李　冰
特约编辑：田学清　罗树利等
印　　刷：涿州市京南印刷厂
装　　订：涿州市京南印刷厂
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　　邮编：100036
开　　本：787×1092　1/16　印张：34　字数：816 千字
版　　次：2017 年 1 月第 1 版
印　　次：2017 年 1 月第 1 次印刷
定　　价：98.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：libing@phei.com.cn。

推荐序

自然语言处理是人工智能领域的一颗明珠，现在已经成为人工智能研究中最为活跃的领域。几十年来，随着计算机技术和人工智能技术的发展，自然语言处理取得了长足的进步。现在，自然语言处理技术正处在一个新的历史转折点，随着可获取信息量的爆炸性增长，信息过载问题愈发严重，以词法分析和词义理解为主的传统自然语言处理技术已经难以满足解决实际问题的需要，句子级乃至篇章级语义理解技术即将成为人工智能技术发展的新趋势。

自然语言处理作为人工智能与语言学的综合学科，理应从两个学科中汲取营养来推动自身的进步。但目前概率和数据驱动的方法在自然语言处理领域占据绝对的主流，加之近几年深度学习的异军突起，语言学知识在自然语言处理领域中受到的重视程度愈发不足。而以我在自然语言处理领域工作的经验来看，越深入研究，越能感觉到语言学知识不足的掣肘。特别是深层次的语义理解，脱离了语言学知识，就会变成无源之水、无本之木。常见的自然语言处理书籍对于解决具体问题的方法讲解已经足够丰富，但对于语言学基础理论的介绍和思考还略显不足。一些前辈虽然一直在思考语言和认知的本质，但其发表出来的内容只限于思考结果的一鳞半爪，较少结集成书。加之现在自然语言处理领域的学习者大多是计算机背景，极少系统地学习过语言学的基础理论。这样造成的现状就是从事自然语言处理的技术人员越来越多，但相互之间的讨论和经验分享多集中于具体的技术手段或算法的数学原理，而极少涉及语言学的基础理论和语义理解的本质问题。

本书作者通过对前人语言学理论和自然语言处理技术的深入梳理，形成了自己对于语义理解，特别是汉语语义理解独特的思考和一整套理论体系，提出了语义理解的系统解决之道。尽管如何才能让计算机理解语义，在学术界还没有定论，但作者系统性的思考和解决思路是非常难能可贵的。本书在内容上保证了理论和技术的平衡，在介绍术的同时，充分展示了作者对于道的思考成果。此书是自然语言处理书籍中的一股新风，希望其可以对语义理解的研究和发展起到积极的推动作用，同时引导自然语言处理领域的研究者，特别是初学者，加强对于语言学的理论的学习，更多地从问题的本源来寻求新的解决思路，而不仅仅满足于在传统解决思路上尝试新的技术手段。

愿每一位有志于从事自然语言处理的研究者，都能从此书中获得一些启示。

贾文杰：早年在富士通研发中心，著名的 1998 年人民日报语料库的研发单位之一，任高级研究员，负责情感分析，后进入 360 搜索引擎自然语言处理部，项目核心成员之一，主持搜索引擎分词，纠错等核心模块研发工作，历时 3 年，对搜索效果的提升起到重要作用。目前，转入移动互联领域，负责猎豹移动的自然语言处理部，任负责人。

前　言

写作本书的动机

自然语言处理（Natural Language Processing，NLP）是人工智能和语言学领域的分支学科，主要研究如何让计算机处理和运用自然语言。自然语言处理广义上分为两大部分，第一部分为自然语言理解，是指让电脑“懂”人类的语言；第二部分为自然语言生成，是指把计算机数据转化为自然语言。本书重点讲解汉语自然语言处理方面的最新理论、技术和进展。

自然语言处理作为一个独立的学科诞生至今，已经半个多世纪了。与绝大多数传统学科的最大不同是，在这半个世纪中，它始终离问题的终结遥遥无期，当人们千辛万苦地获得一次又一次的突破后，又会被新出现的问题无情地阻拦，而再次陷入迷惘之中。在 NLP 中，问题好像没有最终解决方案，甚至连最佳实践也没有，而只有最新现状（State of art）。而近些年，那些历史上的 State of art 正被不断地刷新、不断地超越。

就在十多年前，商业化的人机交互都是人们可望而不可即的目标，但现在智能机器人正逐渐走入市场，走入人们的生活。虽然这些技术还不够成熟，还要解决诸多问题，即便普通大众也能意识到，我们离人工智能的终极目标越来越近了。

面对市场上诸多的人工智能系统，以及背后的各种算法理论，使我想起了一部获奖的英国电影《模仿游戏》。这不是一部艺术上的 State of art，却赢得了第 87 届奥斯卡金像奖最佳改编剧本奖。在肯定这部作品的诸多因素中，我认为最重要的是，它宣誓了现阶段人工智能的本质：模仿。这也是本书自始至终贯穿的主题：模仿→象似性→算法理论。

但从另一个角度，我们希望能够终结一些问题，即便这些问题还未得到百分之百的解决（当然，从概率论的角度而言，没有百分之百），否则，我们很难进入以下阶段的研究，整个学科只会停滞不前。幸运的是，近些年，在序列标注上的全面突破，使我们有幸将目光放到了句子的范畴，最近提出的语义依存理论，更使汉语自然语言处理，无论理论还是实践都迎来了新的曙光。汉语的句子分析，终于跨越了句法的误区，走向了语义解析的道路。相信不久的将来，在语义解析的道路上，汉语 NLP 将会获得更大的突破。

本书的受众与特色

本书是一本研究汉语自然语言处理方面的基础性、综合性书籍，涉及 NLP 的语言理论、算法和工程实践的方方面面，内容繁杂。为此，我们设定本书的读者为如下几种：

- 具有一定计算机编程基础，对自然语言处理感兴趣的非专业人员。
- 希望构建完整的 NLP 应用系统的专业工程技术人员。
- 高校计算机专业和自然语言处理专业的大学生、研究生。
- 高校自然语言处理专业的教师。

需要指出的是，本书是一本系统介绍认知语言学和算法设计相结合的中文 NLP 书籍，并从认知语言学的视角重新认识和分析了 NLP 的句法和语义相结合的数据结构。这也是本书的创新之处。

内容及体系结构

为兼顾各方面的需求，我们对全书各部分做了精心的安排。从结构上，全书分为如下三大部分。

（1）语言理论部分：涉及 4 个章节，第 2 章为汉语的发展历史；第 6 章为传统的句法理论；第 7 章为语料库和知识库的构建理论；第 8 章为认知语言学理论。

（2）算法部分：涉及 4 个章节，第 3 章为中文分词算法；第 4 章为 NLP 中的概率图模型算法体系；第 6 章为句法的自动分析算法，包括转换生成语法的算法原理，以及依存句法的应用；第 9 章系统介绍了神经网络到深度学习算法体系，以及使用 LSTM 实现序列标注和依存句法。本书介绍的算法都提供开源的代码，具体下载地址已在每章介绍算法的时候指出，读者可参考书籍和网址的讲解内容进行调试，快速应用于实践中。

（3）案例部分：涉及 4 个章节，第 1 章为开源 NLP 系统概览及入门代码；第 5 章为使用概率图模型算法进行词性标注、语义组块、命名实体识别等序列标注；第 9 章为使用 Word2Vec 的训练词向量模型；第 10 章为使用 SVM 进行长句切分、使用语义角色标注分析汉语句子等。

基本上每段理论讲解之后都辟出专门的案例讲解，以加深理论认识。对于重要的理论，甚至开辟专门的章节讲解其实现。案例分为两大部分，一部分是程序代码，读者可以参考书中的代码，将其直接应用到实践中；另一部分是语料，读者可以按书中指定的网络链接下载。

目 录

第 1 章 中文语言的机器处理.....1

1.1 历史回顾.....2

1.1.1 从科幻到现实.....2

1.1.2 早期的探索.....3

1.1.3 规则派还是统计派.....3

1.1.4 从机器学习到认知计算.....5

1.2 现代自然语言系统简介.....6

1.2.1 NLP 流程与开源框架....6

1.2.2 哈工大 NLP 平台及其演示环境.....9

1.2.3 Stanford NLP 团队及其演示环境.....11

1.2.4 NLTK 开发环境.....13

1.3 整合中文分词模块.....16

1.3.1 安装 Ltp Python 组件...17

1.3.2 使用 Ltp 3.3 进行中文分词.....18

1.3.3 使用结巴分词模块.....20

1.4 整合词性标注模块.....22

1.4.1 Ltp 3.3 词性标注.....23

1.4.2 安装 StanfordNLP 并编写 Python 接口类.....24

1.4.3 执行 Stanford 词性标注..... 28

1.5 整合命名实体识别模块..... 29

1.5.1 Ltp 3.3 命名实体识别.. 29

1.5.2 Stanford 命名实体识别..... 30

1.6 整合句法解析模块..... 32

1.6.1 Ltp 3.3 句法依存树..... 33

1.6.2 Stanford Parser 类..... 35

1.6.3 Stanford 短语结构树... 36

1.6.4 Stanford 依存句法树... 37

1.7 整合语义角色标注模块..... 38

1.8 结语..... 40

第 2 章 汉语语言学研究回顾.....42

2.1 文字符号的起源..... 42

2.1.1 从记事谈起..... 43

2.1.2 古文字的形成..... 47

2.2 六书及其他..... 48

2.2.1 象形..... 48

2.2.2 指事..... 50

2.2.3 会意..... 51

2.2.4 形声..... 53

2.2.5 转注 54
2.2.6 假借 55
2.3 字形的流变 56
2.3.1 笔与墨的形成与变革 ...56
2.3.2 隶变的方式 58
2.3.3 汉字的符号化与结构 ...61
2.4 汉语的发展 67
2.4.1 完整语义的基本形式——句子 68
2.4.2 语言的初始形态与文言文 71
2.4.3 白话文与复音词 73
2.4.4 白话文与句法研究 78
2.5 三个平面中的语义研究 80
2.5.1 词汇与本体论 81
2.5.2 格语法及其框架 84
2.6 结语 86
第 3 章 词汇与分词技术 88
3.1 中文分词 89
3.1.1 什么是词与分词规范 ...90
3.1.2 两种分词标准 93
3.1.3 歧义、机械分词、语言模型 94
3.1.4 词汇的构成与未登录词 97
3.2 系统总体流程与词典结构 98
3.2.1 概述 98
3.2.2 中文分词流程 99
3.2.3 分词词典结构 103
3.2.4 命名实体的词典结构 105
3.2.5 词典的存储结构 108
3.3 算法部分源码解析 111
3.3.1 系统配置 112
3.3.2 Main 方法与例句 113
3.3.3 句子切分 113
3.3.4 分词流程 117
3.3.5 一元词网 118
3.3.6 二元词图 125
3.3.7 NShort 算法原理 130
3.3.8 后处理规则集 136
3.3.9 命名实体识别 137
3.3.10 细分阶段与最短路径 140
3.4 结语 142
第 4 章 NLP 中的概率图模型 143
4.1 概率论回顾 143
4.1.1 多元概率论的几个基本概念 144
4.1.2 贝叶斯与朴素贝叶斯算法 146
4.1.3 文本分类 148
4.1.4 文本分类的实现 151
4.2 信息熵 154

4.2.1 信息量与信息熵 154
4.2.2 互信息、联合熵、条件熵 156
4.2.3 交叉熵和 KL 散度 158
4.2.4 信息熵的 NLP 的意义 159
4.3 NLP 与概率图模型 160
4.3.1 概率图模型的几个基本问题 161
4.3.2 产生式模型和判别式模型 162
4.3.3 统计语言模型与 NLP 算法设计 164
4.3.4 极大似然估计 167
4.4 隐马尔科夫模型简介 169
4.4.1 马尔科夫链 169
4.4.2 隐马尔科夫模型 170
4.4.3 HMMs 的一个实例 171
4.4.4 Viterbi 算法的实现 176
4.5 最大熵模型 179
4.5.1 从词性标注谈起 179
4.5.2 特征和约束 181
4.5.3 最大熵原理 183
4.5.4 公式推导 185
4.5.5 对偶问题的极大似然估计 186
4.5.6 GIS 实现 188
4.6 条件随机场模型 193
4.6.1 随机场 193
4.6.2 无向图的团（Clique）与因子分解 194
4.6.3 线性链条件随机场 195
4.6.4 CRF 的概率计算 198
4.6.5 CRF 的参数学习 199
4.6.6 CRF 预测标签 200
4.7 结语 201
第 5 章 词性、语块与命名实体识别 202
5.1 汉语词性标注 203
5.1.1 汉语的词性 203
5.1.2 宾州树库的词性标注规范 205
5.1.3 stanfordNLP 标注词性 210
5.1.4 训练模型文件 213
5.2 语义组块标注 219
5.2.1 语义组块的种类 220
5.2.2 细说 NP 221
5.2.3 细说 VP 223
5.2.4 其他语义块 227
5.2.5 语义块的抽取 229
5.2.6 CRF 的使用 232
5.3 命名实体识别 240
5.3.1 命名实体 241
5.3.2 分词架构与专名词典 243

5.3.3 算法的策略——词典与统计相结合 245
5.3.4 算法的策略——层叠式架构 252
5.4 结语 259
第 6 章 句法理论与自动分析.....260
6.1 转换生成语法 261
6.1.1 乔姆斯基的语言观 261
6.1.2 短语结构文法 263
6.1.3 汉语句类 269
6.1.4 谓词论元与空范畴 274
6.1.5 轻动词分析理论 279
6.1.6 NLTK 操作句法树 280
6.2 依存句法理论 283
6.2.1 配价理论 283
6.2.2 配价词典 285
6.2.3 依存理论概述 287
6.2.4 Ltp 依存分析介绍 290
6.2.5 Stanford 依存转换、解析 293
6.3 PCFG 短语结构句法分析 298
6.3.1 PCFG 短语结构 298
6.3.2 内向算法和外向算法 301
6.3.3 Viterbi 算法 303
6.3.4 参数估计 304
6.3.5 Stanford 的 PCFG 算法训练 305
6.4 结语 310
第 7 章 建设语言资源库 311
7.1 语料库概述 311
7.1.1 语料库的简史.....312
7.1.2 语言资源库的分类.....314
7.1.3 语料库的设计实例：国家语委语料库.....315
7.1.4 语料库的层次加工.....321
7.2 语法语料库 323
7.2.1 中文分词语料库.....323
7.2.2 中文分词的测评.....326
7.2.3 宾州大学 CTB 简介...327
7.3 语义知识库 333
7.3.1 知识库与 HowNet 简介.....333
7.3.2 发掘义原.....334
7.3.3 语义角色.....336
7.3.4 分类原则与事件分类.....344
7.3.5 实体分类.....347
7.3.6 属性与分类.....352
7.3.7 相似度计算与实例.....353
7.4 语义网与百科知识库 360
7.4.1 语义网理论介绍.....360
7.4.2 维基百科知识库.....364
7.4.3 DBpedia 抽取原理 365
7.5 结语 368

第 8 章 语义与认知.....370
8.1 回顾现代语义学.....371
8.1.1 语义三角论.....371
8.1.2 语义场论.....373
8.1.3 基于逻辑的语义学.....376
8.2 认知语言学概述.....377
8.2.1 象似性原理.....379
8.2.2 顺序象似性.....380
8.2.3 距离象似性.....380
8.2.4 重叠象似性.....381
8.3 意象图式的构成.....383
8.3.1 主观性与焦点.....383
8.3.2 范畴化：概念的认知.....385
8.3.3 主体与背景.....390
8.3.4 意象图式.....392
8.3.5 社交中的图式.....396
8.3.6 完形：压缩与省略.....398
8.4 隐喻与转喻.....401
8.4.1 隐喻的结构.....402
8.4.2 隐喻的认知本质.....403
8.4.3 隐喻计算的系统架构.....405
8.4.4 隐喻计算的实现.....408
8.5 构式语法.....412
8.5.1 构式的概念.....413
8.5.2 句法与构式.....415
8.5.3 构式知识库.....417
8.6 结语.....420
第 9 章 NLP 中的深度学习.....422
9.1 神经网络回顾.....422
9.1.1 神经网络框架.....423
9.1.2 梯度下降法推导.....425
9.1.3 梯度下降法的实现.....427
9.1.4 BP 神经网络介绍和推导.....430
9.2 Word2Vec 简介.....433
9.2.1 词向量及其表达.....434
9.2.2 Word2Vec 的算法原理.....436
9.2.3 训练词向量.....439
9.2.4 大规模上下位关系的自动识别.....443
9.3 NLP 与 RNN.....448
9.3.1 Simple-RNN.....449
9.3.2 LSTM 原理.....454
9.3.3 LSTM 的 Python 实现.....460
9.4 深度学习框架与应用.....467
9.4.1 Keras 框架介绍.....467
9.4.2 Keras 序列标注.....471
9.4.3 依存句法的算法原理.....478
9.4.4 Stanford 依存解析的训练过程.....483

9.5 结语 488

第 10 章 语义计算的架构 490

10.1 句子的语义和语法预处理 ... 490

10.1.1 长句切分和融合 491

10.1.2 共指消解 496

10.2 语义角色 502

10.2.1 谓词论元与语义角色 502

10.2.2 PropBank 简介 505

10.2.3 CPB 中的特殊句式 506

10.2.4 名词性谓词的语义角色 509

10.2.5 PropBank 展开 512

10.3 句子的语义解析 517

10.3.1 语义依存 517

10.3.2 完整架构 524

10.3.3 实体关系抽取 527

10.4 结语 531

第 1 章

中文语言的机器处理

自然语言处理（Natural Language Processing，NLP）是研究能实现人与计算机之间用自然语言进行有效通信的各种理论和方法，也是人工智能领域中一个最重要、最艰难的方向。说其重要，因为它的理论与实践与探索人类自身的思维、认知、意识等精神机制密切相关；说其艰难，因为每一项大的突破都历经十年乃至几十年以上，要耗费几代人的心血。

近些年，NLP 在中文分词、词性标注、词汇语义、句法解析方面均获得了很大的突破。大量的技术都应用于商业实践，并在商业领域获得了良好的市场和经济效益。文本方面主要有：基于自然语言理解的智能搜索引擎和智能检索、智能机器翻译、自动摘要与文本综合、文本分类与文件整理、自动阅卷系统、信息过滤与垃圾邮件处理、文学研究与古文研究、语法校对、文本数据挖掘与智能决策、基于自然语言的计算机程序设计等。语音方面主要有：机器同声传译、智能客户服务、聊天机器人、语音挖掘与多媒体挖掘、多媒体信息提取与文本转化、对残疾人智能帮助系统等。

这使得从事研究这个行业的科技人员越来越多，本书的目的之一就是为使这些朋友尽快进入这个领域而降低门槛。为此，笔者不想在开篇谈那些艰深的理论知识和数学原理，这些沉重的主题还是留在后面为读者慢慢道来。我们先简要回顾一下 NLP 的历史，然后通过几个开源系统，领略一下当代 NLP 的风采。

1.1 历史回顾

世界上早期的科幻小说作者，英国诗人雪莱的夫人玛丽·雪莱，于 1818 年创作了科幻小说《弗兰肯斯坦——现代普罗米修斯的故事》。小说讲的是，一个叫作弗兰肯斯坦的年轻科学家希望利用所学的生物学知识，制造出一个类人生物。在强烈的科学探索欲望的驱使下，他从停尸房等处取得不同人体的器官和组织，拼合成一个人体，并利用雷电的电能激活了这个人造的生命。经过电击，人造人瞬间被弗兰肯斯坦赋予了生命。刚刚诞生的人造人天性善良并向往美好。不过，由于相貌丑陋，社会并不接纳它，并将其视作怪物。因为他与社会的种种矛盾，终于导致怪物走上了报复之路。故事的结尾是悲惨的，弗兰肯斯坦因病而死，怪物也自焚消失。

1.1.1 从科幻到现实

这里想要探讨的并不是故事在文学上的价值，而是其在科学上的重要意义。自工业革命以来，在生产生活的各个领域，科技都走入了人们的视野：人们创造了铁路系统、公路系统以延伸人的双腿；创造了各种各样的机械臂装置来模仿人的双手；创造了五颜六色的纺织品模仿动物的皮毛；还发展了电影和广播系统来愉悦视听。两百年来，这部作品给我们留下一个发人深省的启示：似乎我们所有的科学技术活动都围绕着一个方向——人类如何创造而发展自身！

时间过得很快，这个问题第二次摆在我们面前的时候，从《弗兰肯斯坦》发表又过了一百多年。经过二次大战的洗礼，人类对自身的认识进入了一个新的阶段。在民主制度与专制制度、民族主义与殖民主义的斗争中，科学技术得到了空前的发展。1950 年，计算理论的先驱阿兰·麦席森·图灵写了一篇著名的论文——《计算机器与智能》，其内容是：如果电脑能在 5 分钟内回答由人类测试者提出的一系列问题，且其超过 30%的回答让测试者误认为是人类所答，则电脑通过测试（来自百度百科）。这就是著名的“图灵测试”。计算机专业的读者对于图灵测试应该并不陌生。这个测试想从实验的角度提出一个假设：“机器能与人类交流吗？”问题听起来似乎有点悬，我们换一种方式来重新描述一下这个问题：“有可能设计出具有类似人类智能的机器吗？”

虽然从现代的视野来看，图灵问题本身显得有些粗糙，但不得不承认，图灵问题的提出是人类科技的一个重要的里程碑。它揭开了科学幻想那遥不可即的面纱，将人工智能最重要的任务赤裸裸地摆在了人们的面前——所谓人工智能的终极任务就是人类要制造出具有人类语言和思考能力的机器。

1.1.2　早期的探索

计算机刚一诞生，人们就开始着手研究用它来解析人类的自然语言。这一需求不仅源于科学家的个人兴趣，而且具有重要的战略意义：20 世纪 50 年代开始，大家都意识到以美、苏两国为首的两大政治集团迟早要进入冷战时代。此时，美国就尝试着利用计算机将大量俄语资料自动翻译成英语，以窥探苏联科技的最新发展。虽然当时的计算机还在襁褓之中，但研究者从破译军事密码中得到启示，简单地认为语言之间的差异只不过是对“同一语义”的不同编码而已，从而想当然地采用译码技术解析不同的语言。这就是最早机器翻译理论的思想。

1954 年 1 月 7 日，美国乔治敦大学和 IBM 公司首先成功地将 60 多句俄语自动翻译成英语。当时的系统还非常简单，仅包含 6 个语法规则和 250 个词。但是，由于媒体的广泛报道，美国政府备受鼓舞，认为这是一个巨大的进步，长期发展将具有重要的战略意义。而实验者声称：

在三到五年之内就能够完全解决从一种语言到另一种语言的自动翻译问题。

当时普遍认为只要制定好各种翻译规则，通过大量规则的堆砌就能完美地实现语言间的自动翻译。1956 年，美国语言学家 N. Chomsky 从 Shannon 的工作中利用了有限状态马尔科夫过程的思想，首先把有限状态自动机作为一种工具来刻画语言的语法，并且把有限状态语言定义为由有限状态语法生成的语言。这些早期的研究工作产生了“形式语言理论”（Formal Language Theory）。它为最初的机器翻译工作提供了理论基础。

经过近十年的努力，机器翻译并未获得本质性的突破。1964 年美国科学院成立了语言自动处理咨询委员会（ALPAC），开始了为期两年的综合调查分析和测试。直到 1966 年年底，委员会公布了一个题为《语言与机器》的报告（简称 ALPAC 报告）。该报告全面否定了机器翻译的可行性，并建议停止对机器翻译项目的资金支持。这一报告的发表终结了自然语言处理的第一个时代——机器翻译时代。

1.1.3　规则派还是统计派

虽然机器翻译时代结束了，但自然语言处理这一新兴学科（NLP）却没有消亡。时间进入 20 世纪七八十年代后，随着经济发展特别是国际市场机制的成熟，国与国之间的语言障碍越来越成为更深层次国际交流的壁垒。传统的人工作业方式已经不能满足需求，这就需要一种自动机器来取代人工。同时，计算机硬件技术大幅度提高，使中等规模的语料（百万级）处理成为可能。经过十多年的发展，自然语言处理逐渐作为人工智能的

一个独立领域而发展起来，此时的自然语言处理也分为两种不同的派别。

一种是以语言学理论为基础，根据语言学家对语言现象的认识，采用规则形式描述或解释歧义行为或歧义特性，称为规则派。规则派的方法通常是基于乔姆斯基的语言理论的。它通过语言所必须遵守的一系列原则来描述语言，以此来判断一个句子是正确的（遵循语言原则）还是错误的（违反语言原则）。规则派首先要对大量的语言现象进行研究，归纳出一系列的语言规则。然后再形成一套复杂的规则集——语言分析或生成系统，对自然语言进行分析处理。

另一种是以基于语料库的统计分析为基础的经验主义方法，也称为统计派，该方法更注重用数学，从能代表自然语言规律的大规模真实文本中发现知识，抽取语言现象或统计规律。统计派来源于多种数学基础，包括通香农（Shannon）的信息论、最优化方法、概率图模型、神经网络、深度学习等。它将语言事件赋予概率，作为其可信度，由此来判断某个语言现象是常见的还是罕见的。统计派的方法则偏重于对语料库中人们实际使用的普通语言现象的统计表述。统计方法是语料库语言学研究的主要内容。

两派曾经一度相执不下。这里不考虑两派之间孰是孰非，而是希望通过一个著名的实验给大家一点启示，这个实验就是著名的约翰·赛尔的中文屋子实验。

一个对中文一窍不通的、以英语为母语的人被关闭在一间只有两个通口的封闭屋子中。屋子里有一本用英文写成、从形式上说明中文文字句法和文法组合规则的手册及一大堆中文符号。屋子外的人不断向屋子内递进用中文写成的问题。屋子内的人便按照手册的说明，将中文符号组合成对问题的解答，并将答案递出屋子。

约翰·塞尔认为，尽管屋子里的人甚至可以做到以假乱真，让屋子外的人以为他是中文的母语用户，然而，他压根就不懂中文。而在上述过程中，屋子外的人所扮演的角色相当于程序员，屋子中的人相当于计算机，而那本手册则相当于计算机程序。

正如屋子中的人不可能通过手册理解中文一样，计算机也不可能通过程序来获得对自然语言（中文）的理解能力。塞尔由此得出结论：图灵测试中机器根本不理解回答的问题，机器根本没有思考，机器也没有智能。（来自网络文摘）

塞尔的中文屋测试本来是针对图灵测试的一个反驳意见，但它所揭示的意义是深刻的。当时所谓的人工智能，特别是对自然语言处理领域的主要任务，不过是使用机器来解析人类的语言符号，将其转换为机器能够处理的形式和结构，在机器内部按照人们已经设定好的逻辑进行处理，最后将处理的结果再转码为人类理解的形式，传输给人类。这与大多数非智能的计算机程序没有本质的不同。