蒙古文信息熵

淑琴[1]，那日松[2]

（1、内蒙古大学 蒙古学学院蒙语所，内蒙古 呼和浩特 010021；2、杭州师范大学，中国 杭州 310012）

摘要：熵的测定是数理语言学的一项基础研究。语言符号的熵是语言符号不肯定性程度的度量，它表示语言符号中所包含信息量的多少。蒙古文字母的熵是蒙古文字母所含信息量大小的数学度量。本文讨论传统蒙古文字母所包含的信息量，即传统蒙古文字母中的熵，并给出了最大熵、平均熵及多余度。

关键词：蒙古文；字母；信息熵

中图分类号码：H212　　**文献标识码：**A

一、定义

熵（Entropy）的概念最先是由德国物理学家克劳伊士（Rudolf Clausius）于1865年所提出，并应用在热力学中。热力学中熵表示的是“系统混乱状态”；信息论中信息熵表示的是信息量（Information content）；生态学中熵表示的是生物多样性。

1984年，信息论的创始人美国数学家申农（C.E.Shannon）引入了一个重要概念：不确定程度。申农把不确定程度 H 称为信息熵，就这样，“信息”这个词进入了科学的领地，它在定量化的进程中又与物理学中的“熵”概念联系起来，信息熵也称为申农熵。人们就从消除了多少不确定程度的角度来定义一个消息中含有的信息量。1951年，他首次运用信息论方法测出了英语中包含在一个字母中的熵。

我们可以这样理解：不确定程度 = 熵（信息熵）= 信息量

熵在信息论中的定义如下：

如果有一个系统S内存在多个事件 $S = \{E_1,...,E_n\}$，每个事件的概率分布 $P = \{p_1, ..., p_n\}$，则每个事件本身的信息为：

$I_e = -\log_2 p_i$

如英语有26个字母，假如每个字母在文章中出现次数平均的话，每个字母的信息量为：

$$I_e = -\log_2 \frac{1}{26} = 4.7$$ ；

而汉字常用的有2500个，假如每个汉字在文章中出现次数平均的话，每个汉字的信息量为：

$$I_e = -\log_2 \frac{1}{2500} = 11.3$$ ；

整个系统的平均信息量为：

$$H_s = \sum_{i=1}^{n} p_i I_e = -\sum_{i=1}^{n} p_i \log_2 p_i$$ 。

这个平均信息量就是信息熵。因为和热力学中描述热力学熵的玻耳兹曼公式形式一样，所以也称为“熵”。平均每个字符或者词汇的熵称为熵率（entropy rate），可以用熵率来定义该语言的熵（entropy of language）。

实际上，每个字母和每个汉字在文章中出现的次数并不平均，因此实际数值并不如同上

述，但上述计算是一个总体概念。使用书写单元越多的文字，每个单元所包含的信息量越大。

信息熵的概念建立，为测试信息的多少找到了一个统一的科学的定量计量方法，奠定了信息论的基础。这里引入的信息熵的概念，既不与热力学过程相联系，也与分子运动无关，但信息熵与热力学熵之间有着密切的关系。

可以证明，信息熵与热力学熵二者之间成正比关系。从某种意义上讲，我们完全可以这样看，熵概念在热力学中即为热力学熵，应用到信息论中则是信息熵。

二、单位

信息论中采用比特(bit)作为信息量的单位，如果某一消息由两个出现概率相等的符号构成，那么，包含在该消息符号中的信息量，叫做 1 比特。由于信息量等于被消除的熵，因而我们也可采用比特作熵的单位。而在计算机述评中常用字节(byte)作为信息量的单位，1 个字节是 8 个比特(1byte=8bit)，它容得下一个 8 位二进制数，或说它可记住 256 个(2^8)可能状态中究竟是哪一个。平常我们说微机的内存为 64k(k 为千——kilo)，是说它供用户任意存放数据的空间 ram 是 64×10^3byte(字节)。

在一种非扩展的无记忆信息源中，字符编码的长度不能小于信息源的熵。这个定理适合所有的语言文字，是计算机和网络通讯的科学技术基础和工程设计的基本依据。汉语文，一开始是用了双字节的（即 16 比特），满存储是 6 万多，现在汉语文用了约 1/3。

国际标准编码中，蒙古文基本字符集（包括蒙古文标点符号、数字、控制符、传统蒙古文、托忒文、锡伯文、满文和阿礼嘎礼字母）占有 176 个位数，其中还没用到的位数有 21 个。

三、语言符号的熵及其计算公式

语言是人类最重要的交际工具，也是社会上传递信息的工具，是信息的载体。因此，语言处理技术中的许多应用，都大量地运用或借鉴了信息论的基本理论与量度。信息论是语言处理的又一数学基础之一。

自然语言中所包含的信息量的大小，也就是自然语言的熵。

语言符号的熵是语言符号不肯定性程度的度量，它表示语言符号中所包含信息量的多少。其计算公式如下：

一般公式：

$H= -\ EPi\log_2(Pi)$

说明：

⑴ H：信息熵；E：取和 [Σ（i=1，n）]；Pi：为每个字母在信息中出现的概率；$\log_2$：以 2 为底的对数。

⑵ ① 因为 $0\leq Pi\leq1$，$\log_2(Pi)\leq0$，所以 $H\geq0$

② 负号是为了保证熵不可能为负值。

⑶ 公式中的 Pi 应满足条件 E Pi=1 [Σ（i=1，n）]，即某一特定语言各符号的概率之和为 1。

⑷ Pi=出现次数/文句总字次，即 fr/N。

特殊形式：

$H_0 = \log_2 n$

这是一般公式的一种特殊情况，在一般公式中若 $P_1=P_2=\cdots=P_n$

则有 $H= - n Pi\log_2(Pi)$,

由于 $E Pi=1 [\Sigma (i=1, n)]$，即 $P_1+P_2+\cdots+P_n=1$,

所以，$P_1=P_2=\cdots=P_n = 1/n$,

这样，$H = - n Pi\log_2(Pi)$

$= - n\cdot 1/n \log_2 1/n = - \log_2 1/n$

$= - \log_2 n^{-1} = \log_2 n$.

n 是文本中不同语言符号的数目。

四、汉字熵值测定及不同语种的熵值比较

近几十年来，国外学者已经陆续测出了一些使用拼音文字的语言（如英语、法语、德语、俄语等）包含在一个字母中的熵。这些语言使用的字母数目非常有限，如英语只有 26 个字母，俄语只有 33 个字母。要测出这些语言包含在一个字母中的熵是比较容易的，而汉字的数量很大，要测出包含在一个汉字中的熵就困难得多。

冯志伟老师多年来进行汉字熵的统计和计算工作，采用逐渐增加汉字容量的办法，初步测出了在考虑语言符号出现概率差异的情况下，包含在一个汉字中的比较稳定的熵值为 9.65 比特，当汉字容量不大时，包含在一个汉字中的熵随着汉字容量的增加而增加；当汉字容量达到 12366 个字时，包含在一个汉字中的熵就不再增加了。由此得出结论：从汉语书面语总体来考虑，在全部汉语书面语中，包含在一个汉字中的熵是 9.65 比特，即每当我们从汉语书面语文句中读到一个汉字时，我们就获得 9.65 比特的信息量。

美国宾夕法尼亚大学梅维恒（V.H.Mair）教授曾经评论冯老师说：“ 如果一个人能够用科技术语和数学方程式来论述他的对于现代标准汉语的观点，那么，这样的论述当然应该是非常雄辩而有说服力的。 ”

不同语种一个字母（汉字）中的熵值比较：

英语：4.03 比特；法语：3.98 比特；德语：4.10 比特；

俄语：4.35 比特；意大利语：4.00 比特；汉语：9.65 比特

由上可知，印欧语系语言都是使用拼音文字，采用的字母数目不多，因而它们的不肯定性程度小；而汉语不是使用拼音文字，汉字数量多，因而不肯定性程度很大。但是，汉字笔画的平均熵为 3.43，比英文字母的 4.03 还要小 0.6。

五、蒙古文字母熵值的测定

熵的测定是数理语言学的一项基础研究。蒙古文字母①的熵是蒙古文字母所含信息量大小的数学度量。

英文、法文、德文、意大利文等文字按其文字系统有个统称叫做“ 拉丁文 Latin ”，俄文、乌克兰文、白俄罗斯文、南斯拉夫文等文字按其文字系统也有个统称叫做“ 西里尔文 CYRILLIC ”。

我们沿用确精扎布教授《蒙古文编码》中的方案，即蒙古文 Mongolian letter (script)

[illegible] 作为统称，它包括：

⑴ 传统蒙古文 *Traditional Mongolian letter (script)* [illegible]

⑵ 托忒文 *Todo letter (script)* [illegible]

⑶ 锡伯文 *Sibe letter (script)* [illegible]

⑷ 满文 *Manchu letter (script)* [illegible]

本文讨论传统蒙古文字母所包含的信息量，即传统蒙古文字母中的熵。传统蒙古文字母（这里参考了《蒙古文编码》中的字母）有：元音 8（e）+辅音 27（有 KHA 〈183B〉，没有 SHI），总共 35 个字母。假如每个字母在文章中出现概率相等的话，包含在一个字母中的熵（最大熵）为：

H_{max} =$\log_2 35$=lg35÷lg2=1.544÷0.301 =5.129 比特

由于蒙古文字母在蒙古语书面语中的出现概率是不相等的，所以用一般公式 H= －EPi$\log_2$（Pi）来计算出现概率差异的情况下，包含在一个蒙古文字母中的熵。 通过统计每个蒙古文名义字符在“ 20 万词级蒙古文拉丁转写语料 ”②中出现的概率，发现蒙古文名义字符中使用频率高的前六个字符分别是 A,I,E,G,N,V,它们的总和占了字符总数的 50%。以每个字符在该语料中出现的概率代入一般公式后得出出现概率差异的情况下蒙古文名义字符信息熵为 H= － EPi$\log_2$（Pi）≈4.165，多余度为（H_{max} -H）/ H_{max}≈0.188（即 18.8%）。[1]

六、语言符号的熵与马尔可夫链

根据上文可知包含在一个汉字中的熵为 9.65 比特，但是，在测定这个熵值的时候，仅只考虑到了汉字在文本中出现概率的差异，而完全没有考虑文本中汉字出现概率之间的相互影响。事实上，在任何一个真实的自然语言文本中，语言符号的出现概率是相关的，是彼此相互影响的。除了文字符号的一定的概率分布，包含了一部分信息外，文字的上下文关系，包含了更多的信息，使某一语言文字的实际信息量更小。因此，按一般公式

$$H_s = \sum_{i=1}^{n} p_i I_e = -\sum_{i=1}^{n} p_i \log_2 p_i$$

来计算出的熵值，显然没有反映出真实的自然语言文本中每一个字母（或音素）的实际信息量的大小，还必须考虑语料中每一个字母（或音素）的共现（cooccurrence）关系。

早在 1913 年，俄国著名数学家马尔可夫（A . A . MAPKOB）就注意到语言符号出现概率之间的相互影响，他通过统计普希金《欧根·奥涅金》中的元辅音字母共现关系，得出在俄语中，元音字母在辅音字母之后出现的概率大于元音字母在元音字母之后出现的概率，确切地说明了元音字母和辅音字母之间出现概率的相互影响。

我们可以把语言的使用看成一个随机过程，在这个随机过程中，所出现的语言符号是随机试验的结局，语言就是一系列具有不同随机试验结局的链。

如果在随机试验中，各个语言符号的出现彼此独立，不相互影响，那么，这种链就是独立链。

如果在独立链中，每个语言符号的出现概率相等，那么，这种链就叫做等概率独立链。等概率独立链中语言符号的熵值计算公式为（即**最大熵**）：

H_0 = $\log_2 n$

如果在独立链中，各个语言符号的出现概率不相等，有的出现概率高，有的出现概率低，则这种链叫做不等概率独立链。不等概率独立链中语言符号的熵值计算公式为（即**平均熵**）：

$$H_s = \sum_{i=1}^{n} p_i I_e = -\sum_{i=1}^{n} p_i \log_2 p_i$$

如果在随机试验中，每个语言符号的出现概率不相互独立，每一个随机试验的个别结局

依赖于它前面的随机试验的结局，那么，这种链就叫做马尔可夫链。如果考虑到前面的语言符号对后面的语言符号出现概率的影响，那么，可得出条件熵，马尔可夫链的熵就是条件熵，其计算公式为（即**极限熵**）：

$$H = -\ E\ P[bi(n-1),\ j]\log_2 P_{bi(n-1)}(j)$$

E：取和 [Σ（i，j）]

其中，bi(n-1)是由 n-1 个结局构成的组合，在它后面有第 j 个结局，P[bi(n-1), j]是这个组合出现的概率，$P_{bi(n-1)}(j)$是在由前面 n-1 个结局构成的组合之后，第 j 个结局出现的条件概率。

在马尔可夫链中，前面的语言符号对后面的语言符号是有影响的，事实上，语言就是马尔可夫链。

如果只考虑前面一个语言符号对后面一个语言符号出现概率的影响，这样得出的语言成分的链，叫做一重（阶）马尔可夫链，依此类推。随着马尔可夫链重数的增大，随机试验所得出的语言符号链越来越接近有意义的自然语言文本。美国语言学家乔姆斯基（N.Chomsky）和心理学家米勒（G.Miller）指出，这样的马尔可夫链的重数并不是无穷地增加的，随着马尔可夫链重数的增加，熵逐渐趋于稳定而不再减少，这个不再减少的熵就是包含在自然语言一个符号中的真实信息量，叫做“极限熵”。它的极限就是语法上和语义上成立的自然语言句子的集合。

随着马尔可夫链重数的增大，根据前面的语言符号来预测下一个语言符号出现的这个随机试验的不肯定性越来越小，因而包含在一个语言符号中的熵值也就越来越小。当这个越来越小的熵值达到极限时，就能反映出真实的自然语言文本中语言符号所包含的实际信息量的大小。

这样就有理由把自然语言的句子看成是重数很大的马尔可夫链了。可通过借助马尔可夫模型和条件概率来计算极限熵值。

国外学者已经求出包含在一个英语字母中的极限熵大约在 0.9296 比特到 1.5604 比特之间，其平均值为 1.245 比特。1995 年，冯老师测定了在充分考虑汉字上下文的影响时，包含在一个汉字中的熵，即汉字“极限熵”。通过英汉文本字符容量的对比来间接地推算极限熵，避免了复杂的测试和计算，汉字的极限熵介于 3.0212 比特与 5.0713 比特之间，其平均值为 4.0462 比特。

测定包含在一个蒙古文字母中的极限熵时，我们可以借鉴冯老师的方法，通过英蒙双语语料库文本字符容量的对比来间接地推算极限熵。

七、熵的其他应用

熵是不确定性的量度。因此对事物了解得越多，它的熵就越小。对于构造语言的统计模型而言，如果一个语言模型更加精确地描述了语言的结构，那么它的熵应该越低。我们能够使用熵作为衡量语言模型的质量的参数。

八、下一步的工作及需要讨论的问题

（一）大规模真实语料基础上计算传统蒙古文 35 个字母在不等概率独立链中的熵值**（平均熵）**和多重马尔可夫链中的熵值**（极限熵）**。

拉丁转写语料和蒙古文语料中的熵值计算，这就涉及到名义字符和显现字形。拉丁转写语料中以字母（名义字符）作为一个计算单位，而在蒙古文语料中应该以字母“外形”（显现字形）作为一个计算单位？一对一、一对多、多对一、多对多的关系该怎么处理（同形异码 T/D,H/G,A/E/N,0/V/O/U 等）？

（二）蒙古文不同语言单位（比较稳定的、数量有限的）的熵值计算：字素（强制性合体字）、音素、音位（字母）、音节、词素（包括构词附加成分和构形附加成分）、单词（词根词、派生词）、固定短语（复合词、成语、习用语、固定词）等，分别求出上述语言单位

所包含的信息量，即熵。

（三）语料文本信息量应该包括文字、标点符号、数字等，蒙古文标点符号、蒙古文数字的熵值计算。

注释

① 这里，蒙古文字母指蒙古文名义字符。

② 该语料由264000词（其中包括数字、标点符号、英文字母）组成，它是由内蒙古大学蒙古语言研究所研制的100万词级《现代蒙古语文数据库》的一部分，现已人工做完并校对了词的切分还原及词性标注。

参考文献

[1]那日松，淑琴.蒙古文信息熵和拉丁转写研究[M]. ICCC2007 会议论文.

[2]冯志伟.计算语言学探索[M].哈尔滨：黑龙江教育出版社，2001.

[3]冯志伟.现代汉字和计算机[M].北京:北京大学出版社，1989.

[4]冯志伟.中文信息处理与汉语研究[M].北京:商务印书馆，1992

[5]王晓龙,关毅.计算机自然语言处理[M].北京:清华大学出版社，2005.

[6]确精扎布.蒙古文编码[M].呼和浩特:内蒙古大学出版社，2000.

[7]互联网物理学选读材料.最大信息熵原理[DB/OL].

[8]互联网物理学选读材料.熵与信息[DB/OL].

[9]互联网信息论选读材料.信息熵（Entropy）到底是用来衡量什么的？与 Philip ZHANG 商榷[DB/OL].

[10]人民网：熵：一种新的世界观[DB/OL].

Mongolian Information Entropy

Shuqin, Narisu

（1、The Institute of Mongolian Studies, Inner Mongolia University, Hohhot 010021,China；2、Hangzhou Normal University, Hangzhou 310012, China）

Abstract: Measuring the information entropy is a basic research of mathematical linguistics. The entropy of linguistic sign is the measurement of linguistic sign's uncertainty. It implies the amount of information contained in linguistic sign. The entropy of Mongolian letter is the information content in Mongolian letter. This paper mainly discusses the information content in traditional Mongolian letter, namely the entropy of traditional Mongolian letter, and gives the maximum entropy, average entropy and redundancy.

Key words: Mongolian; letter; information entropy

收稿日期:2009-01-10;

基金项目:教育部、国家语委民族语言文字规范标准建设及信息化项目《蒙古语语言知识库的建立》（MZ115-038）；国家自然科学基金项目《<蒙古语语义信息词典>的设计与实现》（60873084）；

作者简介:1、 淑琴（1979－），女，内蒙古科左后旗人，内蒙古大学蒙古学学院博士研究生,主要研究方向为蒙古文信息处理；2、那日松（1980—），女，内蒙古扎赉特旗人，杭州师范大学应用语言学研究中心研究人员，主要研究方向为计算语言学。