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

§3 汉语的特点

- 字
- 词
- 句
- 篇章

- 音
- 形
- 义

3.1 字汇

- 所谓字汇就是指汉字的集合，字汇量与计算机对文字处理的方式有很大关系。

甲骨文 3000余个

东汉许慎《说文解字》 9353个

清代张玉书《康熙字典》49030个

3.1 字汇（续）

- 1952年，教育部公布了《常用字表》，其中收录了汉字2000个（包括500个补充用字）。
- 1955年，中国文字改革委员会公布了《通用字表（初稿）》，收录汉字5709个。
- 1965年修订后的《印刷通用汉字字形表》，收录汉字6196个。
- 1988年公布的《现代汉语通用字表》收录汉字7000个。

3.1 字汇（续）

- 对300万字语料的检测结果：
  - 2500个常用字的覆盖率为97.97%，1000个常用字的覆盖率为91.51%，3500字合计覆盖率达到99.48%。

3.1 字汇（续）

- 随着汉字文化和历史的演变，有些字出现了很多异体字，很多字成为“死字”而不再使用。
- “回”字六种写法：

回囲囘迴迵廻

龔 龏 yǎn

3.1 字汇（续）

- 70年代末，我国专门成立专家组，确定了计算机中汉字的字符集，其中最常用的是GB2312－80：

  一级汉字　　3755个

  二级汉字　　3008个

- 辅助集包含16000余个汉字。

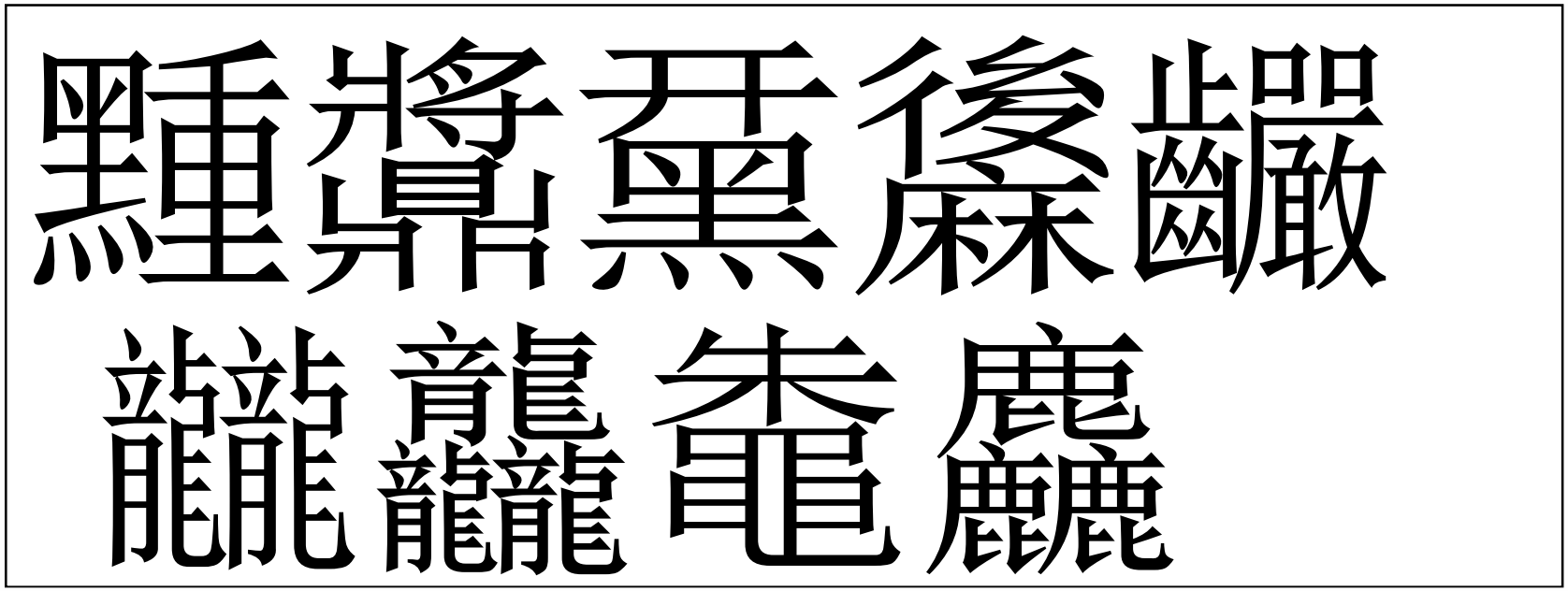
3.1 字汇（续）

- GB2312-80不足：
  - 收录了2个不规范字：“浖”；“镟”
  - 人名、地面用字少：旻(min) 喆(zhe) 赟（yun）
  - 动物名用字多 貜 鹛 鸸 鼍
- GBK，GB18040
- ISO10646-2001定义了5万多个汉字

3.1 字汇（续）

你看到过这些汉字吗？

你认识这些汉字吗？



字汇的国际标准

- 国际标准ISO 10646(Unicode)收录了更多的汉字
  - ISO 10646:1993(Unicode 2.0)
    - 收录20902个汉字
  - ISO 10646:2000(Unicode 3.0)
    - 收录27484个汉字
  - ISO 10646:2003(Unicode 4.0)
    - 收录70195个汉字

3.2 字形

- 汉字是象形文字，其每个字符都具有特定的形状和构造，这是其与各种拼音文字的最大区别。
- 目前对汉字字形的分解方法和分解标准尚未统一，现在的分解方法大体上可以分为**单字**、**字根**、**笔画**（笔形）和**形素**四个层次。

3.2.1 单字

- 单字分成多种结构类型，大体上可以分为**独体型**、**上下结构型**、**左右结构型**和**内外结合型**四种。
- 如果对单字结构进行更精细的划分，可以分为如下十二种：

3.2.2 字形结构

左右	“朋”	左中右	“彻”
上下	“吕”	上中下	“意”
全包围	“国”	右开口	“区”
上开口	“函”	下开口	“向”
左下开口	“句”	右上开口	“达”
右下开口	“库”	重叠	“巫”

3.2.3 字根

- 字根是组成单字的**基本结构单元**，它本身由**笔画组成**。它的基本要求是组字能力强，组成的单字字形匀称。
- 目前实际常用的字根为100-300个。康熙字典中规定的214个部首。
- 字根的划分不是绝对的，目前还没有相关的强制性标准，只有指导性标准：

  例如：土　旦　王

3.2.4 笔形

- 每一次从落笔到提笔，便构成一个**笔画**。
- 一个**笔画所形成的轨迹**就是**笔形**。

36画　　30画　　48画

3.2.5 “札”字笔形

- 汉字常用的笔形有五种：
  横、竖、撇、捺、折

札

- 各种笔形在汉字中使用的频度为：
  横28%　　竖18%　　撇15%
  捺13%　　折17%　　其他19%

3.2.6 笔画数

- 汉字笔画数

  最少的仅1画

  多的可达30余画

  少数可达60画以上

  平均每字约11画

- 据说是笔画最多的汉字？

3.3 字频

- 字频：汉字的**出现频率**，即某个汉字在一定语料中使用(出现)的次数与样本总字数的比率。
  - 比如在一个一万字的文本中，“的”字一共出现过400次，那么“的”字在该文本中的字频便是：4%。

3.3.2 字的使用覆盖率

- 汉字有五、六万个，一般的人仅掌握三千到五千个常用汉字，不会出现文字交流的障碍吗？
- 汉字的**使用覆盖率**
  - 164个汉字使用覆盖率占50%
  - 1000个汉字使用覆盖率占90.4%
  - 2500个汉字使用覆盖率达97.97%

3.3.3 字频的特点

	政治		文化		新闻		科技		综合	
编号	字	频度	字	频度	字	频度	字	频度	字	频度
1	的	0.0536	的	0.0324	的	0.0375	的	0.0320	的	0.0384
2	是	0.0165	一	0.0218	一	0.0132	一	0.0097	一	0.0125
3	一	0.0136	了	0.0196	了	0.0120	在	0.0092	是	0.0098
4	在	0.0115	不	0.0165	和	0.0086	用	0.0079	在	0.0095
5	这	0.0109	是	0.0141	在	0.0086	有	0.0073	了	0.0082
6	主	0.0108	说	0.0130	人	0.0083	是	0.0070	不	0.0081
7	不	0.0101	他	0.0130	大	0.0083	不	0.0069	和	0.0075
8	和	0.0098	这	0.0119	主	0.0083	中	0.0066	有	0.0069

3.3.3 字频的特点

- 字频有明显的局部性

  字频统计的结果与字频统计时使用的文本的性质有关

- 字频也有一定的时间性

  在不同的历史时期同一历史时期的不同阶段，某些特定字的使用频度可能会出现较大的波动

  例如：镕

3.3.4 字频的应用

- 汉字的输入系统
- 汉字的字形信息存储
- “动态”字频

3.4 字音

- 汉字是单音节文字。
- 注音字符包括了**注音符号**和**拼音符号**。注音符号创建于五四运动前后，它对汉字注音和推广国语起到很好的作用。目前台湾地区还在继续使用。汉字注音法也有多种，包括：威妥玛式方案、国语罗马字拼音法、北方话拉丁化新文字和《汉语拼音方案》等。

3.4.1 汉语拼音方案

- 《汉语拼音方案》是20世纪50年代制定出来的一个汉字标音系统。它用**26个西文字母**作为拼音字母，用**21个声母**、**35个韵母**、**4声调**以及**1个隔音符**来记录汉语和标注汉字。

3.4.2 拼音规则

- 绝大多数的汉字音节由**一个辅音音素**和**一个（或多个）元音音素**构成。
- 现代汉语有**417个基本音节**。
- 加上阴平、阳平、上声、去声、轻声五个声调，共有约1330个音节。

3.4.3 一字多音

- 六万多个汉字一共1330种读音，所以，汉语中**同音字**是很多的。
- 例如：GB2312收录的6763个汉字而言，

  没有同音字的读音有25个

  如：佛给能您耨暖日森僧贼抓

  同音字最多的读音是yi4（55个）
- 由于一般的人掌握一千多个常用汉字是没有困难的，所以，出现了用常用字注音的方法，非常实用。例如：

  赟 同 晕 或 赟 同 云（阴平）

一音多字的极端例子

- 语言大师赵元任先生全部用同音字(shi)创作了著名的《施氏食狮史》
  - 石室诗士施氏，嗜狮，誓食十狮。氏时时适市视狮，十时，适十狮适市，是时，适施氏适市，施氏视是十狮，拭矢试，使是十狮逝世，适石室，石室湿，氏使侍拭石室，石室拭，始食是十狮尸，始识是十狮尸，实十石狮尸，试释是事。

3.4.4 一字多音

- 在汉语中除了一音多字现象以外，还有一字多音的现象，如：如：行、重、厦、会、血、参
- GB2312收录的6763个汉字中：其中多音字有866个，占12.8%。

一字多音现象

- 通过对“辞海”（1979年版）中的16296个字进行统计，得到：
  - 单音字13663个，占83.84%
  - 二音字2112个，占12.96%
  - 三音字422个，占2.59%
  - 四音字81个，占0.5.%
  - 五音字18个，占0.11%

3.5 字义

- 字义是汉字属性中最复杂的属性，对字义很难做客观的量化。
- 汉字原来是一种望文生义的文字，汉字的形与义之间有着千丝万缕的联系。

3.5.1 一字多义

- 现在一个汉字并不只是一个字义，据统计，**一个汉字平均约有四个字义**。所以，现代汉语中汉字的表义能力明显下降，尤其是简化汉字。如：
  - “困”—穷苦、包围、疲乏
  - “就”—立刻、只有、靠近、成功
  - “记”—记得、想念、写下来、量词、拥有

3.5.2 字义与语义

- 汉字的**字义**往往和上下文环境密切相关，由此上升为**语义**：

跑（跑步）（逃跑）；
好（好评）（好赌）；
认（辨认）（认罪）。

3.5.3 字义与字音

- 汉字的字义还会影响汉字的读音，如：

行动、行业
调动、调音
照相、相好

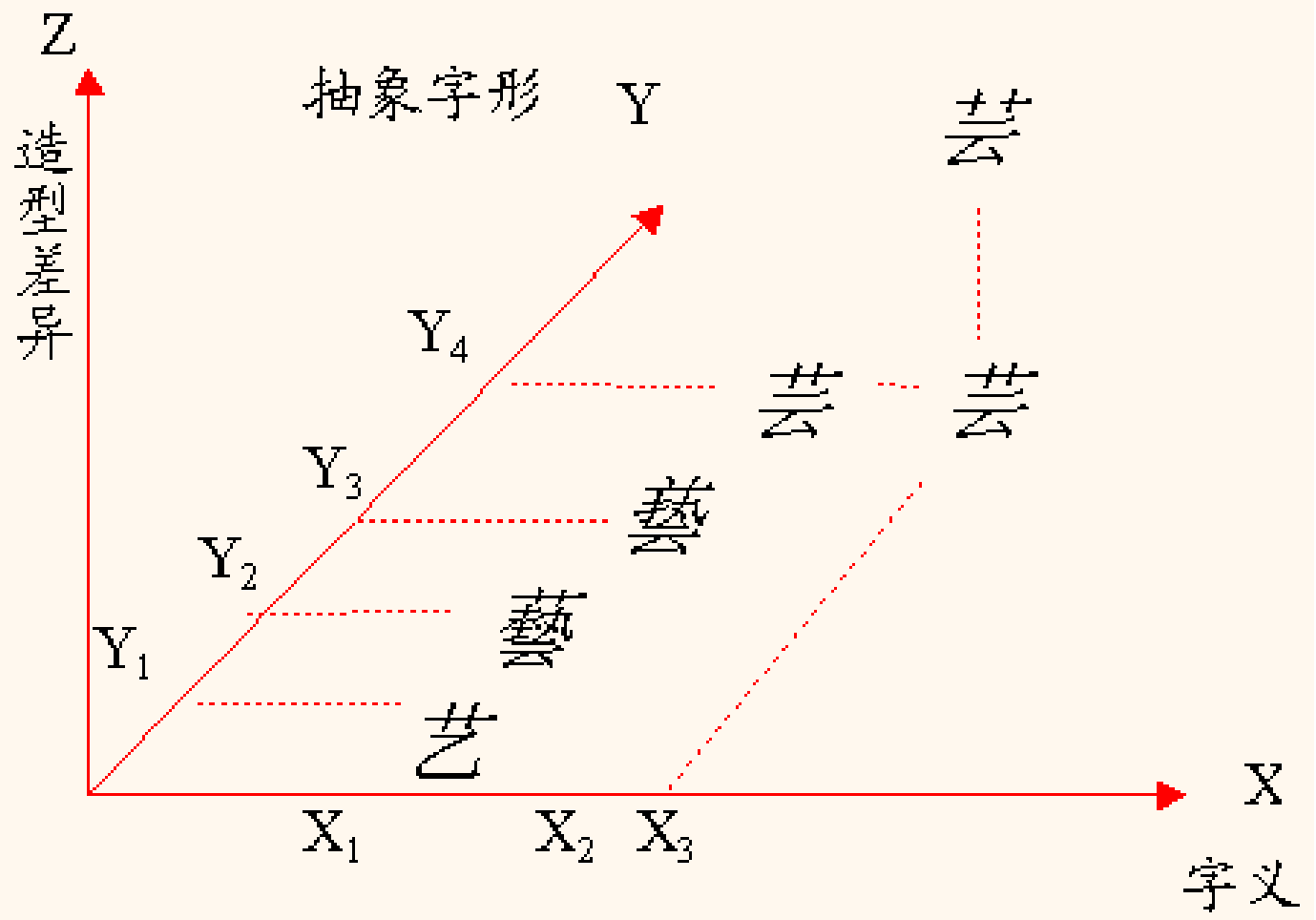
3.5.4 计算机中文信息处理的特点

- 汉字的字汇问题
- 汉字在计算机内部的表示方法问题
- 汉字的输入问题
- 汉字字形信息的存储问题
- 汉字的输出问题
- 多种文字共存问题

3.5.5 汉字的认同

- 在中国（包括台湾和香港地区）、日本、韩国和越南（简称CJKV），在文字信息处理中，都会遇到大量的汉字。为了实现统一的表示。国际标准组织和Unicode集团提出了进行汉字认同的基本概念。即汉字的XYZ模型（XYZ mode for Ideograph），其中，X—表示字义，Y表示字形（抽象字形），Z表示字型（具体造型）。

汉字的认同（续）



3.6 词汇

- **词汇**是语言中**所有的词和短语的总和**。
- 词是由**语素**构成，是句子中最小的能够独立运用的语言单位。
- **单音节语素**在书面上用**单个的汉字**书写。
- 古汉语中由一个单音节语素构成的词占绝对优势，所以书面上基本一个汉字也就是一个词（只有极少数连绵词例外）。这就形成了汉字**连篇书写**的传统。
- 20世纪20年代开始，文章开始分段，并使用新式标点符号，不再连篇书写，基本上为**按句连写**。

3.6.1 语素

- **自由语素**则是指能够独立成词的语素，例如：“水、木、金、心、火”等。自由语素能够单独成词，也可以与其他语素组合成词
- **粘着语素**是指一般不单独构成词的语素，例如：“伟、型、丰、咐”等。粘着语素必须跟别的语素组成词。

3.6 .2 词根和词缀

- 汉语中的词有词根和词缀
  - **词根**是指词内**意义实在的语素**，它是词的核心部分，词根在词内的位置不固定。
  - **词缀**是指词内**意义不实在的粘着语素**，它在词内的位置固定在前或后，词缀是词的辅助部分。

例如：

“筷子”中的**“筷”是词根**，“子”是词缀。

2.6.3 单纯词

- 汉语中由**一个语素构成的词**叫做单纯词。
  - 包含一个语素构成的词（例如，“人、走、红、天”等）
  - 双音节连绵词（例如，“鸳鸯、垃圾、葡萄、琳琅、吩咐”等）
  - 音译词（例如，“沙发、咖啡、巧克力、巴黎、逻辑”等）
  - 译自少数民族的地名（例如，“哈尔滨、呼和浩特、吐鲁番”等）。

2.6.4 合成词

- 由**两个或两个以上语素构成的词**称为合成词。
- 合成词包括三类：**重叠**、**附加**和**复合**。
  - **重叠式**：由两个相同的词根相叠构成的词。例如：

哥哥、姐姐、刚刚、星星、整整齐齐

- **附加式**的词是由词根和词缀构成。词缀在词根之前称为前缀，在词根之后则称后缀：
  - **前加式**（前缀+词根）：老虎、老乡，阿姨、阿毛，微处理器、微笑。
  - **后加式**（词根+后缀）：刀子、饼子、胖子、桌子，石头、木头、苦头，作者、读者、科技工作者、唯物主义者，芦花、规范化、现代化.

- **复合式**词是由两个或两个以上词根成分组成的附加式合成词。
- 汉语复合词的内部结构基本上是和句法结构一致的，有主谓、述宾、补充、偏正、联合等。例如：

  年轻、民主、自动，司机、站岗、美容

  提供、推广、改进，气功、腾飞、火红

  体制、开关、质量。

3.7 语境

- **语境**是语言单位出现时的环境。一般分为**上下文**语境和**情景**语境。
- 词、短语、句子等在文本中出现时，它前面或后面出现的其他语言单位都是该单位的**上下文语境**。
- “上下文”是一个宽泛的概念，在一段话或一篇文章中凡出现在某语言单位之前的词、短语、句子等都是该语言单位的上文，出现在其后的都是其下文。

§4　自然语言处理的难点

- 自然语言处理
- 自然语言处理研究的内容
- 自然语言处理面临的困难

4.1 自然语言处理

- NLP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。要研究表示语言能力和语言应用的模型，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完善这样的语言模型，根据这样的语言模型设计各种实用的系统，并探讨这些使用系统的评测技术。

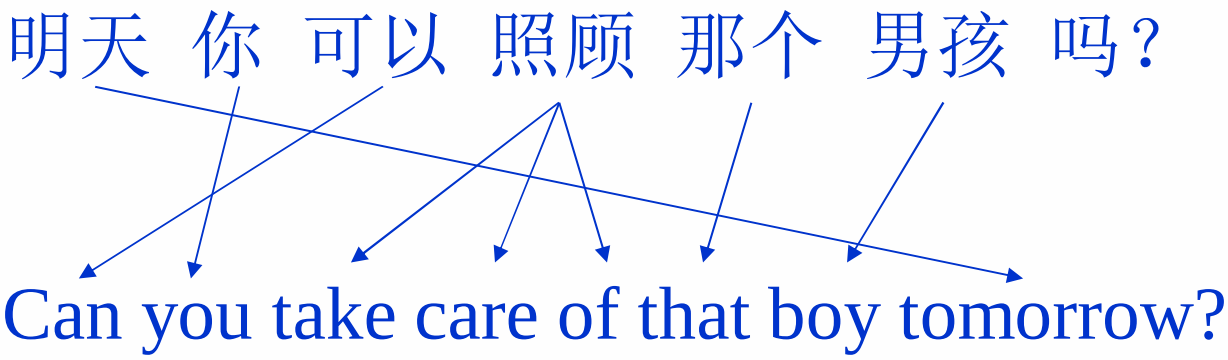
4.2 自然语言处理研究的内容

- 机器翻译
- 语音技术
- 文字识别
- 信息检索
- 分档分类
- 自动文摘
- 问答系统

4.2.1 机器翻译—分析转换生成

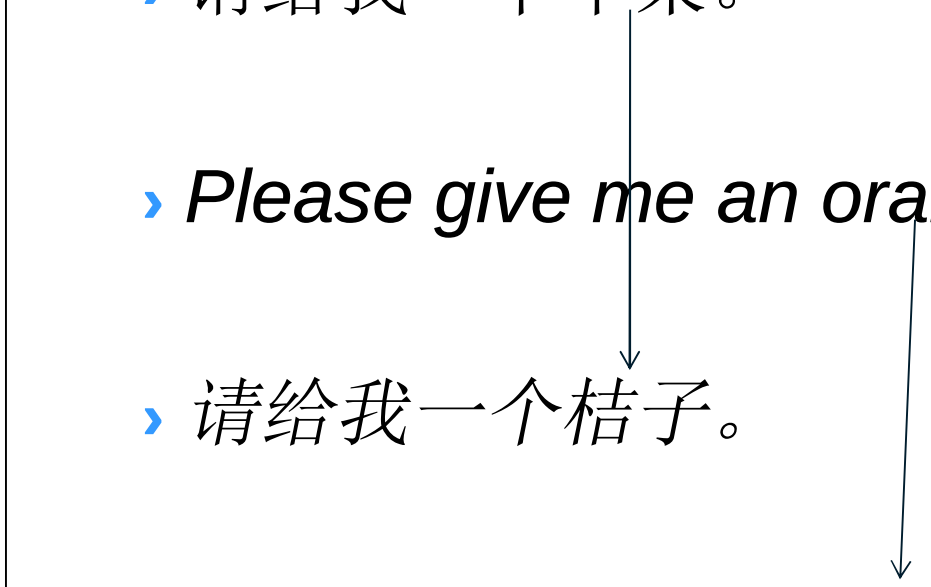
- 源语言分析
- 源语言到目标语的转换
  - 转换规则
    - A 的 B => B of A
- 目标语的生成
  - 可读？

4.2.2 机器翻译—双语对齐



4.2.3 基于实例的机器翻译

- 请给我一个苹果。
- *Please give me an orange.*
- *请给我一个桔子。*
- Please give ma an apple.



4.2.4 统计机器翻译

- 破译密码
  - P(e|c) = P(c|e)*P(e)/P(c)
- 翻译模型
  - P(银行|bank)
  - P(河岸|bank)
- 语言模型
- 双语对齐语料库

4.2.5 语音技术

- 从语音到拼音，从拼音到文字
- 最自然的人机接口
- 研究
  - 李开复，非特定人连续语音识别
    - 语音识别推动了自然语言处理的发展
  - IBM ViaVoice
- 语音合成
  - TTS(Text to Speech)
    - 重音
    - 间隔
    - 句法结构和韵律结构

4.2.6 文字识别

- 手写体
- 汉王笔
- 如意笔
- 清华文通
- 豪文笔
- 商务通

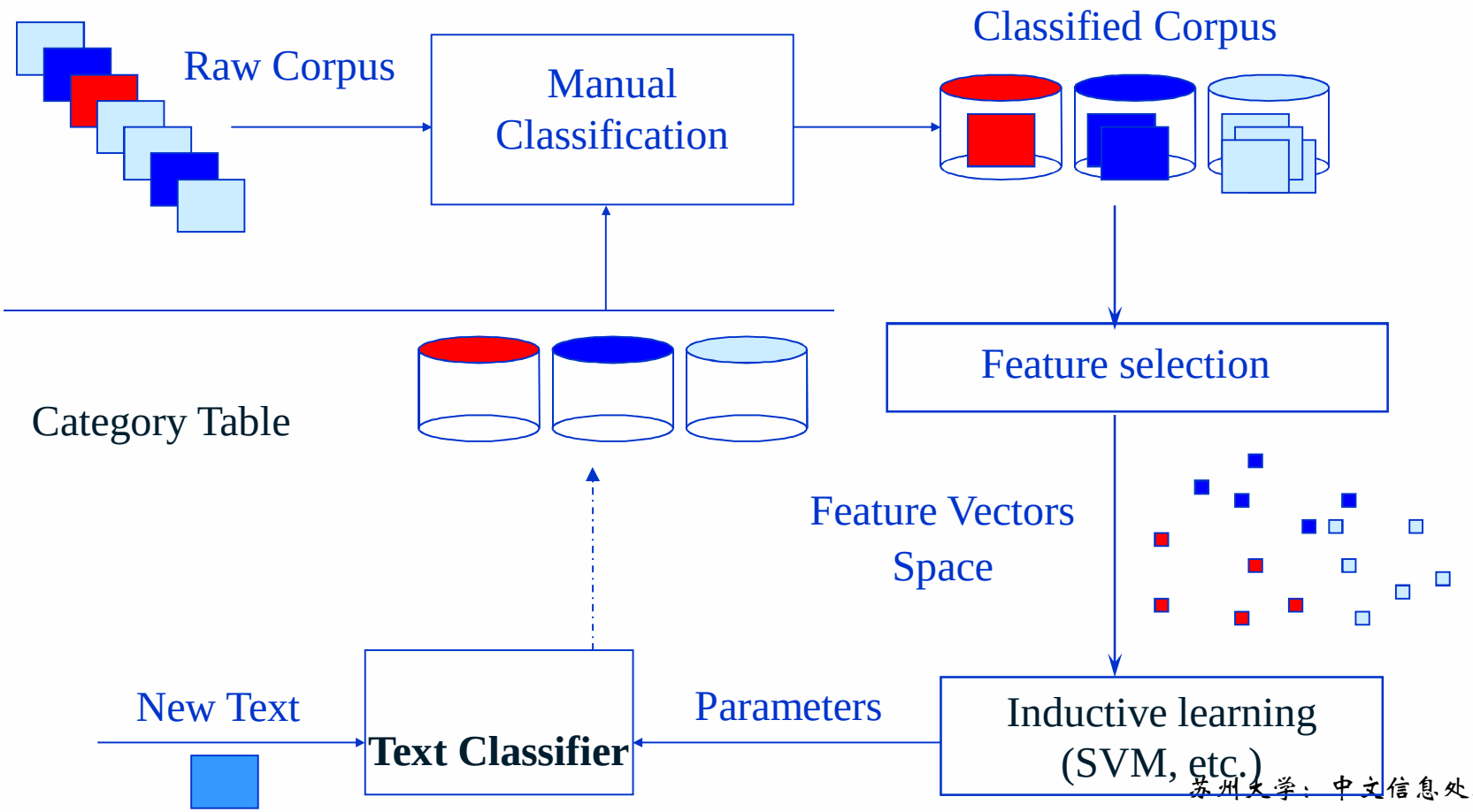
4.2.7 信息检索

- 在信息海洋中找到What you want
- 图书馆情报检索
- 全球图书馆----Internet
- 搜索引擎

4.2.8 文本分类

- Yahoo起家靠什么？
- 分类的层次体系Hierarchy
- 文本分类是模式识别问题
  - 特征提取
  - 统计机器学习
- 信息过滤
- 有害信息过滤

Text Categorization



4.2.9 自动文摘

- 中心思想
- 统计方法
- 理解方法
- 文本结构

4.2.10 信息抽取

- 地震
  - 何时？
  - 何地？
  - 几级？
  - 死亡人数？
- 信息提取，框架填充，文本生成。

4.3 自然语言处理面临的困难

- 语言中的歧义的识别
- 命名实体识别问题
- 词性的标注
- 大规模语料库的建设

4.3.1 汉语分词

- 分词(text segmentation, word segmentation)就是把一个句子按照其中词的含义进行切分。
- 分词也就是将连续的字串或序列按照一定的规范重新组合成词序列的过程。

4.3.2 汉语歧义

- 切分歧义
  - 伪歧义
    - 他从马上下来
    - 美国会宣布
    - 结合成分子时
  - 真歧义
    - 乒乓球拍卖完了
    - 解除了
    - 张三演好戏
- 理解歧义
  - 咬死猎人的狗

4.3.3 英语中的歧义

- 英文：

  Put the block in the box on the table.

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

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

I saw a man in the park with a telescope.

4.3.4 语境的理解

- 他说，“她这个人真有**意思**（funny)。”
  她说：“他这个人怪有**意思**（funny)。”
  于是以为他们有了**意思**（wish），并让他向她**意思意思**（intention）。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）？”时候有人说：“真**有意思**（funny)。”也有人说，“真**没意思**（nonsense）”。

4.3.5 词性标注

- 发展 体育 运动

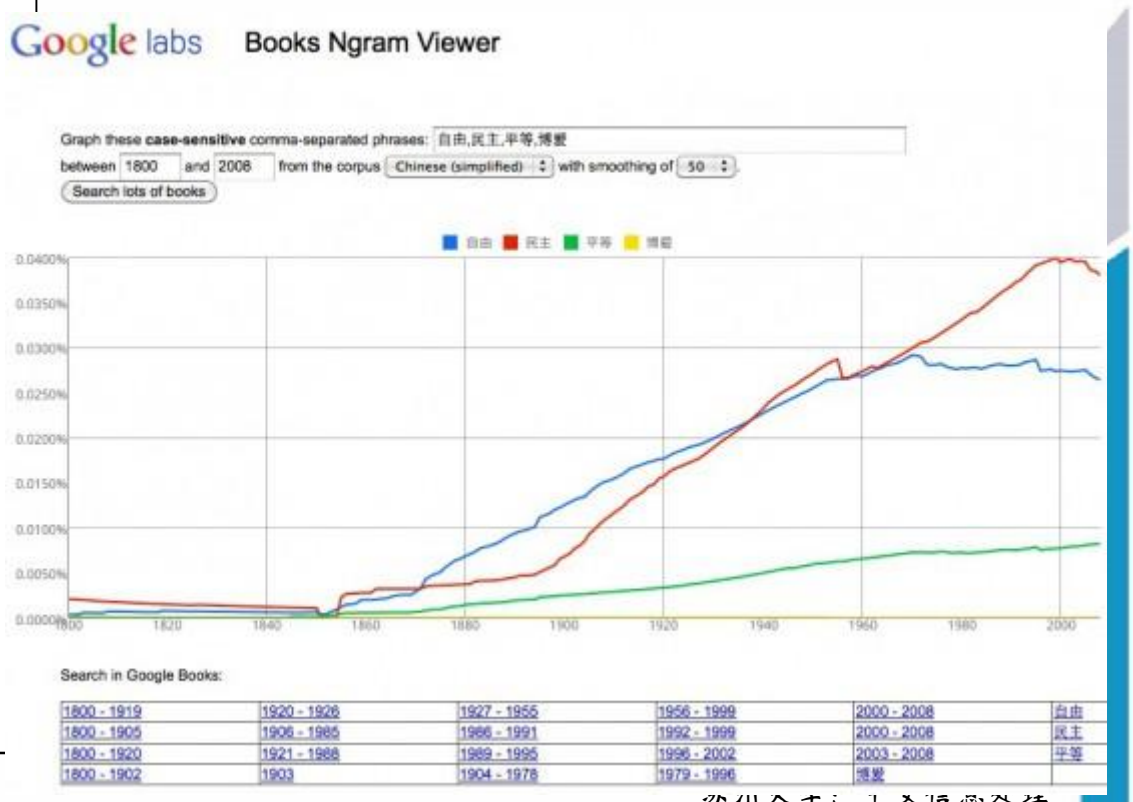
  N N N

  V V

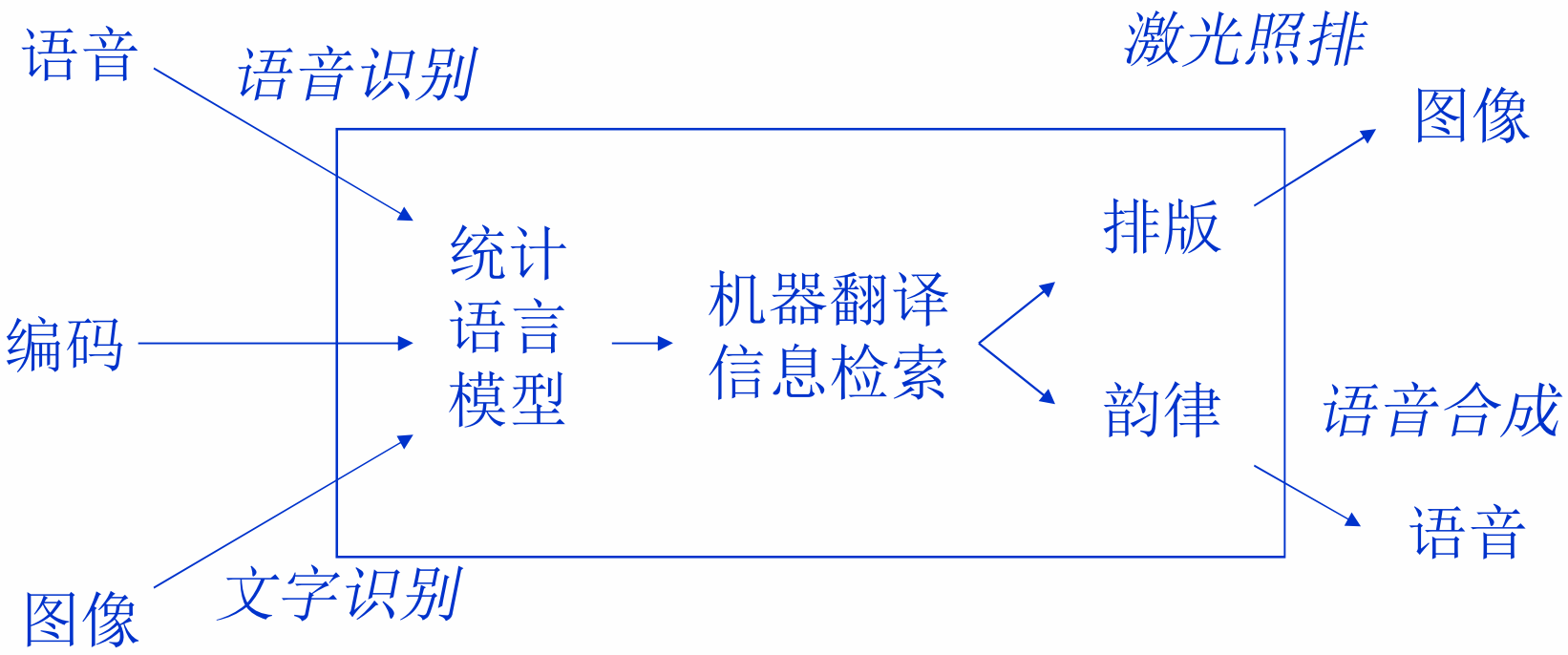
- 发展/V 体育/N 运动/N

4.3.6 统计语言模型

- N-gram
- 词的统计模型
  - Bigram
  - Trigram
- Google Ngram Viewer



4.3.7 统计语言模型—应用



5. 基本方法和发展方向

- 自然语言处理的基本方法
  - 规则
  - 统计
  - 规则+统计(Hybird)

- 自然语言处理的发展方向
  - 大数据时代的中文信息处理

5.1 规则方法

- 句法+语义规则：生产语言学
  - 理性主义者：书面语是一组规则的符合集合。通过语言学家进行归纳总结。
  - 例如：早期的机器翻译系统。
  - 特点：特定的小范围的（较为封闭的）领域具有优势。

5.2 统计方法

- COLING90：“处理大规模真实文本的理论、方法和工具”
- 语料库
- 机器学习
- 语言模型
  - 隐马尔可夫模型（简称**HMM**）
  - 概率上下文无关语法（简称 PCFG）
  - 基于决策树的语言模型（Decision-Tree Based Model）
  - 最大熵语言模型（MaximumEntropy Model）

5.3 规则+统计

- 规则语言模型
  - 60%的语言现象有规律，适用于规则描述。
- 统计语言模型
  - 词、短语、句子等不同颗粒的统计模型

5.4　网络环境下的NLP特点

- 人机关系控制模糊
- 信息真伪难以辨别
- 口语化文本与多语言并存
- 语言成分大量缺省、重复、修正、停顿等。

本章小结

- 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

作业

- P20: 3-7