第一章 中文信息处理概述

本章主要内容

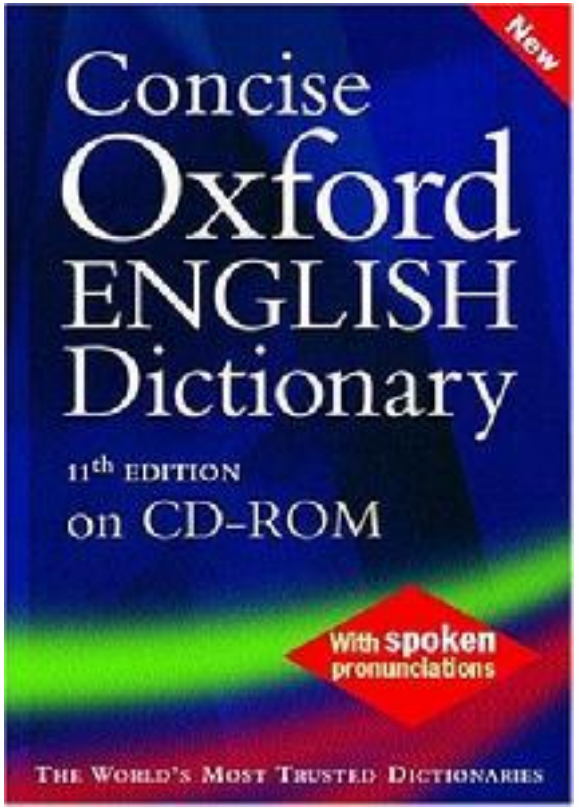
- 绪言
- 中文信息处理的发展简史
- 汉语的特点
- 自然语言处理的难点
- 自然语言处理的基本方法及发展方向

§1 绪言

- 信息
- 信息与数据
- 信息处理
- 中文信息处理
- 中文信息处理的必要性
- 中文信息处理的实现途径
- 中文信息处理系统的组成
- 国际化和本地化

1.1 什么是信息？

- 信息作为一个科学术语被提出和使用，可追溯到1928年R.V.Hartly在《信息传输》一文中的描述。他认为：**信息是指有新内容、新知识的消息。**



1.1 什么是信息？

- 数学家、控制论创始人维纳(Norbert Wiener)认为，从控制论看：“信息是我们在适应外部世界、控制外部世界的过程中，同外部世界交换内容的名称”----1948年（信息论）。

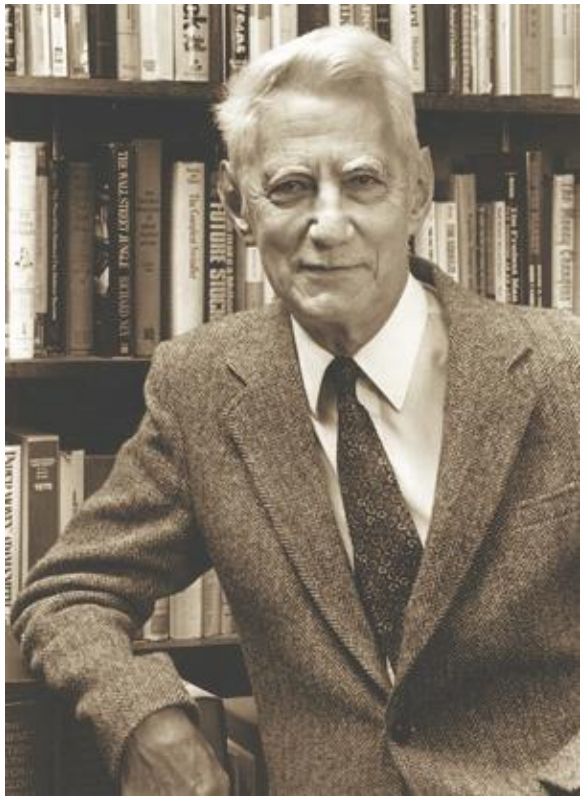
- “信息就是信息，既非物质，也非能量”。

NORBERT WIENER
1894–1964

1.1 什么是信息？

- 信息论奠基者香农(Clause Shannon)认为：信息就是**能够用来消除不确定性的**东西，是一个事件发生概率的对数的负值。

- 信息**是一个可以度量的概念，且可度量**。



小结：

- 信息（Information）是各种事物所发出的消息、情报、指令、数据和信号中所包含的表征该事物的内容。
- 信息、质量和能量三者是构成客观世界的三大要素。
- 除了可再生资源（如动、植物等）和非再生资源（如矿物等），信息是维持人类生产活动、经济活动和社会活动的第三种资源。
- 人类社会创造的最重要信息形式就是语言文字。

1.2 信息与数据

- 数据：**记录客观事物的、可鉴别的符号。具有客观性，但本身无意义。**

- 关系：
  - 既有区别又有联系

1.3 信息处理

- 信息处理就是对**信息的接收、存储、转化、传送和发布**等。
  - **信息的接收**包括信息的感知、信息的测量、信息的识别、信息的获取以及信息的输入等；
  - **信息的存储**就是把接收到的信息进行转换、传送或发布中间信息通过存储设备进行缓冲、保存、备份等处理；

1.3 信息处理

- **信息转化（加工）**就是把信息根据人们的特定需要进行分类、计算、分析、检索、管理和综合等处理；
- **信息的传送**把信息通过计算机内部的指令或计算机之间构成的网络从一地传送到另外一地；
- **信息的发布**就是把信息通过各种表示形式展示出来。

1.4 中文信息处理

- 什么是中文信息？
  - 原先定义为：由中文组成的信息
  - 后来发展为：含有中文的信息
- 什么是中文信息处理？
  - 从**广义来说**，由我们祖先创立中文开始，就一直在进行；
  - 从**狭义来说**，从第一部中文字典产生以来，就一直在进行中文信息的分析和综合处理。

1.4 中文信息处理

- 研究我国**语言文字**的信息处理问题，是自然语言处理的一个重要分支。
  - 它由中国语言文字学、数学、计算机科学等组成的，一种以计算机为主要工具，以中国语言文字为处理对象，在上世纪七十年代才发展起来的**多学科交叉**的综合性学科。

1.5 中文信息处理的必要性

- 汉语正成为一种新的强势语言而被世人瞩目，汉语理解所涉及的科学问题让国际计算语言学界无法回避。
- 汉语使用者拥有的巨大市场令国际企业界不敢轻视。
- 中文信息处理所面临的困难是其他任何一种自然语言处理都会遇到的共性问题，但也有其个性问题，因此中文信息处理更具挑战性。

1.6 实现中文信息处理的途径

- 早期：
  - 计算机的中文化
    - 通过改造计算机使它适合中文信息的处理。
  - 中文的计算机化
    - 通过改造我国的文字环境，使它适合计算机的处理。
- 当前：
  - 大数据时代的中文信息处理

1.7 中文信息处理系统的组成

- 狭义：
  - 硬件
    - 计算机硬件
    - 字库
    - 输入设备
    - 输出设备
  - 软件
    - 系统软件
    - 应用软件
- 广义：
  - 以Internet为平台的中文信息处理

1.8 国际化和本地化

- **国际化是把原来只为英文设计的计算机系统或应用软件，改造为可以支持世界上多种语言的过程**
  - 在系统层可以处理世界上多种文字符号
  - 在I/O层可以输入输出世界上多种文字符号
- **本地化是把原来只为英文设计的计算机系统或应用软件，改造为可以支持某个国家或地区的语言的过程**
  - 实现本地文字和英文的处理
  - 用本地文字和习惯显示各种窗口和信息

1.8 国际化和本地化

- **国际化只是一种机制**，使软件在支持一个新的语言（本地化）时，不需要修改软件的源程序；
- **本地化是一种产品**，如果在国际化的基础上进行本地化，只需要翻译与语言习惯相关的数据即可；
- 早期没有形成国际化时，对软件进行本地化是比较困难的，因为需要其源程序。

1.8 国际化和本地化

- 国际化和本地化的相关问题
  - 国家标准和国际标准
  - 不同字符集问题
  - 简繁问题
  - 操作系统问题
  - 。 。 。 。

§2 中文信息处理的发展简史

- 从时间上看：
  - 起步实验阶段
  - 创新探索阶段
  - 系统研究阶段
  - 蓬勃发展阶段
  - 广泛应用阶段
  - 互联网阶段

- 从技术上看：

汉字信息处理

⇩

汉语信息处理

⇩

真实大规模文本的信息处理

2.1 起步实验阶段

- 50年代，起步阶段：
  - 103、104计算机的诞生，1956年的开始了俄汉机译研究，并于1959年取得成功。
  - 当时的技术主要是词与词翻译和模式匹配，缺乏句法和语义分析俄汉翻译。

苏州大学：中文信息处理

2.2 创新探索阶段

- 邮电部6401国家级大会战的科研项目之一：汉字电报译码机
- 由邮电部邮电研究院第三研究室303组负责：将手工操作的，由传统机电设备承担的电报业务，变换为自动运行。

2.3 系统研究阶段

- 70年代，系统地探索和发展阶段：

- 1974年周恩来总理亲自批准“七四八”工程。

2.3.1 “748”工程内容

- “748”工程：
  - 汉字通信----汉字如何进入计算机
  - 汉字情报检索----中文信息处理技术应用
  - 汉字精密照排----重大行业应用

2.3.2 汉字通信

- “七四八”工程查频组于1976年12月完成了**《汉字频度表》**。
  - 字频统计使用的语料时间范围为1973-1975年，语料内容包括科学技术、文学艺术、政治理论和新闻通讯四类，统计方式为手工操作。备选语料3亿多字次，选用语料2160多万字次，统计得出**6376**个字种。

2.3.3 汉字激光照排

- 王选，江苏无锡人，1937年2月生于上海，1958年毕业于北京大学数学力学系。北大计算机研究所所长、教授、博士生导师，中国科学院院士、中国工程院院士、第三世界科学院院士。
- 我国计算机汉字激光照排技术的创始人。

2.3.3 汉字激光照排

2.3.3 汉字激光照排

- 汉字字形信息压缩及快速复原的技术
- RIP技术：Raster Image Processor。逐线扫描方式设计的控制器。
- 潍坊计算机厂生产主机和控制器。负责华光I到华光IV的总成。

2.4 蓬勃发展阶段

- 80年代，应用开发和基础研究蓬勃发展阶段：
  - 汉字键盘 输入编码：1986年3月，有关部委举办全国汉字编码方案评测。
  - 86年底创刊的《中文信息学报》。
  - 1988年初，北航等承担国家“七五”科技攻关项目《信息处理用规范现代汉语词库》，最终制定了《信息处理用规范现代汉语分词规范》。
  - GB2312-80

2.5 广泛应用阶段

- 90年代，
  - 字、词处理阶段 **à** 句子、篇章处理阶段。
  - 语料库收集、整理研究。
  - 在借鉴国外的自然语言语义理论的基础之上，先后提出了一系列符合汉语特点的语义分析方法和语义表示理论。
  - 自动分词工具、搜索引擎、Hownet、HNC理论等等。
  - 北京大学计算语言所的《现代汉语语法信息词典》完成。

2.5.1 语料库

- Corpus：文本的集合，指经科学取样和加工的大规模电子文本库。
- 语料库特征：
  - 语料库中存放的是在语言的实际使用中真实出现过的语言材料。
  - 语料库是承载语言知识的基础资源，但并不等于语言知识；
  - 真实语料需要经过加工（分析和处理），才能成为有用的资源。

2.5.2 语料库的类别

- 按照语料的语种，语料库也可以分成**单语**的（Monolingual）、**双语**的（Bilingual）和**多语**的（Multilingual）。
- 双语和多语语料库：**平行语料库**和**对照语料库**两种。

2.5.3 英文语料库

- Brown Corpus：60年代初由美国布朗大学研发而成的第一个机读语料库。
- 英国Lancaster大学与挪威Oslo 大学与Bergen 大学联合建立了 LOB 语料库。
- LDC 语言数据联合会 (Linguistic data Consortium): 设在美国宾州大学，实行会员制，有163 个语料库(包括Text 的以及 speech 的)，共享语言资源。

2.5.4 中文语料库

- 汉语现代文学作品语料库(1979 年)，527 万字，武汉大学。
- 现代汉语语料库（1983 年），2000 万字，北京航天航空大学。
- 现代汉语词频统计语料库（1983 年），182 万字，北京语言学院。
- 《人民日报》光盘数据库。
- 香港理工大学LIVAC(Linguistic variety in Chinesecommunities)语料库。

2.6 互联网真实文本智能处理

- 21世纪，以Internet为主要应用的大规模真是文本的智能处理：
  - 信息分类
  - 信息提取
  - 自动问答
  - 基于内容的快速信息检索
  - 基于个性的信息推送
  - 数字化图书馆和信息网格

小结：

- 中文信息处理技术研究的对象：
  - 中国语言文字
  - 计算机技术
  - 信息技术
  - 信息产业
- 研究方法：
  - 从系统到软件
  - 从应用技术到理论创新

作业

- P20: 1-3