1.概述

词是自然语言中能够独立运用的最小单位，是自然语言处理的基本单位。

自动词法分析就是利用计算机对自然语言的形态进行分析，判断词的结构和类别

词性是词汇最重要的特性，是连接词汇到语法的桥梁

英语的形态分析

基本任务：

- 单词识别
- 形态还原

形态分析的一般方法：

- 查词典，如果词典中有这个词，直接确定该词原形
- 根据不同情况查找相应规则对单词进行还原处理
- 进入未登录词处理模块

汉语自动分词概要

自动分词是汉语句子分析的基础

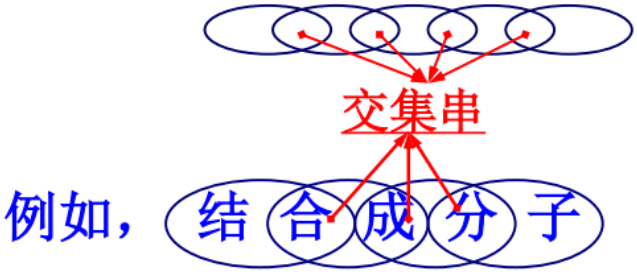
主要问题：汉语中什么是词?

- 单字词与词素
- 词与短语

歧义切分字段处理

交集型歧义

➢ **定义：链长** 一个交集型切分歧义所拥有的交集串的集合称为交集串链，它的个数称为链长。



“结合”、“合成”、“成分”和“分子”均构成词，交集串的集合为｛合，成，分｝，因此，链长为3。

其实就是一个句子去头去尾（去掉首部一个字和尾部一个字）

组合型歧义

一词多义

分词与词性标注结果评价方法

两种测试

- 封闭测试/开放测试
- 专项测试/总体测试

评价指标

- 正确率
- 召回率
- F-测度值

$$F - measure = \frac{(\beta^2 + 1) \times P \times R}{\beta^2 \times P + R} \times 100\ \%$$

一般地，取 $\beta = 1$，即

$$F1 = \frac{2 \times P \times R}{P + R} \times 100\ \%$$

自动分词基本算法

- 有词典切分/无词典切分
- 基于规则的方法/基于统计的方法

最大匹配法（MM）

- 正向最大匹配算法（FMM）
  - 逆向最大匹配算法（BMM）
- 双向最大匹配算法

正向最大匹配算法

- $i = 0$，当前指针 p_i 指向输入字串的初始位置
- 计算当前指针到字串末端的子树 n，如果 n = 1，输出切分结果并结束。否则，令 m = 词典中最长单词的字数，若 n < m，令 m = n
- 从当前位置起，取 m 个汉字作为词 w_i

(a) 如果 w_i 确实是词典中的词，则在w_i 后添加一个切分标志，转(c)；当前词在词典中，加一个划分标志。修改指针的位置（当前词的下一个位置）

(b) 如果 w_i 不是词典中的词且 w_i 的长度大于1，将 w_i 从右端去掉一个字，转(a)步；否则（w_i 的长度等于1），则在w_i 后添加一个切分标志，将w_i 作为单字词添加到词典中，执行 (c)步；长度等于1直接切分，因为不能再短了

(c) 根据 w_i 的长度修改指针 p_i 的位置，如果 p_i 指向字串末端，转(4)，否则，$i=i+1$，下一个 返回 (2)；

- 输出切分结果，结束程序

优点：

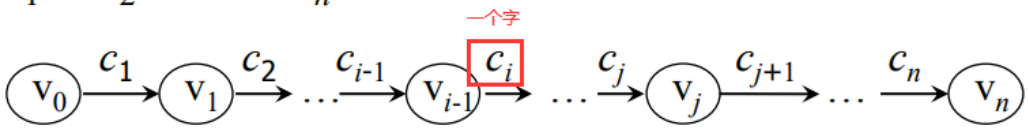
- 程序简单，开发周期短
- 需要很少的语言资源，不需要语法，句法，语义资源

弱点：

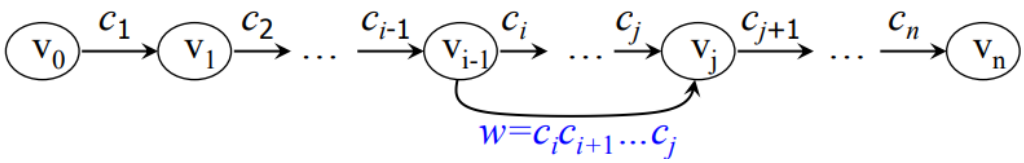
- 歧义消解能力差
- 切分正确率不高

最少分词法（最短路径法）

设待切分字串 $S=c_1\,c_2\ldots c_n$，其中$c_i(i=1,2,\ldots,n)$ 为单个的字，n 为串的长度，$n\geq 1$。建立一个节点数为$n+1$的切分有向无环图G，各节点编号依次为V_0，V_1，V_2，…，V_n。



- 相邻节点之间建立有向边，边对应的词默认为 c_k
- **(2)** 如果 $w=c_ic_{i+1}\ldots c_j$ $(0<i<j\leq n)$ 是一个词，则节点v_{i-1}, v_j 之间建立有向边 $<v_{i-1}, v_j>$，边对应的词为 w。



- 重复步骤 2，知道没有新路径（词序列）产生
- 从产生的所有路径中，选择路径最短的（词数最少的）作为最终分词结果

- 比如节点 1 到 10 之间有一条边，这就证明这是个字是一个词，就可以跳过 1---10，选择最短路（词数最少）

优点：

- 切分原则符合汉语自身规律
- 需要的语言资源（词表）不多

弱点：

- 对许多歧义字段难以区分，最短路径有多条时，选择最终输出结果缺乏应有的标准
- 字串长度较大和选取的最短路径数增大时，长度相同的路径数急剧增加，选择最终正确的结果困难越来越大

基于语言模型的分词方法

➢ **方法描述：**

设对于待切分的句子S，$W = w_1 w_2 \ldots\ldots w_k$ $(1 \le k \le n)$ 是一种可能的切分。

$$W^* = \arg\max_W p(W \mid S)$$

$$= \arg\max_W p(W) \times p(S \mid W)$$

详见第5章举例。

语言模型

生成模型

优点：

- 减少了很多手工标注的工作
- 在训练语料规模足够大喝覆盖领域够多时，可以获得较高的切分正确率

弱点：

- 训练语料的规模和覆盖领域不好把握
- 计算量较大

基于HMM的分词方法

➢ **基本思想：**

把输入字串(句子)S 作为HMM μ 的输入；切分后的单词串 S_w 为状态的输出，即观察序列 $S_w = w_1 w_2 \cdots w_n$，$n \geq 1$。词性序列 S_c 为状态序列，每个词性标记 c_i 对应 HMM 中的一个状态 q_i，$S_c = c_1 c_2 \cdots c_n$。

$$\hat{S}_w = \arg\max_{S_w} p(S_w \mid \mu)$$

详见第6章举例。

优缺点和上面的一样

由字构词的分词方法

➢ **基本思想**：将分词过程看作是字的分类问题。该方法认为，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)。假定每个字只有4个词位：词首(B)、词中(M)、词尾(E)和单独成词(S)，那么，每个字归属一特定的词位。

优点：能够平衡地看待词表词和未登录词的识别

生成式方法与区分式方法相结合

大部分基于词的分词方法采用的是生成式模型，而基于字的分词方法采用区分式模型

生成式模型

对 $P(o|q)$ 进行建模

假设 o 是观察值，q 是模型。如果对$p(o|q)$进行建模，就是生成式模型。其基本思想是：首先建立样本的概率密度模型，再利用模型进行推理预测。要求已知样本无穷多或者尽可能地多。该方法一般建立在统计学和 Bayes 理论的基础之上。

从统计的角度表示数据的分布情况，能够反映同类数据本身的相似度

优点：所带的信息比判别式模型丰富，研究单类问题比判别式模型灵活，模型可以通过增量学习得到，且能用于数据不完整情况

缺点：学习和计算过程比较复杂

判别式模型

如果对后验概率 $p(q|o)$ 进行建模，就是判别式模型

模型。基本思想是：有限样本条件下建立判别函数，不考虑样本的产生模型，直接研究预测模型。表性理论为统计学习理论。

特点：寻找不同类别之间的最优分类面，反映的是异类数据之间的差异

优点：判别式模型更容易学习

缺点：黑盒操作，变量间关系不清楚

基于字的区分模型有利于处理集外词，基于词的生成模型更多考虑词汇之间以及词汇内部字与字之间的依存关系

结合方法1

✧ 结合方法1：将待切分字串的每个汉字用$[c,\ t]_i$ 替代，以$[c,\ t]_I$ 作为基元，利用语言模型选取全局最优(生成式模型)。

其中 c 是原本的汉字，t 是判别式模型中的 $B.M.E.S$ 符号

- 优点：
  - 充分考虑了相邻字之间的依存关系
  - 相对于区分模型，对集内词有较好的鲁棒性
- 弱点：难以利用后续的上下文信息

结合方法2

✧ 结合方法2：插值法把两种方法结合起来

$$Score(t_k) = \alpha \times \log(P([c,t]_k \mid [c,t]_{k-2}^{k-1})) + (1-\alpha) \times \log(P(t_k \mid c_{k-2}^{k+2}))$$

$$(0.0 \le \alpha \le 1.0)$$

Generative score

Discriminative score

优点：充分结合了基于字的生成模型和基于字的区分式模型的优点

方法比较

- 最大匹配分词算法是一种简单的基于词表的分词方法。只需要最少的语言资源（仅一个词表，不需要任何词法，句法，语义知识），程序实现简单，但是对歧义字段的处理能力不够强大
- 全切分方法首先切分出与词表匹配的所有可能的词，然后运用统计语言模型和决策算法决定最优的切分结果。这种方法的优点是可以发现所有的切分歧义，但是解决歧义的方法很大程度上取决于统计语言模型的精度和决策算法，需要大量标注语料。且分词速度会因搜索空间的增大而变慢
- 最短路径分词方法的切分原则是使切分出来的词数最少。这多数情况下符合汉语的语言规律，但无法处理例外情况。而且，如果最短路不止一条，往往不能确定最优解
- 统计方法具有较强的歧义区分能力，但需要大规模语料库的支持，需要的系统开销也大

未登录词识别

命名实体

其他新词

中文姓名识别方法

- 姓名库匹配，以形式作为触发信息，寻找潜在名字
- 计算潜在姓名的概率估值及相应姓氏的姓名阈值，根据姓名概率评价函数和修饰规则对潜在姓名进行筛选

流程

- 计算概率估计值
- 确定阈值
- 设计评估函数
- 使用修饰规则
    - 如果姓名前是一个数字，或与 . 字符的距离小于2个字节，否定此姓名
    - 确定潜在的姓名边界
        - 左界规则：若潜在姓名前面是一个称谓，标点符号，或潜在姓名在句首，或姓氏使用率为100%，则姓名左界确定
        - 右界规则：姓名后面是称谓，指界动词，标点，或潜在姓名在句尾，尾字使用频率为100%，则姓名右界确定
    - 校正潜在的姓名
        - 依据：含重合部分的潜在姓名不可能同时成立

中文地名识别方法

基本资源

- 建立地名资源知识库
- 建立识别规则库

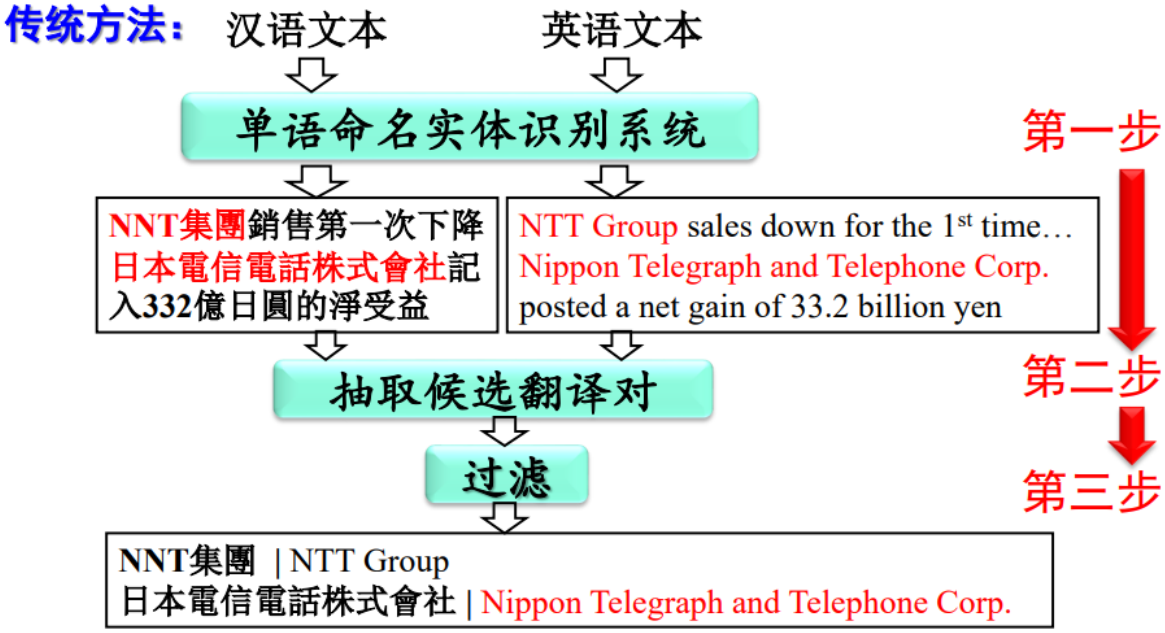
基本方法

- 统计模型
- 通过训练语料选取阈值
- 地名初筛选
- 寻找可以利用的上下文信息
- 利用规则进一步确定地名

中文机构名识别

- 找到一机构称呼词
- 根据相应规则往前逐个检查名词作为修饰名词的合法性，直到发现非法词
- 如果所接受的修饰词同机构称呼词构成一个合法的机构名称，则记录该机构名称
- 统计模型

双语实体自动识别与对齐的联合模型（没记，感觉不考）



基于神经网络的命名实体识别方法

把 NER 看作序列标注任务，输入输出均为序列

- 在序列标注任务上，RNN(LSTM)优于CNN；
- LSTM无需使用外部词表资源，效果依然很好；
- 可同时应用到多种语言，多种序列标注任务上；
- LSTM变种结构多、参数多、调参过程困难。

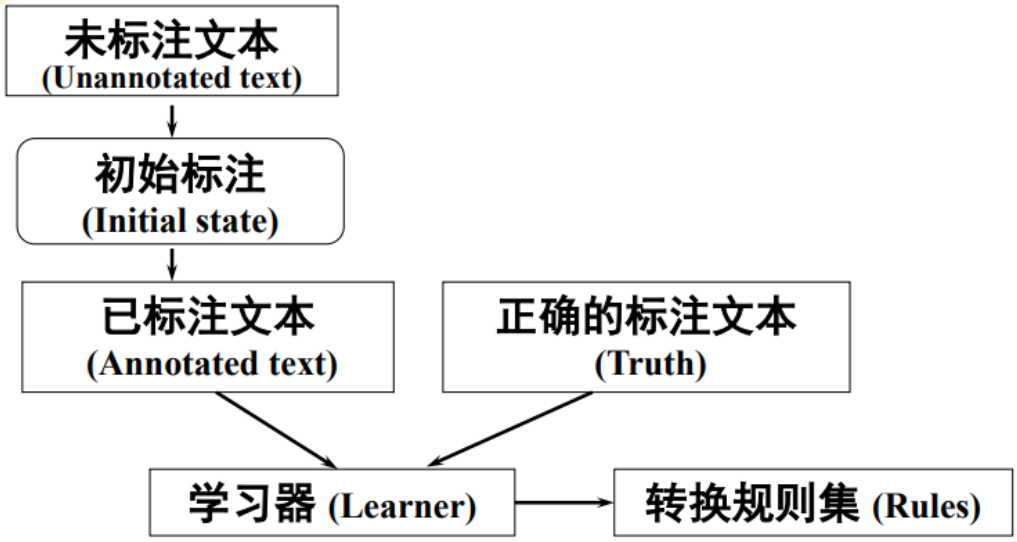
词性标注概述

词性标注的主要任务是消除词性兼类歧义（一个词可以表达多个意思，多种词性）

词性标注方法

基于规则的词性标注方法

- 手工编写词性歧义消除规则
- 机器自动学习规则：基于错误驱动的机器学习方法
  - 初始词性赋值
  - 对比正确标注的句子，自动学习结构转换规则
  - 利用转换规则调整初始赋值



基于转换规则的错误驱动的机器学习方法

基于HMM的词性标注方法

规则和统计相结合的词性标注方法

- 规则消岐，统计概率引导
- 统计方法赋初值，规则消岐

分词与词性标注技术水平

当前分词技术存在的主要问题

- 分词模型过于以来训练样本，而标注大规模训练样本费时费力，且局限于个别领域，导致对新词的识别能力差
- 现有的训练样本主要在新闻领域，而实际应用千差万别