中 文 信 息 学 报
第 12 卷 第 3 期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol. 12 No. 3

中文文本自动校对技术现状及展望①

张仰森　丁冰青

山西大学计算机科学系

摘要　本文概述了中文文本自动校对技术的产生背景,分析了预校对文本常见的错误类型及文本自动校对(自动查错和确认纠错)的难点,探讨了当前商品化的文本校对软件的校对策略和发展趋势。

关键词　中文文本自动校对　自动查错　确认纠错

一、引　言

“信息爆炸”,“数字化生存”是当今社会发展的总趋势,电子书、电子报纸、电子邮件、办公文件等文本电子出版物不断涌现,如何保证这些文本的正确性,显得越来越重要。中文文本自动校对系统(包括文本的自动查错和确认纠错)的研究已成为一项亟待解决的紧迫课题。

二、文本自动校对技术的产生背景

随着汉字编码输入技术理论研究和应用开发的不断成熟,中文文字处理系统日益走向实用化、商品化,计算机作为一种写作、编辑和排版的工具越来越频繁地出现在机关办分室、编辑部和出版印刷行业,而使用计算机进行文字录入编辑,不可避免地会出现一些文字错误,例如丢字、多字、别字、英文单词拼写错误、不规范标点等。校对(proofread)工作是出版前审核把关的重要环节,而目前大多采用人工校对的方法,校对工作单调,劳动强度大,效率低,人工校对越来越成为印刷出版自动化的瓶颈。为改变传统的人工校对模式,实现录入文稿的自动校对,提高校对质量,把校对员从大量枯燥细致的工作中解放出来,提出这样一个课题:开发实用化的自动文字校对软件。

汉字识别(Chinese Character Recognition)和语音识别(Phonetics Recognition)[3][4][5][6][7]可以使汉字的输入实现自动高速,是具有远大发展前景的输入技术[2]。近年来,利用 OCR (Optical Character Recognize)技术和语音识别技术解决汉字的计算机自动录入问题已取得了很大的成就。基于统计识别和结构识别相结合等方法,国内推出了多套印刷体汉字识别、联机手写体汉字识别及脱机手写体汉字识别系统,如清华文通 TH - OCR 高性能中英文印刷文本自动识别输入系统;北京中自汉王笔手写体识别系统等。基于模式匹配法、隐马尔可夫模型法和人工神经网络法,推出了多套语音识别系统,如中科院声学所研制的实时语音识别系统等。汉字识别和语音识别的正确率是汉字识别最重要的指标之一,目前上述汉字识别系统的正确识别率可达 85 %～95 %左右,这样所得到的文件质量与出版要求相距甚远,令人难以满意。

① 本文 1997 年 9 月 18 日收到

这主要是由于汉字的类别数量极为巨大,汉字字形变化剧烈,汉字识别的实际文本图像随机噪声和干扰(如文字模糊,笔划粘连,断笔,黑白不均,纸张噪声,油墨反透,字形大小,书写用笔,笔尖粗细,写字质量等)严重,语音识别受说话环境干扰,语调、语速等影响,易造成汉字拒识、误识,降低了汉字识别的正确率。作为识别后处理的中文文本自动校对系统(利用实际汉字文本的相关信息,对识别结果代码文件进一步加工,以提高系统识别率,降低误识率)便成了实用的汉字识别系统不可缺少的一个重要环节。

三、预校对文本常见的错误类型分析

1. 按出错来源分,预校对文本错误可分为录入错误、识别错误和原稿错误三类。

(1) 录入错误:指键盘录入文稿时所产生的错误。录入错误主要是由于输入过程中的疏忽造成的,且与输入法有很大的关系,如五笔字型输入法造成的错字,其形相似,而拼音输入法造成的错字,其音相同或相似等。

①错别字:如击键错、重码或者联想输入中误选,可形成一些错别字。(以五笔字型输入为例。《中国科技期刊研究》97.1;《编辑之友》96.4)

Ⅰ. 击键先后次序颠倒;

放(yt) →入(ty)

Ⅱ. 少击键位错;

齿(hwb) →具(hw)

Ⅲ. 手型不规范 d、f、g、h 弄错;

居(nd) →导(nf)

Ⅳ. 键位分析错误或键位击错;

垂(tgaf) →和平共处(tgat)

Ⅴ. 重码字选错;

fcu →去云支

Ⅵ. 词组误操作。

老奸巨猾(fva) →地区(fba)

②丢字、多字、英文单词拼写错误、串行等。如录入时跳过一个或几个字,或多余删字操作造成漏字;而多余的击键操作动作,同一字重复,联想输入可能造成多字错误。

【例 1】原句为:文稿中仍会遗留许多错误。

录入为:文稿中仍会遗留许多误。 (丢字)

【例 2】原句为:维护建筑市场秩序。

录入为:维护建筑市场场秩序。 (多字)

【例 3】原句为:95 年 4 月推出新版联想 office 办公软件。

录入为:95 年 4 月推出新版联想 office 办公软件。 (拼写错误)

③同音或同形字带来的错误;

【例 4】原句为:我们这些华夏子孙。

录入为:我们这些化夏子孙。 (同音)

【例 5】原句为:油价很可能突破 10 元每桶大关。

录入为:油价很可能实破 10 元每桶大关。 (同形)

④标点、数字错误或括号不匹配;

【例 6】原句为:1992 年 9 月 1 日,我上了大学。

录入为:19921 年 9 月 1 日,我上了大学。　　(数字错误)

【例 7】原句为:“我是中国人民的儿子,我深爱着我的祖国和人民。”

录入为:我是中国人民的儿子,我深爱着我的祖国和人民。”　　(括号不匹配)

(2) 识别错误(包括拒识、误识两种情况)

在语音识别,光字符汉字识别(OCR)系统中,由于识别系统识别的字数有限,对些生僻字会拒识;由于形近、音近汉字较多,还可能产生误识错误;又由于噪音影响,按信噪模型

W →噪音信道 →Y

识别结果 Y 可能不同于待识别文本 W,如:

【例 8】原句为:“校雠学”鼻祖刘向

识别为:“校 X 学”鼻祖刘向　　(拒识)

【例 9】原句为:我觉得可以从两个方面进行考察。

识别为:我觉得可以从两个方面速行老察。　　(误识)

(3) 原稿错误:文稿形成过程中由于作者疏忽而形成的错误。如写错别字,搭配不当,结构残缺等。

2. 按预校对文本句子语法、语义分,一个字或词与其上下文环境不相适应会造成语法、语义搭配不当,表现为以下几种错误:

(1) 构词错误:错字、缺字或多字破坏了原文词的结构,出现了所谓的“非词现象”。

【例 10】原句为:漫长的等待过后 ……

预校文本:漫长的等侍过后 ……

(2) 句法错误:虽不违背构词法,但可能破坏句子整体结构,造成语法错误。

①词性搭配错误:

【例 11】原句为:只有一条记录是当前记录。

预校文本:只有一第记录是当前记录。

②关联词语搭配错误:(如原稿出错)

【例 12】原句为:因为你用功学习,所以你成绩好。

预校文本:虽然你用功学习,所以你成绩好。

③句型错误:如句子成分残缺,成分位置不当,造成搭配错误和语法错误。

【例 13】原句为:联合国安理会决议

预校文本:联合同安理会决议

【例 14】原句为:使用吸尘器后,大大延长了电机的寿命。

预校文本:使用吸尘器后,大大处长了电机的寿命。

(3) 语义错误:语法正确,但语意离奇。

【例 15】原句为:本人不慎将脸盆丢失

预校文本:本人不慎将脸丢失

四、中文文本查错、纠错的难点分析

“校雠学”鼻祖刘向在《别录》中说:“校雠,人读书,校其上下,得谬误,为校;一个持本,一人读书,若怨家相对为雠。”意思是说校对有两部分工作,第一是发现原稿中可能存在的差错 ——自动查错;其二,提出自己的修改意见,帮助编辑或作者校正 ——确认纠错。

利用计算机进行中文文本自动校对的基础是自然语言理解,自然语言理解特别是计算机处理汉语的障碍和困难也就成了文本校对中自动查错(automatic checking)和错认纠错(confirmative correction)的瓶颈。计算机处理汉语的障碍主要表现有[21]:

(一)中文文本的侦错与纠错较西文校对困难

1. 汉语和西文本身的差异

大多西文是表音文字,其书面语言词与词之间有空格,而汉语是表意文字,词与词之间无空格,且西文多有形态变化(时、数、量)等而汉语缺少形态变化。因此,词的切分问题成了计算机处理汉语的首要问题,且计算机对汉语的处理无法利用形态,只能在语法、语义上找出路。汉语和西文本身的差异决定了中文文本校对较西文校对困难得多。

2. 工作方式差异

大多英文编辑器如 MS - WORD 都带有一个拼写检查程序(Spelling Checker),帮助用户检查英文文本中出现的拼写错误。这种程序依靠一部在线的英文词典和一些词法规则,采用字符串匹配的工作方式完成拼写错误的检查。而汉语是表意文字,这种工作方式对汉语并不适用,因为不论用何种输入方式(如键盘输入,汉字识别输入或语音识别输入等),显示在计算机屏幕上的每个汉字都必须是汉字编码字符集中的一个单字,它可能是别字或冗余字,但不会是缺一点少一捺的错字。因些要实现汉语的侦错与纠错,就要对文本进行分词,如果一个字在某些词(语)中的出现是不合理的,机器就判定是别字,而判定不合理的实现要涉及到语法、语义等问题,是很复杂的。

(二)中文文本自动校对技术上的难点

1. 分词处理的难点

汉语自动分词是中文信息处理的基础课题之一,也是中文文本自动校对的首要问题之一。歧义切分、生词处理等问题是分词的重点和难点,分词处理的好坏直接影响到文本校对时对文本进行的语法、语义分析的质量和查错率及校正率[23]。

(1)歧义切分(Ambiguity)[17]:

通过句式搭配把词安排好以后,有一些句子还可以作两种以上解释,称为歧义现象(包括交集型字段,包字型歧义字段)。如:"援助的是中国"有两种解释:"援助国是中国"或"受援国是中国"。如何结合上下文处理歧义现象,是分词的难点。

(2)生词、新词及专用名词(中文姓名、英译姓名、地名等)处理[12][18][19]:

这部分词占有比例小,但如果不加处理就会导致为数可观的分词错,使得语法、语义分析受影响,查错和纠错不准确。如:"德索托她帮忙办件事",人物是"德索"还是"德索托"?

2. 语法(Grammar)、语义(Semantic)分析的难点

线图、文本规则、语义树、格语法等语法分析理论尚不太完善,其中词性的兼类("非常序")、省略("空位")[25]及不规范语法处理等问题成为语法、语义分析的难点。

3. 纠错的难点

纠错是校对的另一个重要环节,即要根据查错的结果,并结合上下文知识给出有效的修改建议,而计算机自动校对不同于人工校对,由于缺乏历史背景知识及对特定语言环境的理解等启发性知识造成纠错建议的有效性远远不够甚至不着边际。

总之,汉语词与词之间没有明显的标志,汉语词类缺乏形式标记;汉语词类与句法成分之间不存在某种简单的对应关系,这给汉语词语、语法、语义分析带来了很多困难。但实际文本中,由于大多数单字和它相邻的文字受词、句法、语义的约束,因而是相关的,利用这些相关性进行查错和纠错成为一条行之有效的途径。

五、自动中文文本校对软件的校对策略及发展现状

由于文本自动校对理论和技术尚不太成熟,该领域的论述还不多见,就目前现有的与中文自动校对相关的文献来看,由于词的切分问题尚未得到很好的解决,一般只能进行比较简单的处理。当前自动校对软件的校对策略一般如下[20]:

①机器自动查错和人工确认纠错相结合;

②词法、语法、语义多层次查错、纠错策略;

③规则方法和语料库统计方法相结合;

④面向查错的"粗分析"方法。

它们的校对技术深度一般有以下三个层次[3]:

①简单上下文匹配:在查出的错误字词前后一定范围内匹配,用词库和后补信息或文本特征来判别;

②词切分上下文匹配:对文本中的句子自动切分,用词库和后补字信息来纠正错误字;

③自然语言理解上下文匹配:用词、语法、语义等知识,逐句对文本进行分析、理解,由此选择正确的代替字。

例如,方正金山校对软件主要是在现代汉语语法规则制导下,利用模糊分词及多遍扫描技术,将一个句子分为若干词或词组,对照系统提供的通用词库、专业词库及用户自定义词库进行校对。清华大学利用汉字二元同现概率的大规模汉语语料文本的统计分析,互信息、同现信息等结果进行分词、词性标注、词性排歧,用语料库知识指导文本校对;基于语料库统计的MARKOV语言模型,利用上下文相关信息指导文本校对。杭州大学基于N联字的汉字识别后处理研究,以及北京师范大学基于句法结构和特征信息的词法、句法语义校对系统的实现[30]等都取得了很大的进展,对文本自动校对极具指导意义和创造性启示[10][11][15][23][24]。

当前,有不少优秀的校对软件如"黑马文字校对"、"方正金山中文校对"、"三欧"、"文捷"、"人工智能校对通"以及台湾的"啄木鸟"等系统已走向市场,实现了商品化,可实现对语句不通、丢字多字、打字错误、错别字、重句、英文单词拼写错误、不规范标点、年月日错误、数字及章节错误的自动校对。校对速度可达50万字/小时。但以上系统基本上是基于"词组"校对的辅助校对系统,存在以下不足之处[8]:

1. 虽然系统能够查出许多字、词错误来,但表现出同样错误症状的一些字、词错误却不能被发现。如:"报纸杂志以挥舆论监督功能"一句不能发现错误。

2. 校对能力基于词法层次上,对涉及句法、语义的许多错误无法发现。如:"他们不能新闻记者有关材料"一句不能发现错误。

3. 误判率较高。如:"有记者共约300人"误判"共约"为"公约"。

4. 对许多错字串未给出建议信息,即使给出也离正确的字、词相差甚远。

5. 没有提出校对软件查错的召回率(recall ratio)、查准率(accurate ratio)和误判率(error correction ratio)指标。但据有关资料统计,这些系统的召回率一般在70%左右,查准率一般约为2.5%-30%。显然,召回率还比较高,但查准率却比较低。本文中所指的召回率、查准率、误判率之计算公式如下:

$$\text{召回率(recall ratio)} = \frac{\text{查出预校正文本真正错误的个数}}{\text{预校正文本中实际错误的个数}}$$

$$\text{查准率(accurate ratio)} = \frac{\text{查出预校正文本真正错误的个数}}{\text{查出的预校正文本中错误的个数}}$$

$$\text{误判率(error correction ratio)} = \frac{\text{预校正文本中正确的词判错的个数}}{\text{查出的预校正文本中错误的个数}}$$

我们在中文文本自动校对领域进行了初步的探讨,在文本校对理论研究和技术实现上进行了有益的尝试。基于句子中词与词的义类组合规律(《同义词词林》上海辞书出版社),我们希望在语义平面上对中文文本自动校对理论的研究有所突破。为此,我们建立了《词林》大类词典及其索引,生成了一个小型的义项三元同现矩阵,提出了基于语义组合规律的校对算法,初步实现了一个基于语义的中文校对系统(SBPS),基中对生词、新词、专用名词等进行了专站处理,并在校对中采用了"并列类推","同境类推"等侦错技术和策略。在封闭语料下进行测试,本系统召回率为72.5%,查准率为35.2%。由于该系统的测试是在封闭语料下进行的,在开放环境下还存在着一些技术和策略上的问题,有待于进一步改进和完善。

六、自动中文文本校对软件的发展前景

随着自然语言处理特别是计算语言学理论的发展及这些理论在自动文本校对软件开发实践中的应用,自动校对软件性能(包括查错能力,纠错建议有效性,在保持较高召回率的基础上进一步提高查准率,降低误判率等)将进一步提高。笔者认为,今后自动文本校对软件的研究与开发,应着力于以下几个方面:

1. 加强校对软件对歧义字段、生词、新词以及专用名词(中文姓名、英译姓名、地名等)的辨识能力。

召回率、查准率、误判率是自动校对软件的三个重要的性能指标。在词法分析时,对歧义字段、生词、新词以及专用名词的辨识错误将直接影响到语法、语义分析,最终导致查准率下降,误判率上升;而且歧义现象和生词以及专用名词在现代汉语中占相当大的比重,因此为了提高自动校对软件的性能,加强歧义字段、生词和专用名词的辨识能力,不仅是一条行之有效的途径,而且是必须解决的基本问题。

国家语委冯志伟教授提出了潜在歧义论(Potentional Ambiguity)[17],利用PT结构(词组类型结构)实例化的方法,插入词汇单元以消除歧义的方法,可以应用在校对排歧上。另外,通过建立姓氏名字用字频率表、称谓表、指界动词,如"说、指出、认为"等信息,考虑姓名在句中出现的位置及分布情况,可提高生词的辨识率。校对前对生词进行预分析并加以人工确认,也是一条行之有效的途径。

2. 进一步利用词的上下文匹配和基本句法语义的上下文匹配,提高校对系统的召回率和查准率,降低误判率,提高纠错建议的有效性,进行词法、语法、语义三级校对。

人工校对和自动校对有很大差异。人工校对是基于一定背景知识和语感的启发性校对;而自动校对是基于规则和词典信息的机械校对,自动校对还不能真正达到人工校对的认知智能水平,因此,在校对时增加自动校对系统的背景知识和上下文关联信息,显得尤为重要。清华大学在基于语料库统计的研究方面做了大量工作,提出了字字同现概率矩阵理论,这对根据上下文进行预测校对是有力的理论支持。

自动校对是基于错误文本的词法、语法、语义分析过程,要尽可能多地给校对系统提供更多的启发性信息以弥补缺乏背景知识造成的不足。如增加校对时人机交互功能,提高校对系统对人工修改及提示信息的分析和自学习能力。充分利用上下文进行并列类推和同境类推,例如,根据一些关联词和标点符号"既…又…","边…边…"(表并列),"是…还是…","或者…或者…"(表选择),"不仅…而且…"(表递进)(并列类推),"有理想,有道德,有文化,有纪律"(同境类推)等信息推断相关成分的合理性及进行相邻搭配合理性检查,进行词法、语法、语义三级校对,是一种值得尝试的方法。

3. 采用规则和语料库统计相结合的方法,并根据具体的应用目标,采用针对性策略,是今后自动校对软件的发展方向。

计算语言学的研究表明:大规模语料库(Corpus)能够反映一定的语言规律,用统计的方法从中提取有用的信息并结合一定的规则处理自然语言,是当前计算语言学发展的主要趋势,笔者认为:基于规则和语料库统计的方法也是自动校对软件开发的主流方法。

4. 获取预校对文本的输入方式,体裁范围等背景信息,采取针对性的校对策略,不但能提高校对召回率、查准率、纠错有效性,而且能够缩小查错范围,提高校对速度。如:建立五笔字型输入易错库,汉字识别易错库,语音识别易错库,校对常见错误库,专业词分类库,用户自定义词库等,以增强校对系统的自学习能力和查错、纠错有效性的能力。

参 考 文 献

[1]张炘中、沈兰生 “印刷体汉字识别技术在我国的发展和应用” 中文信息学报 Vo1.6 No.1 P49 - 53

[2]赵明 “手写印刷体汉字识别方法综述” 计算机研究与发展 93.4 P59 - 64

[3]郭繁夏、丁晓青 “中文 OCR 的发展现状及其最新技术” 电子与电脑 96.3 P28 - 30

[4]许嘉璐 “中文信息处理技术现状及相关语言文字研究” 软件世界 96.4 P13 - 20

[5]黄昌宁 “中文信息处理中的分词问题” 语言文字应用 97.1

[6]李京葵、周明、黄昌宁 “统计与规则相结合的汉语句法分析研究” 计算语言学进展与应用 1995

[7]冯志伟 “论歧义结构的潜在性” 中文信息学报 Vo1.9 No.4

[8]张俊盛、陈舜德、郑萦 “多语料库作法之中文姓名辨识” 中文信息学报 Vo1.6 No.3 P7 - 12

[9]孙茂松、黄昌宁、高海燕、方捷 “中文姓名的自动辨识” 中文信息学报 Vo1.9 No.2

[10]吴蔚天、罗建林 “汉语计算语言学 —汉语形式和形式分析” 电子工业出版社 1994

[11]慕勇、孙才、罗振声 “汉语文本自动查错与确认纠错系统的研究” 计算语言学研究与进展 1995。

[12]夏莹、常新功、马少平、朱小燕、金奕江 “利用上下文相关信息的汉字文本识别” 中文信息学报 Vo1.10 No.1 P23 - 30

[13]苗兰芳、张森、周昌乐 “基于 N 联字的汉字识别后处理研究” 中文信息学报 Vo1.8 No.2. P39 - 46

[14]文韬 “校对人员的好帮手 —方正金山中文校对系统 VCorrect 试用报告” 中国计算机用户 96.10. P23 - 25

[15]易蓉湘、何克抗 “计算机汉语文稿校对系统” 计算机研究与发展 97.5. Vo1 34. No.5.

Present Condition and Prospect of Chinese Text Automatic Proofread Technology

Zhang Yanysen Ding Bingqing

Dept.of Computer Science Shanxi University

Abstract This paper gives a general statement of emerging background of Chinese text automatic proofread technology, analyses common error type of pre - proofread text and the difficulties of Chinese text automatic proofread (inculding automatic checking and confirmative correction), and discusses proofread strategies and developing trend of commercialized text proofread software today

Keywords Chinese text automatic proofread Automatic checking Confirmative correction