中文信息处理丛书

姚天顺
朱靖波
张　琍
杨　莹
编　著

自然语言理解
——一种让机器懂得人类语言的研究
第2版

清华大学出版社
http://www.tup.tsinghua.edu.cn

中文信息处理丛书

自然语言理解

——一种让机器懂得人类语言的研究

（第 2 版）

姚天顺　朱靖波
张　琍　杨　莹　编著

清华大学出版社

(京)新登字 158 号

内 容 简 介

自然语言理解是人工智能的一个重要分支，主要研究如何利用计算机来理解和生成自然语言。本书重点介绍了自然语言理解所涉及的各个方面，包括语法分析、语义分析、概念分析、语料库语言学、词汇语义驱动、中间语言、WordNet、词汇树邻接文法、链接文法、基于语段的机器翻译方法、内识别与文本过滤、机器翻译的评测等，既有对基础知识的介绍，又有对最新研究进展的综述，同时还结合了作者多年的研究成果。

本书可作为高等院校计算机、人工智能等专业的高年级本科生或研究生的教材及教学参考书，也可供从事中文信息处理、数据挖掘以及人工智能研究的相关人员参考。

书　　名：自然语言理解——一种让机器懂得人类语言的研究(第 2 版)
作　　者：姚天顺　朱靖波　张琍　杨莹　编著
出 版 者：清华大学出版社(北京清华大学学研大厦，邮编 100084)
　　http://www.tup.tsinghua.edu.cn
责任编辑：薛慧
印 刷 者：清华大学印刷厂
发 行 者：新华书店总店北京发行所
开　　本：787×1092 1/16　印张：30.5　字数：704 千字
版　　次：2002 年 10 月第 2 版　2002 年 10 月第 1 次印刷
书　　号：ISBN 7-302-05435-5/TP・3203
印　　数：0001～4000
定　　价：39.80 元

中文信息处理丛书编委会

中文信息处理丛书

序　言

中文信息处理技术在我国现代化及信息化建设中，越来越起着重要的作用，作为一个高新技术的重点，它已经列入国务院批准的“国家中长期科学技术发展纲领”。我国的中文信息处理事业正在不断向前推进，在技术研究、产品开发以及产业化发展等方面都取得了显著的成绩。现在有必要把这些方面的成果加以综合、提炼，以便更好地推广应用，并且作为一个起点，再上一个新台阶。中文信息处理在从汉字信息处理进入汉语信息处理之后，在从单机信息处理进入网络信息处理之后，已经面临着新的更大的挑战和机遇，需要我们重新对中文信息处理进行全面的审视与整合，这就是我们组织编写并出版这套中文信息处理丛书的目的。

在这套丛书出版之际，我愿向读者介绍以下几点：

第一，为什么我们要把中文信息处理技术作为高新技术的一个重点来发展呢？

语言文字是信息的首要载体。我们日常工作中的信息，绝大部分是以语言文字表达、记载、传播和交换的。因此随着计算机和因特网的推广应用，由数据处理、信息处理发展到知识处理，对语言文字处理要求的深度和广度越来越高，可以认为一个国家的语言文字的信息处理水平和处理量基本上代表了这个国家进入信息社会的程度，其语言文字信息的处理能力直接关系到它在网络社会和网络经济中的国际竞争能力。目前，网络社会和网络经济正以我们难以预料的速度在全世界发展，其阻碍发展的首要瓶颈问题就是自然语言的处理问题。网络社会也是人类社会，网络经济也是人类经济，需要以自然语言作为社会交际工具，一旦基于网络的自然语言处理问题得到突破，网络社会和网络经济将会突飞猛进。我们要在下一个世纪成为世界强国，就不能不把语言文字信息处理技术作为高新技术的一个重点来发展。在世界一流高新技术企业纷纷在中国设立“中国研究院”，争先把“中文信息处理”作为研究的重中之重的时候，我们当然要抢占中文信息处理这个高新技术发展的制高点。

第二，中文信息处理与印欧语系的语言信息处理的不同之处是什么？

计算机从诞生之日开始，就是以处理印欧语系的语言为基础的。换言之，计算机对于印欧语系的自然语言处理具有较好的支撑能力，计算机的推广应用在语言文字信息处理方面受到的阻力较小。我们的汉语却与印欧语系的语言差别很大，能够处理那些语言的计算机，面对汉语汉字，却显得无能为力。例如：

- 印欧语系为拼音文字，所使用的字符仅二十余个，而汉语是意音文字，常用的汉字就有六七千个，总数超过五万。这是一个根本性的问题。仅这一个差异就引起了处理汉语的计算机与处理印欧语言的计算机一系列的差异，需要我们自己去解决。包括键盘输入、汉字打印与显示、内部代码、汉字识别、程序语言的数据类型、

数据库的检索和排序等等。

- 印欧语系的书写，词与词之间有空格，而书面汉语的词与词之间无空格，于是词的机器自动切分问题就成了计算机处理汉语的首要问题。
- 印欧语系的同音词较少，而汉语的同音词较多。例如，仅在《现代汉语词典》中 JI 音汉字就有一百多个，辨析同音词就成了汉语语音处理的关键。
- 印欧语系多有形态变化（例如：复数、单数，过去、现在，阴性、阳性等等），而汉语缺少形态变化。计算机对汉语的处理（例如机器翻译、人机接口等）无法利用形态变化，只能在句法、语义上找出路。
- 汉语的语法研究尚未形成规范化，而且人们习惯于约定俗成的语法。于是语义研究显得尤其重要。例如，“吃饭”、“吃大碗”和“吃食堂”的理解只能靠语义来解决。
- 汉语的自动（计算机）处理是多学科和跨学科的研究工作，特别需要计算机科学与语言学、认知科学等学科的密切结合，而且要依靠长期积累的语言学的研究成果。但我国语言学界过去的研究多着重汉语教学，对象是人，而不是机器，因此对其丰硕的研究成果要经过改造、深化、量化、形式化，甚至要从头开始。要清醒地认识到面向机器的汉语研究的艰巨性，要持续不懈地抓下去。

以上只是几个突出的问题，还有一些其他问题，不再赘述。这些语言上的特点造成了计算机处理汉语的众多障碍，每前进一步都会遇到新问题，我们不得不花费比印欧语系的信息处理多得多的力量去解决。

再就计算机的发展趋势而言，计算机产业面临转型期，多媒体和笔记本式计算机成为热门产品，计算机从单机进入网络，网上的汉语信息处理正在成为强势和主流，并对语言文字的信息处理提出新的要求。这些产品的核心技术无不与中文信息处理技术有关。因此，加强中文信息处理的研究，取得网络化的自然语言处理的突破更为必要。

第三，中文信息处理技术包括哪些科目呢？

大体上包括下列一些科目：

- 词的切分和频率统计
- 汉语句型和短语的研究及频率统计
- 汉语语义的研究
- 键盘和非键盘汉字输入技术及处理系统
- 汉语语料库的开发及应用
- 汉字的机器代码，程序设计语言的数据类型
- 汉语开放系统的接口规范
- 语声输入与合成
- 汉字识别
- 字形生成
- 汉语分析及篇章理解
- 汉语生成
- 人机接口
- 汉外外汉机器翻译

- 信息检索
- 自动标引和抽词，自动文摘、文本自动分类与网站自动分类、信息自动提取与知识挖掘
- 全文检索
- 电子印刷出版系统
- 汉语辅助教学与现代远程教学
- 电子词典等

以上这些科目，有些是基础研究，有些是技术研究，也有些可以直接转化为产品。这些科目的分类并非学科分类，不过是按照编者本人日常接触的项目，把它们罗列出来而已。其分类的科学性、正确性和完整性尚待商榷。必须指出，有些基础性研究虽然看不到直接的经济效益，但它的研究成果则是其他研究工作所必需的，而且要先行。

到目前为止，在上述这些项目中，有些已经产业化，例如电子印刷出版和少数几个汉字输入系统；有些项目已经商品化，正向产业化迈进；很多项目已经实用化。但每个领域都有很多问题等待我们去解决。今后的工作只能加强，不能削弱，使我们中文信息处理的每个领域，每个项目都沿着实用化、商品化和产业化的道路奋勇前进。我相信我们这套丛书必将在促进中文信息处理技术的发展方面发挥它应有的作用。这套丛书将陆续出版。

最后，感谢“计算机学术著作出版基金评审委员会”把出版中文信息处理丛书列入了出版计划。感谢清华大学出版社和广西科学技术出版社给予出版基金的支持。

中国工程院院士　陈力为

1992 年 5 月于北京序

2000 年 4 月于北京修订

序　言

语言学是一门古老的科学，是一个民族相互交际的最重要工具。长期以来都是以手工方式进行研究的。然而进入 20 世纪 20 年代以来，语言学在现代科学体系中的地位有了急剧的变化。人们认为语言是哲学和人文科学发展的突破口，是社会科学、自然科学与思维科学的接合部，成了一门带头的科学。所以会发生这种变化，固然由于人们对语言所具有的文化本原性，也是和当前科学技术发展的影响密切相关的。到了 50 年代，一门新的利用计算机研究语言的学问，计算语言学（即自然语言理解）问世了。它不但极大地推动了语言学本身的发展，而且形成了一门深入到人类活动的各个领域，具有广泛应用价值的语言工程学。本书就是一本介绍这门学科的少有的好书。

自然语言理解真正成为一种实用的学科，那是 60 年代以后的事。1962 年国际上成立了计算语言学协会，使得研究走上了有组织的阶段，并形成一门以计算语言学理论为基础的语言工程学科。它广泛地应用于智能计算机人机接口；机器人语音对话；电话翻译系统；大型数据库自然语言查询；专家系统自然语言接口；CAD、CAI 和 OA 的人机交互系统；计算机自动书写，摘要提取，文档自动分类和文书管理系统；大型工业操作过程的自动化语言；机器翻译和机助翻译；自然语言语音通信；国际互联网上的信息分类、浏览、过滤，文学与社会科学的文档和语料计算机自动处理等等。它成为了当前最热门的研究课题之一。

但是对于这样一门重要的学科，比较深入地介绍这方面的专业书籍却十分缺乏，介绍汉语理解方面的就更少了。该书作者把自己八年来从事计算语言方面的研究和研究生教学过程中的经验编写成书，从多方面收集该领域当代最重要的理论和方法，包括有形式语言和短语结构文法、上下文无关文法、转换文法、扩充的上下文无关文法、语义网络、命题逻辑语言、概念依从理论、故事表示、集聚理论、特性与集合、词汇功能文法、合一文法、语料库语言学等等，并特别注意汉语的计算机处理问题。与此同时，他们还把自己关于计算机的汉语理解，以及汉语理解的“概率词汇语义驱动”理论和方法介绍给大家。这是十分难能可贵的。该书的最后，还介绍了如何利用这种方法实现汉语分析和机器翻译等等。确实是一本极为需要的书籍。相信它的出版，必将为中文信息处理和计算语言学的理论和技术在我国的普及推广发挥积极的作用。

陈力为

1994.02.26

目　录

引　言

NLU 是人工智能领域的一个分支

人工智能的研究已成为当前十分重要的研究学科之一，而自然语言理解（NLU，natural language understanding）与自然语言处理（NLP，natural language processing）是同义词，都是人工智能的一个分支。它们之间的关系可以用著名的智慧树（图 0.1）表示。

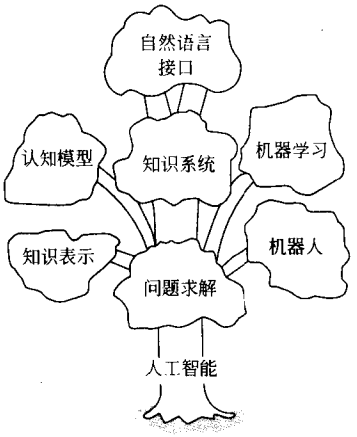


图 0.1　智慧树

从这棵树可知，人工智能的理论和方法是基础，就像是一棵树的树根。问题求解是人工智能中基础性的研究方法，用以开发认知模型、知识表示、机器学习、机器人和知识系统等，它们都是这棵树的树梢和叶子，而这棵树的顶部则是 NLP 的接口，它吸收所有树根、树干和树叶的营养，形成一门独立的学科。由此可知 NLU 的重要性。智慧树的所有树梢都是人工智能领域的某个分支，都是很有价值的应用系统。

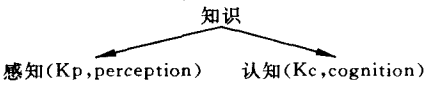
知识处理问题

在这棵智慧树上，作为处理系统的核心信息是知识。什么是知识？知识和数据有什么不同？这是非常难以回答的问题。一种观点是：

数据是明显的信息表示，而知识则是信息的含蓄表示。

知识是这样的一种信息，它是可以描述的，但又是不可能完全描述清楚的，常常是用自然语言表示的。

对于知识，我们可以分为两种类型："感知的"和"认知的"。



认知是一种能形式化的知识，是能用语言描述的。而感知的并不一定都能用语言表达。

比如，人可以感知到七百多万种颜色。但是，能用人类语言描述的只不过几十种。人可以感知千千万万张人的脸，但能用语言表达出其差异的极有限。形式化的表达，又可分成两种：用自然语言形式表达(lm)和用机器语言形式表达的(km)。这样，在实际运用过程中，存在着这样一种转换：

$$kp \Rightarrow Kc \Rightarrow km$$

感觉的东西，并不一定认识它，认识的东西还不一定理解它，这是两种不同的层次。

自然语言理解的研究起始于机器翻译。1946 年，美国宾州大学摩尔工学院 J. P. Eckert 和 J. M. Mauchly 的第一台 ENIAC 计算机问世，引起世界震惊。差不多在同一年(1946)英国的 A. Donald Booth(布斯)，美国的 W. Weaver(韦弗)就开始了机器翻译的研究。

Weaver 在《翻译》备忘录里讲：当我阅读一篇用俄文写的文章时，我可以说，这篇文章实际上是用英语写的，只不过它是用另外一种奇怪的符号编了码而已。当我在阅读时，我是在进行解码，即

$$\text{原文} = \text{译文}$$

$$\text{A 语言} \Rightarrow \text{universal language} \Rightarrow \text{B 语言}$$

$$\text{(interlingua)}$$

想象中好像不是很困难的。为此，得到财政界很多支持。例如，在麻省理工学院组织的第一次机器翻译会议上，1954 年乔治敦大学在 IBM 公司的支持下，在 IBM-701 机器上进行了世界上第一次自动翻译并取得初步成功，引起了国际上的机器翻译热。以后，美国的华盛顿大学、麻省理工学院、兰德公司、Bunker-Ramo 公司、哈佛大学、密执安大学、宾夕法尼亚大学、国家标准局(NBS)、加州大学、得州大学、美空军国家技术处(FDT)、苏联语言研究所、苏联科学情报研究所、列宁格勒大学、日本的防卫厅第一研究中心、京都大学、九州大学、意大利、比利时、英国欧洲原子能联营、捷克、匈牙利、德国等都掀起了一股研究热潮。

但是机器翻译的问题很复杂。由于低估了它的困难程度，初步的成功形成了一种假象，以至于又走向了它的反面，出现了低潮。

1963 年 10 月，由于来自各方面对机器翻译的看法，美国国家基金负责人 L. Haworth(霍沃恩)向美国科学院提出："建议调查国防部、中央情报局及国家基金会在开展关于用机器来翻译一般领域外语的研究和发展情况"的报告。美国科学院为此专门成立了一个《自动语言处理咨询委员会》(Automatic Language Processing Advisory Committee)，写了一个报告，简称 ALPAC 报告。报告中说："尽管在机器翻译上投入了巨大的努力，但使用开发这种技术，在可预期的将来是不会成功的……"

报告出笼以后，很多资助都停止了。机器翻译的研究从此十年不振，之所以造成这样的后果，有理论上的原因，也有机器设备、条件上的问题。但终究不是个好现象，在很大程度上影响了机器翻译的研究进程。

自然语言理解及其研究内容

自然语言理解(NLU)，有时也称为计算语言学(computational linguistics)，它是研究如何利用计算来理解和生成自然语言的。一般我们把人工设计的像 BASIC 语言、FORTRAN 语言、ADA 语言等统称为人工语言，而自然语言就是我们日常使用的语言，以示区别。NLP 是新一代计算机的三大突破口之一，人机接口系统问题，正受到世界各国广泛的重视。由于它在发展的道路上曾经走过了“兴起—判死刑—又复兴”的阶段，有过相当一段沉默的时期，到了 20 世纪 70 年代，情况才开始发生变化。到了 90 年代，就大不相同了。国际资源发展公司总裁预测：自然语言软件，在 1983 年大约有一千万美元产值，每年大约增长一倍，前途是相当远大的。

我们知道，乔姆斯基(N. Chomsky)[18,19,20]的转换生成文法(transformational generative grammar，简称 TG)在语言学界引起了一场“乔姆斯基革命”，同时也使得程序设计语言学得到了极大的发展，出现了如 BASIC、FORTRAN、ADA、PL/1 等等的上千种语言。乔姆斯基的转换生成文法的出现，使得语言学引进了定量的概念，成为人文科学和自然科学的交叉性学科，推动了语言学的进一步发展。

但是乔姆斯基的转换生成文法所描述的虽然是基于语句生成、十分严格的过程，然而对于人类自然形成的极其复杂的语言现象来说，乔姆斯基的理论还没能具备足够的能力处理自然语言问题。为此，乔姆斯基和他的学生们开展了很多进一步的研究。

形式语法、语义和语用研究

乔姆斯基的转换生成语法对语言学的研究影响很大，使得语言学和计算机科学之间有可能相互渗透。但是使用转换语法作句法分析并不成功。随着研究的深入，逐渐出现了一批更适宜的语法理论。如 ATN(扩充转换网络)语法、系统功能语法、格语法、语义语法、LSP(语言串处理)语法、对话语法以及各种经过扩充改进的短语结构语法等等。

到了 20 世纪 80 年代，乔姆斯基又提出了管辖与约束理论(The theory of government and binding)，简称 GB 理论[20]；1982 年 Bresnan[11]和 Kaplan[43]提出的词汇功能文法(LFG，lexical functional grammar)和 1980 年 Gazdar[39]等人提出的扩充短语结构文法(GPSG，generalized phrase structure grammar)，总的来说，从语言学的语法、语义出发，开发各种不同的有意义的理论和方法。

语义研究也在深化，随着 1965 年乔姆斯基提出的语法理论若干问题。60 年代初卡特(J. Kat)和福根(J. Fodor)等人接受 TG 理论的影响，专门考虑语义问题，把语义部分深入 TG 并论述了它的性质和作用。乔姆斯基接受了这个建议，他在《深层结构、表层结构和语义解释》(Deep Structure，Surface Structure and Semantic Interpretation，1970) 一文