关于中文数字化几个问题的思考

Wang Xiuli [1;1)]

1(Anhui University Hefei 23009)

摘要 本文考察中文数字化的几个问题：繁简转换的问题、古文字编码的问题、文字统一问题、汉字简化等问题，就这几个问题提出了自己的看法和一些对策。

关键词 繁简转换、信息论，汉字字体，汉字结构，汉字编码。

1 繁简问题

首先，要明确区分语言和文字两个层面“激光”转换为“鐳射”是语言翻译或者方言翻译的问题，不是繁简转换的问题。文字的分歧在文字层面上解决，语言的分歧要在语言层面上解决。有关汉语变体之间如何翻译或者统一或者做其他处理，尤其是书面语之间如何翻译或者统一或者做其他处理，需要在语言层面上解决。一个办法就是，不翻译，不处理，让几种平行或者对应的词语等等竞争从而自然统一，或者任由其存在。一种办法是动用行政或者法律。显然，目前，这在两岸四地是不现实的。不过将来则未必不可能。将文字层面的问题揉合到语言层面来解决，我们没有看出有什么益处。

1.1 繁简转换

一般认为，繁简转换是两岸四地中文信息化的一个重要方面。大家就简繁转换碰到的问题做了很多探讨。我们从信息论的角度和机器翻译的角度来看一下这个问题。

先给出信息论的几个概念和公式：

熵公式的一般形式：

$$H=\lim_{n->\infty}\frac{1}{n}\Sigma p(B_n)\log_2 p(B_n)$$

互信息(Mutual Information)是另一有用的信息度量，它是指两个事件集合之间的相关性。两个事件X和Y的互信息定义为：

$$I(X,Y)=H(X)+H(Y)-H(X,Y)$$

其中H(X,Y) 是联合熵(Joint Entropy)，其定义为：

$$H(X,Y)=-\sigma_{x,y}p(x,y)logp(x,y)$$

从信息论的角度看，自然语言的各个单位之间都是有关联的，作为记录汉语语素的汉字，在记录自然语言时自然也记录或者映射了汉语语素之间的关联，这样的关联可以用几种数学模型来刻画，如n阶马尔可夫链可以较好地刻画汉语文字之间的关联，互信息（mutual information）也可以刻画汉语语素进而汉字之间的互信息。考虑汉字之间的互信息或者n阶马尔可夫链，可以看出，繁简转换出现一定比例的错误，并不一定导致交流或者交际的困难，试举一例，“令人发指”即使转换为“令人發指”.人们也一定能够明白而不影响交际。换言之，繁简转换允许一定的出错率，这个出错率当然需要计算或者实验确定。在没有准确的计算结果前，要求越准确越好。但这其实除了社交礼仪等非信息交流方面而外没有必要。所以，浅见认为，找出或者计算出容许的出错率，是目前一项需要做的工作。

繁简转换可类比机器翻译，只是机器翻译是对两种语言的，繁简转换是对一种语言的两种文字符号的。机器翻译自动评测问题。繁简转换既然不要求准确率为100%，同样有评测或者准确率是否达到要求的问题。一种机器翻译自动评测方法是基于平行句对。作为繁简转换评测的一种方法，可以借鉴机器翻

2008 – 10 –收稿

1) wangxiuli123cn@yahoo.com.cn

译方法。因此有必要建立一些评测用的繁简对齐的平行语料。这尤其容易检测一些容易出错的繁简转换，比如跟词语或者句子相关联的繁简转换。

同样，机器翻译的一个方法就是利用平行语料库的翻译。同样，繁简转换，也可以基于繁简平行语料库。这一资源的建立有助于解决一些难题，比如跟词语或者句子相关联的繁简转换。但是这类难题涉及的范围相对小，建立这一资源的代价和所解决的问题之间如何折中，需要考虑。

顺便说一下，类比机器翻译，我们也可以明白，为什么繁简转换事实上也不需要100%的准确率。

2 文字统一问题

汉字废立之争和繁简统一之争一直在文改界存在。这些争论起初并没有牵涉到信息技术，但现在却与信息技术纠结在一起。我们从信息论和目前信息技术角度考虑这一问题。

如果文字统一，则繁简转换自然不需要。换言之，在信息交换和处理过程中，繁简转换其实起了一种类似统一文字的作用。作为一种自动技术，繁简转换的开销并不大。而在做中文信息处理时，对于繁简，目前的编码技术，确实需要做转换，使用unicode之后，这部分所要做的开发就相对小些。就是说，事实上在信息处理和交换或者中文数字化方面讨论文字统一或者繁简统一，没有以前那么严重或者重要（意义不大？）。

同样，拼音文字和汉字废立之争，在信息交换和处理中的意义需要重新考虑。同样，拼音-文字转换软件使得这一问题的意义变弱，尤其是新技术的应用使得原先以为很严重的问题（如输入法重码率高的问题）现在弱得多了。

3 古文字编码问题

古文字到底是独立编码，还是跟现代汉字统一编码？

给文字编码是方便信息处理和交换，换句话说，文字在信息处理和交换中出现的概率很高。降低信息交换和处理的开销，而给文字编码。这本是基于信息论作出的决策。所以给一种符号编码还是直接用图像（比如使用图形编码）表示，是考虑其在信息处理和交换中出现的概率的。显然，古文字就整个社会来讲需要交换和处理的概率很低，恐怕不足以编码；但就古文字或者文字学界而言，需要交换和处理的概率又高得多。和现代文字统一编码，会带来很多问题。考虑信息论和古文字的使用范围，应该单独编码。

4 信息论

一般认为，香农的《通信的数学理论》[3]为信息论奠定了理论基础。现代有关信息的各个方向几乎都受到其影响。而汉字作为一种符号，编码了汉语中的语素或者词，因此也可以从信息论的角度加以考察。学者们从信息论角度考察汉字，大多是计算汉字的熵[1]，亦即将汉字作为一个不可分割的符号来处理。学者们从此角度研究汉字信息熵，得到了一系列结论，对于汉字与汉字信息处理做出了贡献.

然而事实上，汉字在一个层面上可以作为不可分割或者分析的符号分析，在另外一个层面上则可以作为一个结构来分析。就作者所见，尚未有研究者从信息论角度分析汉字的结构和汉字的演化，作者从信息论的角度研究有关汉字的结构和演化等等问题，力图探索得出汉字演化的规律和编码的准则以及制定汉字字体等等问题的原则

5 信息论和Huffman编码

香农在其《通信的数学理论》以及其他的信息论著作中[3–6]给出了数据压缩的公式，并且证明，无损压缩存在上限，即现在为大家所熟知的熵率（entropy rate）。当然香农也建立了有损压缩的理论，亦即大家所熟知的扭曲速率理论（rate-distortion theory）并且建立了有关最优有损压缩理论。显然，这两种压缩理论有密切关系，有损压缩理论可以看作无损压缩理论的推广。这两种理论给出了各种信源编码算法的上限。

无损压缩

熵公式的一般形式：

$$H = \lim_{n->\infty} \frac{1}{n} \Sigma p(B_n) \log_2 p(B_n)$$

其中B_n表示前n个符号.

定理 1 *:令R_n^*为最优n阶无损压缩编码，那么*

$$-\frac{1}{n}\Sigma p(B_n)\log_(B_n) \le R_n^* < \frac{1}{n}\Sigma p(B_n)\log_2 p(B_n) + \frac{1}{n}$$

令

$$n \to \infty$$

，*则*

$$\lim_{n\to\infty} R_n^* = H$$

有关有损压缩的公式（暂略）

最优无损压缩和有损压缩仅仅给出了理论上限，并没有给出实现的方式。具体实现的技术则要考虑其他因素，比如计算效率或者复杂性等等。

Huffman编码即为信息压缩或数据压缩的一种方案。其基本的思想是，根据符号的权值，给符号一编码，使得所得码长在统计意义上为最短。具体编码方案可参看有关著作。为研究方便，我们用huffman编码而非其他编码方案如算术编码、前缀编码等等，所得结论可以容易地推广到其他编码或者一般信息压缩的情景。

6 汉字

如果从甲骨文出现算起，则汉字经历甲骨文、金文、小篆、现代汉字的演变。学者对汉字结构、形体及其演变做了深入研究，这些研究基本上是文字学意义上的，有关信息处理方面的研究则是计算汉字熵、编码汉字等等。当然得到不少有价值的结论。我们不打算在此文中做类似的研究，而是从另外一个角度来分析其结构形体演变，力图得出一些有价值的结论。。

6.1 汉字演化

关于汉字的演化，大家都认为，基本是从图形意味或者绘画意味浓厚的古文字变为少有绘画意味的现代文字，从书写困难变到书写相对容易，从笔画繁多、结构复杂逐步趋向简单。这些都是定性的描述，而且多为印象式的描述。

而演化的发生，往往是首先出现在非正式的场所，往往是由民间而到官府或者政府。比如，隶书的产生，大致可以认为先是在非正式场合或者民间产生使用，然后成为正式的官方标准问题。即以现代简体字而言，也多可在民间用字上找到根源. 显然，这一现象是由几方面原因造成的。首先，汉字作为记录语言的符号，是辅助交际的，必须满足交际的需要。其次，理论上，只要达到交际目的，文字便可不受其他限制，而民间或者非正式场合的限制就较少，比如正式的形体结构等等都可以变通。所以，如果不考虑一些强制的标准(官方)，则非正式场合或者民间更多表现出一种相对自然的演化过程。自然演化的特点更容易在这样的环境中探索出来。

草书为书写省力而快捷，连笔很多，不少一气呵成。在很大程度上，只要各个字形能够相互区别，或者在具体上下文（互信息）中相互区别，就可以。因而出现字体和字形的很多变化。这些变化如果经常出现，则作为一种新字形，可以影响到官方认可的文字。例如（汉字演化举例）。

书信中的文字（王羲之等人的便笺）

民间汉字，可看到很多俗体，即使现在，也可在好多地方看到。

而官方则因为有较多的限制，这样的变化就难出现。

初看起来，汉字只有在文本中，才能计算其熵。这是因为汉字编码了语素或者词，而且我们把汉字作为不可分析的符号。这样分析并无错误。然而如果我们把汉字看作一个特殊的文本，并进而把通常意义上的文本当作文本集合，则笔画或者偏旁部首等等被当作不可分析的符号。这样就可以从信息论角度分析汉字的结构和笔画以及笔画和结构的演化

从信息论的角度来看，出现频率较高的汉字容易简化，即以信息论中的数据压缩定理或Huffman编码看，一个频率较高的汉字出现省笔、简化，可以缩短书写时间或者编码长度（如果我们把笔画算作编码如现代计算机中的1、0的基本单位的话）.上述草书、书信文字、民间文字正是如此。那些很生僻的汉字，即使结构非常复杂，也没有简化，便也是这个原因。

6.2 汉字结构、演化与形式语言

把汉字看作一个特殊的文本，笔画或者偏旁部首等等被当作不可分析的符号。这样就可以从信息论的角度分析汉字的结构和笔画，也可以从形式语言的角度分析汉字的结构和笔画以及笔画和结构的演化

笔画、偏旁和构造组合的方式区分开汉字。因此可以把一个汉字看作形式语法意义上的文本或者形式语法意义上的一个句子。给其构造重写规则，则得到一部文法。如够– >句+多；多– >夕+夕。如此等等。显然，结构的演变和笔画的增减变化首先导致文字系统的变化，都可导致文法的变化。文字系统可以看作形式语言意义上的语言，亦即文字的集合是一语言，变化前后的语言显然往往不同，而文法也不同，前后的文法和语言都不等价。如果结构演变和笔画增减变化前后的文字是一一对应的关系，或者基本是一一对应的关系，则从文字记录自然语言这个角度看，文字未变，而刻画文字的文法有了变化。这样，我们便有了从文法的角度衡量两种文法效率（或

者复杂程度）的问题。引入两种文法的Kolmogorov complexity[2]，便可度量两种文法的复杂度。换言之，从形式语言和Kolmogorov complexity角度看，汉字的演化实在是有优劣之分的。当然，这仅仅是从当代交际的角度看，并没有计入传承古代文化这一因素。

7 余论

从信息论和形式语言等角度考察了汉字形体和结构的一些问题，可以确认：

汉字由繁变简 是因为遵循信息论类似Huffman编码之类的数据压缩规律。

汉字部件笔画的熵 把汉字看作可分析的符号串，可以计算各个笔画或者部件的熵值。进而为汉字笔画和部件演化提供一个新的观察视角而探索其规律。

汉字结构的形式语言视角 据此，可以度量汉字演化（包括简化）的效率问题（复杂性）

这些看法往往是仅仅考虑信息论或者形式语言而得出的。这些观察所得，也没有进行试验研究。进一步的研究需要做较大规模的实验，以期完善，并考虑其他因素。

参考文献(References)

1 冯志伟. 汉字的极限熵. *中文信息*, 2, 1996.

2 Ming Li and Paul M. B. Vitanyi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer-Verlag, Berlin, second edition edition, 1997.

3 C. E. Shannon. A Mathematical Theory of Communication. *Bell System Tech. J.*, 27:379–423, 623–656, 1948.

4 Claude E. Shannon. General treatment of the problem of coding. *IEEE Transactions on Information Theory*, 1:102–104, 1953.

5 Claude E. Shannon. Certain results in coding theory for noisy channels. *Information and Control*, 1(1):6–25, 1957.

6 Claude E. Shannon, Robert G. Gallager, and Elwyn R. Berlekamp. Lower bounds to error probability for coding on discrete memoryless channels. ii. *Information and Control*, 10(5):522–552, 1967.

Some random thoughts on Chinese digitization

wang xiuli[1;1)]

1 (Anhui University,Hefei 23009, Anhui China)

Abstract The article discusses Chinese digitizing questions transformation between simple Chinese character and complex Chinese character , character of archaic Chinese ,unification of Chinese character and simplification of Chinese character presents our opinions on the questions and measure to solve the problems

Key words transformation between simple Chinese character and complex Chinese character ,information theory, font of Chinese character ,structure of Chinese character ,Chinese character encoding

Received date Oct. 2008

1) wangxiuli123cn@yahoo.com.cn