

文本文件修复工程初探

李飞 宋柔

北京工业大学计算机学院

摘要 在大规模汉语语料库中, 不可避免的存在乱码, 为了尽量不影响对语料库的统计和分析, 应尽量修复带有乱码的文本。我们用汉字编码分区识别和接续关系识别的办法对乱码进行了处理, 收到了初步成效, 研制了只需极少人工干预的文本文件修复软件。

关键词 乱码 字法级乱码 字义级乱码 乱码修复

Research of Repairing Text Files

Li Fei Song Rou

Computer Institute of Beijing Polytechnic University

Abstract: In the large scale Chinese corpus, it is inevitable that there are disordered codes in it. In order to ensure the results of the statistics and analysis for the corpus, the disordered codes should be restored to the full. We use the recognition of Chinese character code section and the recognition of continuation relations to process the disordered codes and develop the software to restore the text file with little human working.

Keywords: disordered code non-character disordered code meaningless disordered code disordered code restoration

一、引言

在中文处理领域的研究中, 通常要用到大规模汉语语料库。然而在汉语语料库的搜集、建立和维护中会碰到乱码问题。打开文本文件, 有时会发现文本中夹杂了一些根本读不通的汉字字符串, 其中有的是“怪”汉字, 也可能有一些空字符。这些字符串有的是几个字, 有的占一行, 有的占好几行。这就是通常所说的中文文本中的“乱码”了。

汉字编码的特点之一是利用两个字节(分别存储汉字的区码和位码)存储一个汉字, 如果汉字文本中有错误或干扰信息, 就可能扰乱原来的汉字区位码排列关系, 造成乱码。

造成乱码的几种原因和后果是:

1. 用程序处理汉字文本或用不完善的编辑器对汉字文本编辑的过程中, 汉字文本中意外丢失或被更改了一个字节(半个汉字), 例如:

现有汉字串“南北战争”, 内码与汉字对照关系如下表:

内码:	“0xc4 0xcf 0xb1 0xb1 0xd5 0xdb 0xd5 0xf9”
对应汉字:	南 北 战 争

假如汉字串丢失了“北”字的第一个字节“0xb1”，内码与汉字的对照关系变成下表：

内码：	“0xc4 0xcf 0xb1 0xd5 0xdb 0xd5 0xf9”
对应汉字：	南 闭 桔

“北”字之后的汉字错位显示。

2. 汉字文本文件中插入了一些排版系统的排版符号，例如：

①中文处理软件 WPS6.0 中的排版符号‘软回车’，在 WINDOWS95/98 系统中会被显示为汉字“崐”。②有的排版软件的回车符号不规范，例如把回车符号 0x0d, 0x0a 存储成了 0x0d, 0x8a，使 0x8a 成了下一行第一个汉字的区码，WINDOWS95/98 系统中的大多数汉字处理软件显示这种文本时，会出现以“嫫”、“娟”、“姘”、“始”等字开头的乱码行，乱码范围仅限于当前一行。

3. 有的汉语语料是数据库或超文本格式，在收集语料时，把这些语料转为纯文本格式时，一些数据库文件头、字段分隔符和多媒体链接指针等多余字符，可能与正常文本混在一起，产生乱码。若把 ASCII 码为 0x7f、0xff 的字节加入文本，就会使大部分字处理软件误认为文本文件已结束，造成后面的文本丢失。

4. 在文件传输的过程中（如拷贝、收发 E-MAIL），意外丢失或被更改了多个字节，这不同于第 1 种乱码中所说的意外，这里所说的意外通常是大量的汉字文本丢失或被更改。例如：拷贝多年不用的软盘中所存的文本文件时，如果在碰到软盘坏道时仍硬性拷贝，操作系统通常会用随机生成的一些字符填满目标文本文件的相应区域，形成乱码。

有的全文翻译软件或字处理软件因为受乱码干扰而造成软件非正常退出或死机。

乱码不仅妨碍对文本的处理，而且从根本上影响汉语处理系统的性能。汉语处理系统开发时需要做统计工作，如词频、接续关系、词语间转移概率等。统计用语料如有乱码，统计结果会失真。用这种统计结果开发的汉语处理软件，其性能显然不可靠。

但是，修复乱码时会遇到各种困难。

1. 人工识别及修复的困难：

首先是工具问题，由于目前常用的中文处理软件（例如 MICROSOFT WORD、WPS97 等等）都是对汉字进行整字处理，无法操作半个汉字。如果在文本文件中发现乱码则很难纠正。即使能操作半个汉字，原文已经丢失或更改的乱码也不能完全修复。上面第 3 点和第 4 点中所说的非汉字字符都混在正常文本文件中，要人工分辨哪些是汉字，哪些是多余或随机生成字符，还有困难。且有的字符是不可见的 ASCII 码字符（如小于 0x20 的字符），中文处理软件一样不能处理甚至不能发现而留下隐患。更甚的是如果 ASCII 码为 0x7f、0xff 的字符掺入文本，使字处理软件误认为文本结束，甚至人工也不能察觉后面文本已丢失。

其次是工作量问题，近年来汉语研究领域扩大、深度加深，大规模汉语语料库的字数有几亿甚至几十、几百亿，这样大规模的语料，完全用人工来维护和修复是不可想象的。

2. 机器识别及修复时遇到的困难：

上文中所举的几种原因造成的乱码可分为两类：即字法级乱码和字义级乱码。

字法级乱码即不符合规定范围汉字编码规范的乱码，即文本中出现的规定字符集以外的符号或非合法的 ASCII 码控制字符，这种乱码在进行汉字编码检查的时候绝大部分能够被发现，但修复时不能确定应该如何修复、修复成什么。

字义级乱码是指符合汉字编码规则，但仍是读不通的汉字乱码，就像上面举的例子

一样，“南北战争”变成了“南闭粘”。这种乱码因为符合汉字编码规则而无法用汉字编码检查来识别，虽然人工可能一眼就看出有乱码，但机器没有汉语知识而不能发现乱码，使得这种乱码的自动识别、自动改正成了问题。

我们利用汉字的编码分区识别技术和汉语的词语接续检查技术编写了一个对纯文本文件中存在的乱码进行检出、修复的软件工具《工智文本修复器》，并设置了人工干预修复和自动修复两种模式，取得了较好的效果。

二、处理方法

一、方法简介

(一)、汉字的编码分区识别技术

在处理简体汉字文本时，除了极少数汉字外（如“朱镕基”的“镕”、“彭珮云”的“珮”、“瞭望”的“瞭”、“睡眼蒙眈”的“眈”等）其余汉字均在 GB 字符集以内。

目前常用的计算机操作系统是简体中文 Windows95、98，所用汉字字符集是 GBK 字符集，GBK 字符集包含并兼容 GB 字符集，是 GB 字符集的扩展。

GBK 字符集编码的范围是： 区码：0x81~0xfe 位码：0x40~0xfe

GB 字符集编码的范围是： 区码：0xa1~0xf7 位码：0xa1~0xfe

汉字字符集中的汉字与符号的排列很有规律，如果把 GBK 字符集按照简繁汉字与符号区域划分的话，可得图 1。有了 GBK 编码的分区图后，就可以按照汉字编码或符号编码的所在区域来判断汉字是否合法。

判断文本中汉字是否合法分为如下两步：

第一步，由用户指定要检查的文本中所包含的汉字和全角符号的类型，尤其是要确认是否包含繁体汉字（除“镕”、“珮”、“瞭”、“眈”等以外的繁体汉字）。

第二步，由图 2 所示算法来确定文本中汉字与全角符号的合法性。

通常，在找到乱码后，应该提交给用户修改，但是，如果找到的乱码是非法的控制字符，也可以用程序把乱码删除或更改成可显示字符（例如空格）。

用汉字的编码分区识别技术，可以发现以下几种错误：

1. 发现大部分乱码的“尾巴”，由于大部分乱码的形成是由于在文本行的行头或中间有干扰信息，使汉字错位显示，所以，只要这行汉字的末尾是汉字或全角符号，那么，在这行文本的末尾，即回车符的前一个字节，通常会和回车符的第一个字节（ASCII 是 0x0d 的字符）构成一个新“汉字”，这个“汉字”的编码是区码至少为 0xa1，位码是 0x0d，这是 GBK 字符集以外的编码，是乱码的“尾巴”，现在只要向行头方向找，就可以找到乱码的“头”（机器不能自动识别乱码的“头”，需要用户干预）。
2. 发现 ASCII 码小于 0x20 的非法控制字符。
3. 发现在 GBK 字符集以内但不在用户指定汉字区域范围内的汉字，如特殊排版符号等。
4. 发现 ASCII 码为 0x7f 或 0xff 的非法字符。

(二)、词语接续关系检查技术

汉字的编码分区识别技术有一些不能识别的明显的乱码，原因是有一部分乱码是由

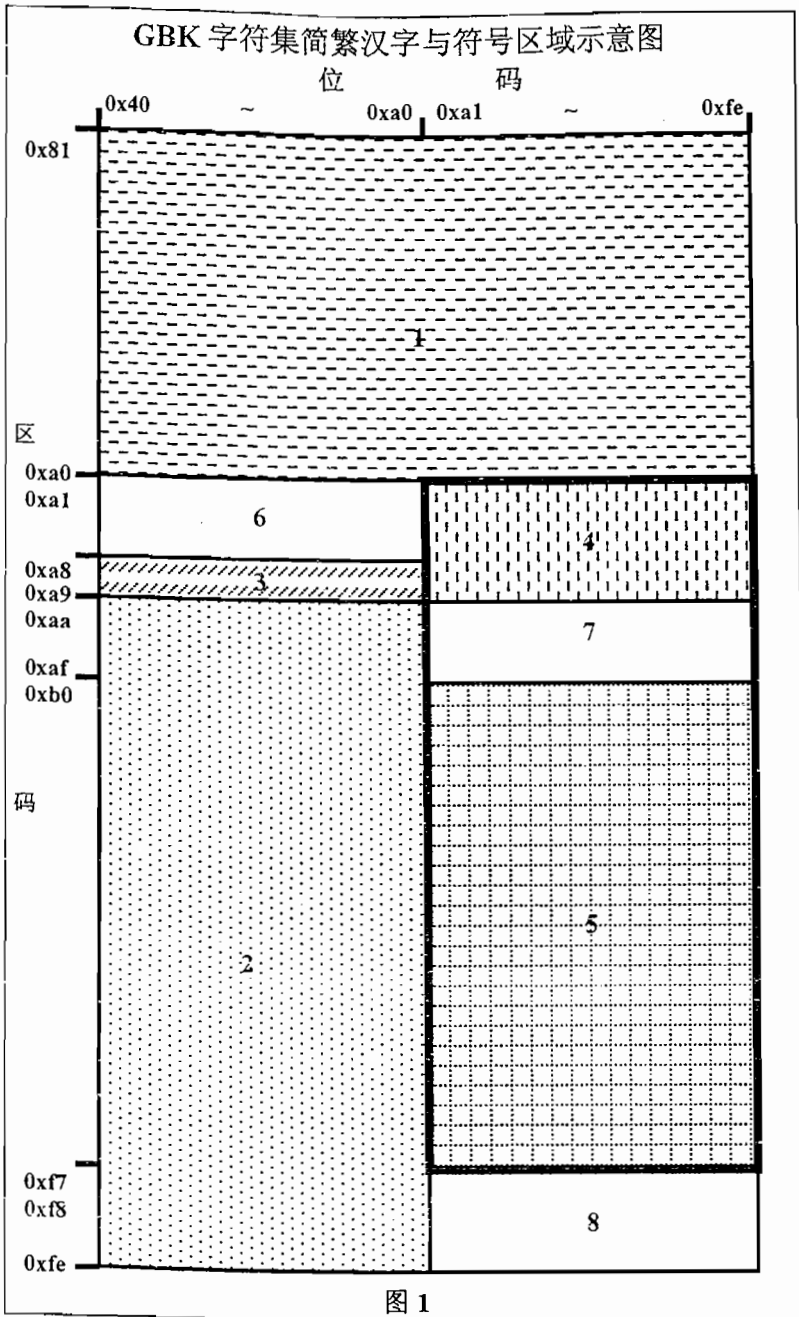


图 1 中区域注解:

1. 繁体汉字 I 区
区码: 0x81-0xa0
位码: 0x40-0xfe
 2. 繁体汉字 II 区
区码: 0xaa-0xfe
位码: 0x40-0xa0
 3. 全角符号 I 区
区码: 0xa8-0xa9
位码: 0x40-0xa0
 4. 全角符号 II 区
区码: 0xa1-0xaf
位码: 0xa1-0xfe
 5. 简体汉字区
区码: 0xb0-0xf7
位码: 0xa1-0xfe
 6. 主要空白区 I 区
区码: 0xa1-0xa7
位码: 0x40-0xa0
 7. 主要空白区 II 区
区码: 0xaa-0xaf
位码: 0xa1-0xfe
 8. 主要空白区 III 区
区码: 0xf8-0xfe
位码: 0xa1-0xfe
- (在图中用粗线框包围的区域是 GB 字符集的汉字及符号区域)
- 另: 图中的“4”(全角符号第 II 区)还可以分得很细, 比如全角英文区、全角日文区、全角俄文区等。

于一般的中文文本中的字符都在 GB 字符集范围内, 而 GB 字符集的编码范围是区码 0xa1~0xf7、位码 0xa1~0xfe, 可以说是个基本对称的矩阵, 所以如果汉字错位显示, 编码范围仍有可能在 GB 码范围内, 这时乱码没有被发现。

图 3 中, 第 3 区是第 2 区汉字错位存储与显示后汉字编码的区域, 且第 3 区的编码在 GB 编码的范围之外。第 1 区汉字的位码落在 0xa1~0xf7 之间, 区位错位之后, 仍是 GB 之内的汉字, 这部分汉字的字数占 GB 汉字总数的比例为

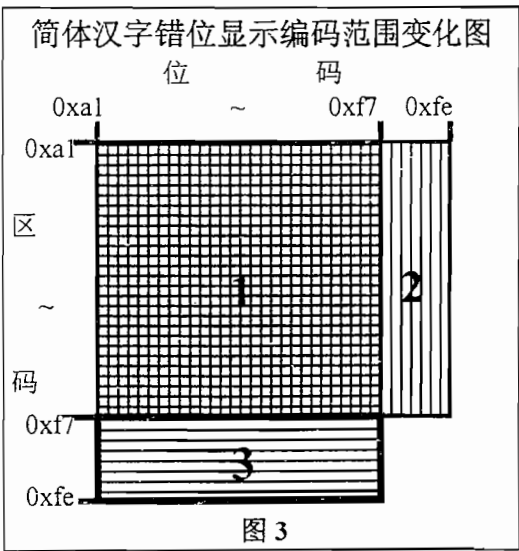
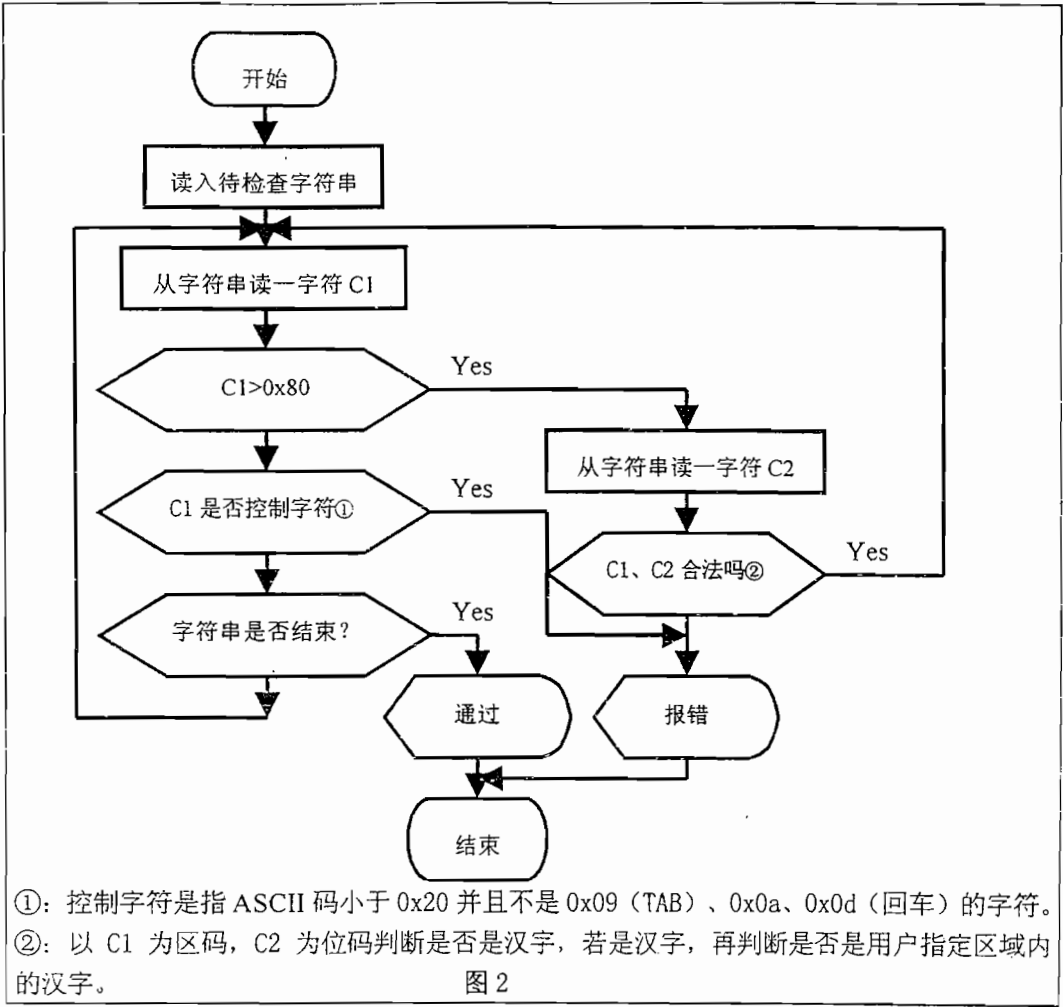


图 3 中区域注解:

1 U 2: 正常存储与显示的简体汉字编码区
 区码: 0xa1-0xf7
 位码: 0xa1-0xfe

1 U 3: 错位存储与显示的简体汉字编码区
 区码: 0xa1-0xfe
 位码: 0xa1-0xf7

2: 正常存储与显示的简体汉字特有区
 区码: 0xa1-0xf7
 位码: 0xf8-0xfe

3: 错位存储与显示的简体汉字特有区
 区码: 0xa1-0xfe
 位码: 0xa1-0xf7

$$(0xf7-0xa0) / (0xfe-0xa0) = 87/94 = 92.55\%$$

用程序统计一亿字左右的简体文本文件（被确定为基本无乱码的文本文件）中汉字的频率，其中，第1区的汉字占了所统计文本中总汉字数的90.98%。

以上数据说明，GB中的汉字错位后，绝大部分仍然是GB内的汉字，此时用汉字编码分区技术无法识别，只能寻求其它技术。

在汉语处理中常常需要分析和利用字间、词间、短语间、句间的相关关系。在汉语处理的最低层次即分词时，所能利用的只是字词间的相关关系。多个字词之间的相关关系显然比两个字词之间的相关关系更能反映汉语的语言规律，但所引起的组合爆炸使目前的微机难以承受，故一般只利用两个字词间且只考虑相邻两个字词间的相关关系。因为相邻关系有前后顺序，故称之为二元接续关系。利用二元接续关系解决分词中的歧义、专名识别、新词提取等问题已有许多成功的例子，校对、整句拼音输入、文字识别输入等应用系统的基本技术之一也是利用了二元接续关系。

我们把一个二元接续关系中的两个词语叫做一个词语接续对。利用二元词语接续关系的应用系统有一个词语接续对的数据库（简称接续库）。建立接续库的常规办法是对大规模语料库进行分词，然后收集其中的全部接续对，并统计同一接续对出现的次数，换算成接续强度，登记在库中。不在接续库中的字词对，或虽然在接续库中但接续强度太低的字词对，被认为是不接续的。一般来说，非语素字和部分非词语素与任何词语都不接续。实际运行时用接续库检查被处理对象中的词语接续强度如何，从而决定处理策略。例如：

①“吃”是及物动词，“苹果”是普通名词，在词语接续对统计过程中，“吃”和“苹果”有过多次连续出现，所以，在今后被检查文章中，“吃苹果”被认为是接续词语。

②在词语接续对统计中，“租赁”和“设备”多次连续出现，所以“租赁设备”被认为是接续词语，但若被检查文本串中少了“设”字而变成“租赁备”，接续检查程序就会因为“租赁”和“备”没有连续出现过而在相应位置报错。

汉字错位形成的乱码虽然绝大多数情况下还是合法的汉字，但往往字词间不接续。如果待查文本中有一串字词互相之间不接续，但错掉半个汉字后就能够接续了，那么原来那一串字词很可能就是乱码。在具体识别时，汉语的词语接续检查技术可以检查出大部分乱码的开头（即汉语词语开始不接续的地方）。如果漏掉，则可以被汉字的编码分区识别技术发现（但因不能确定乱码出现的原因，所以，这种乱码要交给用户处理）。

众所周知，汉语词语接续检查技术还不很成熟，有可能在文本中误报，也就是把接续的词语对判断为不接续；针对这种情况，我们用以下方法来判断报错的位置是否是乱码：

第一步，从报错的位置向后搜索直到找到回车符为止，统计这段文本的词语接续检查结果（一般来说，这段文本长度至少要四个汉字长，否则词语接续检查的结果就不可靠了。我们在后面讨论此问题），如果不接续词语达到一定比例（比如30%或50%以上），就可以比较肯定地认为这段文本是乱码；这里解释一下，从报错位置起只搜索到回车符为止是因为回车符是两个内码小于0x80的ASCII码，即使出现乱码，影响范围也只涉及回车符的第一个字节，由于第二个字节内码也小于0x80，所以在回车符之后的第一个字节如果大于0x80，那么，它就顺理成章成为汉字的区码，乱码的影响范围就只限于当前段的汉字。

第二步，如果第一步统计出的不接续词语的比例达到了标准，那么，取出这段文本，去掉头一个字节，再对这段文本进行词语接续检查，如果检查结果要大大好于原来结果甚

至没有不接续现象，就可以证明原文中存在乱码，把去掉一个字节的文本回填即可。

用这个算法，绝大部分乱码可以被发现并修复，但也有个别的情况，主要是乱码的长度过短的情况。例如位于行尾的字符串“（作者名）”中的右括号错成了“i”，由于乱码的长度只有一个汉字长，而且是合法的英文字母，故无论用汉语的词语接续检查技术或汉字的编码分区识别技术都无法识别。此外，还可能使一些正常的文本被误改。比如“林、牧、副、”这种“单字+顿号+单字+顿号+…”的模式中，单字是非词语素，通常与其前后都不接续，于是成为一个不接续的串。将这个字符串头一个字节去掉，原文本变成了“帧20.痢8.薄”。由于接续库统计结果中阿拉伯数字与任何词语几乎都能接续，机器就修改了原文本串，从而产生了误改的现象。为了解决这个问题，我们把它作为一种特殊模式特殊处理，再碰到像“单字+顿号+单字+顿号+…”模式的文本就不作接续检查了。

三、处理结果及其讨论

一、处理结果

我们对未经过处理的大规模语料进行了整理修复工作，语料情况如下：

- ①经济日报 1992 年语料（约 1820 万字）；
- ②经济日报 1994 年语料（约 2150 万字）；
- ③人民日报 1993 年语料（约 2340 万字）；
- ④人民日报 1994 年语料（约 2180 万字）；
- ⑤人民日报 1996 年语料（约 2510 万字）；
- ⑥人民日报 1997 年语料（约 2570 万字）；
- ⑦新华社 1994 年语料（约 693 万字）；
- ⑧新华社 1995 年语料（约 2500 万字）；
- ⑨新华社 1996 年语料（约 600 万字）；
- ⑩市场报 1994 年语料（约 844 万字）。

因为语料的数量较大，我们从其中随机抽取将近 1 亿字的语料进行试验。

疑错阈值设置为：被怀疑串长度不小于 8 个字节（4 个汉字）、不接续词长占被怀疑串长度 50%以上时尝试修复；修改后不接续词长占被怀疑串长度 30%以下时认为修复成功，否则不修改；由人工干涉修改。

程序运行完毕后，根据文本修复的记录文件显示：

- ①文本中有 ASCII 码小于 0x20 的非法字符和 ASCII 码为 0x7f 或 0xff 的字符约有 150,000 个，主要原因是有一千五百万字左右的文本残存着数据库文件头和字段分隔符。
- ②有超长不接续串 12000 多个（无非法控制字符和 ASCII 码为 0x7f 或 0xff 的字符）。去掉头一个字节，再检查，有 820 个串恢复正常接续状态，而被系统提示给用户有乱码存在。其中，760 多处是乱码，并已修复；50 多处不是真正的乱码，例如原文是“林、牧、副、”这种“单字+顿号+单字+顿号+…”的字符串被提示成“帧20.痢8.薄”等等，未修改。

造成上述 760 处乱码的原因主要有：

- a. 排版软件的特殊排版符号、回车符号不规范；360 多处。
- b. 没能被汉字的编码分区识别技术发现的数据库格式或超文本格式的数据库文件头、字段分隔符和多媒体链接指针等多余字符；380 多处，这些乱码没能被汉字编码分区识别技术发现的原因是它们的编码符合汉字编码的规范。

我们还搜集了一些在文件传输过程中（如拷贝、收发 E-MAIL）造成的乱码，在没有变动程序及疑错阈值设置情况下，《工智文本修复器》查出 15 处因意外丢失或被更改若干字节造成的乱码，无漏报。因有的字符已经丢失或被更改，所以修复后产生了丢字现象。

最后，为了检验该软件的性能，我们在上述语料中集中了近 50 万汉字的乱码较多的

文本文件，人工找出所有乱码，再用软件进行集中修复，共发现乱码 305 处，没有漏报。

二、乱码修复的讨论

1. 如果在《工智文本修复器》运行时，把被怀疑乱码串长度设置得很短，有何利弊。

首先，目前的汉语处理技术不能分辨所有的接续与不接续对，误报时有发生。其次，正常文本中，不接续词语的数量应该很小，很少会出现一串词语全不接续的情形。如果文本中出现了一串乱码，那么，就会出现一串不接续词语的情形，且不接续词语占乱码的绝大多数。所以，我们把被怀疑文本串中不接续词语长度是否占到一定比例作为判断乱码的依据。当文本串总长度较大时，这个比值就能比较真实地反映这段文本的词语接续情况，而在文本串总长度较小时，一个词语的接续或不接续就能极大地改变这个比值（一般来说，被怀疑串长设为 3 时系统所提出的被怀疑串数量几乎是串长设为 4 个汉字时数量的 10 倍），更因为目前的汉语处理技术还不完善，一个误报就可能使一段正常文本被误判为乱码。

所以，一般在文本修复时，我们规定被怀疑串长要大于 4 个汉字。

2. 当汉字丢失一个字节（半个汉字）时，能否用剩下的字节来提示被丢失的汉字。

理论上，一个汉字由两个字节描述，丢失一个字节后，可根据所剩一个字节缩小丢失汉字的搜索范围，用汉语词语接续检查技术“猜”出丢失字节可能是什么值。在实际操作中，情况复杂得多，首先，要确定下来乱码的种类，是丢、多字节还是错字节造成的，甚至还要判断乱码是否特殊排版符、是否数据库文件头，等等。就目前技术来看还有困难；其次，就算能确定乱码是丢字节造成的，像“的”、“其”、“和”等字的接续对非常多，有可能给出多个建议，使用户无所适从。如何提示所丢汉字，是我们今后的工作之一。

本论文得到以下项目支持：

- 1 国家 863 计划（重点项目）现代汉语通用分词系统 863-306-ZD03-04-2 1999.1-2000.12
- 2 国家自然科学基金 现代汉语通用分词系统研究 69882001 99.1-2001.12
- 3 北京市自然科学基金（重点项目）实用化短语级计算机辅助汉语校对系统 4971001 97.9-99.12
- 4 国家自然科学基金 汉语词语接续关系的柔性系统及其应用研究 69682001 97.1-99.12

参考文献

- [1]朱德熙，《语法讲义》，商务印书馆，1982
- [2]邱超捷，宋柔，欧阳龙根 大规模语料库中词语接续对的统计与分析
1997 年计算语言学全国联合学术年会，北京，1997.8
- [3]宋柔，邱超捷，欧阳龙根，二元接续关系及其在汉语分词和校对中的应用
1996 International Conference of Chinese Computing(ICCC'96).
- [4]宋柔，关于分词规范的探讨，
《语言文字的应用》，1997 年第 3 期
- [5]北京语言学院语言教学研究所，现代汉语频率辞典，北京语言学院出版社