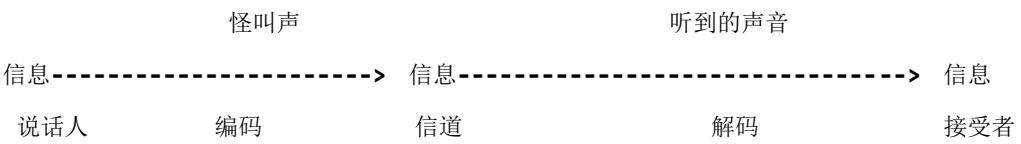
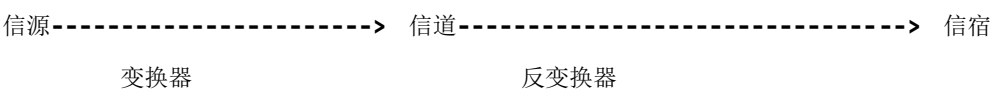
第 1 章 文字和语言 vs 数字和信息

数字，文字和自然语言一样，都是信息的载体，它们之间原本有着天然的联系。

通信模型



在通信领域 通信系统的模型



聚类： 随着文明的进步，信息量的增加，文字的数量便不会随文明的进步而增多。

因为没有人能够记得那么多的文字。所以概念的第一次概括和归类就开始了。

将物理或抽象对象的集合分成由类似的对象组成的多个类的过程叫做聚类。

百度百科 http://baike.baidu.com/view/31801.htm

书中讨论更多的是数据聚类。 http://zh.wikipedia.org/wiki/%E6%95%B0%E6%8D%AE%E8%81%9A%E7%B1%BB

上下文区分：文字按照意思来聚类，最终会带来一些歧义性。区分这些聚类的最好的方法就是 上下文区分。

罗塞塔石碑： 它是公元前 32 世纪历史的记录者。而从它的身上可以得到几点指导意义。

1. 信息的冗余是信息安全的保证。它的内容 重复了三次。
2. 语言的数据（语料），尤其是双语或者多语的对照语料对翻译至关重要，它是从事机器翻译的基础。

罗塞塔石碑上用三种文字记录了托勒密的诏书。

基于罗塞塔的意义，今天很多的翻译软件都叫 罗塞塔。

最短编码原理：常用字短，生僻字长。

信道传输原理： 宽信道，无需压缩。窄信道，需压缩。

如： 在古代，两个人说话，宽信道，不压缩。 书写慢是窄信道，需压缩。信道压缩就是把白话文写成文言文

在今天，wap 是窄信道，需压缩，页面设计较小。宽带互联网则是宽信道，页面设计较大。

校验码：　　为了抄圣经的时候 检查抄写错误。他们把 每一个希伯来字母对应于一个数字，这样每行文字加起来便得到一个特殊的数字这个数字便成为这一行的校验码。同样，对应每一列也是这样处理，当抄完一页的时候，他们需要把每一行的文字加起来，看看新的校验码是否和原文的相同。然后对每一页进行同样的处理。

百度百科 http://baike.baidu.com/view/243582.htm 不同的校验码有不同的检验方法。

注：本章提到的<<从一到无穷大>>书中的原始部落和酋长，但是在原书中第一章大数开始的部分讲的是匈牙利的贵族。