

文本处理的内容

- 采集——分词——清洗——规范——标引——摘要——聚类

为什么要了解文本信息的处理？

- CNNIC统计——中国网页数

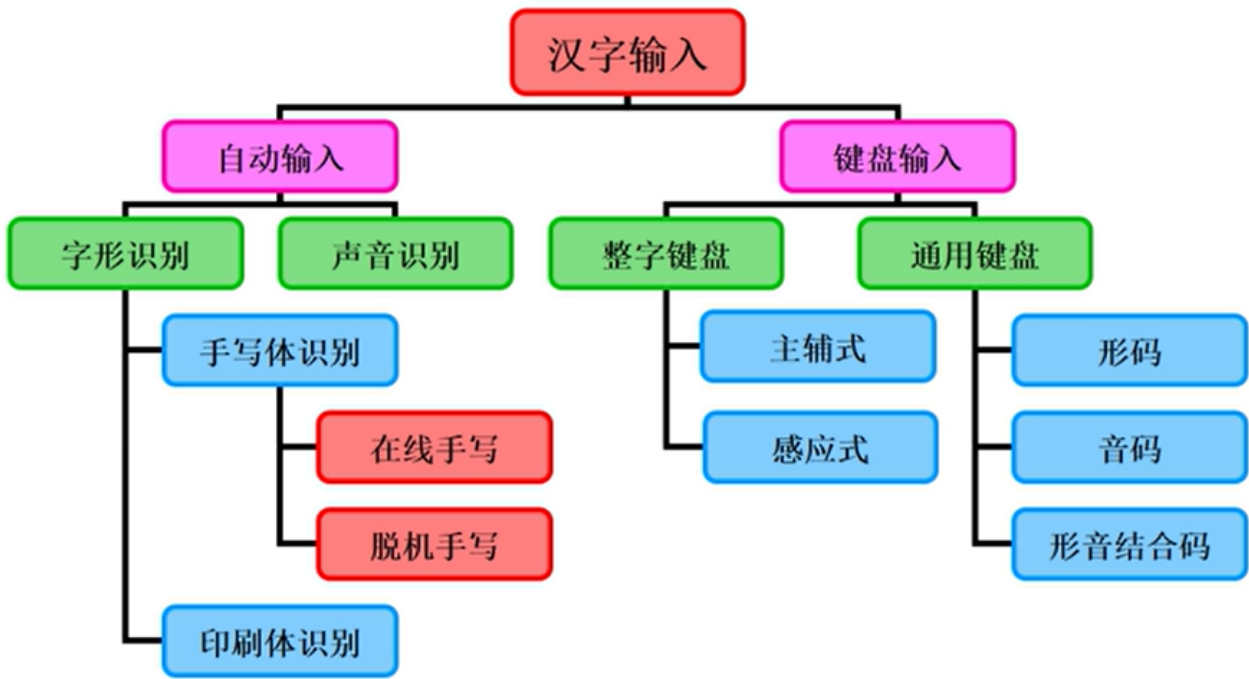
中文信息处理

中文信息处理发展史

- IBM公司，文本自动处理的思想，掀起信息处理的理论方向
- 1. 学习和理论探索的萌芽阶段
- 2. 汉字信息处理为主的早期阶段
- 3. 字、词等表层处理为特征的初级阶段
- 4. 句法和语义等深层处理为代表的中期阶段
- 5. 语料库统计方法兴起的近期阶段
- 6. 以Web为主要应用对象，大规模真实文本、智能信息访问的现阶段
- 自然语言信息处理的分支

中文信息处理/=中文文本的信息处理

- 信息的两个层次
 - 符号层——中文/汉语/汉字
 - 内容层——符号所承载的意义
- 中文信息处理的两个层次
 - 字符处理（输入、存储、输出等）
 - 内容处理（词语切分、词性标注、结构分析、意义理解、推理、翻译……）
- 符号层的汉字处理技术



内容层的信息处理

- 内容识别：分词、标引、分类、聚类
- 内容生成：文摘、翻译、理解、推理、创作

标引

- 对少数文本对概念进行标注
- 分类、自动问答、画像、摘要
- (人工) 标引流程
 - 主题分析、排重、标引、审核、记录结果
 - 受控词、主题词进行标引

自动标引

- 文献标引：对所收集的文献给出标识导引，包括文献标题、作者名、分类号和主题词
- 文献标引作业流程
 - 文献文本分析
 - 特征信息（主题词、关键词或其他标识）的提取与描述
 - 建立索引或倒排档

- 自动标引：就是用机器**抽取或赋予索引词**，一旦编制好程序和规则，就**不需要人工干预**
- 意义
 - 适应信息资源快速增长的需要
 - 效率高、成本低
 - 稳定性好、一致性好

词

- 最小的能够独立运用的语言单位——*缺乏操作标准*

分词

- 中文（自动）分词：由机器在中文文本中词与词之间加上标记

汉字和汉语

- 汉语文本是基于单字的，汉语的书面表达方式也是以汉字作为最小单位的，词与词之间没有显性的界限标志，因此分词是现代汉语文本分析处理中首先要解决的问题

分词的意义

- 正确的机器自动分词是正确的中文信息处理的基础
 - 文本检索
 - 文语转换
 - 词频统计
 - 句法分析、语义分析、机器翻译、语音合成、自动分类、自动摘要、自动校对

英语也需要分词

- 不能仅凭空格和标点符号解决切分问题
 - 缩写词
 - 连写形式以及所有格结尾
 - 数字、日期、编号
 - 带连字符的词

- 英语的切分较为容易

分词的方法

- 基于词典
 - 基于字符串（词）匹配的分词方法：按照一定的策略将待分析的汉字串与一个“充分大的”机器词典中的词条进行匹配，若在词典中找到某个字符串，则匹配成功。可以切分，否则不予切分
 - 实现简单，实用性强，但机械分词法最大的缺点就是词典的完备性不能得到保证
- 无词典
 - 基于统计
 - 基于规则
 - 基于字标注

分词的难点

歧义消解

- 切分歧义类型
 - 交集型歧义：对于汉字串AJB，AJ、JB同时成词
 - 结合/成，结/合成
 - 组合型歧义：对于汉字串AB，A、B、AB同时成词
 - 门/把手/坏/了，请/把/手/拿/开
 - 混合型歧义：同时包含交集型歧义和组合型歧义
 - 这样的/人/才能/经受住考验
 - 这样的/人才/能/经受住考验
 - 这样的/人/才/能/经受住考验
- 切分歧义真伪
 - 真歧义：歧义字段在不同的语境中确实有多种切分形式
 - 伪歧义：歧义字段单独拿出来看有歧义，但在所有真实语境中，仅有一种切分形式可接受
 - 对于交集型歧义字段，真实文本中伪歧义现象远多于真歧义现象

未登录词识别

- 虽然一般的词典都能覆盖大多数的词语，但有相当一部分的词语不可能穷尽地收入系统词典中，这些词语称为未登录词或新词
- 类别
 - 专有名词：中文人名、地名、机构名称、外国译名、时间词

- 重叠词
- 口语
- 派生词：一次性用品
- 与领域相关的术语

词典分词

正向最大匹配（FMM）

1. 设自动分词词典中最长词条所含汉字个数为l
2. 取被处理材料当前字符串序数中的l个字作为匹配字段，查找分词词典。若词典中有这样的一个l字词，则匹配成功，四配字段作为一个词被切分出来，转6
3. 如果词典中找不到这样的一个l字词，则匹配失败
4. 匹配字段去掉最后一个汉字，l--
5. 重复2-4，直至切分成功为止
6. l重新赋初值，转2，直到切分出所有词为止

- 特点
 - 对交叉歧义和组合歧义没有什么好的解决方法
 - 往往不单独使用，而是与其他方法配合使用

逆向最大匹配分词（BMM）

- 分词过程与FMM方法相同，不过是从句子（或文章）末尾开始处理，每次匹配不成功时去掉的是前面的一个汉字
- 比最大匹配法更有效

双向匹配法

- 比较FMM和BMM法的切分结果，从而决定正确的切分
- 可以识别出分词中的交叉歧义
- 算法时间、空间复杂性较高

常用规则

- 颗粒度越大越好：词长度越长越好
- 非词典词越少越好
- 总体词数越少越好

无词典分词

基于理解的分词

- 通过让计算机模拟人对句子的理解，达到识别词的效果。其基本思想就是在分词的同时进行句法、语义分析，利用句法信息和语义信息来处理歧义现象
- 由于汉语语言知识的笼统、复杂性，难以将各种语言信息组织成机器可直接读取的形式，因此目前基于理解的分词系统多处在试验阶段

基于统计的分词

- 基于统计的分词方法：基本原理是根据字符串在语料库中出现的统计频率来决定其是否构成词
- 无词典分词法也有一定的局限性，会经常抽出一些共现频度高、但并不是词的常用字符串，如“这一”、“之一”以及“提供了”等等
- 在实际应用的统计分词系统中都要使用一部基本的分词词典（常用词词典）进行串匹配分词，即将字符串的词频统计和字符串匹配结合起来，既发挥匹配分词切分速度快、效率高的特点，又利用了无词典分词结合上下文识别生词、自动消除歧义的优点

基于字标注的分词

- 把分词过程视为字在字串中的标注问题。由于每个字在构造一个特定的词语时都占据着一个确定的构词位置（即词位），假如规定每个字最多只有四个构词位置：即B（词首），M（词中），E（词尾）和S（单独成词），那么下面句子甲的分词结果就可以直接表示成如乙所示的逐字标注形式：
- 基于字标注的方法通过改进未登录词识别能力，提升了分词系统的总体性能
- “基于字标注的方法+机器学习”成为中文分词主流技术，算法复杂度较高

常见分词工具

NLPIR

Jieba

THULAC

语料库在线

停用词

- stop words/禁用词、非用词

- 在信息检索中，为节省存储空间和提高搜索效率，在处理自然语言数据（或文本）之前或之后会自动过滤掉某些字或词
- 停用词表是一种特殊的词表，在这个词表中含有冠词、虚词、叹词、连词、介词以及语义泛泛的词等一切在上下文中没有检索意义的词

(英文) 词干化

- Stemming
- 英语单词有时态、单复数等变化
- 有成熟工具：coreNLP、SnowballStemmer(Python)

自动标引

半自动标引（标引工作的自动化程度）

1. 文献纪录（题目等著录项目）键入终端后显示在屏幕上
2. 操作人员移动光标从题目中抽取关键词
3. 利用人机对话方式输入与标题内容有关的隐含概念词，以保证主题标引的全面性：同时删除计算机程序错误组配的词
4. 根据词库中的参照系统将关键词转换成标准主题词，进行上位登录。词库（主题词表）是计算机辅助标引的核心


标引词的来源

抽词标引（自由词标引）

- 利用计算机直接从文献题名、文摘或正文中自动抽出能表达文献主题的词作为标引词，并自动生成关键词索引或倒排档。抽词标引的标引词只能来源于文献本身的文内关键词
- 类别
 - 主关键词标引：计算机从抽出的全部关键词中选出少量主要关键词做索引词
 - 全关键词标引：把除停用词以外的全部关键词抽出，直接做索引词
- 优点：无需主题切换，接近自然语言
- 缺点
 - 标引用词不规范，影响查全率

- 同义词检索降低系统的时间效率
- 难以找出词和词之间的相互关系，很难进一步利用语义信息

赋词标引/受控词标引

- 让计算机模仿人的赋词标引方法，分析文献的内容，选取与文献主题相符或密切相关的语词符号作为索引词
- 其标引词是由描述词组成的，这些词不一定来源于文献本身所用的词，而是选自预先编制的词表
- 优点
 - 规范化用词
 - 词表可以反映词的“类—属”关系
- 缺点
 - 受控词标引往往有一定的标引误差
 - 词典面临老化的问题
 - 主题词表对用户来说往往是一个负担
- 自动赋词标引是在自动抽词标引的基础上发展起来的
- 最合理的标引方法：混合标引方法？
- 自动标引流程

标引源

- 全文：数据量大，处理麻烦
- 标题：主要标引源；信息量少，歧义多，标引质量差
- 文摘：主要标引源，大部分情况下够用
- 首尾章节、章节的首尾段、段落的首尾句

确定关键词

统计法

绝对词频统计法

- 以词在文章中出现的绝对频次为根本依据确定文章的中心关键词，理论基础是 齐夫定律
- 词在文献中的出现频率是该词对该篇文献重要性的有效指标，文献中只有词频 介于高频和低频之间 的那部分中频词最适合作为标引词

- 齐夫定律/词频分布定律
 - 如果把一篇较长文章(>5000)中每个词出现的频率统计起来，按照高频词在前、低频词在后的递减顺序排列，并用自然语言给这些词编上等级序号，即频次最高的词的等级为1,频次次高的等级为2, ..., 频次最小的词等级为D(或L),若用 f 表示等级为 r 的词在文献中出现的 相对频次 则有： $fr=C$ (C 是一个常数，大约等于0.1)
- Luhn标引算法
 - 用试错法确定高频词和低频词的阈值
 - 去掉高频词和低频词后，将余下的中频词选作标引
- Zipf第二定律/低频词定律
 - 文章中词频为 n 的词与词频为1的词数量上有数学关系
- 改进的标引方法
 - 存在一个词由高频行为转为低频行为的临界区，只有处于临界区内的词才最适于描述文献的主题
 - 以 n 为临界区的中点，以最高词频处为临界区的上界，取与 n 到上界之间等级距离相等的另一端为临界区的下界，位于临界区内的词经过筛选即可选为标引词

词频权重法

- 除考虑词频外，还考虑词的位置、词的词性、词本身的价值、词的长度等因素，对词进行加权，然后根据权值大小确定关键词
- 标引词加权：词的权值一般表示该词的重要程度
- 位置加权法
- 机器学习标引（统计学习）

语言法

- 句法分析法
- 语义分析法

人工智能法

标引结果

- 抽词标引：直接标注标引结果
- 赋词标引

- 关键词与受控词 主题词、副主题词、特征词 之间存在着一定的关系
如同义词关系、上位关系、下位关系等
- 使用一定的方法将以上提取的关键词转换为受控词
 - 使用关键词-受控词对照表
 - 利用词汇相似度

单汉字标引

- 自然语言标引
- 在标引时将概念词拆为单汉字，以单汉字为处理单位，利用汉字索引文件实现自动标引和逻辑检索
- 处理过程：计算机对处理的文本逐一抽字，并去掉无意义的虚字；对剩下的字建立单字索引文件
 - 搜索引擎就是单汉字标引，且不去虚词
- 优点
 - 不分词，简单容易
 - 字匹配，查全率高
- 缺点
 - 索引规模大
 - 速度慢