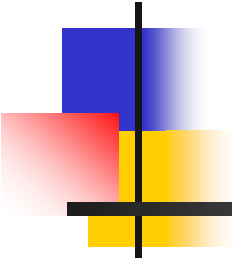
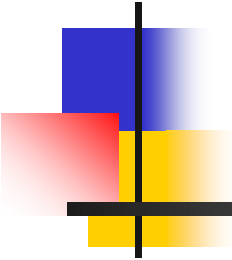


# 第2章 语言学及数学基础知识



# 一、语言学基础知识



# 1.1 语音

# 1.1 语音-语音的划分

人类是通过语音来感知语言、理解语言的。我们听到一种完全陌生的语言时，会感到它只是一连串的不断变化的声音，理解不了它的意义。如果听到了的是自己的母语或者是学习过的语言，就能很快地将连续不断的语言之流分成长短不一各种语音片段，理解它的意义。

音流可以分解为语调上和意义上的完整音段，也就是通常说的句子。句子和句子之间可以有较大的停顿，每个句子又可以进一步分成若干个较小的音段，叫做**节拍群**。例如：**五星红旗/飘扬在/天安门广场上**，可以分成

# 1.1 语音-语音的划分

三个**节拍群**，中间有较小的停顿。节拍群中有**重音**和**轻音**，还可以分成更小的音段---称为**音节**，它是在听感上可以自然感到的最小的语音片段，**是语音的基本结构单位**。比如“五星红旗”在语音上就是四个音节。

“红”和“同”对比一下读音就可以发现，一个音节又可以分为更小的片段，“红”的语音有h和ong两部分构成，分别称为**“声母”**和**“韵母”**。对汉语来说，还有一个**声调**，比如轰（hong1）和红（hong2）意义是不一样的。



# 1.1 语音-语音的划分

声母一般不可再分，韵母可以进一步分为**音素**，例如“豪”的韵母是ao，可以分成两个明显不同的声音a和o，这是音流中最小的语音单位，称为**音素**。

**音素**根据不同的性质分为**元音**和**辅音**。

**语调、音节、轻重音、声母、韵母、声调、音素、元音、辅音**等都是语音学里的基本概念。

我们研究自然语言处理，尤其是涉及语音处理的，就要弄清这些概念，了解他们所指的语音成分及在语音中的作用，**是怎样构成表达语言的语音系统的。**



# 1.1 语音-语音的符号

语音一发即逝，不留踪迹，必须有一套符号记录下来，才便于学习和分析研究。

通行最广的是汉语拼音方案和国际音标。

汉语拼音方案是1958年2月由第一届全国人民代表大会第五次会议批准，推行全国。与以往的直音、反切以及笔画式的注音符号相比，有以下优点：

- (1) 符号数量少，26个；
- (2) 采用国际通行的拉丁字母
- (3) 字母音素化，用来记录或分析语音准确灵活。

汉语拼音方案主要设计人是周有光教授。

# 1.1 语音-语音的符号

周有光，原名周耀平，1906年1月13日出生于江苏常州，是中国著名的语言学家，被誉为“汉语拼音之父”。他在1955年被调至北京，进入中国文字改革委员会，专职从事语言文字研究。



主要成就包括参与设计“汉语拼音方案”并主持制订了《汉语拼音正词法基本规则》。2017年1月14日去世，享年112岁。2018年4月被评为“逝世的十位国家脊梁”之一。2019年12月6日获评第七届“中华之光——传播中华文化年度人物”致敬奖。



次 雌 田 笔字排作 zhi、chi、shi、ri、zi、ci、si。

# 1.1 语音-语音的符号

## 4. 声调符号

阴平	阳平	上声	去声
—	/	∨	\

声调符号标在音节的主要母音上。轻声不标。例如：

妈 mā	麻 má	马 mǎ	骂 mà	吗 ma
(阴平)	(阳平)	(上声)	(去声)	(轻声)

## 5. 隔音符号

a, o, e 开头的音节连接在其他音节后面的时候,如果音节的界限发生混淆,用隔音符号(′)隔开,例如 pi'ao(皮袄)。

# 1.1 语音-语调和语气

从声波的振幅、周期、频率等属性可以说明语音的四个要素：音高、音强、音长、音色（音质）

· 一句话的词汇意义加上语调意义才算是完全的意义。  
。这就涉及的语调的问题

语调是指说话的腔调，即一句话里声调高低抑扬轻重的配置和变化。语调的意义在于表达说话人的态度或口气，使一句话的意义更加完整。 例如：升调-他来了 和 降调-他来了 的意义就不同

# 1.1 语音-语调和语气

语气则是指在某种特定思想感情的支配下，语句的声音形式。语气通过声音和气息来表达不同的语意和感情，是“声气传情”的技巧。

语调和语气的主要区别在于它们的表现形式和功能。语调主要关注的是说话时声音的高低、强弱、快慢等变化，这些变化能够传达说话人的情感 and 态度。而语气则更侧重于通过声音形式来表达特定的情感 and 态度，如喜、怒、哀、乐等。语气以内心感情的色彩和分量为灵魂，以具体的声音形式为躯体。

# 1.1 语音-声母和韵母

声母容易混淆的:

zh z; ch c; sh s

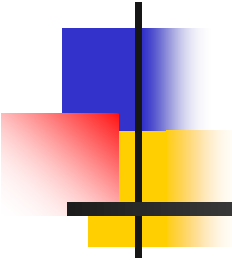
l n; h f;

韵母容易混淆的:

eng en; ing in

ai an; ei en; ui un

uai uan; iai ian



## 1.2 文字



## 1.2 文字-汉字的特点

语音是语言的物质外壳，人类的语言是有声的语言。但语音是一发即逝，为了把语言记录下来，发明了文字。文字是语言的视觉形式，突破了语言的时空限制，扩大了语言交际范围。

拼音文字的字母只有形和音，没有意义，而汉字不但有形和音，而且也有意义，所以，汉字是形音义的统一体，是表意文字。

## 1.2 文字-汉字的结构

汉字的结构可分为独体字和合体字两类。由笔画直接连接组合形成的字叫做独体字，如：人、山、养、口；由较小的结构单位组合而成的字叫做合体字，合体字的字形具有可分析性：林、休、体、知、炙、癸、炎。

构成字的部件之间有如下7中组合方式：

- (1) 左右型； (2) 左中右型； (3) 上下型；
- (4) 上中下型； (5) 全包围型； (6) 半包围型
- (7) 穿插型



## 1.2 文字-汉字的简化与整理

- 1956年国务院公布《汉字简化方案》，515个简化汉字和54个简化偏旁。1959年7月推行完毕。
- 1964年公布《简化字总表》。
- 1986年公布调整后的《简化字总表》，共包括2235个简化字。

	1-10画	11-20画	21-27画	合计
简化前	917字	1030字	53字	2000字
简化后	1395字	570字	2字	1967

## 1.2 文字-字量、字音、字序

### 一、字量

汉字的总字数：

年代	书名	收字量
公元100年	说文解字	9353
1008年	广韵	26194
1716年	康熙字典	47043
1915年	中华大字典	48000多
1968年	中文大辞典	49905
1986年	汉语大字典	54678

## 1.2 文字-字量、字音、字序

### 一、字量 **常用字：**

汉字序号 (按频率高 低进行排列)	10	40	160	950	2400	3800	5200
累计频度	11%	25%	50%	90%	99%	99.9%	99.99%

1988年，国家教委和国家语言文字工作委员会联合公布了《现代汉语常用字表》，包括常用字2500，次常用字1000，共3500字。覆盖率为99.48%

**通用字：**1988年3月，现代汉语通用字表包含7000字，常用和非常用各3500字。



# 1.2 文字-字量、字音、字序

## 二、字音

汉语字音的确定是按照北京话的语言系统确定的。

### (1) 多音字:

《新华字典》中共有多音字828个，1857个读音，占字典总数的10%

1979年出版的《辞海》有多音字2641个，占字典总数的22%，其中一字二音的2112个，一字三音的422，一字四音的81个，一字五音的18个，一字六音的7个，一字八音的1个。比如“和”就有5中读音，“参”有三种



# 1.2 文字-字量、字音、字序

## 二、字音

### (2) 同音字：

普通话共有1300多个音节，汉字字数如果按1万字计算，每个音节负载7.5个汉字。

常用汉字大概有400多个音节。

## 三、字序

(1) 部首法

(2) 笔画法

(3) 四角法

(4) 音序法

## 1.2 文字-汉字的前途

以前的国家落后，很多人把落后的原因归结为汉字的使用。鲁迅、瞿秋白等进步文化人士都积极推动拉丁化运动。他们激烈地抨击汉字，认为汉字“是劳动大众身上的结核，病菌都潜伏在里面，倘不首先除去它，结果只有自己死”。

鲁迅甚至提出了“汉字不灭，中国必亡”。

新中国成立后，推行文字改革，提出了三项任务：简化汉字、推广普通话、制定和推行汉语拼音方案。周总理明确指出，汉语拼音方案是为汉字注音和标注普通话的，不是代替汉字的拼音文字。

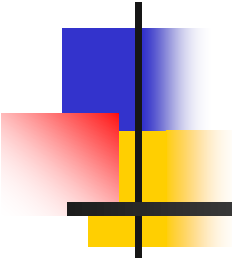


## 1.2 文字-汉字的前途

信息化时代，我国科研工作者积极进取，取得很大的成绩：

- (1) 操作系统的汉化和输入法的研制
- (2) 汉字的定量分析得到许多重要的数据
- (3) 计算机处理汉字的研究取得很大的成绩
- (4) 从神经心理学的方面对汉字进行研究，提出了很多值得深入探讨的问题
- (5) 汉字和拼音文字的比较研究正在深入，人们认识到汉字和拼音文字各有优缺点

**我们计算机学科学生肩负着汉语自然语言处理的重任，为国家的科技发展做出重要贡献。**



## 1.3 词汇



## 1.3 词汇

### 一、什么是词

词是最小的能够独立运用的、有意义的、最小的语言单位。

词分为**实词**和**虚词**

**实词**是指那些意义比较实在的词。如，“山”就是一个实词，它是指“地面形成的高耸部分”，“耸立”、“险峻”等都是实词。

**虚词**是指意义比较抽象的词。如“的”、“和”、“但是”、“虽然”等。虚词一般在句子中表示语法意义。

# 1.3 词汇

## 二、语素和词

**语素** (morpheme) 是语言中最小的音义结合体，必须同时满足“最小、有音、有义”三个条件。语素是语言中能独立表示意义的最小部分，可以通过它组成更长的语言单位，如词、短语、句子等。**语素定义包括三个意思：**语素具有一定的语音形式；这个语音形式表示一定的意义；这个语音语义结合体必须是“最小的”。

例如：“耸立”这个词可以进一步分为“耸”和“立”，“耸”——高高直立，“立”——竖立，它们就是语素。

## 1.3 词汇

### 语素的分类：

**自由语素：**可以独立成词，也能与其他语素组合成词。  
例如，“人”可以独立成词，也可以与其他语素组合成“人们”、“人民”等。

**粘着语素：**不能独立成词，只能粘附在其他语素之后。  
例如，“者”、“第”、“员”等。

语素的判断可以通过替换法进行，即用已知语素替换有待确定的语言单位，看其是否还能保持原有的意义和功能。

## 1.3 词汇

语素按音节划分，可以分为单音节语素、双音节语素和多音节语素。汉语中，单音节语素如“人”、“电”“耸”、“立”，双音节和多音节语素如“葡萄”、“徘徊”“白兰地”、“尼古丁”、“盘尼西林”等。多音节语素中的每个音节没有任何意义。

语素的概念不仅适用于口语，也适用于书面语言。在书面语言中，语素同样以文字形式存在，与语音形式一起实现其表达功能。因此，**语素也可以被认为是语言中的“最小的音形义结合体”**。

## 1.3 词汇

语素和词的关系：词是由语素构成的。分为两种情况：

- (1) 一个词由一个语素构成
- (2) 一个词由两个以上的语素构成。

在两个语素组成的语言单位中，不成词语素+不成词语素的是词：具体、特殊、思索、介绍、感性

不成词语素+成词语素的是词：普通、习惯、考虑、阅读

成词语素+成词语素的有的是词组：大楼、小树、红笔、好米

如果意义不是两个语素的简单组合又不能拆开的就是词：钢笔、火车、小米、马路

# 1.3 词汇

## 三、构词法---英语构词法

英语中的词汇大致可以分为两类：一类词汇是基本词，又称为原生词（primitives）；另一类词汇是新词。

基本词大都是单音节的，这种词是不可拆分成为两个独立词汇的，如fish、sun等。英语中的原生词数量有限，随着社会的发展，语言自身的扩充以及不同语言之间的融合，使得语言变得越来越复杂，有限的原生词已经不能满足人们的需求。为了表示新鲜事物与概念，一些新词随着社会的发展逐渐产生，这些新词并不是随意创造的，而是按照语言的一定规律进行创造的，**这种创造新词的方法就称为构词法，或称为构词理论。**



# 1.3 词汇

## 三、构词法---英语构词法

英语构词法主要有派生法、合成法、转化法、缩略法和混合法等五种方法。

### (1) 派生法

首先介绍英语构词法中派生法涉及的两个基本概念：**词根**和**词缀**。**词根**是表示实际意义的单词，大家经常用到的很多单词都可认为是词根，例如work、aware等。词缀是通常不能单独作为单词使用，只能附加在一个具有实际意义的单词上来使用。

# 1.3 词汇

## 三、构词法---英语构词法

词缀附加在单词前面就称为**前缀**，附加在单词后面就称为**后缀**。比如unaware中的un就是前缀，而worker中er则是后缀。**所谓的派生法**就是在词根的前面或后面加上词缀所以形成新的单词，所构成的新单词在词义上与原词根的词义有所变化或是其词义与词根的词义截然相反。

◆ **前缀**。单词加上前缀之后，一般单词的词性不会改变，只改变词义。前缀主要分为两种，一种是表示否定的前缀，另一种是表示其他意义的前缀。



## 1.3 词汇

**①表示否定的前缀：**主要有anti-， counter-， dis-， im-， in-， ir-， il-， mis-， non-和un-等,在单词的前面加这类前缀通常构成与该词意义完全相反的新词，如表1。

前缀	用法	例子
anti-	加在名词、形容词前	antibiotic(抗生素)
counter-	加在名词、动词前	counterattack(反击)
dis-	加在名词、动词和形容词前	disadvantage(缺点)
im-	加在首字母为m,b,p的单词前	impossible(不可能的)
in-	加在名词、形容前	incorrect(不正确的)
ir-	加在首字母为r的单词前	irregular(不稳定的)
il-	加在首字母为l的单词前	illegal(不合法的)
mis-	加在名词、动词前	misunderstand(误解)
non-	加在名词、形容词前	non-alcoholic(不含酒精的)
un-	加在名词、形容词和副词前	unaware(未察觉到的)

## 1.3 词汇

### ②表示其他意义的前缀:

还有很多前缀可以表示具体含义，如fore-表示“在.....之前”，forecast（预报）；re-表示“再、又”，如rewrite（重写）；a-表示“的”，多构成表语形容词，如alone（单独的）；tele-表示“远程的”，如telephone（电话）；en-表示“使”，构成动词，如enable（使能够）；inter-表示“关系”，如international（国际的）。

## 1.3 词汇

◆ **后缀。** 单词在加上后缀之后，一般会构成一个新的单词，这个单词的词义和原来单词的词义相同，但词性可能会发生变化。也有少数单词在加上后缀之后，不但词性会发生变化，而且词义也有所变化。后缀的种类主要依据要加后缀的单词词性来划分，主要包括五大类，分别是名词后缀、动词后缀、形容词后缀、副词后缀和数词后缀等。

## 1.3 词汇

①**名词后缀**：常用的名词后缀有-ment/-ness, -tion, -(e)r/-or, -ese, -ian, -ess, -ist等，如表2。

后缀	含义	例子
-ment/-ness	用来表示性质或状态	agree→agreement（协议） happy→happiness（幸福）
-tion	表示动作或是过程	explain→explanation（解释）
-(e)r/-or	表示从事某事的人	write→writer（作家）
-ese	表示某地人	China→Chinese（中国人）
-ian	表示精通.....的人	music→musician（音乐家）
-ess	表示雌性	act→actress（女演员）

## 1.3 词汇

②**动词后缀**：常见的动词后缀有-ate, -en, -ify, -ise/ize等如表3。

后缀	含义	例子
-ate	使.....成为	valid→validate(使合法化)
-en	使.....; 变得.....	rich→richen(使富)
-ify	转为; 变为	beauty→beautify(变美)

## 1.3 词汇

③**形容词后缀**：常用的形容词后缀有-(a)n/-ese, -able, -ern, -less, -y等，如表4。

后缀	含义	例子
-able	表示有能力的	eat→eatable, 能吃的
-ern	表示方向的	east→eastern (东方的)
-less	表示否定的	home→homeless (无家可归的)
-y	表示天气	snow→snowy (雪的)

## 1.3 词汇

④**副词后缀**：常见的副词后缀有-er, -est, -ly, -ward(s)和wise等，如表5。

后缀	含义	例子
-er	形容词比较级，更……地	hard→harder(更努力地)
-est	形容词最高级，最……地	hard→hardest(最努力地)
-ly	以……方式	easy→easily(容易地)
-ward(s)	表示方向	back→backward(向后)
wise	以……方式	clock→clockwise(顺时针方向地)

## 1.3 词汇

⑤**数词后缀**：常用的数词后缀主要有-teen, -ty和-th, 数词后缀表如表6。

后缀	含义	例子
-teen	表示十几	five→fifteen十五
-ty	表示几十	nine→ninty九十
-th	构成序数词	six→sixth第六



## 1.3 词汇

### (2) 合成法

合成法就是将两个或两个以上有独立意义的单词进行合成，构成一个新词。合成法所能构成的新词主要包括合成名词、合成动词、合形成形容词、合成副词、合成代词、合成介词等。

**①合成名词：**通常由名词和其他词性的词构成一个新的名词  
名词+名词，如class+room→classroom（教室）；  
形容词+名词，如loud+speaker→loudspeaker（演讲人）；  
名词+动词，如day+break→daybreak（黎明）。

## 1.3 词汇

②**合成动词**：通常由动词和其他词性的词构成一个新的动词

副词+动词，如under+stand→understand（理解）；

形容词+动词，如broad+cast→broadcast（广播）；

名词+动词，如sleep+walk→sleepwalk（梦游）。

③**合形成形容词**：通常由形容词和其他词性的词构成一个新的形容词，但是也有由非形容词构成的情况，同时，在构成新词时常使用连字符。如：

名词+现在分词，如English-speaking（讲英语的）；

## 1.3 词汇

名词+过去分词，如man-made（人造的）；

名词+形容词，如day-long（整天的）；

名词+to+名词，如one-to-one（一对一的）；

形容词+名词，如high-quality（高质量的）；

形容词+现在分词，如good-looking（相貌较好的）。

常由名词和其他词性的词构成一个新的名词。

名词+名词，如class+room→classroom（教室）；

形容词+名词，如loud+speaker→loudspeaker（演讲人）；

名词+动词，如day+break→daybreak（黎明）。

## 1.3 词汇

④**合成副词**：通常由副词和其他词性的词构成一个新的副词。

介词+副词，如for+ever→forever（永远）；

形容词+副词，如any+where→anywhere（任何地方）；

副词+副词，如how+ever→however（尽管如此）。

⑤**合成代词**：通常由代词和其他词性的词构成一个新的代词，但是也有由非代词构成的情况。

物主代词+self，如my+self→myself（我自己）；

代词宾格+self，如her+self→herself（她自己）；

形容词+名词，如any+thing→anything（一切）。

## 1.3 词汇

⑥**合成介词**：一般由介词和其他词性的词构成一个新的介词，但是也有由非介词构成的情况。

副词+名词，如out+side→outside（在……外面）；

介词+副词，如with+in→within（在……之内）；

副词+介词，如in+to→into（进入）。

随着科技进步，许多新的合成词在不断涌现，上面所列举例子只是常用的一些情况，并不能穷举出合成法的所有情况。另外，一个单词一般会有多个词性，这里的举例只是就某一词性来说明的，目的是帮助读者理解合成法的意图，请不要深究上文的词性判断。

## 1.3 词汇

### (3) 转化法

词语使用中，英语单词词性转化是非常活跃，人们把名词用作动词、动词转化为名词、形容词用作动词的现象非常多见，这种把一种词性用作另一种词性的方式就叫做词性的转化。词性转化后的单词，只要知道单词的原意，就可根据上下文判断出转化词的意义。

**①动词转化为名词。**在把动词转化为名词时，有时单词的词义并无变化，如Let me have a try。try本来是动词，在这里用作名词；有时单词的词义有一定的变化，如He was about the same build as his brother。build本来是动词，词义为建筑、建造，但在本句中用作名词时，词义却为体型、体格。

## 1.3 词汇

②**名词转化为动词**。在英语中有很多名词都可以用作动词，特别是有许多表示物体的名词可以用作动词来表示动作，例如，It can **seat** 1000 people. 其中的seat本是名词。词义为座位，但在这里用作动词，词义为容纳；表示某类人的名词也可做动词，He insisted on staying up to **nurse** the child. Nurse本为名词，词义为护士，但在这里用作动词，表示照顾、护理等词义。

③**形容词转化为动词**。在英语中也有少数形容词可以用作动词，如 Don't **dirty** your clothes. Dirty本为形容词。词义为脏的，在这里用作动词，表示的词义为弄脏；Please **warm** up the cold meat. Warm本为形容词，词义为暖和的，在这里作为动词使用，词义为加热。

## 1.3 词汇

④**介词转化为动词**。英语中还有少数介词也可以转化为动词，如Murder will out。out介词意为向、离去，在这里用作动词，表示败露的词义。

⑤**形容词转化为名词**。英语中表示颜色的形容词一般可转化为名词，例如，You should be dressed in black at the funeral。其中的Black为形容词，词义为黑色的，但在这里用作名词，意为黑色的衣服；The old in our village are living a happy life。其中的The old在此处用作名词，意为老年人。



## 1.3 词汇

### (4) 缩略法

缩略法本质上并不能产生新的词汇，只是对原有的词汇或短语进行缩短或是简写来代表该词或短语。缩略法将复杂的词语变得简单，便于记忆与书写。缩略法主要分为两类：

一类是对原来词汇进行裁剪，例如，telephone→phone；automobile→mobile；influenza→flu。

另一类是将一个短语中的多个单词的首字母提取出来，用这些字母的组合来表示这个短语的含义，例如，Internet Protocol→IP；very important person→VIP；National Basketball Association→NBA；World Trade Organization→WTO。

# 1.3 词汇

## 三、构词法---汉语构词法

在汉语构词理论中有一个很重要的概念就是语素，语素是最小的语音、语义结合体，是最小的、有意义的语言单位，但语素不是能独立运用的语言单位，它的主要功能是作为构成词或短语的材料，也就是说它是词的构成成分。例如“我们”有“我”和“们”2个语素。语素有单音节的，也有多音节的，例如，“我”、“们”等就是单音节语素，而“徘徊”、“雷达”、“尼古丁”等就是多音节语素。多音节语素中的各音节并没有意义，它们结合起来才能表示一个意义。

## 1.3 词汇

因此，语素是构成词语的单位，所谓的汉语构词理论就是汉语语素构成词语的方法。

语素主要分为两大类：词根和词缀。词根是词语结构体的基本构成部分，意义比较实在，如“电灯”中的“电”和“灯”。词缀是词语结构体的附加成分，没有具体的意义，主要起构词的作用。词缀根据其在构词时出现的位置，可以分为前缀、中缀和后缀三种形式，如“阿哥”中的“阿”属于前缀，“来得及”中的“得”属于中缀，“嫂子”中的“子”属于后缀。由语素构成的汉语词主要有单纯词、合成词和复合词三类。

## 1.3 词汇

因此，语素是构成词语的单位，所谓的汉语构词理论就是汉语语素构成词语的方法。

**(1) 单纯词。**指只由一个语素构成的词。语素可以分为单音节和多音节，多音节的语素构成的单纯词又可分为连绵词、口语词和音译词。

单音节单纯词：人、鸟、挑、美、三、百

多音节单纯词：枇杷、参差、吩咐（双声词）

骆驼、馄饨、膀胱（叠韵词）

饽饽、姥姥、奶奶、太太（叠音词）

## 1.3 词汇

多音节单纯词：枇杷、参差、吩咐（双声词）

骆驼、馄饨、膀胱（叠韵词）

饽饽、姥姥、奶奶、太太（叠音词）

叭，扑通、轰隆隆、叽叽喳喳（拟声词）

加仑、布拉吉、布尔什维克（译音词）

## 1.3 词汇

**(2) 合成词。**指由两个或是两个以上的语素构成的词。根据语素的种类不同，合成词可分为两类：**重叠词和派生词**。**重叠词**是指由词根重叠而成的词，而**派生词**则是指由词根和词缀组合而成的词。

**重叠词**在我们的生活中比较常见，有单字重叠的情况，也有双字重叠的情况。

单字重叠：“爸爸”、“妈妈”、“爷爷”、“奶奶”等

双字重叠：“马马虎虎”、“形形色色”、“啰啰嗦嗦”、战战兢兢等。

## 1.3 词汇

**派生词**根据词根和词缀的组合位置的不同可以分为“前缀+词根”、“词根+后缀”和“词根+中缀+词根”这三种形式。

前缀+词根：阿爸、阿福、老周、大哥、阿姨；

词根+后缀：嫂子、妗子、工程师、理发员；

词根+中缀+词根：来不及、来得及、糊里糊涂、土里土气、古里古怪、妖里妖气、微乎其微。



# 1.3 词汇

## (3) 复合词

复合词是指由词根和词根组合而成的词，根据不同的组合方式主要分为下列六种类型：

**①联合式。**指由两个意义相近、相关或相反的词根并列组合而成的词。根据两个词根之间的意义关系，可以分为：

- A、两个词根意义相近，例如“糊涂”和“思想”这类词；
- B、两个词根意义并列，但是构成一个新的词义，如“江湖”、“河海”这类词；
- C、两个词根的意义相反，如“上下”和“接送”这类词



## 1.3 词汇

②**偏正式**。指在构成词的两个词根中，用前一个词根修饰或限制后一个词根，而所构成的词的词义是以后一个词根的词义为主，前一个词根的词义为辅，主要可以分为三类：

A、名词性的，如“茶杯”和“黑板”这类词；

B、谓词性的，如动词“嚎哭”和形容词“崭新”这类词

C、副词性的，如“立即”和“何必”这类词。

③**述宾式**。指构成词的两个词根之间的关系是支配和被支配的关系，其中前一词根为表示支配关系的动作或是行为，而后一词根为表示被支配关系支配对象。这种形式主要构成动词，如动词“打雷”、“唱歌”、“跳舞”等。

## 1.3 词汇

④**述补式**。又称动补式复合词，是指构成词的两个词根中前面一个为动词性词根，表示中心成分，后面一个词根作为一种结果状态对前面的动词性词根进行补充说明。这类形式构造的复合词一般都是动词，如“揭露”和“改正”等就是述补式复合词。

⑤**主谓式**。指构成词的两个词根前后是陈述和被陈述的关系，例如“心虚”和“地震”这类词就是主谓式复合词。这里“虚”陈述“心”的状态，“震”陈述“地”的状态。

## 1.3 词汇

⑥**量补式**。指构成词的两个词根中前面一个为名词性词根，后面一个词根为计量单位，对前面的名词性词根进行补充说明。这类形式构造的复合词都是名词，其不受数量词的修饰，如“船只”和“马匹”等就是量补式复合词。

# 1.3 词汇

## 四、词语与语素义

词义是词的内容，是对客观事物现象的反映。例如，“自行车”的词义反映了自行车的4个特点。

词的附属色彩

(1) 感情色彩---指词义所附带的表示褒贬态度的色彩。

(2) 语体色彩---指有些词只适用于某一种交际范围、场合、文体当中，而不适合于另外的场合、文体。有些是适合于书面语，有些适合于口语。



# 1.3 词汇

## 四、词语与语素义

语素义是指存在于语素所构成的合成词或固定结构总的语素的意义。

不成词的语素意义只存在与构成的词中。

成词语素的意义有两种：一种是它的某个意义既是词义又是语素义；另一种是成词语素的意义只是语素义。例如“空”有一个意思是“天空”，这个意思只是语素义，只能存在于“空”所构成的词中，如“高空”“太空”“空降等。

”

# 1.3 词汇

## 四、多义词与同音词

多义词是指同一个词在不同的上下文中有不同的意义。每个意义称为“义项”。下面的例子中“低”就有3个义项。

- (1) 飞机飞得很低。
- (2) 这个球队技术水平比较低。
- (3) 弟弟低着头，一句话也没说。
- (4) 文化程度低会影响工作。
- (5) 低低的围墙把院子围了起来。

# 1.3 词汇

## 四、多义词与同音词

同音词是指声母、韵母、声调都相同的词。有些词同音同形，但意义不同。例如：

拼：合在一起；

拼：不顾一切地干；

草：草本植物的总称；

草：草率，粗劣不细致；

大家：著名的专家；

大家：代词，指一定范围内的所有人。

# 1.3 词汇

## 五、同义词、反义词、上下位词、术语、熟语

(1) 同义词和反义词

(2) 上下位词

农民	瓦工
人-----工人-----木工	
商人	花匠

著名的词典：同义词词林、知网 (HowNet)

WordNet、中文概念词典CCD

(3) 术语和熟语



# 1.3 词汇

## 六、词类的划分

现代汉语中，词可以划分成15类：

- (1) 名词：书、水、桌子、椅子
- (2) 动词：看、走、学习、调查
- (3) 形容词：红、好、干净、伟大
- (4) 状态词：通红、雪白、绿油油、黑咕隆咚
- (5) 区别词：男、女、公、母、棉、单
- (6) 数词：一、二、十五、百、千、万
- (7) 量词：个、条、斤、尺、头、口
- (8) 副词：不、很、都、忽然、简直

# 1.3 词汇

## 六、词类的划分

现代汉语中，词可以划分成15类：

(9) 代词：你、我、他、你们、那样

(10) 介词：把、被、对于、关于、从

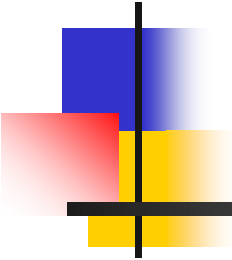
(11) 连词：和、或、并、而且、不但、因为

(12) 助词：了、着、过（这三个为动态助词）；  
的（地）、得、所、似的（这些是结构助词）

(13) 语气词：啊、吗、呢、呗、吧、嘛、了

(14) 感叹词：啊、哎、喂

(15) 拟声词：啪、咚、哗啦、叮铃当啷



# 1.4 语法

# 1.4 语法

## 一、语法的性质及作用

**语法就是语言的组词造句的规则。** 每一句话中的词并不是任意的组合，而是按照一定的规则的组织的。

语法不是语法学家主观确定的，而是从实践中、实际语言的使用中归纳、总结、提炼出来的，又反过来指导和影响语言。

语法有指导语言实践的作用，这无可质疑，但也不可死搬教条。比如有语法规则：

**x 把 y 怎么样了**

# 1.4 语法

## 一、语法的性质及作用

按照这个规则，以下这些语句都是合法的：

- (1) 风把门吹开了
- (2) 风把吹开了门
- (3) 门把风吹开了
- (4) 太阳是黑的

这些从逻辑上显然是不合适的。当然，有些话表面上看似乎违背事理或不合逻辑，但大家都这样说且懂得什么意思，就应该承认是合法的。比如，“吃什么”、“买什么”，“都八点了，你还睡什么！”

# 1.4 语法

## 二、句子

从表达的角度看，句子是一个基本的表达单位。由句子组成段落，由段落组成篇章。

从语法的角度看，句子是最大的语法单位。

从结构上看，许多句子都包含主语和谓语两部分：

母亲和宏儿都睡着了（鲁迅《故乡》）

这种句子称为主谓句。也有非主谓句的情况：

蛇！（单词），集合！（单词），下雨了（述宾结构），禁止吸烟（述宾结构）



# 1.4 语法

## 二、句子

从句子所表达的内容看或说话人要达到的目的看，句子又可分为以下几类：

**陈述句：**今天星期五。

**疑问句：**今天星期几？

**祈使句：**你们要认真听课！

**感叹句：**你们的成绩真好！

**呼应句：**呼唤或应答的句子，例如：

喂，士老，贾大夫要的方子呢？（曹禺《明朗的天》）

赵铁生同志，你愿意听听我们的分析吗？

# 1.4 语法

## 句子的成分结构

**主语-谓语：**主谓结构表示陈述关系，主语是说话人所要陈述的对象，谓语是对于主语的陈述。

**述语-宾语：**述宾结构表示支配关系。宾语是对述语而言的，通常述语由动词充任，宾语是受述语动词支配、制约的对象。

**述语-补语：**补语放在动词或形容词之后做补充说明的成分。例如“洗干净”，“洗”是述语，“干净”是补语。

**定语-状语：**修饰语分为定语和状语两大类。名词前头的修饰语一定是定语，动词和形容词前面的修饰语可以是状语也可以是定语



## 1.4 语法

### 三、语气

一句话的语气主要取决于语调。例如“有人不同意”，用普通的语调就是陈述语气，如果句尾语调上扬就变成疑问语气了。语气有以下四种语气

**陈述语气：**今天星期五。

**疑问语气：**今天星期几？

**祈使语气：**你们要认真听课！

**感叹语气：**你们的成绩真好！

## 1.4 语法

### 四、虚词

虚词在数量上比实词少的多，但其作用是非常重要的。在语言应用中，缺少几个实词可能到不了不能说话的地步，但缺少了“不、了、的、和、呢、把、才、就”几个虚词，可能话说不了了。

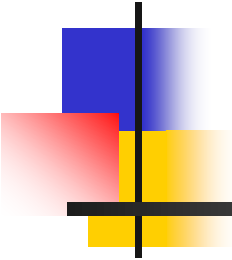
有观点认为现代汉语虚词大约有700多个，而常用的虚词大约是350多个。另一观点指出，高考要求掌握的文言虚词大约有18个左右。虚词包括副词、介词、连词、助词、语气词等，它们在句子中不表示实在的意义，主要作用是组合语言单位，帮助构成句子的结构。

## 1.4 语法

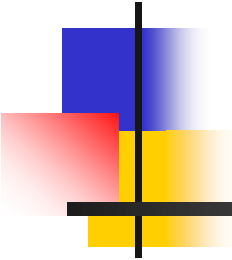
### 四、虚词

虚词的个性很强，同一类里各个虚词在用法上差别可以很大，因此，对虚词一定要一个一个地学，一个一个地掌握。常用且容易用错的介词和连词如下：

把、被、对于/对、在、比、连、和/跟/同/与/及/以及  
或者/还是、与其/宁可（宁肯、宁愿）、何况/况且，  
虽然/尽管/固然，即使/纵使，不管/不论/无论、  
只要/只有、既然/既、 因为/由于、 进而/从而



## 二、数学基础知识



## 2.1 概率论基础

## 2.1 概率论基础

### □ 概率 (probability)

概率是从随机实验中的事件到实数域的函数，用以表示事件发生的可能性。如果用  $P(A)$  作为事件  $A$  的概率， $\Omega$  是实验的样本空间，则概率函数必须满足如下公理：

公理1：  $P(A) \geq 0$

公理2：  $P(\Omega) = 1$

公理3： 如果对任意的  $i$  和  $j$  ( $i \neq j$ )，事件  $A_i$  和  $A_j$  不相交 ( $A_i \cap A_j = \Phi$ )，则有：

$$P\left(\bigcup_{i=0}^{\infty} A_i\right) = \sum_{i=0}^{\infty} P(A_i) \quad (1)$$

## 2.1 概率论基础

### □ 最大似然估计

(Maximization likelihood estimation, MLE)

如果一个实验的样本空间是  $\{s_1, s_2, \dots, s_n\}$ , 在相同情况下重复实验  $N$  次, 观察到样本  $s_k$  ( $1 \leq k \leq n$ ) 的次数为  $n_N(s_k)$ , 则  $s_k$  的相对频率为:

$$q_N(s_k) = \frac{n_N(s_k)}{N} \quad (2)$$

由于  $\sum_{k=1}^n n_N(s_k) = N$  因此,  $\sum_{k=1}^n q_N(s_k) = 1$

## 2.1 概率论基础

当 $N$ 越来越大时，相对频率  $q_N(s_k)$  就越来越接近 $s_k$ 的概率 $P(s_k)$ 。事实上，

$$\lim_{N \rightarrow \infty} q_N(s_k) = P(s_k) \quad (3)$$

因此，相对频率常被用作概率的估计值。  
这种概率值的估计方法称为最大似然估计。



## 2.1 概率论基础

### □ 条件概率 (conditional probability)

如果  $A$  和  $B$  是样本空间  $\Omega$  上的两个事件， $P(B) > 0$ ，那么在给定  $B$  时  $A$  的条件概率  $P(A|B)$  为：

$$P(A | B) = \frac{P(A \cap B)}{P(B)} \quad (4)$$

条件概率  $P(A|B)$  给出了在已知事件  $B$  发生的情况下，事件  $A$  发生的概率。

一般地， $P(A|B) \neq P(A)$ 。



## 2.1 概率论基础

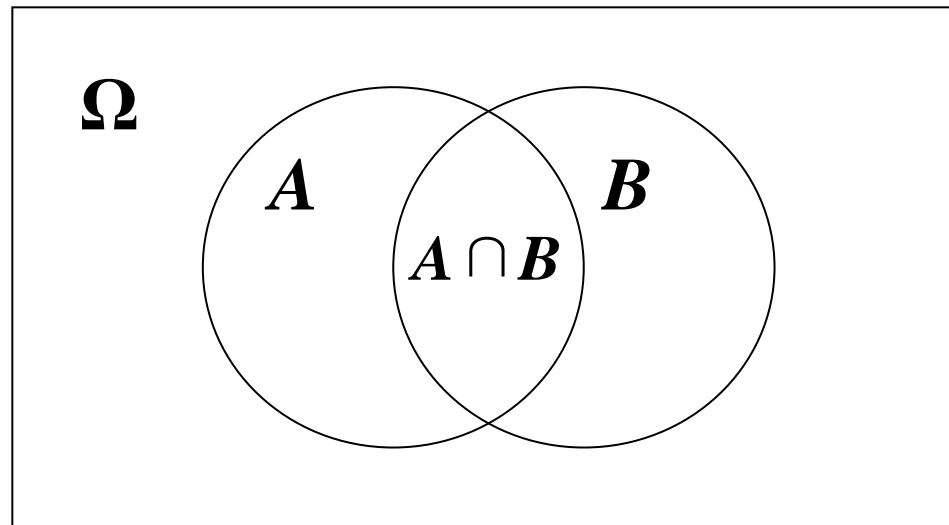


图2-1. 条件概率示意图

## 2.1 概率论基础

### □ 全概率公式

设  $\Omega$  为实验  $E$  的样本空间,  $B_1, B_2, \dots, B_n$  为  $\Omega$  的一组事件, 且他们两两互斥, 且每次实验中至少发生一个。即:

$$(1) B_i \cap B_j = \Phi \quad (i \neq j; i, j = 1, 2, \dots, n) \quad (5)$$

$$(2) \bigcup_{i=1}^n B_i = \Omega \quad (6)$$

则称  $B_1, B_2, \dots, B_n$  为样本空间  $\Omega$  的一个划分。

## 2.1 概率论基础

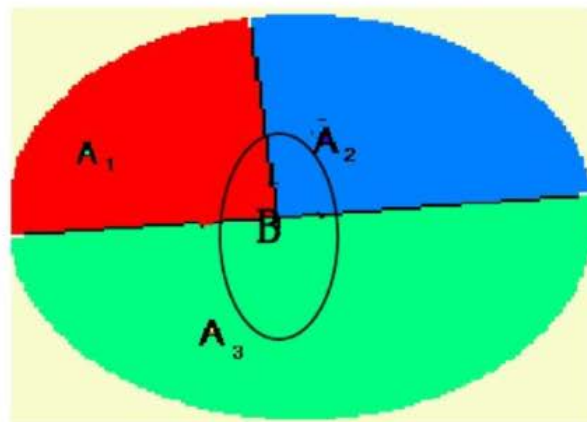
设 $A$ 为 $\Omega$  的事件,  $B_1, B_2, \dots, B_n$  为 $\Omega$ 的一个划分, 且  $P(B_i) > 0$  ( $i=1, 2, \dots, n$ ), 则 全概率公式为:

$$P(A) = P\left(\bigcup_{i=1}^n AB_i\right) = \sum_{i=1}^n P(AB_i) = \sum_{i=1}^n P(B_i)P(A|B_i) \quad (7)$$

## 2.1 概率论基础

市场上有甲、乙、丙三家工厂生产的同一品牌产品，已知三家工厂的市场占有概率分别为 **1/4**、**1/4**、**1/2**，且三家工厂的次品概率分别为 **2%**、**1%**、**3%**，试求市场上该品牌产品的次品概率。

设： $B$ ：买到一件次品  
 $A_1$ ：买到一件甲厂的产品  
 $A_2$ ：买到一件乙厂的产品  
 $A_3$ ：买到一件丙厂的产品



$$\begin{aligned} P(B) &= P(BA_1) + P(BA_2) + P(BA_3) \\ &= P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + P(B|A_3)P(A_3) \\ &= 0.02 \times \frac{1}{4} + 0.01 \times \frac{1}{4} + 0.03 \times \frac{1}{2} \approx 0.0225 \end{aligned}$$

@蒙蒙么圈圈

Baidu 文图

## 2.1 概率论基础

■ 高射炮向敌机发射三发炮弹，每弹击中与否相互独立且每发炮弹击中的概率均为0.3，又知敌机若中一弹，坠毁的概率为0.2，若中两弹，坠毁的概率为0.6，若中三弹，敌机必坠毁。求敌机坠毁的概率。

解：设事件 $A_k$ 表示“敌机中 $k$ 弹”，其中 $k=0,1,2,3$ 。  
事件 $B$ 表示“敌机坠毁”。

首先，计算每个 $A_k$ 的概率：

- $P(A_0) = (1-0.3)^3 = 0.7^3 = 0.343$
- $P(A_1) = C_3^1 \times 0.3 \times 0.7^2 = 3 \times 0.3 \times 0.49 = 0.441$
- $P(A_2) = C_3^2 \times 0.3^2 \times 0.7 = 3 \times 0.09 \times 0.7 = 0.189$
- $P(A_3) = 0.3^3 = 0.027$

## 2.1 概率论基础

- 接下来，根据全概率公式，计算敌机坠毁的概率  $P(B)$ ：
- $P(B) = P(A_1)P(B | A_1) + P(A_2)P(B | A_2) + P(A_3)P(B | A_3)$
- 其中，
- $P(B | A_1)$  是敌机中一弹时坠毁的概率，即 0.2
- $P(B | A_2)$  是敌机中两弹时坠毁的概率，即 0.6
- $P(B | A_3)$  是敌机中三弹时坠毁的概率，由于题目已给出中三弹必坠毁，所以  $P(B | A_3) = 1$
- 代入上述值，得：
- $P(B) = 0.441 \times 0.2 + 0.189 \times 0.6 + 0.027 \times 1 = 0.0882 + 0.1134 + 0.027 = 0.2286$
- 故敌机坠毁的概率为 0.2286。

## 2.1 概率论基础

### □ 贝叶斯法则 (Bayes' theorem)

如果  $A$  为样本空间  $\Omega$  的事件,  $B_1, B_2, \dots, B_n$  为  $\Omega$  的一个划分, 且  $P(A) > 0$ ,  $P(B_i) > 0$  ( $i = 1, 2, \dots, n$ ), 那么

$$P(B_i | A) = \frac{P(B_i)P(A | B_i)}{\sum_{j=1}^n P(B_j)P(A | B_j)} \quad (8)$$

当  $n=1$  时,

$$P(B | A) = \frac{P(A | B)P(B)}{P(A)} \quad (9)$$



## 2.1 概率论基础

例2-1: 给定语音信号 $A$ , 找出对应的语句 $S$ , 使得 $P(S|A)$ 最大, 那么,

$$\hat{s} = \arg \max_S P(S | A)$$

根据贝叶斯公式,

$$\hat{s} = \arg \max_S \frac{P(S)P(A | S)}{P(A)}$$

由于  $P(A)$  在  $A$  给定时是归一化常数, 因而,

声学模型

$$\hat{s} = \arg \max_S \underbrace{P(A | S)}_{\text{声学模型}} \underbrace{P(S)}_{\text{语言模型}}$$

语言模型

## 2.1 概率论基础

例2-2：假设某一种特殊的句法结构很少出现，平均大约每100,000个句子中才可能出现一次。我们开发了一个程序来判断某个句子中是否存在这种特殊的句法结构。如果句子中确实含有该特殊句法结构时，程序判断结果为“存在”的概率为0.95。如果句子中实际上不存在该句法结构时，程序错误地判断为“存在”的概率为0.005。那么，这个程序测得句子含有该特殊句法结构的结论是正确的概率有多大？

## 2.1 概率论基础

解：假设 $G$ 表示事件“句子确实存在该特殊句法结构”， $T$ 表示事件“程序判断的结论是存在该特殊句法结构”。那么，我们有：

$$P(G) = \frac{1}{100000} = 0.00001 \quad P(\bar{G}) = \frac{100000 - 1}{100000} = 0.99999$$

$$P(T | G) = 0.95$$

$$P(T | \bar{G}) = 0.005$$

求：  $P(G|T)=$

$$\begin{aligned} P(G|T) &= \frac{P(T|G)P(G)}{P(T|G)P(G) + P(T|\bar{G})P(\bar{G})} \\ &= \frac{0.95 \times 0.00001}{0.95 \times 0.00001 + 0.005 \times 0.99999} \approx 0.002 \end{aligned}$$

## 2.1 概率论基础

### □ 期望 (expectation)

期望值是一个随机变量所取值的概率平均。

设  $X$  为一随机变量，其分布为  $P(X = x_k) = p_k$ ， $k = 1, 2, \dots$  若级数  $\sum_{k=1}^{\infty} x_k p_k$  绝对收敛，那么，

随机变量  $X$  的数学期望或概率平均值为：

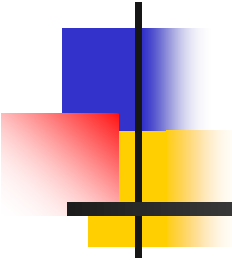
$$E(X) = \sum_{k=1}^{\infty} x_k p_k \quad (12)$$

## 2.1 概率论基础

### □ 方差 (variance)

一个随机变量的方差描述的是该随机变量的值偏离其期望值的程度。设  $X$  为一随机变量，其方差为：

$$\begin{aligned} \text{Var}(X) &= E((X - E(X))^2) \\ &= E(X^2) - E^2(X) \end{aligned} \tag{13}$$



## 2.2 信息论基础

## 2.2 信息论基础

### □ 熵 (entropy)

香农 (Claude Elwood Shannon) 于1940年获得 MIT 数学博士学位和电子工程硕士学位后，于1941年加入了贝尔实验室数学部，并在那里工作了15年。1948年6月和10月，由贝尔实验室出版的《贝尔系统技术》杂志连载了香农博士的文章《通讯的数学原理》，该文奠定了香农信息论的基础。

熵是信息论中重要的基本概念。

## 2.2 信息论基础

如果  $X$  是一个离散型随机变量，其概率分布为：  $p(x) = P(X = x)$ ,  $x \in X$ 。  $X$  的熵  $H(X)$  为：

$$H(X) = - \sum_{x \in X} p(x) \log_2 p(x) \quad (14)$$

其中，约定  $0 \log 0 = 0$ 。

$H(X)$  也可以写为  $H(p)$ 。通常熵的单位为二进制位比特 (bit, binary unit的缩写)。以e为底单位为奈特 ( nat, natrue unit的缩写),以10为底,单位为哈特 ( hart, hartley的缩写)



## 2.2 信息论基础

根据对数换底关系可算出:1奈特=1.44比特,1哈特=3.32比特。

信息量单位中的比特,与计算机术语中的“比特”的含义有所不同,计算机中的比特代表二元数字(binary digits)。二者的关系是每个二元数字所能提供的最大平均信息量为1比特。



## 2.2 信息论基础

熵又称为自信息 (self-information) , 表示信源  $X$  每发一个符号 (不论发什么符号) 所提供的平均信息量。熵也可以被视为描述一个随机变量的不确定性的数量。一个随机变量的熵越大, 它的不确定性越大。那么, 正确估计其值的可能性就越小。越不确定的随机变量越需要大的信息量用以确定其值。

## 2.2 信息论基础

**例 2-3** 计算下列两种情况下英文（26个字母和空格，共27个字符）信息源的熵：(1)假设27个字符等概率出现；(2)假设英文字母的概率分布如下：

字母	空格	E	T	O	A	N	I	R	S
概率	0.1956	0.105	0.072	0.0654	0.063	0.059	0.055	0.054	0.052

字母	H	D	L	C	F	U	M	P	Y
概率	0.047	0.035	0.029	0.023	0.0225	0.0225	0.021	0.0175	0.012

字母	W	G	B	V	K	X	J	Q	Z
概率	0.012	0.011	0.0105	0.008	0.003	0.002	0.001	0.001	0.001

## 2.2 信息论基础

解：（1）等概率出现情况：

$$\begin{aligned} H(X) &= - \sum_{x \in X} p(x) \log_2 p(x) \\ &= 27 \times \left\{ -\frac{1}{27} \log \frac{1}{27} \right\} = \log 27 = 4.75 \text{ (bits/letter)} \end{aligned}$$

（2）实际情况：

$$H(X) = - \sum_{i=1}^{27} p(x_i) \log p(x_i) = 4.02 \quad (\text{bits/letter})$$

说明：考虑了英文字母和空格实际出现的概率后，英文信源的平均不确定性，比把字母和空格看作等概率出现时英文信源的平均不确定性要小。

## 2.2 信息论基础

法语、意大利语、西班牙语、英语、俄语字母的熵[冯志伟, 1989]:

语言	熵 (bits)
法语	3.98
意大利语	4.00
西班牙语	4.01
英语	4.03
俄语	4.35

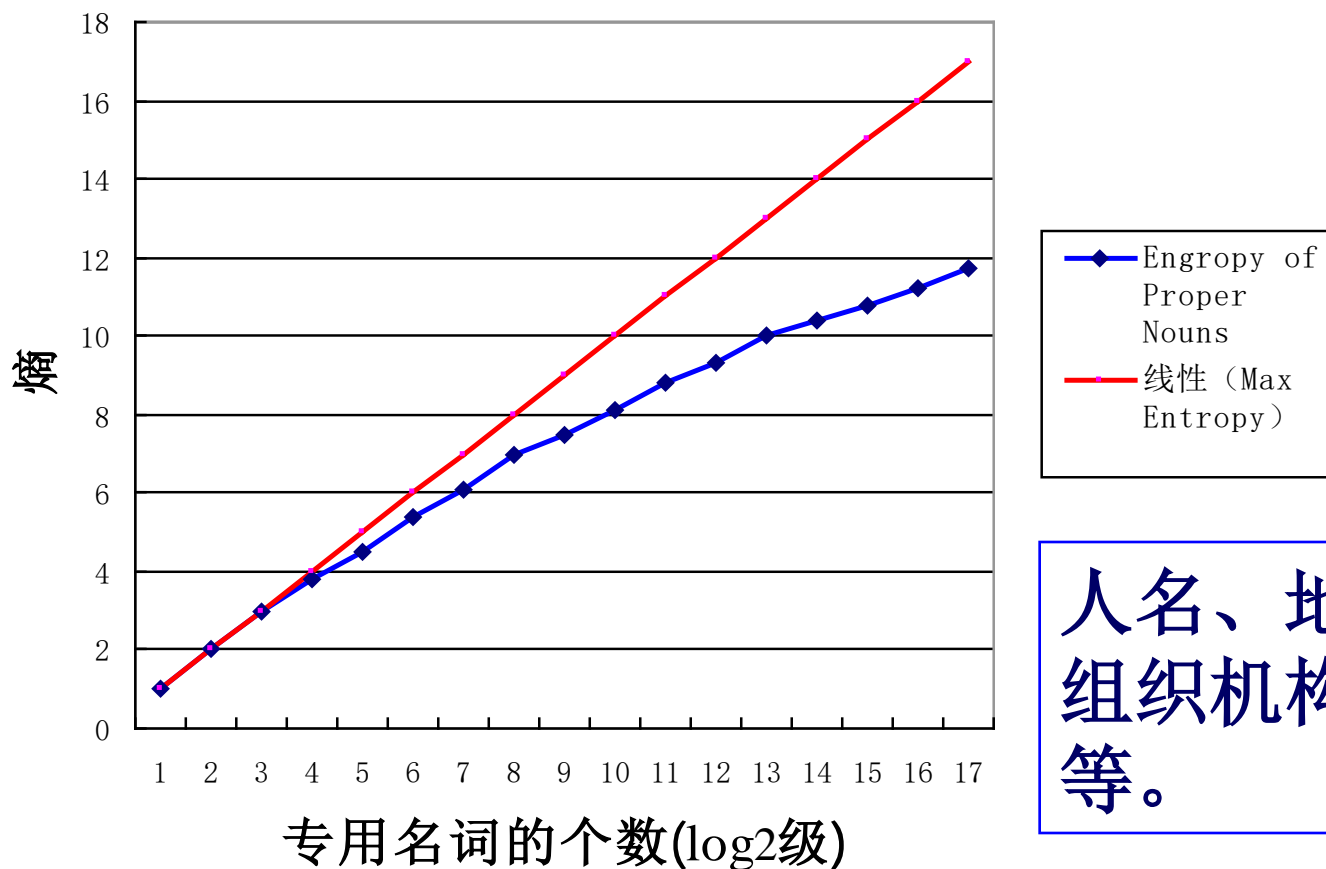
英语词的熵约为10 bits。

## 2.2 信息论基础

1970年代末期冯志伟首先开展了对汉字信息熵的研究，经过几年的语料收集和手工统计，在当时艰苦的条件下测定了汉字的信息熵为9.65比特(bit)。1980年代末期，刘源等测定了汉字的信息熵为9.71 比特，而汉语的词熵为11.46比特。

## 2.2 信息论基础

### 专用名词的熵[Tsou, 2001]



人名、地名、  
组织机构名  
等。

## 2.2 信息论基础

### 北京、香港、台北三地汉语词的熵[Tsou,2003]

北京5年		台北5年		香港5年		京、港、台5年	
A1	A2	B1	B2	C1	C2	D1	D2
11.45	11.11	11.69	11.36	11.96	11.64	11.96	11.60

其中，A1, B1, C1 分别是从小LIVAC 语料库中北京、台北、香港三地5年各约1000万字语料中所提取的数据；A2, B2, C2 为三地语料剔除专用名词之后的数据。D1, D2分别为三地语料合并之后剔除专用名词前后的数据。



## 2.2 信息论基础

### □ 联合熵 (joint entropy)

如果  $X, Y$  是一对离散型随机变量

$X, Y \sim p(x, y)$ ,  $X, Y$  的联合熵  $H(X, Y)$  为:

$$H(X, Y) = - \sum_{x \in X} \sum_{y \in Y} p(x, y) \log_2 p(x, y) \quad (15)$$

联合熵实际上就是描述一对随机变量平均所需要的信息量。

## 2.2 信息论基础

### □ 条件熵 (conditional entropy)

给定随机变量  $X$  的情况下，随机变量  $Y$  的条件熵定义为：

$$\begin{aligned} H(Y | X) &= \sum_{x \in X} p(x) H(Y | X = x) \\ &= \sum_{x \in X} p(x) \left[ - \sum_{y \in Y} p(y | x) \log p(y | x) \right] \\ &= - \sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(y | x) \end{aligned} \quad (16)$$

## 2.2 信息论基础

将 (15) 式 中的  $\log_2 p(x, y)$  根据概率公式展开:

$$\begin{aligned} H(X, Y) &= - \sum_{x \in X} \sum_{y \in Y} p(x, y) \log[p(x)p(y | x)] \\ &= - \sum_{x \in X} \sum_{y \in Y} p(x, y) [\log p(x) + \log p(y | x)] \\ &= - \sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(x) - \sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(y | x) \\ &= - \sum_{x \in X} p(x) \log p(x) - \sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(y | x) \\ &= H(X) + H(Y | X) \end{aligned} \quad (17) \quad \text{(\underline{连锁规则})}$$

## 2.2 信息论基础

例2-4. 简单的玻利尼西亚语(Polynesian) 是一些随机的字符序列，其中部分字符出现的概率为：

**p: 1/8, t: 1/4, k: 1/8, a: 1/4, i: 1/8, u: 1/8**

那么，各字符的平均信息量即熵为：

$$\begin{aligned} H(P) &= - \sum_{i \in \{p, t, k, a, i, u\}} P(i) \log P(i) \\ &= -[4 \times \frac{1}{8} \log \frac{1}{8} + 2 \times \frac{1}{4} \log \frac{1}{4}] = 2\frac{1}{2} \quad \text{(bits)} \end{aligned}$$

## 2.2 信息论基础

这个结果表明，我们可以设计一种编码，传输一个字符平均只需要2.5个比特：

<b>p</b>	<b>t</b>	<b>k</b>	<b>a</b>	<b>i</b>	<b>u</b>
<b>100</b>	<b>00</b>	<b>101</b>	<b>01</b>	<b>110</b>	<b>111</b>

这种语言的字符分布并不是随机变量，但是，我们可以近似地将其看作随机变量。如果将字符按元音和辅音分成两类，元音随机变量  $V=\{a, i, u\}$ ，辅音随机变量  $C=\{p, t, k\}$ 。

## 2.2 信息论基础

假定所有的单词都由  $CV$  (consonant-vowel) 音节序列组成, 其联合概率分布  $P(C, V)$ 、边缘分布  $P(C, \cdot)$  和  $P(\cdot, V)$  如下表所示:

$V \backslash C$	p	t	k	
a	1/16	3/8	1/16	1/2
i	1/16	3/16	0	1/4
u	0	3/16	1/16	1/4
	1/8	3/4	1/8	

边缘分布

## 2.2 信息论基础

注意，这里的边缘概率是基于每个音节的，其值是基于每个字符的概率的两倍，因此，每个字符的概率值应该为相应边缘概率的1/2，即：

**p: 1/16   t: 3/8   k: 1/16   a: 1/4   i: 1/8   u: 1/8**

现在我们来求联合熵为多少？

## 2.2 信息论基础

求联合熵可以有几种方法，以下我们采用连锁规则方法可以得到：

$$\begin{aligned} H(C) &= - \sum_{c=p,t,k} p(c) \log p(c) = -2 \times \frac{1}{8} \times \log \frac{1}{8} - \frac{3}{4} \times \log \frac{3}{4} \\ &= \frac{9}{4} - \frac{3}{4} \log 3 \approx 1.061 \text{ (bits)} \end{aligned}$$

同理，可以计算得到

$$H(V) = 1.5 \text{ (bits)}$$



## 2.2 信息论基础

### 困惑度 (perplexity)

用熵评价语言模型，并不是模型本身具有熵值，而是用它去近似地计算语言的熵，得到的语言熵越小，说明模型表述语言的性能越好。**困惑度**则是对应用语言模型预测语言成分出现时的难度进行度量，也称为复杂度。复杂度越大，表明可选择的范围越大，选择的难度也越大。

如果语言L是平稳的、各态遍历的随机过程，语言模型PM是一个n元模型， $PM(w_{i-n+1}^i)$ 表示语言模型PM对L中的词串 $w_{i-n+1}^i$ 的概率估计，假设LN为语言模型训练语料的容量，

## 2.2 信息论基础

$$H(P_M) \approx -\frac{1}{LN} \left( \sum_{i=1}^{n-1} \log_2 P_M(W_i | W_1^{i-1}) + \sum_{i=n}^{LN} \log_2 P_M(W_i | W_{i-n+1}^{i-1}) \right)$$

$H(P_M)$ 的物理意义是，当给定一段历史信息  $W_{i-n+1}^{i-1}$  后，利用所建立的语言模型  $P_M$  预测当前语言成分  $W_i$  出现的可能性只有  $2^{H(P_M)}$  种，如何利用语言模型  $P_M$  从这  $2^{H(P_M)}$  种选出  $W_i$ ，确实是很困难的，也是非常让人感到困惑的，这也可能是将  $2^{H(P_M)}$  称为语言模型  $P_M$  的困惑度（Perplexity）的原因吧。困惑度记为 PP：

$$PP = 2^{H(P_M)} \quad (4.35)$$

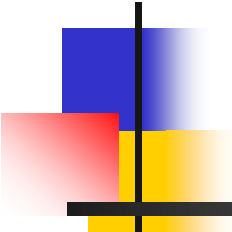
PP 越小，说明利用模型  $P_M$  预测出现  $W_i$  的选择范围越小，即不确定性越小，进而说明语言模型表述语言的能力越强。

因为

$$H_\infty \leq H(P_M) \leq H_0 = \log_2 |V| \quad (4.36)$$

因此，

$$2^{H_\infty} \leq PP \leq |V| \quad (4.37)$$



困惑度也可以根据模型被解释为语言的几何平均分枝因子，表明如果应用分支图描述各语言成分间的关系，可能会出现的分枝数量。

值得指出的是，困惑度既是文本的函数，又是模型的函数。在文本和词典相同的情况下，比较两个语言模型才有意义。较小规模的词典由于通常不包括低频词，所得的 PP 值就小，即使词典相同，不同的文本也会导致比较上的差错。有人想设计出比困惑度更好的语言模型评价标准，但都没有取得太好的结果，目前情况下，困惑度仍是对语言模型构造的一个较好的度量。

**语言模型设计的任务就是寻找困惑度最小的模型，使其最接近真实的语言。**

## 2.2 信息论基础

### 互信息 (mutual information)

根据熵的连锁规则, 有

$H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$ , 因此有

$$H(X) - H(X|Y) = H(Y) - H(Y|X)$$

这个差称为X和Y的互信息, 记作 $I(X; Y)$ , 或者定义为

如果  $(X, Y) \sim p(x, y)$ ,  $X, Y$  之间的互信息  $I(X; Y)$  为:

$$I(X; Y) = H(X) - H(X|Y)$$

## 2.2 信息论基础

根据定义，展开  $H(X)$  和  $H(X|Y)$  容易得到：

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log_2 \frac{p(x, y)}{p(x)p(y)} \quad (20)$$

互信息  $I(X; Y)$  是在知道了  $Y$  的值以后  $X$  的不确定性的减少量。即， $Y$  的值透露了多少关于  $X$  的信息量。

## 2.2 信息论基础

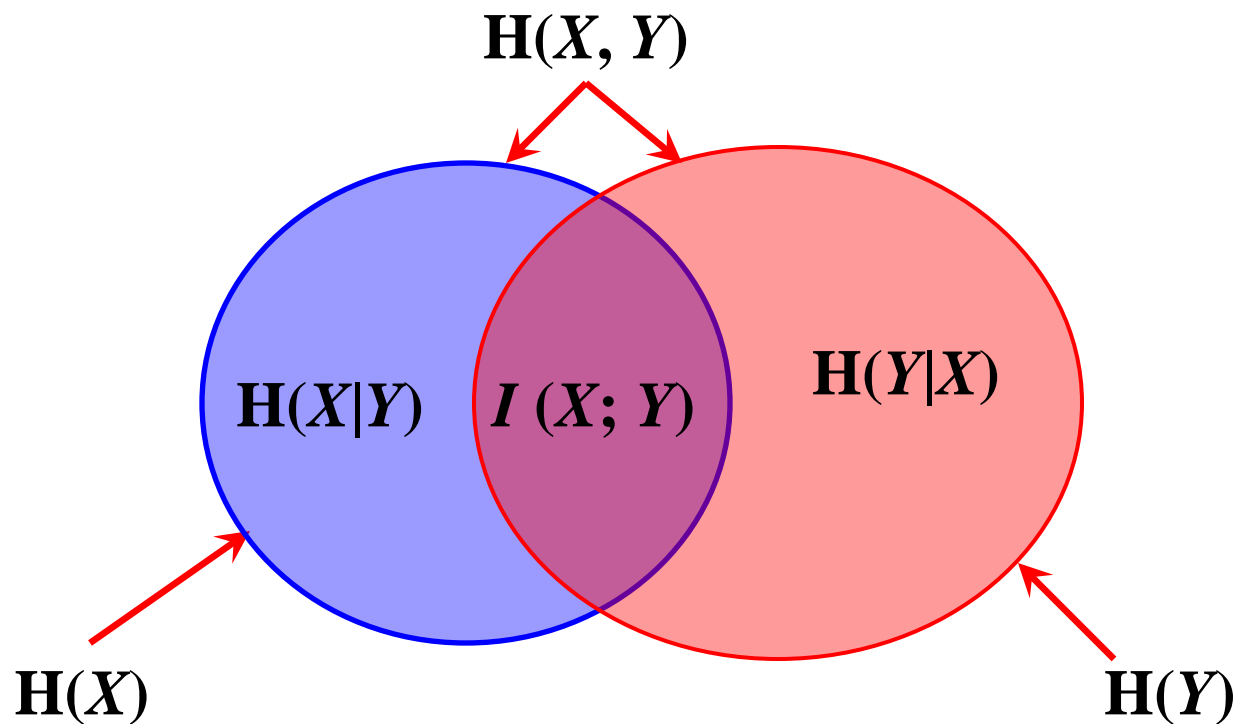


图 2-2 互信息、条件熵与联合熵

## 2.2 信息论基础

由于  $H(X|X) = 0$ , 所以,

$$H(X) = H(X) - H(X|X) = I(X; X) \quad (21)$$

这一方面说明了为什么熵又称自信息, 另一方面说明了两个完全相互依赖的变量之间的互信息并不是一个常量, 而是取决于它们的熵。

## 2.2 信息论基础

例如：汉语断词问题

句子：为人民服务。  
?

利用互信息值估计两个汉字结合的程度：

$$I(x; y) = \log_2 \frac{p(x, y)}{p(x)p(y)} = \log_2 \frac{p(y|x)}{p(y)}$$

互信息值越大，表示两个汉字之间的结合越紧密，越可能成词。反之，断开的可能性越大。



## 2.2 信息论基础

当两个汉字  $x$  和  $y$  关联度较强时，其互信息值  $I(x, y) > 0$ ； $x$  与  $y$  关系弱时， $I(x, y) \approx 0$ ；而当  $I(x, y) < 0$  时， $x$  与  $y$  称为“互补分布”。

在汉语分词研究中，有学者用双字耦合度的概念代替互信息：

设  $c_i$ ， $c_{i+1}$  是两个连续出现的汉字，统计样本中  $c_i$ ， $c_{i+1}$  连续出现在一个词中的次数和连续出现的总次数，二者之比就是  $c_i$ ， $c_{i+1}$  的双字耦合度：

## 2.2 信息论基础

$$Couple(c_i, c_{i+1}) = \frac{N(c_i c_{i+1})}{N(c_i c_{i+1}) + N(\dots c_i | c_{i+1} \dots)}$$

$c_i, c_{i+1}$  是一个有序字对，表示两个连续汉字，且  $c_i c_{i+1}$  不等于  $c_{i+1} c_i$ 。 $N(c_i c_{i+1})$  表示字符串  $c_i c_{i+1}$  构成的词出现的频率， $N(\dots c_i | c_{i+1} \dots)$  表示  $c_i$  作为上一个词的词尾且  $c_{i+1}$  作为相邻下一个词的词头出现的频率。例如：“为人”出现5次，“为人民”出现20次，那么， $Couple(\text{为}, \text{人}) = 0.2$ 。

## 2.2 信息论基础

之所以在这里选择双字耦合度，而不是常用的互信息，是因为一个汉字对(A, B)的双字耦合度比其互信息更适合用来在交叉歧义中判断连续出现的AB属于同一个词的概率大小。互信息是计算两个汉字连续出现在一个词中的概率，而两个汉字在实际应用中出现的概率情况共有三种：(1) 两个汉字连续出现，并且在一个词中；(2) 两个汉字连续出现，但分属于两个不同的词；(3) 非连续出现。

## 2.2 信息论基础

互信息计算的是两个汉字连续出现的可能性大小，实际上，有些汉字在实际应用中出现虽然比较频繁，但是连续在一起出现的情况比较少，一旦连在一起出现，就很可能是一个词。这种情况下计算出来的互信息会比较小，而实际上两者的结合度应该还是比较高的。而双字耦合度恰恰计算的是两个连续汉字出现在一个词中的概率，并不考虑两个汉字非连续出现的情况。

## 2.2 信息论基础

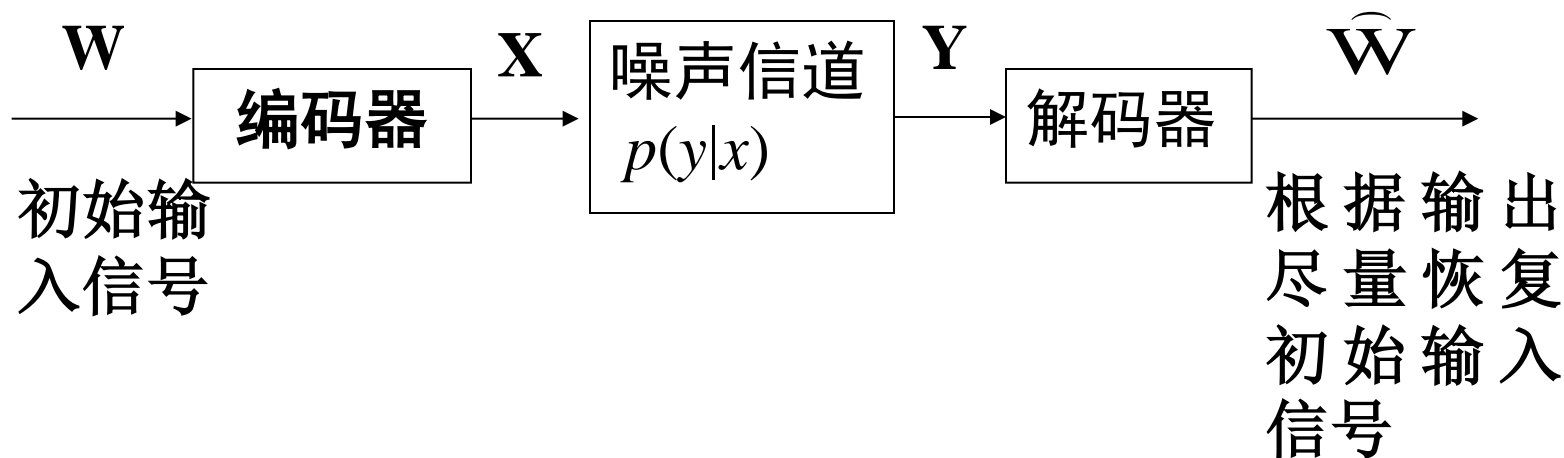
### □ 噪声信道模型 (noisy channel model)

在信号传输的过程中都要进行双重性处理：一方面要通过压缩消除所有的冗余，另一方面又要通过增加一定的可控冗余以保障输入信号经过噪声信道后可以很好的恢复原状。信息编码时要尽量占用少量的空间，但又必须保持足够的冗余以便能够检测和校验错误。接收到的信号需要被解码使其尽量恢复到原始的输入信号。

噪声信道模型的目标就是优化噪声信道中信号传输的吞吐量和准确率，其基本假设是一个信道的输出以一定的概率依赖于输入。

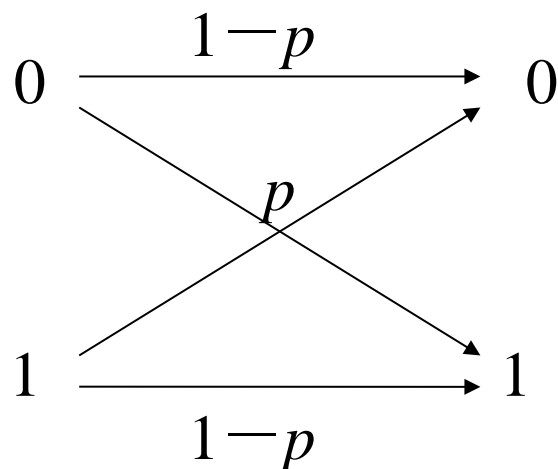
## 2.2 信息论基础

过程示意图：



## 2.2 信息论基础

一个二进制的对称信道(binary symmetric channel, BSC)的输入符号集 $X:\{0, 1\}$ , 输出符号集 $Y:\{0, 1\}$ 。在传输过程中如果输入符号被误传的概率为 $p$ , 那么, 被正确传输的概率就是 $1-p$ 。这个过程我们可以用一个对称的图型表示如下:



## 2.2 信息论基础

信息论中很重要的一个概念就是信道容量 (capacity)，其基本思想是用降低传输速率来换取高保真通讯的可能性。其定义可以根据互信息给出：

$$C = \max_{p(X)} I(X; Y) \quad (27)$$

根据这个定义，如果我们能够设计一个输入编码  $X$ ，其概率分布为  $p(X)$ ，使其输入与输出之间的互信息达到最大值，那么，我们的设计就达到了信道的最大传输容量。



## 2.2 信息论基础

在自然语言处理中，我们不需要进行编码，只需要进行解码，使系统的输出更接近于输入。

例如，法语翻译成英语：



根据贝叶斯公式：

$$P(e | f) = \frac{P(e)P(f | e)}{P(f)}$$

## 2.2 信息论基础

求该式的最大值相当于寻找一个使得右边分子的两项乘积  $P(e) \times P(f|e)$  最大，即：

$$\hat{e} = \arg \max_e P(e)P(f|e) \quad (28)$$

语言模型

翻译模型(translation model)

## 2.2 信息论基础

统计翻译系统框架：



法语句子  $f$   $\Rightarrow$  英语句子  $\hat{e}$

## 2.2 信息论基础

也就是说，如果我们要建立一个源语言 $e$ 到目标语言 $f$ 的统计翻译系统，我们必须解决三个关键的问题：

- (1) 估计语言模型概率 $P(e)$ ;
- (2) 估计翻译概率 $P(f|e)$ ;
- (3) 设计有效快速的搜索算法求解  $\hat{e}$  使得  $P(e) \times P(f|e)$  最大。

# 语言学基础小结

## □语音

- 语音的划分
- 语音的符号
- 语调和语气
- 声母与韵母

## □文字

- 文字的特点
- 文字的结构
- 字量、字音、字序

## □词汇

- 什么是词汇
- 构词法
- 词语与语素义
- 多义词及同音词
- 术语、术语
- 词类的划分

## □语法

- 句法的性质及作用
- 句子
- 语气
- 虚词



# 数学基础小结

## □ 概率论基础

- 概率
- 条件概率
- 期望
- 最大似然估计
- 贝叶斯公式
- 方差

## □ 信息论基本概念

- 熵
- 互信息
- 交叉熵
- 噪声信道模型
- 联合熵
- 相对熵
- 困惑度

# 习题

- 2-1. 任意摘录一段文字，统计这段文字中所有字符的相对频率。假设这些相对频率就是这些字符的概率，请计算其分布的熵。
- 2-2. 任意取另外一段文字（与上题中文字的用字一样），按上述同样的方法计算字符分布的概率，然后计算两段文字中字符分布的 **KL** 距离。
- 2-3. 举例说明（任意找两个分布  $p$  和  $q$ ），**KL** 距离是不对称的，即  $D(p \parallel q) \neq D(q \parallel p)$ 。

# 习题

2-4. 设  $X \sim p(x)$ ,  $q(x)$  为用于近似  $p(x)$  的一个概率分布, 则  $p(x)$  与  $q(x)$  的交叉熵定义为  $H(p, q) = H(p) + D(p \parallel q)$ 。请证明:

$$H(p, q) = -\sum_x p(x) \log q(x)$$





---

# *Thanks*

谢谢!