- content
  {:toc}

序列标注

NLP四大任务

1. 序列标注：分词、词性标注、语义标注和命名实体识别
2. 分类任务：文本分类、情感分析和文本聚类
3. 句子对关系判断：自然语言推理、 问答QA、文本语义相似度
4. 生成任务： 机器翻译、文本摘要、IR等

序列标注

1. CRF
   条件随机场，CRF，
   wait
2. BI-LSTM
   双向LSTM网络
   wait
3. BI-LSTM + CRF
   标注过程条件随机场
   wait
4. Viterbi 算法
   动态规划算法，
   wait

分词算法现状

NLP的底层任务由易到难大致可以分为词法分析、句法分析和语义分析。分词是词法分析（还包括词性标注和命名实体识别)中最基本的任务。成熟95% 以上，说复杂是剩下的5% 很难突破。

1. 粒度，不同应用对粒度的要求不一样
2. 歧义，
3. 未登录词，比如“skrrr"，

使用分词包，也要对这些基础技术有了解，必要时对分词器做调整，介绍常用的分词算法，以及其核心思想进行介绍。

2. 常用分词算法

分词算法根据其核心思想主要分为两类，

3. 基于字典的分词，先把句子切分成词，再寻找词的最佳组合方式；
4. 基于字的分词，即由字构词，先把一个句子分成一个个字，再将字组合成词，寻找最优的切分策略，同时也可以转化为序列标注问题。

２.1 基于词典的分词

2.1.1 最大匹配分词算法

最大匹配分词寻找最优组合的方式，是将匹配到的最长词组合在一起。主要的思路是先将词典构造成一棵Trie树，也称为字典树，如下图：

Trie树由词的公共前缀构成节点，降低了存储空间的同时提升了查找效率。最大匹配分词将句子与Trie 树进行匹配，在匹配到根节点时由下一个字重新开始查找。比如“正向匹配”和“反向匹配”。

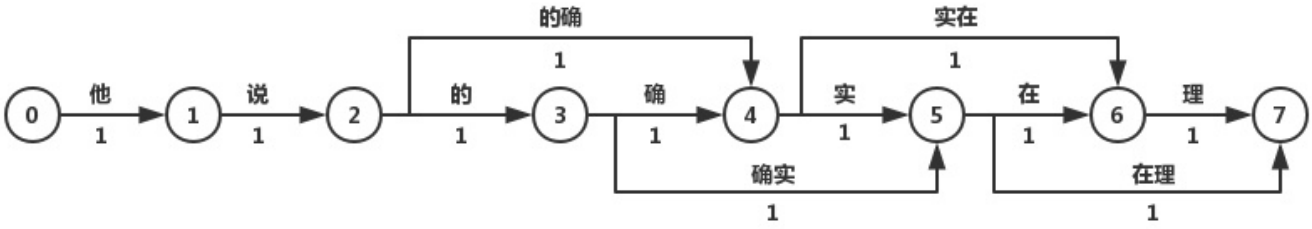
这种分词是最早的分词方法，O(n)时间对句子分词，但是效果很差。

2.1.2 最短路径分词算法

先将一句话中所有词匹配出来，构成词图 (有向无环图DAG) , 之后寻找从起始点到终点的最短路径作为最佳组合方式。

2.1.2.1 最短路径分词算法

每个图的边为字，节点为开始和结束。找到开始到结束的最短路径。



2.1.2.2 N-最短路径分词算法, 保留N条最短路径，回溯

2.1.3 基于n-gram model的分词算法

在前文的词图中，边的权重都为１。而现实中却不一样，常用词的出现频率较高。原有问题转化为最大概率路径问题，即分词结果为D"最有可能的词的组合" 。充足的语料，因此分词任务已经从单纯的"算法" ,上升到"建模"，即利用统计学方法结合大数据挖掘，对"语言"进行建模。

P(他说的确实在理) ＝ p(他) p(说|他) p(的｜他说) p(确｜他说的)... p(理｜他说的确实在)

二元模型～作为概率图的概率

对二元概率图求最大概率路径，即可得到分词结果。

２.2 基于字的分词

与基于字典的分词不同的是，基于字的分词事先不对句子进行词的匹配，而是将分词看成序列标注问题，把一个字记成

B (Begin), I(Inside),O(Outside),E(End),S(Single)。因此也可以看成是每个字的分类问题，输入为每个字及其前后字所构成的特征，输出为分类标记。对于分类问题，可以用统计机器学习或神经网络的方法求解。

输入 X，得到f(x) = Y。另外，机器学习中一般将模型分为两类：生成式模型和判别式模型，两者本质区别在于 X 和 Y 的生成关系。
生成式模型："输出 Y 按照一定规律和输入 X 生成"
判别式模型："直接对后验概率P(Y|X)进行建模"

2.2.1 生成式模型分词算法

生成式模型主要有 n -gram模型，HMM 隐马尔科夫模型，朴素贝叶斯分类等。

HMM模型?

HMM模型认为在解决序列标注问题时存在两种序列，一种观测序列，即人们显性观察到的句子，而序列标签是隐状态序列，即

S - > S -> S - >B -> E -> B -> E