汉字种子混淆集的构建方法研究

施恒利[1] 刘亮亮[1,3] 王 石[2] 符建辉[2] 张再跃[1] 曹存根[2]
(江苏科技大学计算机科学与工程学院 镇江 212003)[1]
(中国科学院计算技术研究所 北京 100190)[2] (中国科学院大学研究生院 北京 100049)[3]

摘 要 汉字混淆集是错别字识别中的重要资源之一。在本项研究中,首先手工整理了11935个汉字的可能的错别字,然后以这些汉字为节点、"可错成"关系为边,将混淆集构造成一个错别字混淆集图。由于人工总结错别字具有很大的局限性,因此在种子错别字混淆集图的基础上,设计了自扩展算法和开源外部补充算法来对错别字混淆集图进行扩展,以发现新的错别字对。根据实验,新发现了15133组错别字对。经过随机抽样校对,准确率达到87.35%。

关键词 错别字混淆集,自扩展,开源数据,基于规则和统计
中图法分类号 TP311 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2014.08.049

Research on Method of Constructing Chinese Character Confusion Set

SHI Heng-li[1] LIU Liang-liang[1,3] WANG Shi[2] FU Jian-hui[2] ZHANG Zai-yue[1] CAO Cun-gen[2]
(College of Computer Science, Jiangsu University of Science and Technology, Zhenjiang 212003, China)[1]
(Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China)[2]
(Graduate University of Chinese Academy of Sciences, Beijing 100049, China)[3]

Abstract The set of Chinese characters which is easily confused is one of the important sources during the process of identifying wrongly written characters. In the study, firstly we sorted out 11935 possibly-wrongly written characters by hand. Then taking those characters as nodes and "possibly-wrongly written characters" relations as sections, we constructed the set of wrongly written characters which is easily confused into a diagram. Due to the great limitation of manually sorting out wrongly written characters, on the basis of the diagram, we designed the internal-expanding algorithm that expands the set of wrongly written characters and the open source data external-supplementing algorithm that supplements the set of wrongly written characters through large quantity of corpus. In that way, we would expand the diagram and new pairs of wrongly written characters. According to the experiment, we newly found 15133 groups of wrongly written characters pairs. After proofreading samples at random, accuracy reaches 87.35%.

Keywords Wrongly written characters set, Self-expansion, Open source data, Rule and statistics base

1 引言

中文文本自动校对是自然语言处理的一个挑战性的研究课题,也是一个具有重大研究意义的课题,从20世纪90年代开始,国内才开展了对中文文本的自动校对研究。由于中文的输入显示在计算机上为一个个单字,不存在拼写的错误,对中文汉字的自动校对涉及到语法语义等问题,因此中文的自动校对工作面临很大的困难[1]。

随着中文校对工作的发展,汉字的错别字混淆集在错别字识别中越来越重要,对错别字混淆集方法的研究这项工作无论是对于理论研究还是实际应用,都有重要的意义,而就目前现有的文献来看,国内在文本自动校对方面所采用的方法中遇到的难题之一就是汉字混淆集的获取[2]。

目前中文文本校对系统中有关对种子汉字混淆集的获取方法主要有:文献[3]采用同音字词典和形近字词典,将发音相同的所有汉字和形近字组合起来,结合似然匹配算法得到汉字种子混淆集,但对生词和专用名词的识别能力较弱;文献[4]采用手工方法整理出字形、字音、字义和输入码相近字,但难点在于如何获取汉字种子混淆集,且手工工作量大;文献[5]采用模式匹配方法对长词进行纠错处理,但该算法没有充分利用错别字串的特征,算法计算量大。

本文首先对中文汉字(共11935个汉字,包括部分繁体字)在文本中可能出现的各种错误类型构建错别字混淆集,然后根据混淆集建立错别字混淆集图模型,进而根据错别字混淆集图自动对错别字混淆集进行补充和验证,最后通过大规模语料对混淆集进行进一步补充,从而获得一个汉字种子混

到稿日期:2013-10-03 返修日期:2013-12-27 本文受国家自然科学基金重点项目(91224006,61173063,61035004),国家自然科学基金面上项目(61203284)资助。

施恒利(1989—),硕士生,主要研究方向为智能信息处理;**刘亮亮**(1979—),男,博士生,讲师,主要研究方向为自然语言理解、知识工程与知识获取;**王 石**(1981—),男,博士,助理研究员,主要研究方向为知识工程与知识获取;**符建辉**(1983—),研究员,主要研究方向为知识工程;**张再跃**(1963—),教授,主要研究方向为计算机逻辑和知识工程;**曹存根**(1964—),研究员,主要研究方向为知识工程。

淆集字典。

2 种子错别字混淆集和错别字混淆集图的构造

2.1 错别字混淆集的人工创建

通过大批训练语料，我们整理出中文文本错别字出现的类型并对其进行分析，一般有以下几类类型：

①音似、形似、义似替换错误

例 1 错误句子：他表示【爱末能助】

正确句子：他表示爱莫能助

错误原因："末"和"莫"为音相似。

例 2 错误句子：清理已是【满目苍痍】的房屋

正确句子：清理已是满目疮痍的房屋

错误原因："苍"和"疮"为形相似。

例 3 错误句子：当库伯·雷恩在为剧本的初稿而【奋笔急书】的时候

正确句子：当库伯·雷恩在为剧本的初稿而奋笔疾书的时候

错误原因："急"和"疾"义相似。

②相邻键位击键错误、击键顺序颠倒、少击键错误

例 4 错误句子：原本万祥和唐门可以说是井水【比】犯河水

正确句子：原本万祥和唐门可以说是井水不犯河水

错误原因：相邻键位"u"和"i"敲击错误。

③拼音相同的词组误选

例 5 错误句子：现在每个大臣都【指导】是谁在最后做出决定

正确句子：现在每个大臣都知道是谁在最后做出决定

错误原因："指导"和"知道"拼音相同，被误用。

④其他情况。例如：方言、发音习惯（如南方方言中，z 和 zh 不分，h 和 f 不分）。

整理得到种子错别字混淆集，建立混淆集字典。

错别字混淆集的人工构建步骤如下：

Step1 得到形似字和音似字。通过汉字的拼音表和汉字的笔顺相似度和点阵相似度获取[6]。

Step2 设计一个键位冲突函数，用于类型②中的错误。

Step3 收集常见的易混淆字词，根据词典获取相同拼音的词组。

Step4 人工添加及校对。

部分结果如表 1 所列。

表 1 部分实验结果

正确字	错别字混淆集
矮	唉挨倭诿哎艾爱哀埃
堵	赌俺掩淹阇埯睹奄唵庵
整	麈鏊敖熬螯熬遨傲赘敫整
摔	婢啤脾俾牌碑卑裨蜱郫埤棰陴
爆	瀑暴曝
傍	旁谤帮邦榜磅
帮	帮

通过手工构建错别字混淆集，我们共发现 48865 组错别字对，平均每个汉字有 4.9 个错别字。

2.2 错别字混淆集图的构造

根据错别字混淆集来创建错别字图，用 $G=(V,E)$ 表示，其中 G 为错别字图，V 是图 G 中顶点的集合，E 是图 G 中顶点之间边的集合。错别字图中，以每个汉字作为顶点，从顶点 v_i 指向顶点 v_j 的有向边表示 v_i 是 v_j 的错别字，用 $\langle v_i, v_j\rangle$ 来表示，v_i 为弧尾，v_j 为弧头。

例：错别字字典中，汉字"桉"的混淆集为{按，案，安}，则表示为〈按，桉〉、〈案，桉〉、〈安，桉〉。

通过对创建的错别字图进行分析，定义一些统计量及名称来刻画错别字图，以便更加清楚地反映错别字混淆集的内部联系。

定义 1（错别字入度） 错别字图中，以该汉字为弧头，通过有向边指向该汉字的弧的条数，用 $indegree(V)$ 表示。

例：汉字"桉"为弧头的边有〈按，桉〉、〈案，桉〉、〈安，桉〉，故"桉"的错别字入度为 3。

定义 2（错别字出度） 错别字图中，以该汉字为弧尾，通过有向边指向其他汉字的弧的条数，用 $outdegree(V)$ 表示。

例：汉字"噻"为弧尾的边有〈噻，馕〉，故"噻"的错别字出度为 1。

定义 3（双向错别字） 顶点 v_i 和 v_j，如果存在边 $e_1\langle v_i, v_j\rangle$ && 存在边 $e_2\langle v_j, v_i\rangle$，则称 v_i、v_j 为双向错别字。

例：汉字"末"和"未"，在错别字图中，存在边〈末，未〉和〈未，末〉，故"末"和"未"为双向错别字。

定义 4（单向错别字） 顶点 v_i 和 v_j，如果存在边 $e_1\langle v_i, v_j\rangle$，不存在边 $e_2\langle v_j, v_i\rangle$，则称 v_i 为 v_j 的单向错别字。

例：汉字"那"和"娜"，在错别字图中，存在边〈娜，那〉，不存在边〈那，娜〉，故"娜"为"那"的单向错别字。

图 1 为部分错别字图的展示。

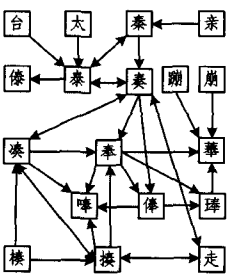

图 1 部分错别字图

图 1 中共有 18 个顶点、32 条边，其中以"奏"为弧头的弧数为 4 条，以"奏"为弧尾的弧数为 5 条，故"奏"的错别字入度为 4，错别字出度为 5，"奏"和"泰"为双向错别字，"奏"是"俸"的单向错别字。

3 错别字混淆集的自扩展

3.1 相关概念及其表示方法

定义 5（汉字权重 λ_v） 汉字 v 在所有汉字中占的权重：

$$\lambda_v=\frac{nF(v)}{\sum_{i=1}^{n}F(a)}$$

其中，$F(v)$ 为汉字 v 的字频，通过大规模训练语料获取，n 为种子汉字的总数，$\sum_{i=1}^{n}F(a)$ 为所有种子汉字的字频总和。

例：汉字"安"，通过大规模训练预料获得其字频为 1.48×108，$n=11935$，$\sum_{i=1}^{n}F(a)=2.417\times10^{11}$，故"安"在汉字中的权重为 7.3。

定义 6（常见字） 汉字 v，如果有 $\lambda_v\times\frac{indegree(v)+outdegree(v)}{2}\geqslant\chi$（$\chi$ 为设定的阈值，经验值为

2.97，$indegree(v)$为汉字 v 的错别字入度，$outdegree(v)$为汉字 v 的出度），则 v 为常见字。

例：通过错别字图，得到汉字“安”的错别字入度和错别字出度分别为 5、24，故通过定义 6 可以验证得出“安”为常见字。

定义 7（生僻字） 汉字 v，如果有 $\lambda_v \times \frac{indegree(v)+outdegree(v)}{2} \leqslant \delta$（$\delta$ 为设定的阈值，经验值为 0.1，$indegree(v)$为汉字 V 的错别字入度，$outdegree(v)$为汉字 v 的出度），则 v 为生僻字。

例：通过错别字图，得到汉字“魆”的错别字入度和错别字出度分别为 7、0，通过定义 5 得到“魆”的汉字权重为 0.004，故有 $0.004 \times \frac{7+0}{2}=0.014<0.1$，故“魆”为生僻字。

定义 8（形似度 α） 根据字的形相似定义以及算法得到的汉字形似度对应表，设定阈值 β（经验值 0.680），当 $\alpha>\beta$ 时，认为这两个汉字形似。

例：根据形相似表，汉字“据”的形相似字和对应的形似度有“倨(0.900)”、“椐(0.900)”、“琚(0.900)”、“锯(0.809)”、“居(0.794)”、“腒(0.782)”、“裾(0.757)”、“踞(0.746)”、“踽(0.709)”。

定义 9（非同音字） 如果两个汉字读音不相同，则它们为非同音字。

定义 10（非同音双向字） 两个汉字读音不相同，且互为双向错别字，则为非同音双向字。

通过对大量语料及错别字图进行分析，我们发现：

一个生僻字，往往很难错写成一个常见字。例如：汉字“硪”很难写成“我”。

3 个汉字，它们互为双向错别字，如果其中任意的两个汉字错写成其他汉字，则另外一个汉字也很可能错写成该汉字。例如：错别字图中有汉字“瑞”、“惴”、“揣”互为双向错别字，且有边〈瑞、端〉、〈惴、端〉，则“揣”错写成“端”的概率很大。

根据发现，我们整理出以下规则：

规则 1（常不错罕规则） 对于一个常见字，它的错别字很难写成一个生僻字。对于汉字 v_i 和 v_j，如果汉字 v_i 的权重 $\lambda_{v_i} \gg$ 汉字 v_j 的权重 λ_{v_j} && 存在边 $e_1\langle v_i, v_j\rangle$，则排除该组错别字对，即从错别字图中删除边 $e_1\langle v_i, v_j\rangle$。

规则 2（同错规则） 设 v_i、v_j、v_k 互为双向错别字，若错别字图中有边 $e_1\langle v_i, v\rangle$、$e_2\langle v_j, v\rangle$，则向错别字图添加边 $e_3\langle v_k, v\rangle$。

3.2 自扩展算法

基于错别字混淆集图内部添加边及验证的算法步骤如下：

Step1 遍历图的每个顶点，获取每个顶点的汉字权重 λ_v。

Step2 获得每个顶点的双向错别字，根据“同错规则”添加边。

Step3 得到每个顶点被指向的边，根据“常不错罕规则”进行验证，若符合规则，则从错别字混淆集图中删除这条边。

Step4 根据汉字的拼音，得到同音字集合，设汉字 v_i 的同音字集合为 $S=\{v_1, v_2, v_3, \cdots v_n\}$，对同音字集合构成的部分错别字混淆集图进行分析：

若同音字集合中 v_i，v_j 形相似，则 v_i，v_j 为双向错别字。

若同音字集合中 v_i，v_j 不是形相似，设 v_i，v_j 共同的错别字集为 $S_1=\{a_1, b_1, \cdots c_1\}$，$v_i$，$v_j$ 同时作为一些汉字集 $S_2=\{a_2, b_2, \cdots c_2\}$ 的错别字，若 $\frac{|(S_1 \cup S_2) \cap S|}{|S|} \geqslant \lambda$（$\lambda$ 为设定的阈值，经验值为 0.25），则 v_i，v_j 为双向字。

Step5 根据汉字的结构、偏旁和笔顺相似度，得到每个汉字的形相似字，并定义形相似度。

Step6 遍历图的每个顶点 v_i，获取 v_i 的非同音相似字 v_j，利用公式得到 v_i 和 v_j 的相似度。公式如下：

$$S_{v_i}(v_j)=\sum_{i=1}^{4} w_i E_i$$

其中，w_i 为权重系数，E_i 为相似度，若 $S_{v_i}(v_j)>\lambda$（λ 为设定的阈值，设为 1.5），则 v_i 和 v_j 为双向错别字。

Step7 遍历错别字图的每个顶点，若两顶点 v_i 和 v_j 的共同错别字集的个数大于 6，则 v_i 和 v_j 为双向错别字。

Step8 遍历错别字图的每个顶点，根据“同错规则”，若顶点 v_i，v_j，v_k 互为双向错别字，且错别字图存在边 $e_1\langle v_i, v\rangle$、$e_2\langle v_j, v\rangle$，不存在边$\langle v_k, v\rangle$，则向错别字图添加边 $e_3\langle v_k, v\rangle$。

Step9 根据规则 1 常不错罕规则，对已添加混淆集进行验证。

以下对相似度公式作两点说明：

1. 权重系数 w_i 的选取。字形似度、入度、出度、字频 4 种因素对总的相似度影响的大小是不一样的，一般情况下为字形似度>入度=出度>字频。因此，通常对 w_i 的选取依照原则：$w_1>w_2=w_3>w_4$，例如，可取 $w_1=1$，$w_2=w_3=0.6$，$w_4=0.3$。

2. 相似度 E_i 的选取：

E_1：与 v_j 的形似度。

E_2：$\frac{indegree(v_j)}{indegree(v_i)}$。如果 $indegree(v_i)$为 0，则取 $E_2=0$。

E_3：$\frac{outdegree(v_j)}{outdegree(v_i)}$。如果 $outdegree(v_i)$为 0，则取 $E_3=0$。

E_4：v_j 和 v_i 的汉字权重之比 $\frac{\lambda_{v_j}}{\lambda_{v_i}}$。

例：汉字“慎”和“填”，相似度为 0.900。“慎”的入度为 5，出度为 9，汉字权重为 0.9268，“填”的入度为 5，出度为 14，汉字权重为 2.3959。则有：$E_1=0.900$，$E_2=1$，$E_3=0.643$，$E_4=0.3868$，则“慎”和“填”的相似度 $S_{慎}(填)=2.002>1.5$，故为非同音双向字。

我们将在第 5 节给出上述算法的实验数据。

4 大数据补充混淆集

4.1 中文分词

本文从某百科中提取 8G 的语料作为训练语料，系统采用中科院计算所开发设计的基于 N 元模型和角色标注的 ICTCLAS 分词系统对语料进行分词处理，如果语料的词含有错别字，则该词在分词时会被分成散串。分词词典也是中文自动分词过程中的一个重要组成部分，本系统选取中科院计算所提供的词典（约 27W 词）作为分词词典。由于分词时查询词典的速度直接关系到分词系统的全局速度，因此，建立一种高效而快速的词典机制，对中文分词而言有很重要的意义。本系统采用双数组 Trie 树（Double-Array Trie）实现对字典的快速查询，用于词条的快速检索[7,8]。

4.2 散串合并与模糊匹配

相似词模糊匹配的目标是在字典中查找与相似词相似的词组，通常在一般的文本中，连续多字词不会出现词层面错误，如果有错误，则一般是上下文相关错误[9,10]。系统按照以下的规则对分词后的文档进行散串合并：

相邻的两个串为连续多字词，不进行合并。

合并词串中间不包含标点、特殊字符。

合并词串最长长度为 9，最短长度为 3。

例：一个优秀作品是冰动三尺，非一日之寒的。

首先，对句子进行分词，词按空格进行分隔，分词结果为：一个 优秀 作品 是 冰 动 三尺，非 一日 之 寒 的 。

根据规则，“一个”、“优秀”、“作品”为连续多字词组合，不进行合并，对“，”前进行合并，得到【冰动三尺】<——>【冰冻三尺】，对“，”后继续模糊匹配，无匹配的相似词。

最终我们选取合并词与模糊匹配串在头尾处汉字不同长度大于 3 的词组，与中间位置不同的词组，其中头尾汉字不同的有 1.8 万组，中间汉字不同有 6.5 万组。

4.3 规则与统计相结合验证

当前自动校对软件没有给出具体的召回率(recall ratio)、准确率(accurate ratio)，据有关资料统计，这些系统的召回率一般在 70%左右，准确率 30%[11]。本文所指的召回率和准确率的公式如下：

$$召回率=\frac{获取文本真正错误的个数}{文本中实际错误的个数}$$

$$准确率=\frac{获取文本真正错误的个数}{获取的文本中错误的个数}$$

4.3.1 运用统计对结果验证

通过对 8.6G 训练语料进行统计，得到合并词和模糊匹配词的词频，根据模糊匹配结果，用 N_i 表示合并词在训练语料中出现的频次，用 N 表示该合并词的模糊匹配词在训练语料中出现的频次。

定义如下公式：

频次比率：$P_i=N/N_i$

表 2 统计词频的实验结果

P_i(≤)	获取的真正正确个数	获取的错误个数	准确率/%
25	577	2839	20.32
15	403	2479	16.26
10	308	2226	13.84
8	272	2096	12.98
6	229	1968	11.64
5	207	1893	10.94
4	174	1785	9.75

通过表 2 我们发现，准确率随着 P_i 的降低而降低。在 $P_i\leqslant 4$ 的基础上，通过降低合并词的词频 N_i 来进一步研究，实验结果见表 3。

表 3 改变 N_i 得到的实验结果

N_i(≤)	获取的真正正确个数	获取的错误个数	准确率/%
20	152	1509	10.07
12	134	1373	9.762
11	129	1342	9.613
10	121	1320	9.167
9	118	1298	9.09
8	113	1265	8.933
7	110	1232	8.923

通过表 2 我们发现，当 P_i 一定时，准备率随着 N_i 的降低而降低。当 $P_i\leqslant 4$ && $N_i\leqslant 8$ 时，该部分结果集准确率低于 9%，选取 $P_i\leqslant 4$ && $N_i\leqslant 8$ 作为阈值，对该部分结果进行排除。

4.3.2 运用规则进行验证

N 元语法模型(N-Gram Model)是目前最常见的语言模型，所谓 N-Gram，指的是由 N 个词组成的串，基于 N-Gram 建立的语言模型，称为“N 元语法模型(N-Gram Model)”，N 元语法模型(N-Gram Model)中尤其以二元语法(Bigram)和三元语法(Trigram)模型应用最为广泛[12,13]。

N 元语法模型反映的是语言的局部规律，描述给定词序列在语言中的出现概率分布，如果模型构造合理，训练语料足够大，这个局部规律将比较可靠。为了使所建模型符合文本查错时的情况，我们以字、词为统计单元，应用 ICTCLAS 对 8.6G 百科语料进行加工、标记，由此构建字、词二元模型。用 $Freq1$ 表示合并词中的错字或错词与语境搭配的频次，$Freq2$ 表示合并词中错字与相邻字或词搭配的频次。

定义如下公式：

频次比率：$P_f=Freq2/Freq1$

部分结果如表 4 所列。

表 4 上下文语境实验结果

匹配词	合并词	左边语境	右边语境	Freq1	Freq2	P_f
资源管理	姿源管理	人力		1	63	63
中小企业	种小企业	三		67390	1	1.484E-05
独一无二	都一无二	的		25361	10	0.0004
中小企业	中小企也		可在	4737	52	0.011
职业教育	职业教与		学	4346	4220	0.971
服装设计	服装设挤		理念	1	1	1

做如下判断：

判断 1 如果匹配词中 $P_f\geqslant\lambda_1$(λ_1 为设定的阈值，经验值为 0.0208)，$Freq1\leqslant\lambda_2$(λ_2 为设定的阈值，经验值为 15000)，则认为该匹配词在句子中是正确的。

判断 2 对不属于上述情况的词组，进行进一步的验证处理，按照以下规则进行判断：如果匹配词通过分词后得到一个散串和一个词语，则认为该匹配词在句子中是错误的。

5 实验结果及分析

5.1 实验结果

根据自扩展算法、开源外部补充算法，我们新发现了 15133 组错别字对。验证实验选取数据开源获取的 7872 条模糊匹配结果集作为测试集，测试集由人工标记，准确率为 61.128%，对实验文本使用规则验证判断，选取频次比 $P_f\geqslant 0.0208$ && $Freq1\leqslant 14099$ 作为阈值，召回率为77.839%，准确率为 75.185%，根据判断 2，对剩余部分使用判断 2 验证，该部分召回率为 85.8%，准确率为 57.8%，获取的结果集中，包含的种子汉字的个数为 4228，多为常见字，证明通过数据开源补充混淆集是切实可行的。

实验选取通过统计规则排除的部分正确错别字组作为测试集，共 245 组错别字对，经过验证，错别字混淆集包含了 214 组，未包含的错别字对为 31 组，准确率达到 87.35%。

5.2 错别字混淆集的排序方法

通过自动扩展、数据开源外部添加方法获取的混淆集并非都是有效的，根据混淆集的字频和形相似度得到优先值，并

(下转第 253 页)

会价值[11]。FCM算法是聚类分析中应用最为广泛的一种聚类算法，但由于对类样本容量缺乏考虑，因此不适用于不平衡数据集。EFCM算法在目标函数中引入样本容量因素，使得算法对样本容量差异性保持鲁棒，利用PSO算法对模糊隶属度进行估计，避免了EFCM算法梯度法参数估计不能保证迭代过程收敛的弊病，理论分析及实验测试验证了所提算法的有效性。

参考文献

[1] Zhao F, Jiao L C, L H Q. Kernel generalized fuzzy c-means clustering with spatial information for image segmentation[J]. Digital Signal Processing: A Review Journal, 2013, 23(1): 184-199

[2] Kannan S R, Ramathilagam S, Chung P C. Effective fuzzy c-means clustering algorithms for data clustering problems[J]. Expert Systems with Applications, 2012, 39(7): 6292-6300

[3] 孙晓鹏，纪燕杰，李翠芳，等. 三维网格模型增量式聚类检索[J]. 计算机科学，2011，38(11)：248-251

[4] Bezdek J C, Hathaway R J, Sobin M, et al. Convergence and theory for fuzzy c-means clustering: counterexamples and repairs [J]. IEEE Transactions on Systems, Man and Cybernetics, 1987, 17(5): 873-877

[5] Zhang Z H, Liu S, Ji C P. An improved fuzzy C-means based on IPSO[J]. International Review on Computers and Software, 2012, 7(1): 241-245

[6] Chen D H, Liu Z J, Wang Z H. A novel fuzzy clustering algorithm based on kernel method and particle swarm optimization [J]. Journal of Convergence Information Technology, 2012, 7(3): 299-307

[7] Huang M, Xia Z X, Wang H B. The range of the value for the fuzzifier of the fuzzy c-means algorithm[J]. Pattern Recognition Letters, 2012, 33(16): 2280-2284

[8] Wu K L. Analysis of parameter selections for fuzzy c-means [J]. Pattern Recognition, 2012, 45(1): 407-415

[9] 王熙照，崔芳芳，鲁淑霞. 密度加权近似支持向量机[J]. 计算机科学，2012，39(1)：182-184

[10] 文传军，詹永照，柯佳. 广义均衡模糊C均值聚类算法[J]. 系统工程理论与实践，2012，32(12)：2751-2755

[11] Yu Y, Zhang B B, Chen L. An improved fuzzy C-means cluster algorithm for radar data association[J]. International Journal of Advancements in Computing Technology, 2012, 4(20): 181-189

(上接第 232 页)

根据优先值对混淆集进行排序，可提高文本校对的效率[14]。汉字的字频信息在一定程度上反映了被写错的概率，一般情况下，高频字被错写成低频字的概率比低频字错写成高频字的概率要大；形相似度越高，写错的概率越大。设$\{v_1, v_2 \cdots v_m\}$为得到的错别字混淆集合，用 N_i 表示混淆集合$\{v_1, v_2, \cdots, v_m\}$中 $v_i(1 \leqslant i \leqslant m)$与正确汉字的形相似度，按以下步骤进行排序：

如果混淆集中某一候选串 v_i 中 $N_i > \lambda$(λ 为设定的阈值，经验值为 0.68)，根据 N_i 的大小进行排序。对剩余的混淆集集合，根据字频的大小进行排序。

5.3 实验结果分析

本实验系统中，实验结果的好坏与一些因素有关，如：

(1)分词的准确度

由于本实验通过大规模语料获取混淆集是在分词的基础上进行的，因此分词的准确度对实验结果有影响，但是到目前为止还没有一种分词方法可以达到100%的分词准确度。另外，由于中文用法的繁杂，语料选取的领域广泛，故在分词过程中会碰到词典的未登录词，这对分词及最终结果有影响。

(2)召回率和正确率的权衡

在本实验系统中所选取的阈值是一个经验值。它取值的高低直接影响到召回率和正确率。阈值选取过高会导致召回率下降，过低则会导致正确率下降，因此这存在一个权衡的过程。

(3)内部规则的不完备性

完整的规则能对错别字混淆集进行很大程度的补充和完善，由于目前对错别字混淆集补充还没有一个统一的规则，本课题所罗列的规则难免有一定的局限性，还应该继续深入考虑挖掘，使其更加完备。

结束语 本文对种子汉字的错别字混淆集进行了规则自补充、数据开源外部补充，得到一个种子汉字的错别字字典，同时，大规模语料获取的错别词词典不受领域的限制，应用的领域广泛，对文本校对有很大的帮助。今后将从内部规则和语法分析入手，进行更深入的研究，以提高系统的完善性。

参考文献

[1] 刘亮亮，王石，王东升，等. 领域问答系统中的文本错误自动发现方法[J]. 中文信息学报，2013，3：77

[2] 张磊，周明，黄昌宁，等. 中文文本自动校对[J]. 语言文字应用，2001(1)：19

[3] 陈笑蓉，秦进，汪维家，等. 中文文本校对技术的研究与实现[J]. 计算机科学，2003，11(16)：53

[4] Zhang Zhao-huang. A Pilot Study on Automatic Chinese Spelling Error Correction[J]. Communication of COLIPS, 1994, 4(2): 143

[5] 于勐，姚天顺. 一种混合的中文文本校对方法[J]. 中文信息学报，1998，12(2)：31

[6] 丰强泽，曹存根. 语音查询中的辨音方法：中国，CN1514387[P]. 2004-07-21

[7] 戴耿毅，余静涛. 基于双数组 Trie 树算法的字典改进和实现[J]. 软件导刊，2012，11(7)：17

[8] 李慧，杨炳儒，潘丽芳，等. 一种基于双数组 Trie 的 B2B 规则串提取方法[J]. 计算机科学，2013，40(5)：206

[9] 王静帆，邬晓钧，夏云庆，等. 中文信息检索系统的模糊匹配算法的研究和实现[J]. 中文信息学报，2007，21(006)：59

[10] 张仰森，曹元大，俞士汶. 基于规则与统计相结合的中文文本自动查错模型与算法[J]. 中文信息学报，2006，20(4)：1

[11] 张仰森，丁冰青. 中文文本自动校对技术现状及展望[J]. 中文信息学报，1998，12(3)：50

[12] 王贤明，胡智文，谷琼. 一种基于随机 n-Grams 的文本相似度计算方法[J]. 情报学报，2013，32(7)：716

[13] 吴春颖，王士同. 基于二元语法的 N-最大概率中文粗分模型[J]. 计算机应用，2007(12)：2902

[14] 张仰森. 中文校对系统中纠错知识库的构造及纠错建议的产生算法[J]. 中文信息学报，2000，15(5)：33