# Chinese Word Segmentation

中文分词(Chinese Word Segmentation) 指的是将一个汉字序列切分成一个一个单独的词。分词就是将连续的字序列按照一定的规范重新组合成词序列的过程。

为什么英文不需要分词而中文需要分词（为什么中文的词之间没有分隔）？

1. 古代汉语中除了连绵词和人名地名等，词通常就是单个汉字，所以当时没有分词书写的必要。而现代汉语中双字或多字词居多，一个字不再等同于一个词。
2. 现代汉语的基本表达单元虽然为“词”，且以双字或者多字词居多，但由于人们认识水平的不同，对词和短语的边界很难去区分。

分词的分类

1. 基于字符串匹配的分词方法

又叫做机械分词方法，它是按照一定的启发式规则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配，若在词典中找到某个字符串，则匹配成功（识别出一个词）。按照扫描方向的不同，串匹配分词方法可以分为正向匹配和逆向匹配；按照不同长度优先匹配的情况，可以分为最大（最长）匹配和最小（最短）匹配。

常用的机械分词法：

1）正向最大匹配法（由左到右的方向）；

2）逆向最大匹配法（由右到左的方向）；

3）最少切分（使每一句中切出的词数最小）；

4）双向最大匹配法（进行由左到右、由右到左两次扫描）

一般说来，逆向匹配的切分精度略高于正向匹配，遇到的歧义现象也较少。统计结果表明，单纯使用正向最大匹配的错误率为1/169，单纯使用逆向最大匹配的错误率为1/245。但这种精度还远远不能满足实际的需要。实际使用的分词系统，都是把机械分词作为一种初分手段，还需通过利用各种其它的语言信息来进一步提高切分的准确率。

机械分词的改进思路：

1. 特征扫描，优先在待分析字符串中识别和切分出一些带有明显特征的词，以这些词作为断点，可将原字符串分为较小的串再来进机械分词，从而减少匹配的错误率。
2. 另一种方法是将分词和词类标注结合起来，利用丰富的词类信息对分词决策提供帮助，并且在标注过程中又反过来对分词结果进行检验、调整，从而极大地提高切分的准确率。

维护一个不成词字表

一旦涉及到组合型歧义的问题，最 大匹配、最少词数、概率最大等所有策略都不能实现具体情况具体分析。

常用机械分词工具：ikanalyzer,paoding

1. 基于理解的分词方法

这种分词方法是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。它通常包括三个部分：分词子系统、句法语义子系统、总控部分。在总控部分的协调下，分词子系统可以获得有关词、句子等的句法和语义信息来对分词歧义进行判断，即它模拟了人对句子的理解过程。这种分词方法需要使用大量的语言知识和信息。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还处在试验阶段。

1. 基于统计的分词方法。

基于统计机器学习的方法。首先给出大量已经分词的文本，利用统计机器学习模型学习词语切分的规律（称为训练），从而实现对未知文本的切分。这种方法能很好处理歧义和未登录词的问题，但最大缺点是需要有大量预先分好词的语料作支撑，而且训练过程中时空开销极大。

常见的序列标注模型有HMM和CRF。ICTCLAS是基于HMM的分词库。

无词典分词

http://www.matrix67.com/blog/archives/5044

分词的技术难点

1. 歧义

交集型歧义

如：结合成

组合型歧义

如：把手

真歧义

如：乒乓球拍卖完了

1. 新词识别

命名实体、新词、术语称为未登录词

常用分词工具：

LTP

哈工大LTP

中科院计算所NLPIR

清华大学THULAC

Jieba

BosonNLP

IKAnalyzer

SCWS

盘古分词

庖丁分词

搜狗分词

1、一个好的分词工具不应该只能在一个数据集上得到不错的指标，而应该在各个数据集都有很不错的表现。从这一点来看，thulac和ltp都表现非常不错。

2、因为分词是个基础部件，分词速度对于一个分词工具来说也至关重要。从这一点来看，thulac和jieba表现的不错。

3、大家都知道，基本的分词依赖模型，但真正想用分词工具来解决应用层面上的问题，都需要借助于词库，本文测试的4个工具均支持用户自定义词库。

4、特别需要强调的一点是，哈工大的ltp支持分词模型的在线训练，即在系统自带模型的基础上可以不断地增加训练数据，来得到更加丰富、更加个性化的分词模型。