文本自动校对技术研究综述*

张仰森[1,2]，俞士汶[1]

(1. 北京大学 计算语言学研究所，北京 100871；2. 北京信息科技大学 计算机及自动化系，北京 100085)

摘 要：概述了文本自动校对技术的产生背景，分析了中西文文本的各自特点以及它们之间的异同，对中西文文本校对的技术难点和解决方法以及国内外的研究现状进行了回顾和评述，探讨了文本校对技术未来的发展方向及需要解决的问题。

关键词：文本自动校对；孤立词校对策略；上下文相关的校对策略；语言模型

中图法分类号：TP311.52 **文献标识码**：A **文章编号**：1001-3695(2006)06-0008-05

Summary of Text Automatic Proofreading Technology

ZHANG Yang-sen[1,2], YU Shi-wen[1]

(1. *Institute of Computational Linguistics, Peking University, Beijing* 100871, *China*; 2. *Dept. of Computer & Automatization, Beijing Information Science & Technology University, Beijing* 100085, *China*)

Abstract: The emerging background for text automatic proofreading technology is generally stated in this article. Each characteristic of the text of Chinese and Western as well as the similarities and differences between them have been analyzed, the review and comment for domestic and international research present status as well as the technical difficulty and solving methods of the text proofreading of Chinese and Western are given, and the future developing directions and the problems need to be solved of text automatic proofreading technology are discussed.

Key words: Text Automatic Proofreading; Strategy for Isolate-word Error Correction; Strategy for Context-sensitive Error Correction; Language Model

1 引言

文本自动校对是自然语言处理的主要应用领域之一。早在20世纪60年代，国外就开展了英文文本的自动校对研究[1]；IBM Thomas J. Watson研究中心首先于1960年在IBM/360和IBM/370用UNIX实现了一个TYPO英文拼写检查器；1971年，斯坦福大学的Ralph Gorin在DEC-10机上实现了一个英文拼写检查程序Spell[11]。多年来，随着计算机技术的不断发展，新的输入技术不断涌现，如OCR识别、语音识别。开展拼写错误校对的研究更加迫切，这方面的研究也在不断取得进展[10~14]，部分成果已经商品化，目前流行的一些文字处理软件(如Word, Wordpefect等)也都嵌入了英文拼写检查功能。国际互联网上还能见到Expert Ease公司推出的Deal Proof, Newton公司推出的Proofread等英文单词拼写检查系统。

国内在中文文本校对方面的研究始于20世纪90年代初期，但发展速度较快。目前有许多科技公司和高等院校或研究机构都投入了一定的人力和财力开展这方面的研究[15~27]，并取得了一些较好的成果，且有部分成果已经商品化，如黑马校对系统、金山校对系统、工智校对通等。本文就文本自动校对技术的国内外发展状况进行了研究。

收稿日期：2005-08-24；修返日期：2005-10-09

基金项目：国家“973”计划资助项目(2004CB318102)；国家“863”计划资助项目(2001AA114210, 2002AA117010)；中国博士后科学基金项目(2005038026)

2 文本中的常见错误类型分析

2.1 常见的文字录入技术

目前，常见的文字录入技术和方法主要有键盘录入、语音识别、OCR识别、手写识别。其中由于键盘录入和OCR识别速度快、准确率高，成为文字录入的主要手段。对于中文来说，由于字符集太大，人们研究了许多种输入法间接将汉字送入计算机。目前比较流行且影响较大的输入法有五笔字型输入法、微软拼音输入法、智能狂拼输入法、智能ABC等。OCR识别也是一种常用的输入技术，但这种输入技术主要用于书写比较工整的手写稿或印刷稿输入，速度极快，目前字迹清晰的印刷稿的识别正确率已在98%以上，手写体的识别正确率还比较低，识别后处理或校对的任务也比较繁重。

除了原稿中的错误外，电子文本中的错误主要来自输入过程。尤其对中文来说，文中的错误还和所使用的输入法密切相关，因此，有必要对这些输入法产生的错误及形式进行分析。

2.2 键盘录入导致的中英文文本错误分析

在应用键盘录入英文字符时，常见的错误有以下几种：非词错误、真词错误和句法语义错误。非词错误是指文本中那些被词边界分隔出的字符串，根本就不是词典中的词。如下面的输入错误：them→tehm, the→thr, partition→patition, study→studdy等就是非词错误，造成这种错误的原因是由于指法错误或粗心造成的，这些错误可以概括为替换错误、易位错误、丢失错误和插入错误等。真词错误是由于输入人员的粗心或指法错

误所形成的字符串, 虽不是想要的单词, 但却是在词典中能够查到的真正的单词。如在输入 from 时由于发生了易位错误, 使 from 变成了 form, 而 form 是词典中的词; 若在输入 employer 时由于字符 r 和 e 相邻, 很可能将 r 输成 e 就得到 employee, 得到的字符串是词典中的单词, 但词义相反。真词错误往往会导致所输入的词与上下文搭配不当, 不是当前语境中所需要的词, 如 "I come form Beijing "中的 "form "应为 "from "。句法语义错误往往是由于真词错误造成的, 或由于原稿本身存在语法错误, 或输入时丢失了某个单词甚至串行或丢失一整行。通常人们将 "非词错误 "称为单词错误, 而将 "真词错误 "称为上下文相关的文本错误。

在应用键盘录入汉字时, 由于汉字数量远远大于键盘上键的数量, 所以必须采用编码输入法。常用的编码输入法有五笔字型输入法和拼音输入法(包括全拼、双拼、智能 ABC、智能狂拼等) 。与英文不同, 汉语输入不会发生非字错误, 能输入到计算机中的字必在汉字库中, 因此, 汉语文本中只会出现由于替换、易位、丢失、插入而导致的上下文相关错误或句法语义错误。使用五笔字型法输入文字时产生的错字往往与原字形相似, 或者它们的编码相近, 如由于手型不规范将 d, f, g, h 弄错, 导致将 "居(nd) "输成 "导(nf) "; 而使用各种拼音法产生的错误, 其音相同或相似, 如 "计算机用户 "输成 "计算机拥护 "。

2.3　OCR 识别导致的中英文本错误分析

在应用 OCR 技术输入文字时, 常见的错误主要有拒识和误识两种情况。由于识别系统识别的字数有限, 对一些生僻字会拒识, 如 "校雠学 "被识别为 "校 X 学 "。而对于那些形近或形似的英文字符或汉字则容易产生误识, 如英文字母 "D "被识别为 "O ", 字母 "l "被识别为数字 "1 ", "已经 "被识别为 "己经 ", "孔子曰 "被识别为 "孔子日 "等。

除了输入过程中造成的错误以外, 还有一种错误就是在文稿形成过程中由于写作人员的疏忽和大意造成的原稿错误, 如写错别字、搭配不当、结构残缺和标点符号错误等。 "像片 "被作者写为 "象片 ", "为人类做出贡献 "被写为 "为人类作出贡献 "等。

3　英文文本中的错误发现与纠错方法

3.1　单词错误的发现与纠错

英文文本中单词错误的检测发现方法目前主要有两种, 即 N-gram 分析法和查词典法。一般情况下, N-gram 错误检测技术对输入串中的每一个 *n* 元串(*n* 一般取 2 或 3) 在事先编辑好的一个 N-gram 表中进行查找, 看它是否在表中存在或它的出现频次, 那些不存在或出现频次非常低的 *n* 元串被认为是可能的拼写错误, 如 "shj "或 "het "就是错误的三元串。N-gram 分析法通常需要一个词典或大规模的文本语料以便事先编辑 N-gram 表。查词典法主要是检查所输入的 *n* 元串是否在词典或可接受的词表中, 如果不在词典中, 则将该输入串标志为一个拼写错误的词。由于基于查词典法的校对系统查错精度高, 因此, 是目前较为流行的错误检测技术。考虑到存取速度, 当词典规模较大时, 为了提高查错速度, 有效的词典查找算法也是人们研究的重点。

单词错误的纠错方法已经有很多研究, 主要有误拼词典法[10]、词形距离法[10]、最小编辑距离法[1]、相似键法[10]、骨架键法[10]、N-gram 法[12,13]、基于规则的技术[1]、词典及神经网络技术[1]。

(1) 误拼字典法。收集大规模真实文本中拼写出错的英文单词并给出相应的正确拼写, 建造一个无歧义的误拼字典。在进行英文单词拼写检查时, 查找误拼字典, 如命中, 则说明该单词拼写有误, 该词的正确拼写字段为纠错建议。该方法的特点是侦错和纠错一体化, 效率高。但英文拼写错误具有随机性, 很难保证误拼字典的无歧义性和全面性, 因此查准率低、校对效果差。

(2) 词形距离法。这是一种基于最大相似度和最小串间距离的英文校对法。其核心思想是构造单词的似然性函数, 如该单词在词典中, 则单词拼写正确; 否则, 按照似然性函数, 在词典中找到一个与误拼单词最相似的词作为纠错候选词。该方法的特点是节省存储空间, 能反映一定的常见拼写错误统计规律, 是一种模糊校对法。

(3) 最小编辑距离法。通过计算误拼字符串与词典中某个词间的最小编辑距离来确定纠错候选词。所谓最小编辑距离是指将一个词串转换为另一个词串所需的最少的编辑操作次数(编辑操作是指插入、删除、易位和替换等) 。还有人提出了反向最小编辑距离法, 这种方法首先对每个可能的单个错误进行交换排列, 生成一个候选集, 然后, 通过查词典看哪些是有效的单词, 并将这些有效的单词作为误拼串的纠错建议。

(4) 相似键法。相似键技术是将每个字符串与一个键相对应, 使那些拼写相似的字符串具有相同或相似的键, 当计算出某个误拼字符串的键值之后, 它将给出一个指针, 指向所有与该误拼字符串相似的单词, 并将它们作为给误拼字符串的纠错建议。

(5) 骨架键法。通过构建骨架键词典, 在英文单词出现错误时, 先抽取出该错误单词的骨架键, 然后再去查骨架键词典, 将词典中与该单词具有相同骨架键的正确单词作为该单词的纠错建议。

(6) N-gram 法。基于 *n* 元文法, 通过对大规模英文文本的统计得到单词与单词间的转移概率矩阵。当检测到某英文单词不在词典中时, 查转移概率矩阵, 取转移概率大于某给定阈值的单词为纠错建议。

(7) 基于规则的技术。利用规则的形式将通常的拼写错误模式进行表示, 这些规则可用来将拼写错误变换为有效的单词。对于一个误拼字符串, 应用所有合适的规则从词典中找到一些与之对应的单词作为结果, 并对每个结果根据事先赋予生成它的规则的概率估计计算一个数值, 根据这个数值对所有候选结果排序。

3.2　上下文相关错误的纠错方法

上下文相关的文本错误即真词错误, 其校对要比单词拼写错误校对困难得多[1]。上下文相关的拼写校对不仅要修正那些 "经典 "的拼写错误类型, 比如同音词错误(如 peace 与 piece) 和字母排序错误(如 form 与 from) , 而且还要修正那些常见的语法错误(如 among 与 between) 和词边界混淆的错误(如 maybe 与 may be) 。因为真词错误的出错字符串是词典中的正确词, 所以针对单词拼写错误的校对方法在这里不一定适用,

要对这类错误进行校对,必须使用上下文信息来判定哪些词在文本中出现是不合理的,这些词可能就是潜在的错误。上下文相关错误的校对较之单词误拼的校对要困难得多,它与自然语言理解的研究紧密相连。受自然语言理解技术进展的影响,文本错误的校对技术目前还没有大的突破。现有的基于上下文的文本错误校对方法有三类:①利用文本的特征,如字形特征、词性特征或上下文特征[2,7];②利用概率统计特性进行上下文接续关系的分析[13,14,23];③利用规则或语言学知识[24],如语法规则、词搭配规则等。

(1) 利用文本上下文的同现与搭配特征

可以将文本的校对过程描述为词排歧过程。若称待校对的词为目标词,则建立混淆集 $C=\{W_1,\cdots,W_n\}$,其中的每个词 W_i 均与文本中的目标词容易发生混淆或歧义。如假设 $C=\{\text{from, form}\}$,如果在文本中出现 from 或 form 时,就将它看作是一个 from 与 form 之间的歧义,校对的任务就是根据上下文决定哪个词是我们想要的词。上下文相关的校对问题由语句和语句中要被校正的词构成,Bayesian 方法和基于 Winnow 的方法都是将这样的问题表示成有效特征表,每一个有效特征表示目标词的上下文中有一个特殊的语言学模式存在。目前常使用的特征有两种类型:上下文的词和词的搭配。上下文词特征用来检查在目标词周围的 $\pm k$ 个词的范围内是否有特殊词存在;词搭配则用来检测在目标词的周围 l 个相邻词和/或词性标注的状态。如假设目标词的混淆集为{weather, whether},若置 $k=10, l=2$,目标词的可用特征包括:

①目标词前后 10 个词范围内的 cloudy;

②当前词后为 to + 动词。

特征①就预示着当前词应为 weather;而②则用来检查词搭配,它表明当前词后紧接着一个“to + 动词”的结构,表明当前词应取 whether(如 I don't know whether to laugh or cry)。在这种方法中,主要要解决的问题包括混淆集的求取;目标词所在上下文中特征的表示,即如何将语句的初始文本表示转换为有效特征。

基于词语同现与搭配特征的校对方法有很多种,较好的有 Bayesian 方法[1]和基于 Winnow 方法[2]。各种 N-gram 模型,如长距离 N-gram、触发对 N-gram 等模型,都可以利用目标词上下文中的词同现特征或搭配特征,采用最大似然估计法、互信息、相关度等方法检测文本中的错误,并通过相邻词间的转移概率确定纠错候选词,实现对目标词的校正。

(2) 利用规则或语言学知识

这种技术利用语言学家的语言学知识或句法语义规则去纠正文本中出现的错误。在基于语言学知识或规则的技术中,随着分析过程的进展,系统将依据句法、语义和篇章结构知识,建立一个它希望在下一个位置看到的词的列表,如果输入字符串的下一字符不在所期望的字符列表中,则系统就认为检测到了一个错误,并从其期望词表中选择一个词作为对其进行修正的候选词。

4 中文文本中的错误类型与校对技术

4.1 中文与西文的差别及处理难点

大多数西文都是表音文字,而汉语是表意文字,它们之间有着很多的不同:①文本结构不同。英语文本中词与词之间有空格,而汉语文本无空格。②词结构不同。英语的词有形态变化(时、数、量),而汉语缺少形态变化且汉语词类与句法成分之间不存在某种简单的对应关系。③字符进入计算机的方式不同。英文单词进入计算机是按字母一个个地录入,而中文字符进入计算机只能借助汉字编码。这种输入过程不可能产生拼写错误,即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,绝不会是缺一点少一捺的错字。因此,中文输入不会产生“非字错误”,只能产生别字错误,这些错误往往与要输入的字或词音同、音近或形近。④字符集规模的差异。英文的字符集是 26 个字母加标点符号,而汉语字符集则是一个包含了超过 6 763 个汉字符的大字符集,这将导致在应用语言模型时参数计算的极大困难。

正是由于汉语和西文的差异,导致汉语文本的处理要比西文文本复杂得多。由于汉语没有“非字错误”,因此,其校对只能是基于上下文的相关性来实现。汉语处理中的主要难点,如文本的切分、标注的歧义处理以及未登录词的识别等,也会反映到中文文本自动校对技术的研究当中,直接影响着中文文本校对时所进行的语法、语义分析的质量,进而影响召回率与查准率。

4.2 中文校对技术的现状

国内在文本自动校对方面的研究主要是针对汉语文本开展的。因为中文文本校对主要面向的是含有错误的文本,因此,汉语自然语言理解的研究也就成了计算机中文文本自动校对的基础。由于汉语与英语本质上的不同,在对中文文本进行查错/纠错分析时,必须要基于自然语言的理解技术,通过研究上下文间的依存关系才能实现,这显然是比较复杂和困难的,某些适于英文单词校对的技术和方法对汉语文本并不太适用。目前,国内有不少单位开展了中文文本校对理论和技术的研究,除了微软亚洲研究院、IBM 中国研究中心、哈尔滨工业大学、清华大学、东北大学、北京师范大学、北京工业大学、山西大学等科研院所外,一些有实力的高新技术公司,如北京黑马电子新技术公司、北大方正公司、金山公司等都开展了中文文本校对软件的研究与开发。

4.2.1 自动查错的研究状况

就目前现有的与中文校对相关的文献来看,国内在自动文本查错方面主要采用三种方法:①利用文本上下文的字、词和词性等局部语言特征,包括词性特征、同现特征或相互依存特征[2,7],甚至包括字形特征等;②利用转移概率对相邻词间的接续关系进行分析[13,14];③利用规则或语言学知识[24],如语法规则、词搭配规则等。其实,这些方法之间没有严格的界限,甚至一般是混合使用的。

(1) 基于上下文的局部语言特征

微软中国研究院设计实现了一个基于多特征的中文自动校对方法,它综合考虑了汉语文本中字、词和词性的局部语言特征以及长距离的语言特征,并采用 Winnow 方法进行特征学习,利用这些上下文特征对目标词混淆集中的词进行选择。其主要难点是如何将目标语句转换为多元有效特征以及混淆集的获取[7]。哈尔滨工业大学将对被校对的句子中的每个字词寻找其可能的候选,构成句子的字词候选矩阵,在此基础上,利

用语言本身所具有的结构特征与统计特征, 从候选矩阵中选出句子的最佳字词候选序列, 将其与原句对照, 找出错误的字词, 并以第一候选加以改正。语言结构特征的获取则应用 t 元规则对字词候选矩阵中的字词进行捆绑与剪枝, 形成语言结构元素, 并将其构成元素格子图, 然后借助文本统计特征, 应用 Markov 模型从语言结构元素格子图中寻找一条最佳的元素路径, 即为从候选矩阵中寻找的待校对语句的最佳句子。该方法的关键是候选矩阵构造以及语言结构特征的获取, 由于候选矩阵中只选择了同音字, 因而, 目前仅适于校对拼音输入法形成的文本。其主要难点在于特征的统一表示与格子图中的有效候选路径的求取[18]。

(2) 基于规则

北京师范大学利用校正文法规则对文稿进行校对[24], 若句子满足校正文法规则, 则根据规则把相应字词标记错误, 但有限的规则很难覆盖大量难以预料的错误现象, 查错能力有限。哈尔滨工业大学则以小句为单位, 对汉语句子进行三遍扫描, 通过自动分词、自动识别生词、用短语规则将单字词散串合成短语, 逐步把正确的字符串捆扎起来[20], 将不能捆绑的剩余单字符串判定为错误。其不足之处是有限的短语捆扎规则难以覆盖大量的语言现象, 短语的捆扎缺乏定量的判断依据, 查错算法只能查出单字(串) 错误, 不能查出多字词的替换错误, 比如 "用户社会主义制度 "这样的错误就无法查出。吴岩等人[17] 还提出了一种词匹配和语法分析相结合的校对方法。采用规则与统计相结合的方法, 不使用大规模语料库, 通过逆向最大匹配和局部语料统计算法发现散串, 并对散串进行词匹配和语法分析处理, 进而发现候选错误字串, 由人机交互的方法对错误串进行自动校正, 取得了较高的查错率。

(3) 基于统计

张照煌[15] 提出一种利用综合近似字集替换, 并用统计语言模型评分的方法, 其基本思想是以事先整理好字形、字音、字义或输入码相近字的综合近似字集替换待校对句子中的每个汉字, 产生许多候选字符串(或许多路径), 利用统计语言模型对各候选字符串评分, 将评分最高的字符串与待校对文本中的句子进行对照, 即可发现错误之所在并提供相对应的正确字。该方法的难点是如何整理综合近似字集, 且若近似字集较大的话, 计算量是非常大的; 其不足之处是只能校对所谓的别字错误, 对多字、漏字、易位等错误难以发现。东北大学提出了一种混合文本校对方法 HMCTC, 采用模式匹配方法进行最长匹配分词, 发现长词错误; 然后根据类三元语法, 将与前后相邻词同现频率乘积小于一定阈值的词标记为错误; 最后对词进行语法属性标注, 在不可能的语法标注序列字词处作错误标记[25]。其缺点是基于词语同现频率的查错判据受限于训练语料的大小和语料选取的领域, 且词语同现频率数据的获取需要大规模经过切分的熟语料, 而这样的熟语料是难以获得的。清华大学利用语料库统计知识指导文本校对[19], 以句为单位, 把句子看作字段和词段, 对字段计算字段平均字频、字段平均转移概率; 对词段计算词间字转移概率、词性转移概率, 将转移概率作为查错判据, 把转移概率小于阈值的字或词作为查出的错误。其中, 查错判据是自动查错研究的核心, 仍有待于进一步研究。北京工业大学计算机学院在对大规模语料库的统计分析基础上, 构建了二字结构工程并引入人名、地名辨识规则, 利用词语类间的接续关系进行查错, 对人名、地名误报率低[21]。

4.2.2　自动纠错的研究状况

自动纠错是文本自动校对的一个重要组成部分, 它为自动查错时侦测出的错误字符串提供修改建议, 辅助用户改正错误。修改建议的有效性是衡量自动纠错性能的主要指标, 它有两点要求: ①提供的修改建议中应该含有正确或合理的建议; ②正确或合理的修改建议应尽可能排列在所有建议的前面。因此, 纠错修改建议的产生算法及排序算法是自动纠错研究的两个核心课题。

由于中文文本自动校对理论和技术尚不太成熟, 自动纠错研究的论述还不多见。东北大学采用模式匹配方法对长词进行纠错处理[25], 但没有充分利用出错字符串的特征, 算法计算量大。IBM 中国研究中心[26] 提出一种替换字表结合主词典, 通过加字和换字对侦测出来的错误字符串提供修改建议的纠错算法, 但该算法的纠错建议局限于替换字表, 没有考虑上下文启发信息, 主要考虑对错字这种错误类型进行纠错, 对漏字、多字、易位、多字替换、英文单词拼写等错误类型的纠错能力较弱。山西大学[30] 提出了一种基于似然匹配的纠错建议候选集产生算法, 对漏字、多字、易位、多字替换等错误类型的纠错能力有了较大的提高。

5　中文文本自动校对存在的问题与对策

经过多年的研究, 已有一些商品化的文本自动校对软件在出版印刷界得到一定程度的应用, 如黑马校对系统、方正金山校对系统等。但与机器翻译一样, 文本自动校对技术是建立在自然语言理解技术的基础之上的, 是一个难度很大的研究课题, 系统的错误召回率和准确率都比较低(召回率小于 70%, 准确率小于 40%)[6], 纠错建议的有效率或首选正确率也很低, 与用户的要求还有较大差距, 故其技术还有待进一步研究。

造成中文文本自动校对技术召回率和准确率较低的原因有如下几点: ①中文文本中的错误都是 "真字错误 ", 针对英文比较有效的单词查错和纠错技术在中文中不太适用; ②目前基于上下文的自动查错技术主要还是字词级的水平, 使用的查错语言模型是字词级的简单统计模型(如 Bigram 或 Trigram), 利用的语言学知识不够丰富, 对于更高级(如句子级) 的错误很难查出; ③尽管实践已经证明, 有指导的统计方法是建立自然语言应用系统模型的有效手段, 但用于语言模型训练的切分/标注的大规模语料很难获得, 因而由于数据稀疏而导致模型训练得不够充分; ④目前的自动校对技术研究重查错轻纠错, 很多使用中的校对系统对查出的错误给不出纠错建议, 或给出的纠错建议很不准确。

针对中文文本的上述问题, 我们认为自动校对技术的研究应在以下几方面得到加强:

(1) 加强句法、语义层次的校对策略研究, 与目前研究较多的词汇级校对策略相结合, 从而能够检查以往无法查出的错误。汉语的词类没有形态的变化, 词类和句法成分之间不存在简单的对应关系, 汉语的词序又非常灵活, 汉语的这些特点使得汉语的语法分析存在很大的难度, 而面向错误文本的句法分析与错误检查难度会更大, 采用什么样的方法既可以降低句法分析难度又能够满足文本错误检测的要求需要很好地研究。

语义问题是语言学与语言信息处理研究中的薄弱环节,而已公开的利用语义信息实现文本校对的研究成果很少,语义错误检查在中文文本校对系统中仍相当困难,但并不是说语义校对无从着手,如通过义素分析法或语义文法,或许能帮助对文本校对中的语义错误进行检查,但这首先需要对文本进行词义排歧与标注,这方面仍需深入研究。

(2) 查错后的纠错处理是校对系统的重要组成部分,目前对如何产生纠错候选词的研究以及如何对纠错候选词的排序方法的研究还不是很多。这方面的研究也涉及到上下文,况且面对的错误可能是多字、漏字、易位或多字替换等各种类型,而对所生成的多个候选词的排序又依赖于目标词所在的上下文。这方面的研究需要不断深入。

(3) 加强自然语言处理的基础研究。构建信息丰富的综合语言知识库,包括标注完整的大规模语料库的建设,这涉及词语切分、标注和未登录词识别等问题,切分和标注中歧义排除本身就是非常难的问题。

(4) 加强从语言知识库中获取文本自动查错知识的机器学习方法研究。根据中文文本的特点,研究面向中文文本自动查错和纠错的计算语言模型。如何抽取语言知识库中的各种特征,进而获得构建文本查错语言模型的知识,如何将文本中蕴涵的语言学特征与统计特征相结合建立更有效的查错与纠错模型还需进一步研究。

参考文献:

[1] aren Kukich. Techniques for Automatically Correcting Words in Text [J]. ACM Computing Surveys, 1992, 24(4): 377-438.

[2] Andrew R Golding. A Winnow-based Approach to Context-Sensitive Spelling Correction[J]. Machine Learning, 1999, 34: 107-130.

[3] Li Jianhua, Wang Xiaolong. Combing Trigram and Automatic Weight Distribution in Chinese Spelling Error Correction[J]. Journal of Computer Science and Technology, 2002, 17(6): 915-923.

[4] Li Jianhua, Wang Xiaolong. Study on Automatic Spelling Check and Correction[J]. Journal of Chinese Language and Computing, 2003, 1(1): 25-36.

[5] Lei Zhang, Ming Zhou, Changning Huang, *et al*. Automatic Chinese Text Error Correction Approach Based on Fast Approximate Chinese Word-matching Algorithm[C]. Microsoft Research China Paper Collection, 2000. 231-235.

[6] Lei Zhang, Ming Zhou, Changning Huang, *et al*. Automatic Detecting/Correcting Errors in Chinese Text by an Approximate Word-matching Algorithm[C]. Microsoft Research China Paper Collection, 2000. 135-141.

[7] Lei Zhang, Ming Zhou, Changning Huang. Multifeature-based Approach to Automatic Error Detection and Correction of Chinese Text [C]. Microsoft Research China Paper Collection, 2000. 193-197.

[8] Golding A R, San R. Applying Winnow to Context-Sensitive Spelling Correction[C]. Proc. of the 13th ICML, Bari, Italy, 1996.

[9] Golding A R. A Bayesian Hybrid Method for Context-Sensitive Spelling Correction[C]. Boston: Proceedings of the 3rd Workshop on Very Large Corpora, 1995.

[10] Joseph J Pollock. Automatic Spelling Correction in Scientific and Scholarly Text[J]. Communication of the ACM, 1984, (4): 358-368.

[11] James L Peterson. Computer Programs for Detecting and Correcting Spelling Errors[J]. Communication of the ACM, 1980, (12): 676-687.

[12] E M Riseman. A Contextual Postprocessing System for Error Correction Using Binary N-gram[J]. IEEE Trans. on Computer, 1974, 22(5): 480-493.

[13] E D Ward, D M Riseman. Contextual Word Recognition Using Binary Digrams[J]. IEEE on Computers, 1971, 20(4): 397-403.

[14] Golding A R, Schabes Yves. Combining Trigram-based and Feature-based Methods for Context-Sensitive Spelling Correction[C]. Santa Cruz: Proceeding of the 34th Annual Meeting of the Association for Computational Linguistics, 1996. 71-78.

[15] Chao-huang Chang. A Pilot Study on Automatic Chinese Spelling Error Correction [J]. Communication of COLIPS, 1994, 4(2): 143-149.

[16] 李晶皎, 张莉, 姚天顺. 汉语语音理解中自动纠错系统的研究[J]. 软件学报, 1999, 10(4): 377-381.

[17] 吴岩, 李秀坤, 刘挺, 等. 中文自动校对系统的研究与实现[J]. 哈尔滨工业大学学报, 2001, (2): 60-64.

[18] 李建华, 王晓龙, 王平, 等. 多特征的中文文本校对算法的研究[J]. 计算机工程与科学, 2001, (3): 93.

[19] 孙才, 罗振声. 汉语文本校对字词级查错处理的研究[A]. 第四届计算语言学会议论文集(语言工程)[C]. 北京: 清华大学出版社, 1997. 319-324.

[20] 刘挺, 等. 中文计算机辅助校对系统原理[J]. 中文信息, 1997.

[21] 邱超捷, 宋柔, 等. 大规模语料库中词语接续对的统计与分析[A]. 第四届计算语言学会议论文集(语言工程)[C]. 北京: 清华大学出版社, 1997.

[22] 张磊, 周明, 黄昌宁, 等. 中文文本自动校对[J]. 语言文字应用, 2001, 2(1): 19-25.

[23] 王云. 计算机中文文本校对的原理与方法[J]. 教学与科技, 2003, 16(2): 41-44.

[24] 易蓉湘, 何克抗. 计算机汉语文稿校对系统[J]. 计算机研究与发展, 1997, 34(5): 346-350.

[25] 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).

[26] 郭志立, 等. 中文校对系统中的修改建议提供算法[A]. 第四届计算语言学会议论文集(语言工程)[C]. 北京: 清华大学出版社, 1997. 325-330.

[27] 骆卫华, 罗振声, 宫小瑾. 中文文本自动校对的语义级查错研究[J]. 计算机工程与应用, 2003, 39(12): 115-118.

[28] 龚小谨, 罗振声. 中文文本自动校对中的语法错误检查[J]. 计算机工程与应用, 2003, 39(8): 98-100.

[29] 张仰森, 丁冰青. 基于二元接续关系检查的字词级自动查错方法[J]. 中文信息学报, 2001, 15(3): 36-43.

[30] 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, 2001, 15(5): 33-39.

[31] 张仰森, 曹元大. 基于统计的纠错建议给出算法及其实现[J]. 计算机工程, 2004, 30(11): 106-109.

[32] 骆卫华, 罗振声, 宫小瑾. 中文文本自动校对技术的研究[J]. 计算机研究与发展, 2004, 41(1): 244-246.

作者简介:

张仰森(1962-), 男, 教授, 博士, 研究方向为人工智能、自然语言处理; 俞士汶(1938-), 男, 安徽宣城人, 教授, 博导, 研究方向为计算语言学。