计算机汉字语句输入的最短码长

王轩　　王晓龙

（哈尔滨工业大学计算机系 150001）

摘要：本文依据信息论的编码理论提出了计算机汉字编码输入的无噪离散信源模型，并给出了 r 元汉字字、词、句编码输入的最短码长的计算方法。通过对上千万字语料的统计计算出了以字、词为单元的 Unigram、Bigram 和 Trigram 统计语言模型下码元数 r 取不同值时的最短码长。

关键词：无噪离散信源　语句输入　统计语言模型

The shortest coding length of computer Chinese character input by sentence

Wang Xuan　　**Wang Xiaolong**

Dept. Of Computer Science and Engineering, Harbin Institute of Technology, Harbin 150001

Abstract: In this paper, Chinese ~~keyboard~~ input was regarded as a discrete noiseless channel. With this model, a method was put forward to compute the shortest coding length under various statistical language model, such as Unigram, Bigram and Trigram. In the end, the experimental results were given.

Keyword: Discrete channel without noise, sentence input, statistical model

一、引言

计算机文字信息输入技术是计算机智能接口和智能应用领域的重要研究方向。在国内外均作为高技术前沿课题加以研究。在我国一直得到国家八六三高技术计划智能计算机主题的重点资助。

用计算机对信息进行处理其首要问题就是解决怎样把信息输入到计算机中。因此，计算机信息输入问题被看作是计算机信息处理的瓶颈问题。大字符集汉字的计算机输入是其中最困难的问题之一。计算机汉字输入从外设使用形式上看可分为键盘输入、声音输入、文字识别三类。近些年来，虽然语音识别和文字识别在理论和实践上取得了长足进步，但距实用化阶段还有很大距离。目前，汉字输入的主流还是计算机键盘编码输入。自七十年代末到现在，计算机汉字输入技术已经经历了单字编码输入、词组编码输入和智能语句输入三个阶段[1]。输入速度不断提高，用户界面更加完善，智能化程度也有了很大提高。不仅专业录入员的瓶颈已经基本解决，而且，非专职录入人员的瓶颈也得到缓解。

从理论上来讲，汉字编码输入的平均码长不会由于采用各种优化方法被无限缩短，输入

速度也不会被无限地提高。也就是说，汉字编码输入的平均码长和输入速度存在着极限。研究汉字编码输入的最短码长，进而推导出汉字输入速度的上限，不仅对汉字输入系统的评价及确定汉字输入技术的研究方向具有重要的意义[2]，而且，其研究成果在电子通信领域、语音识别、文字识别、机器翻译以及信息压缩等方面均有着重要的理论和应用价值。

文献[11]在汉字的字频分析的基础上，讨论了最佳编码和汉字输入问题的关系。文献[2]运用信息论中 Huffman 最优编码理论，把二元 Huffman 最优编码推广到 N 元最优编码，并在 IBMPC 和工作站上就 52 万字和 180 万字的样本数据对 r 取 24 种不同的码元情况下进行了计算，给出了各种常见的字、词编码输入方案的平均码长下限。并就最短码长精度、码元 r 与编码效率、词数与码长的关系以及录入人员平均输入速度的上限进行了讨论。但是限于当时机器容量和速度的限制，一些耗时巨大的试验无法实现，许多理论上的结果也无法得到验证。近些年来，随着计算机的硬件与软件技术的迅猛发展，使得在更大的数据规模下重新计算和验证以往已得到的和未得到的理论结果提供了可能的试验基础。

本文提出了汉字编码输入无噪离散信道模型，运用信息论中的信道编码理论，得到汉字编码输入最短码长的计算方法，并通过对上百万字语料的统计，计算出了不同码元数 r 下应用字、词一元、二元、三元统计语言模型进行汉字字、词、句编码输入的最短码长。

二、汉字编码输入的无噪离散信道编码模型

设原始信源 S 发出 q 种不同的符号，其符号集 $S:\{s_1,s_2,...,s_q\}$。其信源空间为

$$[S\cdot P]:\begin{cases} S: & s_1 & s_2 & \cdots & s_q \\ P(S): & p(s_1) & p(s_2) & \cdots & p(s_q) \end{cases} \tag{1}$$

其中

$$\sum_{i=1}^{q} p(s_i)=1 \tag{2}$$

传输信息的信道 $\{X \quad P(Y/X) \quad Y\}$ 的输入符号集 $X:\{a_1,a_2,...,a_r\}$。这样，由于信源 S 发出的符号 $s_i(i=1,2,...,q)$ 与信道能传递的符号 $a_i(i=1,2,...,r)$ 不一致，为此，在传输之前必须用信道能传输的符号集 $X:\{a_1,a_2,...,a_r\}$ 中的符号对信源 S 中的每一种不同的符号 $s_i(i=1,2,...,q)$ 进行编码。用符号集 $X:\{a_1,a_2,...,a_r\}$ 中的某种符号序列 $W_i(i=1,2,..,q)$ 代表信源 S 中的不同符号。符号集 $X:\{a_1,a_2,...,a_r\}$ 称为码元集，$W_i(i=1,2,..,q)$ 成为码字。

对于通信来说，除了要求信源 S 能适合于信道传输之外，还要求能无失真传递信源 S 发出的每种不同的符号 $s_i(i=1,2,...,q)$，以及 S 的 N 次扩展信源 $S=S_1S_2...S_N$ 的每一种不同的消息 $a_i=(s_{i1}s_{i2}...s_{iN})$，$s_{i1},s_{i2},...,s_{iN}\in\{s_1,s_2,...,s_q\}$，$i1,i2,...,iN=1,2,...,q$，$i=1,2,...,q^N$。即信源的 N 个符号序列。这就要求，每一种码字 $W_i(i=1,2,..,q)$ 必须与信源 S 发出的每一种不同的消息（符号）$s_i(i=1,2,...,q)$ 一一对应，且与信源 S 的 N（任意有限大）次扩展信源

$S=S_1S_2...S_N$每一种不同消息$a_i(i=1,2,..,q^N)$一一对应，这样在无噪声信道中才能达到无失真传递信源S发出的消息。满足上述条件的码字$W_i(i=1,2,..,q)$称为单义可译码。

对于汉字编码输入来说，可以把键盘看作是一个无噪离散信道，把自然语言看作是信源S，它从已知的有限符号集$V=\{w_1,w_2,...,w_q\}$中以某种概率分布$P(w_i),i=1,2,...,q$输出字符w_i。S的N次扩展信源$S=S_1S_2...S_N$的输出就是长度为N的汉语语句和文本。当符号集$V=\{w_1,w_2,...,w_q\}$中的元素分别为字或词时，且信源S为离散无记忆信源，则无噪离散信道编码模型分别对应单字编码输入和词组编码输入。若信源S为离散有记忆信源时，对应的是语句编码输入。

三、r元汉字编码输入的平均码长

若单义可译码$W:\{W_1,W_2,...,W_q\}$的码字$W_i(i=1,2,..,q)$的长度分别为$n_i(i=1,2,..,q)$，由单义可译码的定义可知，码字$W_i(i=1,2,..,q)$与信源符号$s_i(i=1,2,...,q)$一一对应，所以，其相应的码长$n_i(i=1,2,..,q)$与信源的符号$s_i(i=1,2,...,q)$相应的先验概率$p(s_i)(i=1,2,...,q)$也是一一对应的。也就是说，$p(s_i)=p(n_i)(i=1,2,...,q)$。因此，一个信源符号所需的平均码长为

$$\sum_{i=1}^{q}p(s_i)\cdot n_i=\bar{n} \tag{3}$$

$\bar{n}$称为平均码长。

由平均码长界限定理可知[4]，对于熵为$H(S)$，码符号集为$X:\{a_1,a_2,...,a_r\}$的离散无记忆信源S，平均码长的下界为$\frac{H(S)}{\log r}$。若采用直接对平稳遍历信源S的N次扩展信源$S=S_1S_2...S_N$的每一种不同的消息$a_i=(s_{i1}s_{i2}...s_{iN})$，$i=1,2,...,q^N$直接进行编码，则平均码长的下限为$\frac{H_\infty}{\log r}$。特别当$S$为有记忆m阶马尔克夫信源时，有$H_\infty=H(S_{m+1}/S_1S_2...S_m)=H_{m+1}$，则平均码长的下限为$\frac{H_{m+1}}{\log r}$。

可见，只要求出H_∞就可以得到平均码长的下限。由于自然语言是各态遍历的，可用下式计算H_∞[5][6][7]

$$H_\infty(P)=\lim_{n->\infty}[-\frac{1}{n}\log P(x_1,x_2,...,x_n)] \tag{4}$$

其中，$x_i, i=1,...,n$为信源输出的第i个字符。

通常可以把自然语言看成是 m 阶的马尔克夫信源，则

$$P(x_1,x_2,...,x_n)=\prod_{i=1}^{m}P(x_i/x_1,x_2,...,x_{i-1})\cdot\prod_{i=m+1}^{n}P(x_i/x_{i-m},\cdots,x_{i-1})$$

根据大数定理，可以通过统计大量训练文本中词串$x_{i-m}x_{i-m+1},...,x_{i-1},x_i$的出现次数$C(x_{i-m}x_{i-m+1},...,x_{i-1},x_i)$来计算$P(x_i/x_{i-m},...,x_{i-1})$，其公式如下：

$$P(x_i/x_{i-m},...,x_{i-1})\approx\frac{C(x_{i-m},...,x_{i-1})}{\sum_{x_i}C(x_{i-m},...,x_{i-1},x_i)} \tag{5}$$

当训练样本空间不足够大时，许多合法语言现象在训练集中未出现，使得很多参数为 0，这就是所谓的数据稀疏问题。为此要对参数进行平滑处理，处理的方法很多[8][9][10]。本文采用的是插值平滑算法，以 Trigram 为例，其数据平滑公式为：

$$P(x_i/x_{i-2},x_{i-1})=\lambda_1\cdot P(x_i)+\lambda_2\cdot P(x_i/x_{i-1})+\lambda_3 P(x_i/x_{i-2},x_{i-1}) \tag{6}$$

其中，$\lambda_i\geq 0, i=1,2,3$，$\lambda_1+\lambda_2+\lambda_3=1$。

四、实验与结论

实验中采用现代汉语平衡语料库作为统计模型训练数据，语料库规模为 2000 万字，其中《人民日报》1000 万字，《中国新闻》500 万字，各类书籍 250 万字，文学作品 150 万字（其中小说 100 万字、散文 30 万字、报告文学 20 万字），准口语语料：对白 60 万字（话剧剧本）、独白 40 万字（包括单口相声、演讲词、讲话、故事）。语料覆盖文化、经济、地理、历史、文学、军事、政治、科学等领域。

统计出的 Unigram、Bigram、Trigram 的数据分布如表一所示（由于篇幅所限只给出了词的统计结果）。我们把基于字和词的一、二、三元文法模型分别编号为$C-M_1$、$C-M_2$、$C-M_3$、$W-M_1$、$W-M_2$、$W-M_3$。从语料库中随机抽出 100 万字的文本用来计算各模型的熵值，其结果如表二所示。H_{LM}和H'_{LM}分别为本文和文献[]的统计结果，可以看出，两次统计结果基本相同。只是在模型$W-M_1$下的结果相差较大，这可能是由于两次统计所采用的词典和语料的不同造成的。统计模型所采用的数据平滑算法也会对统计结果造成影响。同时还可以看到，词典和语料等因素对高阶统计语言模型的影响要小于同样情况下对低阶模型的影响，这可从在模型$W-M_2$下两次统计结果非常接近得到验证。模型$W-M_3$的熵值（3.868）大大要低于文献[3]中提出的长距离二元文法模型$W-M_{22}$的熵值（5.17），说明$W-M_3$模型的语言约束能力要比$W-M_{22}$模型的语言约束能力强很多。

在不同码元r下字、词、句编码输入的平均码长如表三所示。从实验结果可以看出，以

字为单元的语言模型在不同码元下的最短码长要小于以词为单元语言模型的最短码长，也就是说，基于词的模型比基于字模型更接近于真实的汉语语言模型。对应于语句输入法的$C-M_2$、$C-M_3$、$W-M_2$、$W-M_3$模型的最短码长要小于对应于字、词输入法的$C-M_1$、$W-M_1$模型的最短码长。把模型$C-M_1$、$C-M_2$的结果与文献[2]的结果相比，可知[2]中的结果在实验误差范围内是可信。

表一 UNIGRAM、BIGRAM、TRIGRAM 统计结果

出现次数 k	$\geq k$ 的 UNIGRAM 个数	$\geq k$ 的 BIGRAM 个数	$\geq k$ TRIGRAM 个数
1	44031	1075213	2488123
2	37287	342324	329084
3	32804	198094	136588
4	29481	138547	80033
5	26881	105777	54051
10	18959	46600	17150
20	12490	20409	5496
30	9426	12388	2786
50	6571	6370	1172
100	3904	2503	366
500	875	235	20
1000	393	94	9
5000	66	8	1
10000	28	1	1
20000	14	0	0
50000	4	0	0
100000	2	0	0
200000	1	0	0

表三 N 元汉字编码的最短码长

码元 N	$C-M_1$	$C-M_2$	$C-M_3$	$W-M_1$	$W-M_2$	$W-M_3$
2	9.630	7.142	6.279	8.090	5.489	3.868
3	6.075	4.504	3.961	5.104	3.463	2.440
4	4.815	3.570	3.139	4.054	2.744	1.934
5	4.147	3.075	2.704	3.484	2.363	1.665
6	3.725	2.762	2.429	3.129	2.123	1.496
10	2.898	2.149	1.890	2.435	1.652	1.164
26	2.048	1.519	1.335	1.721	1.167	0.822
36	1.862	1.381	1.214	1.564	1.061	0.748
50	1.706	1.265	1.112	1.433	0.972	0.685

表二各统计模型的熵

模型	$C-M_1$	$C-M_2$	$C-M_3$	$W-M_1$	$W-M_2$	$W-M_3$
H_{LM}	9.63	7.142	6.279	8.090	5.489	3.868
H'_{LM}	9.64	7.2	6.31	7.03	5.46	

参考文献

[1]王晓龙，王开铸，声音语句输入的研究，《计算机学报》Vol.17. No.2,1994.

[2]王晓龙、王轩，N元汉字字词编码输入的最短码长和速度上限，中文信息学报，1993，7(4)，18-26。

[3]吴军、王作英、余嘉联，语言模型复杂度的研究，第二届中国计算机智能接口与智能应用学术会议论文集。1995，7。

[4]姜丹、钱玉美，《信息理论与编码》，中国科技大学出版社，1992。

[5]C.shannon, Prediction and entropy of printed English, Bell System Technical Journal, Vol.30 , pp. 50-64,1951.

[6]P.F.Brown, etc. An Estimate of an Upper Bound for the Entropy of English, Computational Linguistics, pp31-40,1992.

[7]F.Jelinek, Self-Organized Language Modeling for Speech Recognition, IEEE ICASSP'89, pp587-595,1993.

[8]I.J.Good, The population frequencies of species and estimation of population parameters, Biometrika, Vol.40. No.3,4, pp.237-264.

[9]A.Nadas, Estimation of Probabilities in the Language Mode of the IBM Speech Recognition System, IEEE Trans. Acoustic, Speech, Signal Processing ,Vol.ASSP-32,pp.859-861,Aug.1984.

[10]A. Nadas, On Turing's Formula for Word Probabilities, IEEE Trans. Acoustic, Speech, Signal Processing , Vol.ASSP-33,pp.1414-1416,December,1985.

[11]石贵青，徐秉铮，汉字字频分布，最佳编码和输入问题。《电子学报》，No.4,1984。

[12]王晓龙等，汉字编码方案的择优、统一和发展。《电子学报》，Vol.15,No.1,1987。